International Series of Numerical Mathematics

172

Michael Hintermüller,
Roland Herzog,
Christian Kanzow,
Michael Ulbrich,
Stefan Ulbrich, Editors

Non-Smooth and Complementarity-Based Distributed Parameter Systems

Simulation and Hierarchical Optimization

Birkhäuser

ISNM

International Series of Numerical Mathematics

Volume 172

The *International Series of Numerical Mathematics* is open to all aspects of numerical mathematics, with topics of particular interest including free boundary value problems for differential equations, phase transitions, problems of optimal control and optimization, other nonlinear phenomena in analysis, nonlinear partial differential equations, efficient solution methods, bifurcation problems, and approximation theory. When possible, the topic of each volume is discussed from three different angles, namely those of mathematical modeling, mathematical analysis, and numerical case studies.

All manuscripts are peer-reviewed to meet the highest standards of scientific literature. Interested authors may submit proposals by email to the series editors or to the relevant Birkhäuser editor listed under "Contacts."

More information about this series at http://www.springer.com/series/4819

Michael Hintermüller • Roland Herzog
Christian Kanzow • Michael Ulbrich
Stefan Ulbrich
Editors

Non-Smooth and Complementarity-Based Distributed Parameter Systems

Simulation and Hierarchical Optimization

Birkhäuser

Editors
Michael Hintermüller
Weierstrass Institute for Applied
Analysis and Stochastics
Berlin, Germany

Roland Herzog
Institute for Applied Mathematics
University of Heidelberg
Heidelberg, Germany

Christian Kanzow
Lehrstuhl für Mathematik VII
Julius-Maximilians-Universität Würzburg
Würzburg, Bayern, Germany

Michael Ulbrich
Department of Mathematics
Technical University of Munich
Garching b. München, Bayern, Germany

Stefan Ulbrich
Fachbereich Mathematik
Technische Universitaet Darmstadt
Darmstadt, Hessen, Germany

ISSN 0373-3149 ISSN 2296-6072 (electronic)
International Series of Numerical Mathematics
ISBN 978-3-030-79392-0 ISBN 978-3-030-79393-7 (eBook)
https://doi.org/10.1007/978-3-030-79393-7

Mathematics Subject Classification: 49J52, 49J53, 90C33, 46N10, 93A13, 49J21, 49J20, 49K40

This book is published under the imprint Birkhäuser, www.birkhauser-science.com, by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume of the *International Series of Numerical Mathematics* presents research results obtained in the first funding phase of the Special Priority Program (SPP) 1962 on *Nonsmooth and Complementarity Based Distributed Parameter Systems: Simulation and Hierarchical Optimization* from 2016 to 2019. The program was funded by the *Deutsche Forschungsgemeinschaft (DFG)*. Within the funding period, 24 projects located at many universities and research institutions across Germany, involving also a tandem project which was co-funded by the Swiss National Fund (SNF) with one project partner in Lugano, unfolded various research activities, leading to more than 100 preprints as well as several workshops and exchange activities in particular for young researchers. The SPP 1962 also co-organized the *International Conference on Continuous Optimization (ICCOPT)* which was held in Berlin from August 5 to 8, 2019, preceded by a summer school for early career researchers from August 3 to 4, 2019. The coordination project for the entire program (supported scientifically by Amal Alphonse and Michael Hintermüller) was located at the Weierstrass Institute for Applied Analysis and Stochastics in Berlin.

The main mathematical theme of the SPP1962 is non-smoothness as many of the most challenging problems in the applied sciences involve non-differentiable structures as well as partial differential operators, thus leading to non-smooth distributed parameter systems. The non-smoothness considered in this SPP typically arises:

(i) Directly in the problem formulation
(ii) Through inequality constraints, nonlinear complementarity, or switching systems
(iii) As a result of competition and hierarchy

In fact, very challenging applications for (i) come from frictional contact problems, or non-smooth constitutive laws associated with physical processes such as Bean's critical state model for the magnetization of superconductors, which leads to a quasi-variational inequality (QVI) problem; for (ii) are related to non-penetration conditions in contact problems, variational inequality problems, or

inequality constraints in optimization problems, which, upon proper re-formulation, lead to complementarity problems and further, by means of non-linear complementarity problem (NCP) functions, to non-smooth systems similar to (i); and for (iii) come from multi-objective control systems or leader-follower principles, as they can be found in optimal system design in robotics and biomechanics. Modeling "competition" often leads to generalized Nash equilibrium problems (GNEPs) or partial differential games. Moreover, modeling "hierarchy" results in mathematical programs with equilibrium constraints (MPECs), a class of optimization problems with degenerate, non-smooth constraints. All of these problems are highly nonlinear, lead to QVIs, and represent rather novel mathematical structures in applications based on partial differential operators. In these and related applications, the transition from smoothing or simulation-based approaches to genuinely non-smooth techniques or to multi-objective respectively hierarchical optimization is crucial.

Fundamental difficulties in non-smooth partial differential systems, associated optimization, and hierarchical problems are of analytical as well as algorithmic and numerical nature. For instance, for QVIs, the existence and stability of solutions is a major challenge, whereas MPECs suffer from a lack of existence of Lagrange multipliers due to constraint degeneracy, which hinders the derivation of proper stationarity conditions. Numerical challenges, which are present in all non-smooth problems of this SPP, involve the stability of discretization/model reduction schemes or severe mesh dependence of algorithms. In order to overcome these difficulties, the goals of this SPP are to advance tools from non-smooth and set-valued analysis and to build a basis for stable numerical approximation/discretization schemes that enable the design of algorithms with mesh independent convergence. The SPP 1962 also aims to address the influence of parameters, which enter the above applied problems and which either range within a specified set or result from hierarchy. The former leads to robust optimization in form of deterministic MPECs, which challenge the characterization of stationary points and the development of efficient solvers. Hierarchical problems (or MPECs) contain variables which enter into lower-level problems as parameters. Summarizing, the research program of the SPP leads to a modern treatment of non-smooth problems and will therefore shape future applications in the field.

Corresponding to the above goals, each subsequent section of this volume presents the findings of projects within the SPP.

Berlin, Germany Michael Hintermüller

Contents

Error Bounds for Discretized Optimal Transport and Its Reliable Efficient Numerical Solution

Sören Bartels and Stephan Hertzog

Abstract The discretization of optimal transport problems often leads to large linear programs with sparse solutions. We derive error estimates for the approximation of the problem using convex combinations of Dirac measures and devise an active-set strategy that uses the optimality conditions to predict the support of a solution within a multilevel strategy. Numerical experiments confirm the theoretically predicted convergence rates and a linear growth of effective problem sizes with respect to the variables used to discretize given data.

Keywords Optimal transport · Sparsity · Optimality conditions · Error bounds · Iterative solution

Mathematics Subject Classification (2020) 65K10, 49M25, 90C08

1 Introduction

The goal in *optimal transportation* is to transport a measure μ into a measure ν with minimal total effort with respect to a given cost function c. This optimization problem can be formulated as an infinite-dimensional linear program. One way to find optimal solutions is to approximate the transport problem by (finite-dimensional) standard linear programs. This can be done by approximating the measures μ and ν by convex combinations of Dirac measures, and we prove that this leads to accurate approximations of optimal costs. The size of these linear programs grows quadratically in the size of the supports of these approximations, i.e., if M and N are the number of atoms on which the approximations are supported, then the size of the linear programs is MN. Thus, they can only be solved directly on

S. Bartels (✉) · S. Hertzog
Abteilung für Angewandte Mathematik, Albert-Ludwigs-Universität Freiburg, Freiburg i.Br., Germany
e-mail: bartels@mathematik.uni-freiburg.de

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_1

coarse grids, i.e., typically only for approximations with a few thousand atoms. It is another goal of this article to devise an iterative strategy that automatically identifies the support of a solution using auxiliary problems of comparable sizes. For other approaches to the numerical solution of optimal transport problems, we refer the reader to [1–3, 5, 14]; for details on the mathematical formulation and its analytical features we refer the reader to [8, 17, 18].

Our error estimate follows from identifying convex combinations of Dirac measures supported in the nodes of a given triangulation as approximations of probability measures via the adjoint of the standard nodal interpolation operator defined on continuous functions. Thereby, it is possible to quantify the approximation quality of a discretized probability measure in the operator norm related to a class of continuously differentiable functions.

Using the fact that if c is strictly convex and μ has a density, the support of optimal solutions is contained in a lower-dimensional set, we expect that the linear programs have a sparse solution, i.e., the number of nonzero entries in the solution matrix is comparable to $M+N$. Related approaches have previously been discussed in the literature, cf. [12, 15]. In this article, we aim at investigating a general strategy that avoids assumptions on an initial guess or a coarse solution and particular features of the cost function and thus leads to an efficient solution procedure that is fully reliable.

The optimality conditions for standard linear programs characterize the optimal support using the Lagrange multipliers ϕ and ψ which occur as solutions of the dual problem. Given approximations of those multipliers, we may restrict the full linear program to the small set of atoms where those approximations satisfy the characterizing equations of the optimal support up to some tolerance, with the expectation that the optimal support is contained in this set. If the solution of the corresponding reduced linear program satisfies the optimality conditions of the full problem, a global solution is found. Otherwise, the tolerance is increased to enlarge the active set of the reduced problem, and the procedure is repeated. Good approximations of the Lagrange multipliers result from employing a multilevel scheme and in each step prolongating the dual solutions computed on a coarser grid to the next finer grid.

Our numerical experiments reveal that this iterative strategy leads to linear programs whose dimensions are comparable to $M+N$. The optimality conditions have to be checked on the full product grid which requires $\mathcal{O}(MN)$ arithmetic operations. These are however fully independent and can be realized in parallel. The related algorithm of [12] avoids this test and simply adds atoms in a neighbourhood of a coarse-grid solution. This is an efficient strategy if a good coarse-grid solution is available. Similar approaches have been discussed in [9, 11, 16].

Another alternative is the method presented in [15] where the concept of shielding neighbourhoods is introduced. Solutions which are optimal in a *shielding neighbourhood* are analytically shown to be globally optimal. Strategies to construct those sets are presented for several cost functions. However, each cost function requires a particular strategy to find the neighbourhoods, depending on its geometric structure. Critical for the efficiency of the algorithm is the sparsity of shielding

neighbourhoods for which theoretical bounds and intuitive arguments are given, confirmed by numerical experiments.

The efficiency of our numerical scheme can be greatly increased if it is combined with the methods from [12] or [15]. In this case, the activation of atoms is only done within the described neighbourhoods of the support of a current approximation. This is expected to be reliable once asymptotic convergence behaviour is observed.

The outline of this article is as follows. The general optimal transport problem, its discretization, optimality conditions, and sparsity properties are discussed in Sect. 2. A rigorous error analysis for optimal costs based on the approximation of marginal measures via duality is carried out in Sect. 3. Section 4 devises the multilevel active-set strategy for efficiently solving the linear programs arising from the discretization. The efficiency of the algorithm and the optimality of the error estimates are illustrated via numerical experiments in Sect. 5.

2 Discretized Optimal Transport

We describe in this section the general mathematical framework for optimal transport problems, their discretization, optimality conditions, and sparsity properties of optimal transport plans.

2.1 General Formulation

The general form of an optimal transport problem seeks a probability measure $\pi \in \mathcal{M}(X \times Y)$ called a *transport plan* on probability spaces X and Y such that its projections onto X and Y coincide with given probability measures $\mu \in \mathcal{M}(X)$ and $\nu \in \mathcal{M}(Y)$, respectively, called *marginals*, and such that it is optimal in the set of all such measures for a given continuous *cost function* $c : X \times Y \to \mathbb{R}$. The minimization problem thus reads

$$(\widehat{P}) \quad \begin{cases} \text{Minimize } \widehat{I}[\pi] = \iint_{X\times Y} c(x, y)\, \mathrm{d}\pi(x, y) \\ \text{subject to } \pi \in \mathcal{M}(X \times Y),\ \pi \geq 0,\ P_X\pi = \mu,\ P_Y\pi = \nu. \end{cases}$$

Here, $P_X\pi$ and $P_Y\pi$ are defined via $P_X\pi(A) = \pi(A\times Y)$ and $P_Y\pi(B) = \pi(X\times B)$ for measurable sets $A \subset X$ and $B \subset Y$, respectively. This formulation may be regarded as a relaxation of the problem of determining a *transport map* $T : X \to Y$ which minimizes a cost functional in the set of bijections between X and Y subject to the constraint that the measure μ is pushed forward by T into the measure ν:

$$(P) \quad \begin{cases} \text{Minimize } I[T] = \int_X c(x, T(x))\, \mathrm{d}\mu(x) \\ \text{subject to } T \text{ bijective and } T_\#\mu = \nu. \end{cases}$$

Here, the pushforward measure $T_\# \mu$ is the measure on Y defined via $T_\#\mu(B) = \mu(T^{-1}(B))$ for all measurable sets $B \subset Y$. In the case that μ and ν have densities $f \in L^1(X)$ and $g \in L^1(Y)$, the relation $T_\#\mu = \nu$ is equivalent to the identity

$$g \circ T \det DT = f,$$

which is a Monge–Ampère equation if $T = \nabla\Phi$ for a convex potential Φ. Since the formulation (P) does not provide sufficient control on variations of transport maps to pass to limits in the latter equation, it is difficult to establish the existence of solutions directly. In fact, optimal transport maps may not exist, e.g., when a single Dirac mass splits into a convex combination of several Dirac masses. The linear program $(\widehat{P})$ extends the formulation (P) via graph measures $\pi = (\mathrm{id} \times T)\#\mu$ and admits solutions. In the case of a strictly convex cost function c it can be shown that optimal transport plans correspond to optimal transport maps, i.e., optimal plans are supported on graphs of transport maps, provided that μ has a density. In this sense $(\widehat{P})$ is a relaxation of (P); we refer the reader to [8, 17, 18] for details.

2.2 Discretization

In the case where the marginals are given by convex combinations of Dirac measures supported in atoms $(x_i)_{i=1,\dots,M} \subset X$ and $(y_j)_{j=1,\dots,N} \subset Y$, respectively, i.e.,

$$\mu_h = \sum_{i=1}^{M} \mu_h^i \delta_{x_i}, \qquad \nu_h = \sum_{j=1}^{N} \nu_h^j \delta_{y_j},$$

we have that admissible transport plans π are supported in the set of pairs of atoms (x_i, y_j). Indeed, if $A \times B \subset X \times Y$ with $(x_i, y_j) \notin A \times B$, i.e., $x_i \notin A$ for all $i \in \{1, 2, \dots, M\}$ or $y_j \notin B$ for all $j \in \{1, 2, \dots, N\}$, then one of the inequalities

$$\pi(A \times B) \le \pi(A \times Y) = \mu_h(A) = 0,$$
$$\pi(A \times B) \le \pi(X \times B) = \nu_h(B) = 0,$$

holds, and we deduce $\pi(A \times B) = 0$. By approximating measures μ and ν by convex combinations of Dirac measures μ_h and ν_h, we therefore directly obtain a standard linear program that determines the unknown matrix $\pi_h \in \mathbb{R}^{M\times N}$:

$$(\widehat{P}_h) \quad \begin{cases} \text{Minimize } \widehat{I}_h[\pi_h] = \sum_{i=1}^{M} \sum_{j=1}^{N} c(x_i, y_j)\pi_h^{ij} \\ \text{subject to } \pi_h \ge 0,\ \sum_{j=1}^{N} \pi_h^{ij} = \mu_h^i,\ \sum_{i=1}^{M} \pi_h^{ij} = \nu_h^j. \end{cases}$$

The rigorous construction of approximating measures μ_h and ν_h via duality arguments will be described below in Sect. 3. Weak convergence of discrete transport

plans to optimal transport plans can be established via abstract theories, cf. [12, 18] for details.

2.3 Optimality Conditions

Precise information about the support of an optimal discrete transport plan π_h is provided by the Lagrange multipliers corresponding to the marginal constraints. Including these in a Lagrange functional $\widehat{L}_h$ leads to

$$\begin{aligned}\widehat{L}_h[\pi_h;\phi_h,\psi_h] &= \widehat{I}_h[\pi_h] + \sum_{i=1}^{M}\phi_h^i\Big(\mu_h^i - \sum_{j=1}^{N}\pi_h^{ij}\Big) + \sum_{j=1}^{N}\psi_h^j\Big(\nu_h^j - \sum_{i=1}^{M}\pi_h^{ij}\Big)\\ &= \sum_{i=1}^{M}\sum_{j=1}^{N}\pi_h^{ij}\Big(c(x_i,y_j) - \phi_h^i - \psi_h^j\Big) + \sum_{i=1}^{M}\phi_h^i\mu_h^i + \sum_{j=1}^{N}\psi_h^j\nu_h^j.\end{aligned}$$

Minimization in $\pi_h \geq 0$ and maximization in ϕ_h and ψ_h provide the condition

$$c(x_i,y_j) - \phi_h^i - \psi_h^j \geq 0,$$

and the implication

$$\phi_h^i + \psi_h^j < c(x_i,y_j) \quad \Longrightarrow \quad \pi_h^{ij} = 0,$$

which determines the support of the discrete transport plan π_h.

2.4 Sparsity

The Knott–Smith theorem and generalizations thereof state that optimal transport plans are supported on c-cyclically monotone sets, cf. [18]. In particular, if c is strictly convex and if the marginal μ has a density, then optimal transport plans are unique and supported on the graph of the c-subdifferential of a c-convex function Φ. For the special case of a quadratic cost function, it follows that Φ is a solution of the Monge–Ampère equation for which regularity properties can be established, cf. [7, 17]. Hence, in this case it is rigorously established that the support is contained in a lower-dimensional submanifold. Typically, such a quantitative behaviour can be expected but may be false under special circumstances. We refer the reader to [6] for further details on partial regularity properties of transport maps.

On the discrete level, it is irrelevant to distinguish measures with or without densities since the action of a discrete measure on a finite-dimensional set V_h of

continuous functions can always be identified with an integration, i.e., we associate a well defined density $f_h \in V_h$ by requiring that

$$\int_X v_h f_h \, \mathrm{d}x = \langle \mu_h, v_h \rangle,$$

for all $v_h \in V_h$. The properties of optimal transport plans thus apply to the discrete transport problem introduced above. Asymptotically, these properties remain valid provided that we have $f_h \to f$ in $L^1(X)$ for a limiting density $f \in L^1(X)$.

3 Error Analysis

We derive an error estimate for the approximation of the continuous problem $(\widehat{P})$ by the discrete problem $(\widehat{P}_h)$ by appropriately interpolating measures. For this we follow [13] and assume that we are given a triangulation $\mathcal{T}_h$ with maximal mesh-size $h > 0$ of a closed domain $U \subset \mathbb{R}^d$ which represents X or Y with nodes

$$\mathcal{N}_h = \{z_1, z_2, \dots, z_L\}$$

and associated nodal basis functions $(\varphi_z : z \in \mathcal{N}_h)$. With the corresponding nodal interpolation operator onto the set of elementwise affine, globally continuous functions given by

$$\mathcal{I}_h : C(U) \to \mathcal{S}^1(\mathcal{T}_h), \quad \mathcal{I}_h v = \sum_{z \in \mathcal{N}_h} v(z) \varphi_z,$$

we define approximations $\mathcal{I}_h^* \varrho$ of measures $\varrho \in \mathcal{M}(U) \simeq C(U)^*$ via

$$\langle \mathcal{I}_h^* \varrho, u \rangle = \langle \varrho, \mathcal{I}_h u \rangle = \sum_{z \in \mathcal{N}_h} u(z) \langle \varrho, \varphi_z \rangle,$$

i.e., we have the representation

$$\mathcal{I}_h^* \varrho = \sum_{z \in \mathcal{N}_h} \varrho_z \delta_z$$

with $\varrho_z = \langle \varrho, \varphi_z \rangle$. Standard nodal interpolation estimates imply that we have, cf. [4],

$$\big| \langle \varrho - \mathcal{I}_h^* \varrho, u \rangle \big| = \big| \langle \varrho, u - \mathcal{I}_h u \rangle \big| \le c_{\mathcal{I}} h^{1+\alpha} \|u\|_{C^{1,\alpha}(U)} \|\varrho\|_{\mathcal{M}(U)},$$

for all $u \in C^{1,\alpha}(U)$. Analogously, we can approximate measures on the product space $X \times Y$ with triangulations $\mathcal{T}_{X,h}$ and $\mathcal{T}_{Y,h}$, nodes $\mathcal{N}_{X,h}$ and $\mathcal{N}_{Y,h}$, and

interpolation operators $\mathcal{I}_{X,h}$ and $\mathcal{I}_{Y,h}$, respectively, via

$$\langle \mathcal{I}^*_{X\otimes Y,h}\pi, r\rangle = \langle \pi, \mathcal{I}_{X\otimes Y,h} r\rangle = \sum_{(x,y)\in\mathcal{N}_{X,h}\times\mathcal{N}_{Y,h}} r(x,y)\langle \pi, \varphi_x \otimes \varphi_y\rangle,$$

for all $r \in C(X \times Y)$. In the following error estimate, we abbreviate the optimal values of the minimization problems $(\widehat{P})$ and $(\widehat{P}_h)$ by $\min_{\pi\geq 0} \widehat{I}[\pi]$ and $\min_{\pi_h\geq 0} \widehat{I}_h[\pi_h]$, respectively.

Proposition 3.1 *Assume that $\mu_h = \mathcal{I}^*_{X,h}\mu$ and $\nu_h = \mathcal{I}^*_{Y,h}\nu$. If $c \in C^{1,\alpha}(X \times Y)$ with $\alpha \in [0, 1]$ we then have*

$$\min_{\pi\geq 0} \widehat{I}[\pi] - \min_{\pi_h\geq 0} \widehat{I}_h[\pi_h] \leq c_{\mathcal{I}} h^{1+\alpha} \|c\|_{C^{1,\alpha}(X\times Y)}.$$

If for every $\widehat{\pi}_h$ that is admissible in $(\widehat{P}_h)$ we have that the measure

$$\widehat{\pi} = \widehat{\pi}_h + \mathrm{d}x \otimes (\nu - \nu_h) + (\mu - \mu_h) \otimes \mathrm{d}y$$

is nonnegative, then the converse estimate also holds.

Proof

(i) Assume that $\min_{\pi\geq 0} \widehat{I}[\pi] \leq \min_{\pi_h\geq 0} \widehat{I}_h[\pi_h]$. The interpolant $\widehat{\pi}_h = \mathcal{I}^*_{X\times Y,h}\pi$ of a solution π for $(\widehat{P})$ is admissible in $(\widehat{P}_h)$ since

$$\langle \mathcal{I}^*_{X\otimes Y,h}\pi, v \otimes 1\rangle = \langle \pi, \mathcal{I}_{X,h} v \otimes 1\rangle = \langle \mu, \mathcal{I}_{X,h} v\rangle = \langle \mathcal{I}^*_{X,h}\mu, v\rangle = \langle \mu_h, v\rangle,$$

for every $v \in C(X)$, i.e., $P_X\widehat{\pi}_h = \mu_h$. Analogously, we find that $P_Y\widehat{\pi}_h = \nu_h$. This implies that

$$\begin{aligned}\min_{\pi_h\geq 0} \widehat{I}_h[\pi_h] - \min_{\pi\geq 0} \widehat{I}[\pi] &\leq \widehat{I}_h[\widehat{\pi}_h] - \widehat{I}[\pi] \\ &= \langle \widehat{\pi}_h - \pi, c\rangle \leq c_{\mathcal{I}} h^{1+\alpha} \|c\|_{C^{1,\alpha}(X\times Y)},\end{aligned}$$

where we used that $\|\pi\|_{\mathcal{M}(X\times Y)} = 1$.

(ii) If conversely we have $\min_{\pi\geq 0} \widehat{I}[\pi] \geq \min_{\pi_h\geq 0} \widehat{I}_h[\pi_h]$ we let π_h be a discrete solution and consider the measure

$$\widehat{\pi} = \pi_h + \mathrm{d}x \otimes (\nu - \nu_h) + (\mu - \mu_h) \otimes \mathrm{d}y,$$

which is nonnegative and satisfies

$$\langle \widehat{\pi}, r\rangle = \langle \pi_h, r\rangle + \int_X \langle \nu - \nu_h, r(x,\cdot)\rangle \,\mathrm{d}x + \int_Y \langle \mu - \mu_h, r(\cdot, y)\rangle \,\mathrm{d}y,$$

for all $r \in C(X \times Y)$. We have that

$$\langle \widehat{\pi}, v \otimes 1 \rangle = \langle \mu_h, v \rangle + \int_Y \langle \mu - \mu_h, v \rangle \, \mathrm{d}y = \langle \mu, v \rangle,$$

i.e., $P_X \widehat{\pi} = \mu$. Analogously, we find that $P_Y \widehat{\pi} = \nu$. Therefore, $\widehat{\pi}$ is admissible in the minimization problem $(\widehat{P})$ and hence

$$\begin{aligned}
\min_{\pi \geq 0} \widehat{I}[\pi] - \min_{\pi_h \geq 0} \widehat{I}_h[\pi_h] &\leq \widehat{I}[\widehat{\pi}] - \widehat{I}[\pi_h] \\
&= \int_X \langle \nu - \nu_h, c(x, \cdot) \rangle \, \mathrm{d}x + \int_Y \langle \mu - \mu_h, c(\cdot, y) \rangle \, \mathrm{d}y \\
&\leq c_{\mathcal{I}} h^{1+\alpha} \big(\max_{x \in X} \|c(x, \cdot)\|_{C^{1,\alpha}(Y)} + \max_{y \in Y} \|c(\cdot, y)\|_{C^{1,\alpha}(X)} \big) \\
&\leq c_{\mathcal{I}} h^{1+\alpha} \|c\|_{C^{1,\alpha}(X \times Y)},
\end{aligned}$$

where we used the property $\|\mu\|_{\mathcal{M}(X)} = \|\nu\|_{\mathcal{M}(Y)} = 1$. □

The estimate can be improved if assumptions on the transport plan are made.

Remark 3.2

(i) For the polynomial cost function $c_p(x, y) = (1/p)|x - y|^p$, $1 \leq p < \infty$, we have $c_p \in C^{1,\alpha}(X \times Y)$ for $\alpha = \min\{1, p - 1\}$, so that the derived convergence rate is subquadratic if $p < 2$. If the transport plan is supported away from the diagonal $\{x = y\}$, along which the differentiability of c_p is limited, then quadratic convergence applies.

(ii) The assumption on the nonnegativity of the measure $\widehat{\pi}$ is a condition on the regularity of the marginals μ and ν. It can be rigorously established, e.g., if μ and ν have densities that are piecewise affine or convex. More generally, suitable constructions may be necessary that may decrease the order of convergence.

A similar error estimate is expected to hold if the measures μ and ν are approximated via piecewise affine densities f_h and g_h as this corresponds to a rescaling of coefficients and the use of quadrature in the cost functional.

Remark 3.3 Alternatively to the above discretization, transport plans can be approximated via discrete measures π_h which have densities $p_h \in \mathcal{S}^1(\mathcal{T}_{X,h}) \otimes \mathcal{S}^1(\mathcal{T}_{Y,h})$, i.e.,

$$\langle \pi_h, r \rangle = \iint_{X \times Y} r(x, y) p_h(x, y) \, \mathrm{d}(x, y)$$

with

$$p_h(x, y) = \sum_{i=1}^{M} \sum_{j=1}^{N} p_h^{ij} \varphi_{x_i}(x) \varphi_{y_j}(y).$$

We associate discrete densities $f_h \in \mathcal{S}^1(\mathcal{T}_{X,h})$ and $g_h \in \mathcal{S}^1(\mathcal{T}_{Y,h})$ with the marginals μ and ν via

$$(f_h, v_h)_h = \langle \mu, v_h \rangle, \quad (g_h, w_h)_h = \langle \nu, w_h \rangle,$$

for all $v_h \in \mathcal{S}^1(\mathcal{T}_{X,h})$ and $w_h \in \mathcal{S}^1(\mathcal{T}_{Y,h})$ and with (discrete) inner products $(\cdot, \cdot)_h$ on $C(X)$ and $C(Y)$, e.g., if μ and ν have densities f and g, then f_h and g_h may be defined as their L^2 projections. If the inner products involve quadrature, then we have

$$(f_h, v_h)_h = \int_X \mathcal{I}_{X,h}[f_h v_h] \, \mathrm{d}x = \sum_{i=1}^{M} \beta_i f_h(x_i) v_h(x_i),$$

where $\beta_i = \int_X \varphi_{x_i} \, \mathrm{d}x$ and it follows that

$$f_h(x_i) = \beta_i^{-1} \langle \mu, \varphi_{x_i} \rangle$$

for $i = 1, 2, \ldots, M$. Analogously, we have $g_h(y_j) = \gamma_j^{-1} \langle \nu, \varphi_{y_j} \rangle$. The coefficients are thus scaled versions of the coefficients used above. Using quadrature in the cost functional leads to

$$I[\pi_h] = \iint_{X \times Y} c(x, y) p_h(x, y) \, \mathrm{d}(x, y) \approx \sum_{i=1}^{M} \sum_{j=1}^{N} c(x_i, y_j) p_h^{ij} \beta_i \gamma_j.$$

Again, the coefficients here are scaled versions of the coefficients π_h^{ij} used above.

A reduced convergence rate applies for the approximation using piecewise constant finite element functions.

Remark 3.4 Approximating measures by measures with densities that are element-wise constant, i.e.,

$$\langle \mu_h, v \rangle = \sum_{T \in \mathcal{T}_h} \mu_h^T \int_T v \, \mathrm{d}x,$$

we obtain a reduction of the convergence rate by one order.

4 Active-Set Strategy

For a subset of atoms specified via an index set

$$\mathcal{A} \subset \{1, \dots, M\} \times \{1, \dots, N\}$$

which is admissible in the sense that there exists $\widehat{\pi}_h$ with

$$\sum_{j=1,\dots,N,\,(i,j)\in\mathcal{A}} \widehat{\pi}_h^{ij} = \mu_h^i, \qquad \sum_{i=1,\dots,M,\,(i,j)\in\mathcal{A}} \widehat{\pi}_h^{ij} = \nu_h^j,$$

we restrict to discrete transport plans that are supported on $\mathcal{A}$ and hence consider the following reduced problem:

$$(\widehat{P}_{h,\mathcal{A}}) \quad \begin{cases} \text{Minimize } \widehat{I}_{h,\mathcal{A}}[\pi_h] = \sum_{(i,j)\in\mathcal{A}} c(x_i, y_j)\pi_h^{ij} \\ \text{subject to } \pi_h \geq 0,\ \sum_{j,(i,j)\in\mathcal{A}} \pi_h^{ij} = \mu_h^i,\ \sum_{i,(i,j)\in\mathcal{A}} \pi_h^{ij} = \nu_h^j. \end{cases}$$

The following proposition provides a sufficient condition for the definition of an active set that leads to an accurate reduction.

Proposition 4.1 *Assume that we are given approximations $\widetilde{\phi}_h$ and $\widetilde{\psi}_h$ of exact discrete multipliers ϕ_h and ψ_h with*

$$\|\widetilde{\phi}_h - \phi_h\|_{L^\infty(X)} + \|\widetilde{\psi}_h - \psi_h\|_{L^\infty(Y)} \leq \varepsilon_{as}.$$

If the set of active atoms $\mathcal{A}$ on $X \times Y$ is defined via

$$\mathcal{A} = \left\{(i, j) : \widetilde{\phi}_h^i + \widetilde{\psi}_h^j \geq c(x_i, y_j) - 2c_{as}\varepsilon_{as}\right\}$$

with $c_{as} \geq 1$, then the minimization problem $(\widehat{P}_{h,\mathcal{A}})$ is an accurate reduction of $(\widehat{P}_h)$ in the sense that their solution sets coincide.

Proof Let π_h be a solution of the nonreduced problem $(\widehat{P}_h)$ and let ϕ_h, ψ_h be corresponding Lagrange multipliers. If $\pi_h^{ij} \neq 0$ for the pair $(i, j) \in \{1, \dots, M\} \times \{1, \dots, N\}$, then we have $c(x_i, y_j) = \phi_h^i + \psi_h^j$ and hence

$$\widetilde{\phi}_h^i + \widetilde{\psi}_h^j = \widetilde{\phi}_h^i - \phi_h^i + \widetilde{\psi}_h^j - \psi_h^j + c(x_i, y_j) \geq c(x_i, y_j) - 2c_{as}\varepsilon_{\mathrm{as}}.$$

This implies that $(i, j) \in \mathcal{A}$ and π_h is admissible in the reduced formulation $(\widehat{P}_{h,\mathcal{A}})$. □

Proposition 4.1 suggests a multilevel iteration realized in the subsequent algorithm where the Lagrange multipliers of a coarse-grid solution are used as approximations for the multipliers on a finer grid which serve to guess the support of the optimal transport plan. If the optimality conditions are not satisfied up to a

mesh-dependent tolerance, then the variable activation tolerance is enlarged and the solution procedure repeated. Because of the quasioptimal quadratic convergence behaviour of the employed $P1$ finite element method, a quadratic tolerance is used.

Algorithm 1 (Multilevel Active Set Strategy) Choose triangulations $\mathcal{T}_{X,h}$ and $\mathcal{T}_{Y,h}$ of X and Y with maximal mesh-size $h > 0$. Let $\theta_{act} > 0$, $0 < h_{min} < h$, and $c_{opt} > 0$. Choose functions $\widetilde{\phi}_h \in \mathcal{S}^1(\mathcal{T}_{X,h})$ and $\widetilde{\psi}_h \in \mathcal{S}^1(\mathcal{T}_{Y,h})$.

(1) Define the set of activated atoms via

$$\mathcal{A} = \left\{(i, j) : \widetilde{\phi}_h^i + \widetilde{\psi}_h^j \geq c(x_i, y_j) - \theta_{act} h^2\right\}$$

and enlarge $\mathcal{A}$ to guarantee feasibility.
(2) Solve the reduced problem $(\widehat{P}_{h,\mathcal{A}})$ and extract multipliers ϕ_h and ψ_h.
(3) Check optimality conditions up to tolerance $c_{opt} h^2$ on the full set of atoms, i.e., whether

$$\phi_h^i + \psi_h^j \leq c(x_i, y_j) + c_{opt} h^2$$

is satisfied for all $(x_i, y_j) \in \mathcal{N}_{X,h} \times \mathcal{N}_{Y,h}$.
(4) If optimality holds and $h > h_{min}$ then refine triangulations $\mathcal{T}_{X,h}$ and $\mathcal{T}_{Y,h}$, prolongate functions ϕ_h and ψ_h to the new triangulations with new mesh-size $h \leftarrow h/2$ to update $\widetilde{\phi}_h$ and $\widetilde{\psi}_h$, set $\theta_{act} \leftarrow \theta_{act}/2$, and continue with (1).
(5) If optimality fails, then set $\theta_{act} \leftarrow 2\theta_{act}$ and continue with (1).
(6) Stop if optimality holds and $h \leq h_{min}$.

Various modifications of Algorithm 1 are possible that may lead to improvements of its practical performance.

Remark 4.2

(i) The activation parameter θ_{act} is adapted during the procedure, i.e., the parameter is increased if optimality fails on a given level. To avoid activating too many atoms initially, θ_{act} is decreased whenever a new level is reached.
(ii) The quadratic tolerance in the verification of the optimality conditions turned out to be sufficient to obtain a quadratic convergence of optimal costs and of the Lagrange multipliers in our experiments.
(iii) The initial parameter θ_{act} can be optimized on the coarsest mesh by repeatedly reducing it until optimality fails.

5 Numerical Experiments

In this section, we illustrate our theoretical investigations via several experiments. We implemented Algorithm 1 in MATLAB and used the optimization package GUROBI, cf. [10], to solve the linear programs. The experiments were run on a

standard personal computer. Integrals were evaluated using a three-point trapezoidal rule on triangles. The employed triangulations result from uniform refinements of an initial coarse triangulation and are represented via their refinement level $k \in \mathbb{N}$ so that the maximal mesh-size satisfies $h \sim 2^{-k}$. The number of nodes in the triangulations of the spaces X and Y are referred to by M and N, respectively. The measured CPU times reported below are provided to compare different strategies.

5.1 Problem Specifications

We consider four different transport problems specified via the sets X and Y and the marginals μ and ν together with different polynomial cost functions

$$c_p(x, y) = \frac{1}{p}|x - y|^p,$$

where $p \in \{3/2, 2, 3\}$. These choices are prototypical for subquadratic, quadratic, and superquadratic costs leading to singular, linear, and degenerate cost gradients, respectively. In the special case of a quadratic cost function solutions for the optimal transport problem can be constructed using the Monge–Ampère equation

$$\det D^2\Phi = \frac{f}{g \circ \nabla\Phi}$$

and the relations for the transport map and the multipliers

$$T = \nabla\Phi, \quad \phi(x) = \frac{|x|^2}{2} - \Phi(x), \quad \psi(y) = \frac{|y|^2}{2} - \Phi^*(y),$$

with the convex conjugate $\Phi^*(y) = \sup_x x \cdot y - \Phi(x)$ of Φ, cf. [18] for details. Moreover, we then have the optimal cost

$$I[T] = I[\nabla\Phi] = \int_X c_2(x, \nabla\Phi(x))\,\mathrm{d}\mu(x).$$

The first example is one-dimensional and allows for a simple visualization of the transport map.

Example 1 (One-Dimensional Transport) Let $X = Y = [0, 1]$ and μ and ν be defined via the densities

$$f(x) = \frac{2}{3}(x + 1), \quad g(y) = 1,$$

respectively. For $p = 2$ the optimal transport plan is given by the transport map $T = \nabla\Phi$ with the potential

$$\Phi : [0, 1] \to \mathbb{R}, \quad x \mapsto \frac{1}{9}x^3 + \frac{1}{3}x^2,$$

and the Lagrange multiplier ϕ satisfies

$$\phi(x) = \frac{1}{6}x^2 - \frac{1}{9}x^3.$$

The optimal cost for $p = 2$ is given by $I[T] = 1/540$.

Our second example concerns the transport between two rectangles with a differentiable transport map.

Example 2 (Smooth Transport Between Rectangles) Define $X = [0, 1]^2$ and $\Phi(x_1, x_2) = x_1^2 + x_2^3$ and set $Y = \nabla\Phi(X) = [0, 2] \times [0, 3]$ and $g = 1$. The Monge–Ampère equation determines

$$f(x_1, x_2) = 12x_2,$$

so that the optimal cost value for $p = 2$ is given by $I[\nabla\Phi] = 43/10$.

In order to compare our algorithm to the results from [12], we incorporate Example 4.1 from that article.

Example 3 (Setting from [12]) On $X = Y = [-1/2, 1/2]^2$, let μ and ν be defined by the densities

$$\begin{aligned} f(x_1, x_2) = 1 &+ 4(q''(x_1)q(x_2) + q(x_1)q''(x_2)) \\ &+ 16(q(x_1)q(x_2)q''(x_1)q''(x_2) - q'(x_1)^2 q'(x_2)^2) \end{aligned}$$

and $g = 1$, where

$$q(z) = \left(-\frac{1}{8\pi}z^2 + \frac{1}{256\pi^3} + \frac{1}{32\pi}\right)\cos(8\pi z) + \frac{1}{32\pi^2}z\sin(8\pi z).$$

For $p = 2$ we obtain an exact solution via the Monge–Ampère equation.

The final example describes the splitting of a square into two rectangles.

Example 4 (Discontinuous Transport) Let $X = [-1/2, 1/2]^2$ and $Y = \big([-3/2, -1] \cup [1, 3/2]\big) \times [-1/2, 1/2]$ be equipped with the constant densities $f = 1$ and $g = 1$. For any strictly convex cost function, optimal transport maps T isometrically map the left half of the square to the rectangle on the left side and the

other half to the one on the right, i.e., up to identification of Lebesgue functions,

$$T(x_1, x_2) = \begin{cases} (x_1 + 1, x_2) & \text{if } x_1 > 0, \\ (x_1 - 1, x_2) & \text{if } x_1 < 0. \end{cases}$$

For $p = 2$ we have $T = \nabla\Phi$ with

$$\Phi(x_1, x_2) = \frac{x_1^2 + x_2^2}{2} + |x_1|,$$

with corresponding Lagrange multiplier $\phi(x_1, x_2) = -|x_1|$.

Figure 1 shows characteristic features of the four examples. In particular, in the upper left plot of Fig. 1 the transport plan is the graph of a monotone function and we illustrated an activated set of atoms of a discretization that approximates the graph.

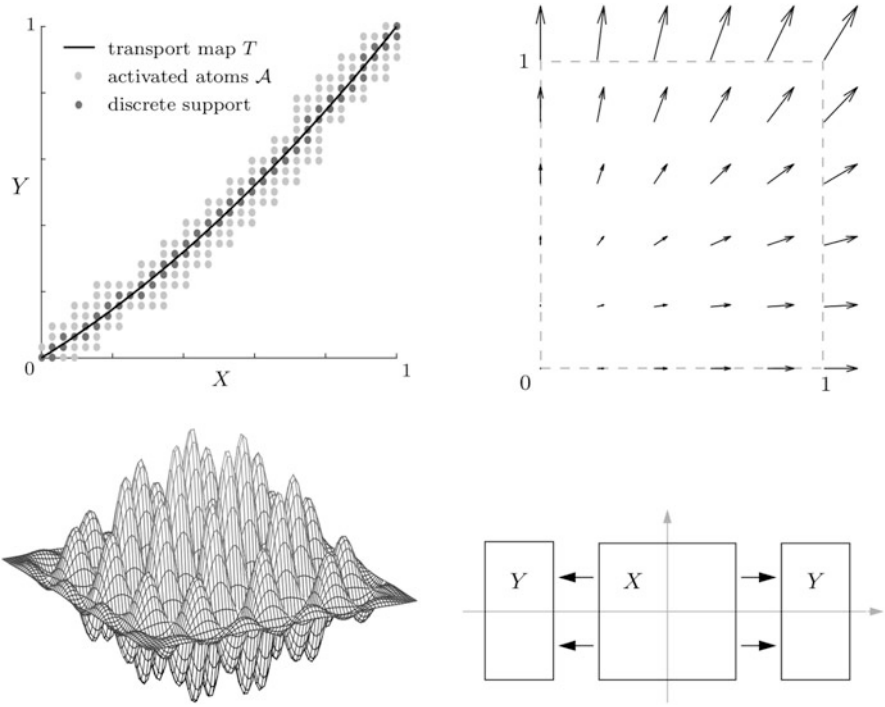

Fig. 1 Characteristic features of the transport problems defined in Examples 1–4 (from left to right and top to bottom): (i) optimal transport plan given by a graph together with activated atoms and discrete support for $k = 5$ in Example 1, (ii) optimal transport map $T = \nabla\Phi$ in Example 2 interpreted as a vector field, (iii) oscillating density f in Example 3, (iv) piecewise affine optimal transport plan T in Example 4

5.2 Complexity Considerations

A crucial quantity to determine the efficiency of our devised method is the growth of the cardinalities of the activated sets. In Table 1 we display for Examples 1–4 the corresponding numbers on different triangulations and for different cost functions. We observe that in all experiments the size of the activated sets grows essentially linearly in strong contrast to the quadratic growth of the theoretical number of unknowns of the corresponding discrete transport problem. A slight deviation of this behaviour occurs in Example 2 for $p = 3$ where the increase of the active-set size is larger than the expected factor 4. We note that we observed a reduction of the active-set sizes by factors of approximately 2^{-d} compared to the sizes obtained with the algorithm from [12] for generic choices of parameters. Because of the very few required redefinitions of the active set, particularly for $p \geq 2$, we conclude that the optimality conditions provide a precise prediction of the supports even if only

Table 1 Total number of nodes $M + N$, number of unknowns in the full optimization problem MN, and cardinalities of activated sets at optimality with number of tolerance increases in brackets in Examples 1–4 on triangulations with refinement level k and different cost functions $c_p(x, y)$

Ex. 1	$k = 7$	$k = 8$	$k = 9$	$k = 10$
$M + N$	258	514	1026	2050
MN	16,641	66,049	263,169	1,050,625
$p = 3/2$	763 (0)	1531 (0)	3067 (0)	6139 (0)
$p = 2$	763 (0)	1531 (0)	3067 (0)	6139 (0)
$p = 3$	763 (0)	1539 (0)	3114 (0)	6442 (0)

Ex. 2	$k = 3$	$k = 4$	$k = 5$	$k = 6$
$M + N$	506	1906	7394	29,122
MN	34,425	467,313	6,866,145	105,189,825
$p = 3/2$	6268 (8)	27846 (1)	179,594 (2)	745,713 (1)
$p = 2$	3929 (0)	15,729 (0)	63,115 (0)	252,951 (0)
$p = 3$	8085 (2)	56,703 (2)	255,965 (1)	1,847,207 (2)

Ex. 3	$k = 3$	$k = 4$	$k = 5$	$k = 6$
$M + N$	162	578	2178	8450
MN	6561	83,521	1,185,921	17,850,625
$p = 3/2$	1389 (0)	20,787 (7)	58,575 (1)	183,465 (1)
$p = 2$	1589 (0)	5755 (0)	24,018 (0)	103,100 (0)
$p = 3$	1495 (0)	6319 (0)	26,205 (0)	106,857 (0)

Ex. 4	$k = 3$	$k = 4$	$k = 5$	$k = 6$
$M + N$	171	595	2,211	8,515
MN	7290	88,434	1,221,858	18,125,250
$p = 3/2$	1346 (0)	6384 (0)	24,135 (0)	95,240 (0)
$p = 2$	1654 (0)	6921 (0)	29,106 (0)	120,153 (0)
$p = 3$	1274 (0)	5602 (0)	21,353 (0)	85,463 (0)

Table 2 Total CPU time in seconds on k-th level in Examples 1–4 with different polynomial cost functions $c_p(x, y)$

Ex. 1	$k = 7$	$k = 8$	$k = 9$	$k = 10$
$p = 3/2$	0.0575	0.1357	0.4281	1.1840
$p = 2$	0.0603	0.1288	0.3294	1.0257
$p = 3$	0.0593	0.1345	0.3943	1.3115
Ex. 2	$k = 3$	$k = 4$	$k = 5$	$k = 6$
$p = 3/2$	0.2063	0.8805	6.9296	49.5964
$p = 2$	0.2187	0.5262	2.4734	21.1738
$p = 3$	0.2584	1.6697	7.3599	97.7239
Ex. 3	$k = 3$	$k = 4$	$k = 5$	$k = 6$
$p = 3/2$	0.1090	0.7655	1.5510	9.5078
$p = 2$	0.1106	0.1718	0.7735	4.6622
$p = 3$	0.1410	0.1886	0.8557	5.6919
Ex. 4	$k = 3$	$k = 4$	$k = 5$	$k = 6$
$p = 3/2$	0.1192	0.1777	0.8014	4.9880
$p = 2$	0.1054	0.1766	0.6771	4.0459
$p = 3$	0.1214	0.1700	0.7194	4.8851

approximations of the multipliers are available, i.e., this property appears to be very robust with respect to perturbations of the multipliers.

In Table 2 we display the total CPU time needed to solve the optimization problem on the k-th level. This includes the repeated activation of atoms, the repeated solution of the reduced linear programs, and the verification of the optimality conditions. We observe a superlinear growth of the numbers. These are dominated by the times needed to solve the linear programs whereas the (non-parallelized) verification of the optimality conditions was negligible in all tested situations.

5.3 Experimental Convergence Rates

In Figs. 2 and 3 we show for Examples 1 and 2 the error in the approximation of the optimal cost, i.e., the quantities

$$\delta_h = \Big| \min_{\pi \geq 0} \widehat{I}[\pi] - \min_{\pi_h \geq 0} \widehat{I}_h[\pi_h] \Big|$$

and the error in the approximation of the Lagrange multiplier ϕ, i.e., the quantities

$$\varepsilon_h = \|\mathcal{I}_{X,h}\phi - \phi_h\|_{L^\infty(X)}.$$

If the exact optimal cost or the multiplier was not known, i.e., if $p \neq 2$, we used an extrapolated reference value or considered the difference $\mathcal{I}_{X,h}\phi_{h/2} - \phi_h$ to define δ_h and ε_h, respectively. We tested different polynomial costs and considered sequences

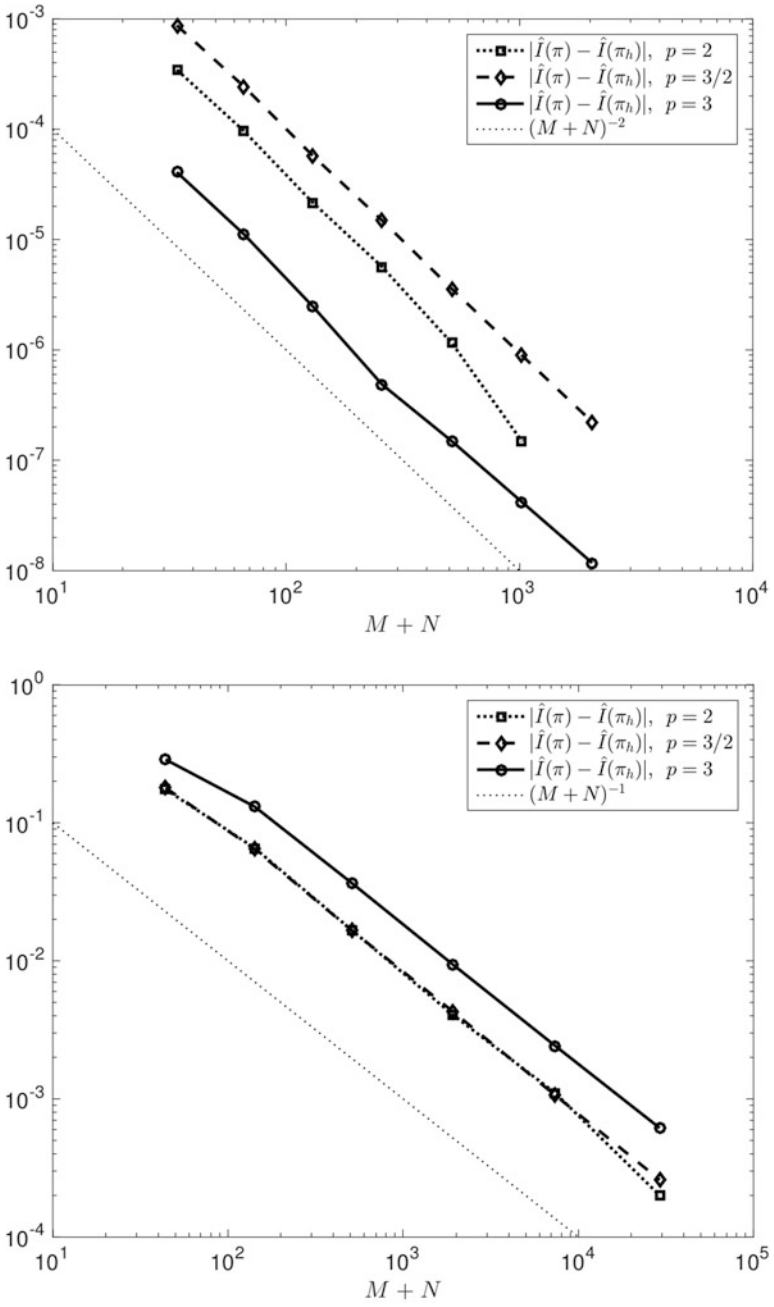

Fig. 2 Experimental convergence of optimal costs in Examples 1 (left) and 2 (right) for different cost functions on sequences of uniformly refined triangulations

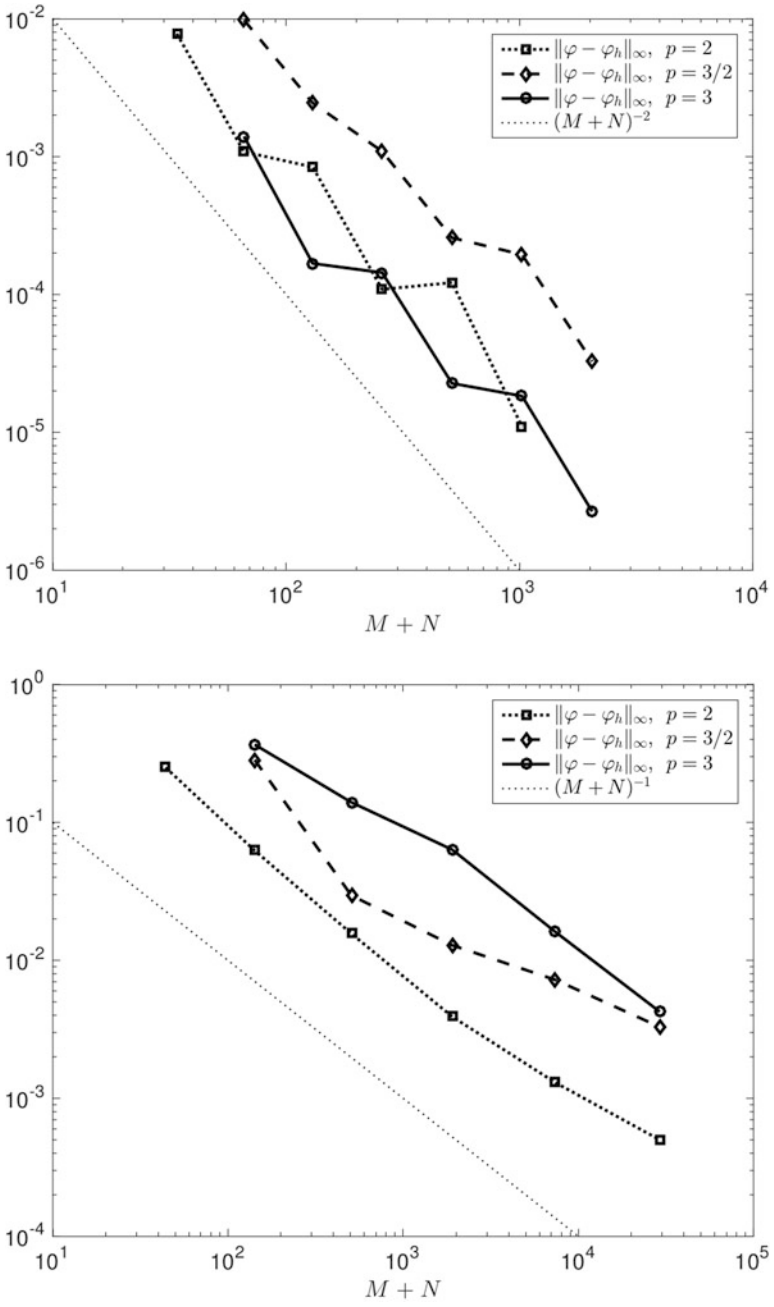

Fig. 3 Experimental convergence rates of the discrete multiplier ϕ_h in Examples 1 (left) and 2 (right) for different cost functions on sequences of uniformly refined triangulations

Table 3 Experimental errors $\varepsilon_h = \|\mathcal{I}_h\phi - \phi_h\|_{L^\infty(X)}$ for discretizations using $P1$ and $P0$ approximations of densities in Example 3 with $p = 2$

ε_h	$h \sim 2^{-5}$	$h \sim 2^{-6}$	$h \sim 2^{-7}$	$h \sim 2^{-8}$	$h \sim 2^{-9}$
$P1$ (Alg. 1)	0.00781	0.00238	0.00086	–	–
$P0$ [12]	0.00721	0.00892	0.00689	0.00241	0.00148

of uniformly refined triangulations. Because of the relation

$$h \sim (M + N)^{-1/d},$$

a quadratic convergence rate $\mathcal{O}(h^2)$ corresponds to a slope $-2/d$ with respect to the total number of nodes $M + N$. Figure 2 confirms the estimate from Proposition 3.1 and additionally shows that the quadratic convergence rate is optimal. The experimental results also reveal that the employed quadratic tolerance in the verification of the optimality conditions in Algorithm 1 is sufficient to preserve the convergence rate of the linear program using the full set of atoms. Figure 3 indicates that quadratic convergence in $L^\infty(X)$ also holds for the approximation of the Lagrange multiplier ϕ provided this quantity is sufficiently regular. In particular, we observe here a slower convergence behaviour for $p = 3/2$.

In [12] an approximately linear convergence rate in L^∞ of the multipliers has been reported for Example 3 which is consistent with the piecewise constant approximation of densities of measures used in that article, cf. Remark 3.4. In particular, discrete duality yields that the Lagrange multipliers occurring in the discretized optimal transport problems are discretized in the same spaces. For our discretization using continuous, piecewise affine approximations we obtain a nearly quadratic experimental convergence rate in this example as well, as can be seen in Table 3 in which we also display the errors from [12].

References

1. Jean-David Benamou and Yann Brenier, *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem*, Numer. Math. **84** (2000), no. 3, 375–393.
2. Jean-David Benamou and Guillaume Carlier, *Augmented Lagrangian methods for transport optimization, mean field games and degenerate elliptic equations*, J. Optim. Theory Appl. **167** (2015), no. 1, 1–26.
3. Jean-David Benamou, Brittany D. Froese, and Adam M. Oberman, *Numerical solution of the optimal transportation problem using the Monge-Ampère equation*, J. Comput. Phys. **260** (2014), 107–126.
4. Susanne C. Brenner and L. Ridgway Scott, *The mathematical theory of finite element methods*, third ed., Texts in Applied Mathematics, vol. 15, Springer, New York, 2008.
5. Sören Bartels and Patrick Schön, *Adaptive approximation of the Monge-Kantorovich problem via primal-dual gap estimates*, ESAIM Math. Model. Numer. Anal. **51** (2017), no. 6, 2237–2261.

6. Shibing Chen and Alessio Figalli, *Partial $W^{2,p}$ regularity for optimal transport maps*, J. Funct. Anal. **272** (2017), no. 11, 4588–4605.
7. Guido De Philippis and Alessio Figalli, *$W^{2,1}$ regularity for solutions of the Monge-Ampère equation*, Invent. Math. **192** (2013), no. 1, 55–69.
8. Lawrence C. Evans, *Partial differential equations and Monge-Kantorovich mass transfer*, Current developments in mathematics, 1997 (Cambridge, MA), Int. Press, Boston, MA, 1999, pp. 65–126.
9. Tilmann Glimm and Nick Henscheid, *Iterative scheme for solving optimal transportation problems arising in reflector design*, ISRN Applied Mathematics **2013** (2013), 12 pages, Id. 635263.
10. Inc. Gurobi Optimization, *Gurobi optimizer reference manual*, 2016.
11. Quentin Mérigot and Édouard Oudet, *Discrete optimal transport: complexity, geometry and applications*, Discrete Comput. Geom. **55** (2016), no. 2, 263–283.
12. Adam M. Oberman and Yuanlong Ruan, *An efficient linear programming method for optimal transportation*, arXiv preprint arXiv:1509.03668 (2015).
13. Tomáš Roubíček, *Relaxation in optimization theory and variational calculus*, De Gruyter Series in Nonlinear Analysis and Applications, vol. 4, Walter de Gruyter & Co., Berlin, 1997.
14. Ludger Rüschendorf and Ludger Uckelmann, *Numerical and analytical results for the transportation problem of Monge-Kantorovich*, Metrika **51** (2000), no. 3, 245–258.
15. Bernhard Schmitzer, *A sparse multiscale algorithm for dense optimal transport*, Journal of Mathematical Imaging and Vision **56** (2016), no. 2, 238–259.
16. Bernhard Schmitzer and Christoph Schnörr, *A hierarchical approach to optimal transport*, Scale Space and Variational Methods in Computer Vision, Lecture Notes in Computer Science, vol. 7893, Springer, Berlin, Heidelberg, 2013.
17. Cédric Villani, *Topics in optimal transportation*, no. 58, American Mathematical Soc., 2003.
18. _____, *Optimal transport: old and new*, vol. 338, Springer Science & Business Media, 2008.

Numerical Methods for Diagnosis and Therapy Design of Cerebral Palsy by Bilevel Optimal Control of Constrained Biomechanical Multi-Body Systems

Hans Georg Bock, Ekaterina Kostina, Marta Sauter, Johannes P. Schlöder, and Matthias Schlöder

Abstract In this chapter, we describe how to use mathematical optimization for diagnosis and treatment planning of cerebral palsy. We give background information on the disease and the medical diagnosis and treatment which is conducted currently and to which we contribute. As common in biomechanics, we model the human body as a rigid multi-body system and give a review on the corresponding dynamics. Assuming that, as a consequence of nature evolutionary process, natural gaits are optimal with respect to a certain performance criterion, the gait itself is modeled as the solution of an optimal control problem subject to state-dependent constrains, where a cost function depends on individual parameters. We present two solution approaches to this problem, which then serves as the lower level of two bilevel problems: An inverse optimal control problem, where we use parameter estimation to extract the patient's individual optimization criteria out of motion tracking data, and a robustified optimal control problem, in which we simulate the effect of interventions, modeled as parameter variations, on a patient's gait while taking into account possible uncertainties.

Keywords Optimal control · Bilevel optimization · Parameter estimation · Robustification · Mathematical programs with complementarity constraints · Rigid multi-body system · Human gait · Treatment planning · Cerebral palsy

Mathematics Subject Classification (2020) 49M37, 65K10

The presented work was carried out in cooperation with Sebastian Wolf (MotionLab, Department of Orthopaedics and Trauma Surgery, Heidelberg University Hospital) and Katja Mombaur (Department of Mechanical and Mechatronics Engineering, University of Waterloo)

H. G. Bock · E. Kostina (✉) · M. Sauter · J. P. Schlöder · M. Schlöder
INF 205, Heidelberg, Germany
e-mail: bock@iwr.uni-heidelberg.de; ekaterina.kostina@iwr.uni-heidelberg.de; marta.sauter@iwr.uni-heidelberg.de; matthias.schloeder@iwr.uni-heidelberg.de

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_2

1 Introduction

The project emerged from the long-standing collaboration with the MotionLab [41] of the Department of Orthopaedics and Trauma Surgery of the Heidelberg University Hospital. The aim of the project is to investigate the involved medical challenges in detail, to transfer them to adequate mathematical tasks, and to develop a mathematical and numerical framework for

- Inverse problems to support proper diagnosis and
- Parameter optimization and optimal control problems (OCPs) to improve the planning of interventions

for patients with CP. The ultimate goal is to provide routinely applicable mathematical tools for the medical doctors.

1.1 Cerebral Palsy

We give some medical background to characterize the problem environment the proposed mathematical methods have to fit in. CP describes a wide range of disorders that are caused by problems before, during, and after birth, such as oxygen deficiency. It is the most frequent cause of motor disability among children in Europe representing 700,000 citizens. The prevalence of CP in Europe ranges between 1.5 and 3.0 cases per 1000 live births [3]. The disease affects muscle tone, posture, and eventually leads to deformed joints and skeleton. The abnormal muscle tones often combined with permanently contracting spastic muscles cause the typical so-called crouched or scissored gait.

There is robust evidence for deterioration in gross motor function and deterioration in walking ability during childhood. Although no specific treatment can remediate the neurological disorder that causes complex functional impairment, there are multiple therapies that aim at improving a patient's activity and participation in daily life. Orthopedic surgery is a key component in the clinical management of musculoskeletal pathologies in CP. Different interventions are applied to avoid progressive worsening of musculoskeletal impairments, improve gait patterns, and thus, a patient's quality of life.

Many CP patients have to undergo multiple surgeries during their treatment. As stated in [3], to avoid the so-called birthday syndrome, which is the correction of one deviation after another occurring with growth, the current treatment approach is to perform a single event multi-level surgery to manage musculoskeletal deformities in CP. This avoids the highly frequent hospitalization of the children, but makes a prediction of the outcome much harder, since many factors—also interacting with each other—play a role.

One crucial question is of course which treatment shall be applied to which patient. In the past, besides looking at the patient's gait visually, doctors used

static data like X-ray photographs to decide which surgery should be applied. Sometimes this resulted in patients whose standing posture looked more "healthy," but who tragically lost their ability to walk. The philosophy nowadays is to consider also dynamic data. Clinical gait analysis provides information on spatiotemporal, kinematics, and kinetic data, but still does not tell what happens inside the patient and how to treat him or her.

Therefore, diagnosis and decision for a treatment still depend strongly on the experience of the medical doctors. Although many patients show improved treatment outcomes now, according to [37] and the references therein there is still a significant fraction of patients (around 23%) who experience negative outcome after surgery despite following the treatment recommendations from clinical gait analysis. Therefore including additional mathematical analysis on 3D gait data to provide direct information on surgically adjustable parameters may help to further improve the clinical decision making.

From the medical point of view, major questions in the current research of CP are:

- Improved understanding of the principles behind a CP patient's gait: Calibrating a dynamic gait model by measurements of the patient's gait will not only provide estimates of kinematic and dynamic parameters but also of state and torque histories to support diagnosis;
- Treatment planning based on a personalized predictive model: Depending on the medical options for interventions and physiotherapy, it is very important to predict and optimize any alteration's effect on the patient's gait based on a theoretical model capturing the main characteristics of a CP patient's gait;
- Criteria to evaluate the success of possible treatments: Currently, one can assess the result of a therapy only by looking at the gait before and after a medical treatment—including CP—and comparing the resulting gait with those of a healthy subject, e.g., as in the *Gait Profile Score* [4]. This favors "good looking gaits," although it is not at all clear, which changes are indeed beneficial for specific patients. As of today, criteria for success like *improved stability* are only a conjecture.

In this project, we contribute to the first two questions using mathematical methods.

1.2 Modeling Approach

Mathematically, tackling the problem results in highly non-smooth bilevel optimization, resp., OCPs, which will be sketched in the following. As a model of a patient's gait, we propose a biomechanical optimal control model, in which the actuator torques are generated by the CP patient according to certain optimization criteria, which need to be identified.

Level 1 The underlying dynamics are given by a multi-body system (MBS) of mechanics, where the torques of the rotational joints are the result of more complex muscle operations. The resulting equations of motion are the solution of a classical variational problem subject to numerous state constraints (e.g. knee stroke, non-self-penetration, collision avoidance, ground contact), and can be written in form of a differential algebraic equation (DAE) system with implicit discontinuities of states and the right-hand side caused by activating and deactivating constraints. This is a dynamical process with complementarity constraints in itself, and complex switching structures are expected. In case of collisions, e.g. when a foot hits the ground, not only the right-hand side of the DAE is discontinuous (type changes occur whenever inequality constraints become active or inactive), but also the velocity states themselves jump at state-dependent switching points. As OCPs constrained by differential equations with implicitly defined, state-dependent discontinuities are notoriously difficult to solve, the modeling and analysis of biomechanical motions under state inequality constraints represents a significant non-smooth problem, that is why we present adequate optimization algorithms for the switched systems. In order to catch the CP gait characteristics, the developed model captures the full 3D motion and can be extended for a sufficiently detailed foot and muscle model.

Level 2 Since the gait is the result of an autonomous decision of a person on the controls generating torques and forces, depending on a combination of optimization criteria (such as stability, energy efficiency, or comfort), state and control inequality constraints, and given—complex—physiological parameters, we model it as an OCP for the underlying dynamics of Level 1. Naturally, the constraints lead to additional non-smoothness and complementarity constraints.

To treat the real medical problem for CP patients, however, requires to include at least one additional optimization level. Two bilevel problems are considered:

Level 3A—Inverse Optimal Control The essential role of this mathematical model for the human gait is to provide a non-invasive diagnosis tool looking into what happens inside the patient, based on measurements of the gait—identifying kinematic parameters such as joint displacements and skeleton deformations, and in particular torque histories as well as a (parameterized) optimization criterion supporting medical diagnosis. Mathematically this results in a parameter estimation problem with an OCP as constraint.

Level 3B—Robustified Optimal Control Once the dynamic model is calibrated to a CP patient's individual data and a sensitivity analysis is performed which indicates the most significant deficiencies compared to healthy subjects, the model can be used to evaluate and eventually improve the effect of a planned surgical intervention or physiotherapy, thus analyzing predictively the effect of altered physiological parameters on the patient's motion. To improve reliability, the OCP needs to be robustified to account for the inaccurate knowledge of the model or the intervention as well as the patient's reaction.

2 Modeling the Human Body

The human musculoskeletal system is very complex, comprising more than 200 bones and 400 muscles, tissue of different solidity and flexibility, etc. This great complexity makes it indispensable to perform (partially drastic) simplifications in order to make the model accessible to the use of numerical methods with an acceptable computational time. Hence, we follow the line of research in [15, 18, 33] and model the human body as a rigid MBS. The body is represented by a set of rigid segments of different sizes, masses, and inertias, connected with joints. As every human is different, this model needs to be personalized. For instance, De Leva [11] reports anthropometric data for segment center locations, segment masses, and inertia depending on gender, weight, and height of a subject. Data like this together with further processing can be used to create individualized models, which are sufficiently close to reality for our purposes. Depending on the task which shall be performed using this model, some parts of the human body demand for a more detailed modeling. This is in particular true for those parts of a patient's body which are most effected by the disease resp. which shall be altered during a medical treatment, but also those, which have a major impact on the patient's gait.

2.1 Rigid Multi-Body Systems

We give a short review on rigid MBS dynamics, based on [13, 14], where the interested reader can find more details. We restrict this review to MBSs of which the topology can be described by a tree. The dynamics of these systems can be expressed by means of generalized coordinates, a non-unique minimal set of coordinates which already comprise the joints constraints. The number of generalized coordinates needed to describe a multi-body system is also called degree of freedom.

We denote the generalized coordinates at time point t by $q(t)$, the generalized velocities $\dot{q}(t)$ by $v(t)$, the generalized accelerations $\ddot{q}(t)$ by $a(t)$, the generalized forces by $\tau(t)$, and body specific parameters by p. When possible without causing confusion, we omit the argument t for the sake of simplicity. The equations of motion for an unconstrained multi-body system can then be expressed by

$$H(q, p)\ddot{q} + C(q, v, p) = \tau, \tag{2.1}$$

where $H(\cdot)$ is the generalized inertia matrix, which agglomerates all segment inertias in a suitable way, depending on the current configuration q of the system, and $C(\cdot)$ is the Coriolis term, also called generalized bias force.

In case of external contacts, in addition constraining forces act on the MBS in order to satisfy the constraint describing the external contact, which can be

expressed in the form

$$g(q, p) = 0. \tag{2.2}$$

The equations of motion then read as

$$\begin{aligned} H(q, p)a + C(q, v, p) &= \tau + G(q, p)^T \lambda, \\ g(q, p) &= 0, \end{aligned} \tag{2.3}$$

where $G(\cdot) = \frac{\partial g}{\partial q}$ is the contact Jacobian, and λ is the constraint force arising from the constraint (2.2). This can be expressed in form of a DAE of differential index 3, which can be reduced to index 1 by differentiating (2.2) twice. The resulting DAE is then given by

$$\dot{q} = v, \tag{2.4a}$$

$$\dot{v} = a, \tag{2.4b}$$

$$\begin{pmatrix} H(q, p) & G(q, p)^T \\ G(q, p) & 0 \end{pmatrix} \begin{pmatrix} a \\ -\lambda \end{pmatrix} = \begin{pmatrix} \tau - C(q, v, p) \\ \gamma(q, v, p) \end{pmatrix}, \tag{2.4c}$$

with $\gamma(q, v, p) = -\left(\left(\frac{\partial G}{\partial q}\right) v\right) v$ being the contact Hessian. If we ensure

$$g(q, p) = 0\,, \quad \frac{d}{dt} g(q, p) = 0 \tag{2.5}$$

for all time instances, the analytical solutions of (2.3) and (2.4) coincide. In fact, because of (2.4c), it is sufficient to satisfy these constraints at the initial time point.

When the external contact changes, e.g. when a foot hits the ground after swinging freely before, a collision impact might occur, which transfers the velocities before the impact, $v(t^-)$, to the velocities after the impact, $v(t^+)$. Throughout our work, we assume this impact to be perfectly inelastic, which has been proven to give a quite good approximation of reality for humans walking on ground with standard feet and soles [34]. The discontinuities in the velocities are then given by

$$\begin{pmatrix} H(q, p) & G(q, p)^T \\ G(q, p) & 0 \end{pmatrix} \begin{pmatrix} v(t^+) \\ -\Lambda \end{pmatrix} = \begin{pmatrix} H(q, p)v(t^-) \\ 0 \end{pmatrix}, \tag{2.6}$$

where $G(\cdot)$ is the contact Jacobian belonging to the constraints acting *after* the impact, and Λ is the contact impulse.

Setting up the equations of motion for a given MBS by hand is cumbersome and error-prone, and practically impossible for half-way complex models. Therefore,

the use of software libraries is advisable. Different formalisms and corresponding computational tools exist, all having their pros and cons. In this project, we decided to use the formalism of Featherstone [13], who proposes to use 6D spatial vectors instead of two 3D vectors, describing the linear and angular aspects of a rigid body's motion. Using this notation, algorithms can be stated more concise, which makes them easier to implement and reduces one possible source of error. A bunch of efficient algorithms, among these the famous *Articulated Body Algorithm*, expressed in the 6D notation is provided in [13], and the software package RBDL [14] implements these algorithms and many more MBS features. This package has proven its efficiency in many optimal control applications, cf. e.g. [16, 30, 31], and is thus well-suited for our purposes.

2.2 Detailed Submodules

In view of the complexity of the human body, it is important to set up a *task-adequate* model, and to focus mainly on the most relevant parts of the body. As we are interested in walking and gait patterns, foot–ground contact is of major concern. Furthermore, since muscle weakness and spasticity are frequently occurring issues in CP, and many medical treatments affect the patient's muscles, a sufficiently detailed muscle model should be included to reflect the disease and possible treatments.

2.2.1 Foot Modeling and Ground Contact

For the proper modeling of CP patient's gait, which may be significantly different from that of a healthy person, it is of central importance to have a *sufficiently* detailed foot model, which reproduces the physics of the ground contact qualitatively and quantitatively correctly because the foot forms the only boundary between the MBS and the ground while walking. Possible models range from finite-element approximations of continuum mechanics models, e.g. Halloran et al. [17], to surrogate models like (3D-) volumetric contact models, e.g. Brown et al. [6], and rigid foot–ground contact models, e.g. Felis at al. [16], Ren et al. [35]. In view of the incorporation into complex bilevel optimization problems, it is also necessary to take the computational complexity of the involved model into account. Therefore, we decided to use point contacts in our first approach where we capture the typical gait of a CP patient with club feet, or *pes equinus*, where the heel never touches the ground while walking. This choice is underpinned by a recent publication of Kleesattel et al. [24] where they used also a foot–ground contact model based on point contacts for sprinting motions supporting only the forefoot.

2.2.2 Muscle Modeling

The generalized forces τ in (2.4)—besides the included gravitational forces—summarize the effect of all involved muscle–tendon complexes as well as passive forces like damping or reset forces caused by stiffness in any sense. The more detailed we model the muscle itself, the better we understand how the resulting generalized forces are generated, and in particular, how muscle specific pathologies influence the gait. Simply speaking, the generalized forces are written as a function of other influencing quantities u,

$$\tau = \tau(u, p)$$

which lie on a deeper level in the formation of the resulting forces, and of course of parameters p. Whatever influencing quantity is used, it will serve as control in the OCP describing the human gait, see Sect. 3. An introduction to muscle modeling can be found in [32] and the references therein.

2.3 Biomechanical Model for CP Patients

To capture the main characteristics of a CP patient's gait, we consider a model with 14 segments including the pelvis as basis segment with six degrees of freedom (three for the translational and three for the rotational motion), see Fig. 1. Furthermore, the upper body consists of a middle and an upper trunk, a head, and two lower and upper arms connected by joints. We decided these joints to be fixed in an average position because the motion capture data from the Heidelberg MotionLab [41] only include motions of the lower body. On the contrary, the rotational joints of the lower

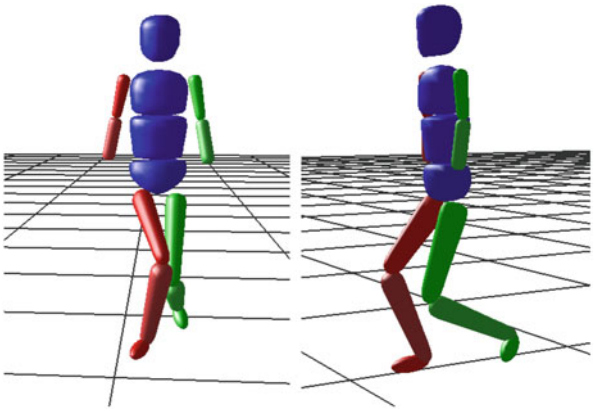

Fig. 1 Biomechanical model for CP patients with 14 segments

body are mandatory for any basic MBS model of a CP patient due to the truly three dimensional pathological gait. Therefore, not only the hip joints connecting the thighs with the pelvis but also the knee and ankle joints connecting the thighs with the shanks and the shanks with the feet have three degrees of freedom. This results in a biomechanical model with 24 degrees of freedom, hence 48 states $x = (q, v)$, and 18 torques as controls $u = \tau$ in the OCP describing the human gait, see Sect. 3. In our first approach, the foot contact is modeled as a totally inelastic collision at a point located at the toe. Together with the three dimensional ankle joints the typical pathological gait of CP patients with club feet, or pes equinus, can be captured. Concerning this deformity for the foot, the heel never touches the ground while walking.

As we are interested in a personalized model the biomechanical model for a specific CP patient includes the inertia of each segment, which is estimated based on anthropometric data for segment locations of center of mass, segment masses, and radii of gyration mainly taken from De Leva [11] depending on gender, height and total mass of this subject, and the segment lengths computed from motion capture data of this patient from the Heidelberg MotionLab.

The prototypical MBS implementation of the basic CP model (Fig. 1) using the RBDL framework allows a flexible incorporation of detailed submodules such as foot modeling and ground contact, and muscle modeling as described in Sect. 2.2.

3 Modeling the Human Gait

A common assumption is that the human gait is the result of an autonomous decision of a person on the controls generating torques and forces, depending on a combination of optimization criteria (such as stability, energy efficiency, or comfort), state and control inequality constraints, and given—complex—physiological parameters. Therefore, optimization serves as a guiding principle of bipedal locomotion of humans [34], and hence it can be mathematically formulated as an OCP where the dynamics of biomechanical MBS are described by the nonlinear differential algebraic equation systems (2.4) and (2.6). Optimization based generation of human walking and running has been studied, e.g., by Ackermann and van den Bogert [1], Felis et al. [15, 16], Schultz and Mombaur [38], Kleesattel et al. [25], Hu [21], Suleiman et al. [40].

3.1 A Multi-Phase Optimal Control Approach

The human gait cycle can be divided into different phases or model stages. Hence, we formulate a multi-phase OCP which generates the actuator torques and incorporates a switched, phase-wise defined differential equation system for one gait cycle of a patient with CP. Assuming a ground contact as in the CP gait model

described in Sect. 2.3 with one contact point at each foot double support phases, single support phases of the right foot and single support phases of the left foot arise. Since double support phases are short compared with single support phases, we consider three single support phases and the transition phases in-between on the time horizon $\mathcal{T} = [t_0, t_f]$ with the time grid $t_0 < t_1 = t_2 < t_3 = t_4 < t_5 = t_f$ and the model stage indices $ms \in \{0, 1, 2, 3, 4\}$: a first single support phase right, $ms = 0$, on the interval $[t_0, t_1]$, transition phase left, $ms = 1$, at time point $t_1 = t_2$, single support phase left, $ms = 2$, on $[t_2, t_3]$, transition phase right, $ms = 3$, at $t_3 = t_4$ and second single support phase right, $ms = 4$, on $[t_4, t_f]$. In each single support phase, $ms \in \{0, 2, 4\}$, the dynamics of the MBS are formulated by the differential algebraic equation system (2.4). At the end of a single support phase of one foot the respective other foot touches the ground and an impact occurs where we have to cope with discontinuities of the velocities. These velocities at the transition phases, $ms \in \{1, 3\}$, can be computed by (2.6). For ease of notation, we write the dynamics in compact notation

$$F(t, x(t), \dot{x}(t), u(t), p, \sigma(x(t), p)) = 0 \quad a.e.\ t \in \mathcal{T}, \tag{3.1}$$

with state variables $x = (q, v)$, controls $u = \tau$, system parameters p, and switching functions $\sigma(x(t), p)$. Possible jumps of the states on impact are expressed by

$$x(t_s^+) = \Delta\left(x(t_s^-), p\right) \quad \text{if } t_s \in \Sigma,$$

where Σ denotes the set of time instances where a component of $\sigma(\cdot)$ changes its sign. Let us note, in this multi-phase optimal control approach the switching structure is known a priori. Furthermore, constraints on the MBS dynamics described by (3.1) have to be considered. At initial time t_0 we have to enforce that the right foot touches the ground $g(q, p) = 0$ and that $\frac{d}{dt} g(q, p) = 0$ holds, see Sect. 2.1. During single support phases, $ms \in \{0, 2, 4\}$, it has to be ensured that the swinging foot does not penetrate the ground. When the foot touches the ground and an impact occurs, i.e. $ms \in \{1, 3\}$, the first condition requires the z-component of the contact point of the foot entering the contact phase to be 0, and the second condition needs the z-component of the velocities in the global frame of the same contact point to be negative. Additional constraints have to be taken into account to describe the initial and terminal orientation and position of the pelvis and the orientation of the joints. Furthermore, simple bounds on the variables x and the controls u have to be considered. All constraints and stage transition conditions can be summarized in nonlinear mixed control-state constraints

$$0 \le c(x(t), u(t), p) \quad a.e.\ t \in \mathcal{T},$$

and nonlinear equality and inequality multi-point constraints

$$0 = r^{eq}(x(t_0), \ldots, x(t_f), p),$$
$$0 \le r^{ieq}(x(t_0), \ldots, x(t_f), p).$$

In the optimal control framework, which describes human locomotion, we assume that the objective function can be formulated as a linear combination of different optimality criteria $\phi_k(\cdot)$. For the CP gait we consider four different optimization criteria

$$\phi(x(t_f), u, p, \alpha) := \sum_{k=0}^{3} \alpha_k \phi_k(x(t_f), u, p), \tag{3.2}$$

including stability, energy consumption, abduction/adduction in the hip, and internal/external rotation in the joints of the lower body, where the third and fourth criterion can be interpreted as convenience criteria. The weights α are not known in advance and, therefore, have to be determined by consideration of motion capture data of a CP patient's gait in an inverse OCP, see Sect. 4.1. For the sake of a clearer presentation of the OCP (3.3) for the CP gait the objective function in (3.2) does not include Lagrange terms. However, in the following we define the criteria for one model stage, $ms \in \{0, 2, 4\}$, in Lagrange form which can be easily transformed in a Mayer term by introducing an additional ordinary differential equation (ODE).

The stability criterion $\phi_0(\cdot)$ minimizes the distance between the y-component of the hip joint position of the supported leg P_h^s and the y-component of the corresponding contact point of the foot P_c^s in global coordinates in each single support phase

$$\phi_0 := \int_{t_{ms}}^{t_{ms+1}} \left([P_h^s(q(t))]_y - [P_c^s(q(t))]_y \right)^2 dt.$$

The second criterion $\phi_1(\cdot)$ minimizes the overall energy consumption on each single support phase. It is defined as the squared and summed up integrals over all torques u_j, $j \in \mathcal{J}$

$$\phi_1 := \int_{t_{ms}}^{t_{ms+1}} \sum_{j \in \mathcal{J}} \left(u_j(t) \right)^2 dt.$$

Because CP patients often have an adduction deformity in the hip joints which is often very painful, we consider an abduction/adduction criterion $\phi_2(\cdot)$ as the third optimization criterion

$$\phi_2 := \int_{t_{ms}}^{t_{ms+1}} \sum_{j \in \mathcal{J}_1} \left(u_j(t) \right)^2 dt,$$

where we sum up and integrate over the squared torques u_j, $j \in \mathcal{J}_1 \subset \mathcal{J}$, which are related to abduction and adduction in the hip joints. Internal and external rotations in the joints are represented in the last criterion $\phi_3(\cdot)$ by minimizing the squared and summed up integrals over the corresponding torques u_j, $j \in \mathcal{J}_2 \subset \mathcal{J}$, in each

single support phase

$$\phi_3 := \int_{t_{ms}}^{t_{ms+1}} \sum_{j \in \mathcal{J}_2} \left(u_j(t)\right)^2 dt.$$

With the four optimization criteria, we have a suitable set which can be extended in the future in close collaboration with our medical partner. The OCP describing the pathological gait of CP patients can now mathematically be formulated as

$$\begin{aligned}
\min_{x,u} \quad & \phi(x(t_f), u, p, \alpha) := \sum_{k=0}^{3} \alpha_k \phi_k(x(t_f), u, p) \\
\text{s.t.} \quad & 0 = F(t, x(t), \dot{x}(t), u(t), p, \sigma(x(t), p)) && a.e.\ t \in \mathcal{T}, && \text{(3.3a)} \\
& x(t_s^+) = \Delta\left(x(t_s^-), p\right) && \text{if } t_s \in \Sigma, && \text{(3.3b)} \\
& 0 \le c(x(t), u(t), p) && a.e.\ t \in \mathcal{T}, && \text{(3.3c)} \\
& 0 = r^{eq}(x(t_0), \ldots, x(t_f), p), && && \text{(3.3d)} \\
& 0 \le r^{ieq}(x(t_0), \ldots, x(t_f), p). && && \text{(3.3e)}
\end{aligned}$$

In Fig. 2 one solution of such an OCP is visualized with the weights $\alpha = (0.999999, 10^{-6}, 0.0, 0.0)$ and some fixed initial and terminal states x using motion capture data of a CP patient's gait. This result was achieved by applying the "direct multiple shooting" method for discretization [8] and an efficient structure exploiting sequential quadratic programming implementation in the software package MUSCOD-II [29].

Augmenting the CP model by a more detailed submodel for the foot–ground contact, more phases have to be taken into account which consequently results in a more complex switching structure of the system. This multi-phase optimal control approach then can still be appropriate at Level 3A considering an inverse OCP where motion capture data is available and hence the structure of the different phases is known. But at least in the case of treatment planning at Level 3B this approach is not sufficient and an optimal control approach which allows an altered switching structure is needed, see Sect. 3.2.

Fig. 2 Visualization of an OCP solution with the weights $\alpha = (0.999999, 1E-6, 0.0, 0.0)$ and some fixed initial and terminal states x using motion capture data of a CP patient's gait

3.2 A Mixed-Integer Optimal Control Approach

The multi-phase approach presented in the last section has the great advantage, that, assuming a fixed switching structure resp. order of phases, the use of a switch-detecting integrator can be avoided. Instead, the stage durations are optimized as well, which is a much easier task. However, there are situations, in which the switching structure is not known a priori, especially when it comes to treatment planning and the prediction of a resulting gait after surgery in CP.

For instance, a common syndrome of CP is pes equinus, a deformity concerning the foot, which has the consequence, that the heel never touches the ground while walking. This behavior is non-desirable, and its remediation is a frequent goal of medical interventions. Despite the impressive progress in surgeries in CP it is a priori not clear, whether the undesired behavior will change or not, and consequently, if the foot–ground contact and in particular the involved contact points will change as a result of a treatment. Different combinations of contact points while walking in turn would imply a different order of model phases and therefore an altered switching structure.

One possibility to overcome this issue is to consider a free-phase formulation, which optimizes the order of model phases along with differential states and controls. For the sake of a handy notation, the model dynamics are assumed to be given by autonomous ODEs. The time horizon $\mathcal{T} = [t_0, t_f]$ is assumed to be fixed without loss of generality. We enumerate all possible model phases and introduce phase-indicator functions $\omega_j : \mathcal{T} \to \{0, 1\}$ with the property

$$\omega_j(t) = 1 \iff \text{model is in phase } j.$$

A change of model phases from a phase j_1 to a phase $j_2 \neq j_1$ is then denoted by $j_1 \to_\omega j_2$, and the set of switching points, which we assume to be finite, by $\mathcal{S}(\omega)$. Whenever the model phase changes, jumps in the differential states are possible, and reflected by the jump-functions Δ_{j_1,j_2}, mapping the differential states before the change to the differential states after the change. The free-phase formulation of our problem can then be stated in form of a mixed-integer optimal control problem (MIOCP)

$$\begin{aligned}
\min_{x,u,\omega} \quad & \phi(x(t_f)) \\
\text{s.t.} \quad & \dot{x}(t) = f^j(x(t), u(t), p) && \text{if } \omega_j(t) = 1 \; a.e.\, t \in \mathcal{T}, && \text{(3.4a)} \\
& x(t_s^+) = \Delta_{j_1,j_2}\left(x(t_s^-), p\right) && \text{if } j_1 \to_\omega j_2 \text{ at } t_s \in \mathcal{S}(\omega), && \text{(3.4b)} \\
& 0 \leq c^j(x(t), u(t), p) && \text{if } \omega_j(t) = 1 \; a.e.\, t \in \mathcal{T}, && \text{(3.4c)} \\
& 0 \leq d(x(t), u(t), p) && a.e.\, t \in \mathcal{T}, && \text{(3.4d)} \\
& 0 \leq r(x(t_0), x(t_f), p), && && \text{(3.4e)}
\end{aligned}$$

where the constraints (3.4b)–(3.4e) are written as multi-dimensional inequality constraints, thus including equality constraints as well. The mode-specific constraints are encoded in $c^j(\cdot)$, and can be used to characterize the model phases.

Now the advantage of the problem formulation (3.4) is, that it optimizes the switching strategy, i.e. the order of phases, as well, and therefore it is theoretically suited for our purposes. Unfortunately, solving MIOCPs is much harder than solving multi-stage OCPs, especially if jumps in the differential states are involved.

In order to tackle problem (3.4), we developed a novel strategy for the solution of MIOCPs with switches, state jumps, and also switching costs. We consider a discretized version of the problem, use partial outer convexification [36] for the ODE, and further also convexify the jump-condition (3.4b), which raises the need for a regularization term, which can be interpreted as the penalized number of switches. The continuous problem is then tackled by solving a sequence of discretized problems, where we refine the grid successively. Details can be found in [22], and an application of the approach to walking motions is described in [23].

4 Two Bilevel Problems for Diagnosis and Therapy Design of Cerebral Palsy

In this section, we present two bilevel OCPs—one for diagnosis and one for therapy design of CP.

4.1 An Inverse Optimal Control Problem for Diagnosis of Cerebral Palsy

One of the aims is to provide a non-invasive diagnosis tool by fitting system parameters p, weights α, and solutions of the gait model of a CP patient to measurement data η from the Heidelberg MotionLab. Mathematically, this results in an inverse OCP of the form

$$
\begin{aligned}
\min_{x,u,p,\alpha} \quad & \Phi[x,u,p;\eta] := \sum\nolimits_{ij} \frac{1}{2}\left(\frac{\eta_{ij} - M_j(t_i^m; x(t^m), u(t^m), p)}{\gamma_{ij}}\right)^2 \\
\text{s.t.} \quad x, u \quad & \text{solution of} \\
\min_{x,u} \quad & \phi(x(t_f), p, \alpha) := \sum\nolimits_k \alpha_k \phi_k(x(t_f), p) \\
\text{s.t.} \quad & 0 = F(t, x(t), \dot{x}(t), u(t), p, \sigma(x(t), p)) && a.e.\ t \in \mathcal{T}, \\
& x(t_s^+) = \Delta\left(x(t_s^-), p\right) && \text{if } t_s \in \Sigma, \\
& 0 \le c(x(t), u(t), p) && a.e.\ t \in \mathcal{T},
\end{aligned}
$$

$$
\begin{aligned}
0 &= r^{eq}(x(t_0), \ldots, x(t_f), p), \\
0 &\leq r^{ieq}(x(t_0), \ldots, x(t_f), p), \\
1 &= \sum\nolimits_k \alpha_k, \qquad 0 \leq \alpha_k \ \forall k, \\
b^l &\leq p \leq b^u,
\end{aligned}
$$

where we consider a parameter estimation problem on the upper level constrained by the lower level OCP (3.3) modeling the human gait described in Sect. 3. The standard choice for the cost function is the l_2 formulation which describes the deviation of the model response $M(\cdot)$ depending on the state and control vectors at measurement times t^m from the measurements η with standard deviation γ. However, l_1 and Huber formulations as in Kostina et al. [5, 27] can be used to take into account possible outliers in the data. The covariance analysis of the parameter estimates can be performed based on methods by Bock et al. [7, 9] and Kostina et al. [28].

In view of the complexity of problems such as the inverse OCP, the solution methods of choice are simultaneous approaches. These methods mainly rely on the reformulation of the lower level OCP of Level 2 by its optimality conditions, which serve as a constraint to the upper level parameter estimation problem. This means that the lower level problem is solved together with the upper level optimization problem. One possibility is to reformulate the lower level OCP, after discretization, by its Karush-Kuhn-Tucker (KKT) optimality conditions. Because inequality constraints are included in the reformulated inverse OCP, it results in the generation of a mathematical program with complementarity constraints (MPCC). Current research in the development of numerical methods for solving the arising MPCC in the context of human locomotion can be found in the work of Albrecht et al. [2] where different regularization and lifting strategies are compared for handling the complementarity constraints in a simple dynamic model moving from a start to a goal position without paying attention to the complex dynamical problem of taking individual steps. Furthermore, Hatz et al. [20] proposed a "direct all-at-once" approach which was successfully applied in a first analysis of a CP patient's gait [18].

Due to the potentially lower computational effort, we follow the "direct all-at-once" approach which can be sketched as follows: First we parametrize and discretize the continuous problem, then replace the discretized and parametrized lower level OCP by its necessary optimality conditions, and solve the resulting nonlinear programming problem (NLP) with a Gauss–Newton-type method resp. a Newton-type method.

For discretization the infinite dimensional controls u are locally approximated by basis functions with finite support on a suitable time grid $t_0 = \tau_0 < \tau_1 < \ldots < \tau_m = t_f$. The discretized control variables are denoted by $w_i,\ i = 0, \ldots, m-1$. Furthermore, we assume that the mixed control-state constraints are only satisfied at the time points $\tau_0, \ldots, \tau_m$. The states x are parametrized based on the "direct multiple shooting" method [8], $s_i,\ i = 0, \ldots, m$ on the same time grid. This results

in the large-scale but structured discrete lower level OCP

$$\begin{aligned}\min_{y} \quad & \phi(s_m, p, \alpha) := \sum\nolimits_k \alpha_k \phi_k(s_m, p) \\ \text{s.t.} \quad & 0 = x(\tau_{i+1}; \tau_i, s_i, w_i, p) - s_{i+1}, \forall i = 0, \ldots, m-1, \\ & 0 \le c(y, p), \\ & 0 = r^{eq}(s_0, \ldots, s_m, p), \\ & 0 \le r^{ieq}(s_0, \ldots, s_m, p),\end{aligned}$$

with $y = (s_0, \ldots, s_m, w_0, \ldots, w_{m-1})$. As a second step, the lower level OCP is replaced by its KKT optimality conditions

$$\begin{aligned} & 0 = x(\tau_{i+1}; \tau_i, s_i, w_i, p) - s_{i+1}, \forall i = 0, \ldots, m-1, \\ & 0 \le c(y, p), \\ & 0 = r^{eq}(s_0, \ldots, s_m, p), \\ & 0 \le r^{ieq}(s_0, \ldots, s_m, p), \\ & 0 = \nabla_y \mathcal{L}(y, p, \alpha, \lambda, \mu), \\ & 0 \le \mu, \\ & 0 = \mu^T c(y, p),\end{aligned}$$

where the Lagrangian is given by $\mathcal{L}(y, p, \alpha, \lambda, \mu)$ and λ, μ are the Lagrange multipliers for equalities resp. inequalities.

The upper level problem is solved using the "all-at-once" approach with a Gauss–Newton-type method for the inverse OCP with the discretized lower level OCP as constraint.

$$\min_{y,p,\alpha,\lambda,\mu} \quad \Phi[y, p; \eta] \tag{4.1a}$$

$$\text{s.t.} \quad 0 = x(t_{i+1}; t_i, s_i, w_i, p) - s_{i+1}, \forall i = 0, \ldots, m-1, \tag{4.1b}$$

$$0 \le c(y, p), \tag{4.1c}$$

$$0 = r^{eq}(s_0, \ldots, s_m, p), \tag{4.1d}$$

$$0 \le r^{ieq}(s_0, \ldots, s_m, p), \tag{4.1e}$$

$$0 = \nabla_y \mathcal{L}(y, p, \alpha, \lambda, \mu), \tag{4.1f}$$

$$0 \le \mu, \tag{4.1g}$$

$$0 = \mu^T c(y, p), \tag{4.1h}$$

$$1 = \sum\nolimits_k \alpha_k,\ 0 \leq \alpha_k\ \forall k, \tag{4.1i}$$

$$b^l \leq p \leq b^u. \tag{4.1j}$$

The inverse OCP (4.1) is indeed an MPCC. The complementarity constraints $0 = \mu^T c(y, p), \mu \geq 0, c(y, p) \geq 0$ of (4.1) fail to satisfy standard constraint qualifications at any feasible point [10]. This leads to degenerate quadratic programs such that appropriate strategies for handling the complementarity constraints have to be considered. To tackle this difficulty, the lifting approach based on [19] has been implemented.

4.2 A Robustified Optimal Control Problem for Therapy Design of Cerebral Palsy

In this subsection, we describe how to design a computational testing environment for ex ante evaluation and assessment of potential surgery plans. Recall the parameterized optimal control models for the human gait described in Sect. 3. We set up problems in a way, that a medical treatment, being a change of the physiology, can be reflected by a change of parameters p in the model. This way, a change of parameters would possibly lead to a different solution of the OCP and thus cause a different gait, which could then be assessed by the medical doctors in charge. Note though, that in this scenario we need to make assumptions on the objective function, which in fact might also depend on p in an unknown fashion.

One major challenge is the adequate modeling of a considered treatment as a change of parameters in the model, reflecting all significant changes, but neglecting that ones, which are of less importance, this way keeping the mathematical model tractable. There exists a large variety of surgeries applied to CP patients in order to improve the gait pattern. These interventions comprise osseous procedures like derotational osteotomies, where tibia and/or femur are rotated in order to remediate torsional deformities, but also soft tissue procedures like tendon transfers or tendon/muscle lengthenings. A list of the main surgical treatments in CP for gait patterns can be found in [3]. The used gait model should be chosen tailored to the considered medical procedure. If we are for instance interested in muscle or tendon lengthenings, the gait model should contain a muscle model appropriate to that task, see Sect. 2.2.2.

Another major issue we need to take care of is the robustness of the mathematical prediction model against uncertainties. Such uncertainties might originate from many sources, like modeling errors, inaccurately performed treatments, but also perturbed muscle activity by the patients itself, caused by the neurological disorder resulting from the disease. Hence, we robustify the parameterized OCP modeling the gait against possible perturbations in parameters *as well as* controls, and the solution of the resulting problem then describes the best possible gait under consideration

of these perturbations. This way we can ensure, that a planned medical treatment still achieves good results in occurrence of expectable disturbances, and thus is reasonable.

As in our case, the decision for a treatment has great impact on the patients' lives, we do not want to gamble, and therefore refrain from using probabilistic robust optimization models, but take a conservative approach. The robustified problem then reads as

$$
\begin{aligned}
\min_{x,u} \max_{(\delta p,\delta u)\in\mathcal{U}} \quad & \phi(x(t_f)) \\
\text{s.t.} \quad & 0 = F(x(t), \dot{x}(t), u(t)+\delta u(t), p+\delta p, \sigma(x(t))) \quad a.e.\, t \in \mathcal{T}, \\
& x(t_s^+) = \Delta\left(x(t_s^-), p+\delta p\right) \qquad \text{if } t_s \in \Sigma, \\
& 0 \le c(x(t), u(t)+\delta u(t), p+\delta p) \qquad a.e.\, t \in \mathcal{T}, \\
& 0 \le r(x(t_0), x(t_f), p+\delta p),
\end{aligned}
$$

where Σ denotes the set of time instances, where a component of $\sigma(\cdot)$ changes its sign, and $\mathcal{U}$ is the uncertainty set, which has to be chosen carefully and in close cooperation with the medical doctors.

In order to solve this problem, we develop a linear approximation, assuming the perturbations are small enough, as in [12, 26]. In case this assumption is not valid, also second order approximations as in [39] need to be taken into account.

5 Conclusions and Outlook

We have given some background information on the disease CP to which we contribute and described the involved medical challenges in detail to be able to transfer them to adequate mathematical tasks. A detailed description of the rigid MBS model for the human body capturing the main characteristics of CP patients and the corresponding dynamics have been given. Assuming that human locomotion can be modeled as the solution of an OCP, we have presented two approaches describing the gait itself: a multi-phase OCP with a fixed sequence of single supporting and double supporting phases tailored to recorded gaits of CP patients, and an MIOCP appropriate for intervention planning with the great advantage that the switching structure does not have to be known a priori. These optimal control formulations can serve as the lower level of two bilevel problems: an inverse OCP for diagnosis and a robustified OCP for intervention planning. For these mathematical and numerical frameworks, detailed descriptions and adequate solution approaches are proposed. The described bilevel OCPs are a first step in supporting the physicians in proper diagnosis and improving treatment planning and therapy measures for patients with CP. An augmentation of the present MBS model

for CP patients, e.g., by a more detailed foot model or a muscle model, affects all levels of the bilevel OCPs with the need of exploiting the resulting structure and thus will be pursued in future research.

Acknowledgments All authors acknowledge funding by Deutsche Forschungsgemeinschaft through Priority Programme 1962 “Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization”.

References

1. Ackermann, M., Van den Bogert, A. J. (2010). Optimality principles for model-based prediction of human gait. Journal of Biomechanics, 43(6), 1055–1060.
2. Albrecht, S., Ulbrich, M. (2017). Mathematical programs with complementarity constraints in the context of inverse optimal control for locomotion. Optimization Methods and Software, 32(4), 670–698.
3. Armand, S., Decoulon, G., Bonnefoy-Mazure, A. (2016). Gait analysis in children with cerebral palsy. EFORT Open Reviews, 1(12), 448–460.
4. Baker, R., McGinley, J. L., Schwartz, M. H., Beynon, S., Rozumalski, A., Graham, H. K., Tirosh, O. (2009). The Gait Profile Score and Movement Analysis Profile. Gait & Posture, 30(3), 265–269.
5. Binder, T., Kostina, E. (2013). Gauss–Newton Methods for Robust Parameter Estimation. In Model Based Parameter Estimation: Theory and Applications, 55–87. Springer, Berlin, Heidelberg.
6. Brown, P., McPhee, J. (2018). A 3D ellipsoidal volumetric foot–ground contact model for forward dynamics. Multibody System Dynamics, 42(4), 447–467.
7. Bock, H. G. (1987). Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen, volume 183 of Bonner Mathematische Schriften. Rheinische Friedrich–Wilhelms–Universität Bonn, Bonn.
8. Bock, H. G., Plitt, K. J. (1984). A Multiple Shooting Algorithm for Direct Solution of Optimal Control Problems. In Proceedings of the 9th IFAC World Congress, 242–247, Budapest. Pergamon Press.
9. Bock, H. G., Kostina, E., Kostyukova, O. (2007). Covariance Matrices for Parameter Estimates of Constrained Parameter Estimation Problems. SIAM Journal on Matrix Analysis and Applications, 29(2), 626–642.
10. Chen, Y., Florian, M. (1995). The nonlinear bilevel programming problem: formulations, regularity and optimality conditions. Optimization, 32, 193–209.
11. De Leva, P. (1996). Adjustments to Zatsiorsky-Seluyanov’s segment inertia parameters. Journal of Biomechanics, 29(9), 1223–1230.
12. Diehl, M., Bock, H. G., Kostina, E. (2006). An approximation technique for robust nonlinear optimization. Mathematical Programming, 107(1–2), 213–230.
13. Featherstone, R. (2008). Rigid Body Dynamics Algorithms. Springer.
14. Felis, M. L. (2017). RBDL: an efficient rigid-body dynamics library using recursive algorithms. Autonomous Robots, 41(2), 495–511.
15. Felis, M. L. (2015). Modeling Emotional Aspects in Human Locomotion. Doctoral dissertation, Heidelberg University.
16. Felis, M. L., Mombaur, K. (2016). Synthesis of Full-Body 3-D Human Gait using Optimal Control Methods. In 2016 IEEE International Conference on Robotics and Automation (ICRA), 1560–1566. IEEE.

17. Halloran, J. P., Ackermann, M., Erdemir, A., Van den Bogert, A. J. (2010). Concurrent musculoskeletal dynamics and finite element analysis predicts altered gait patterns to reduce foot tissue loading. Journal of Biomechanics 43(14), 2810–2815.
18. Hatz, K. (2014). Efficient numerical methods for hierarchical dynamic optimization with application to cerebral palsy gait modeling. Doctoral dissertation, Heidelberg University.
19. Hatz, K., Leyffer, S., Schlöder, J. P., Bock, H. G. (2013). Regularizing Bilevel Nonlinear Programs by Lifting. Preprint ANL/MCS-P4076-0613, Argonne National Laboratory, Mathematics and Computer Science Division.
20. Hatz, K., Schlöder, J. P., Bock, H. G. (2012). Estimating Parameters in Optimal Control Problems. SIAM Journal on Scientific Computing, 34(3), A1707–A1728.
21. Hu, Y. (2017). The role of compliance in humans and humanoid robots locomotion. Doctoral dissertation, Heidelberg University.
22. Kirches, C., Kostina, E., Meyer, A., Schlöder, M. (2018). Numerical Solution of Optimal Control Problems with Switches, Switching Costs and Jumps. Optimization Online Preprint, 6888.
23. C. Kirches, E. Kostina, A. Meyer, M. Schlöder (2019). Generation of Optimal Walking-Like Motions Using Dynamic Models with Switches, Switch Costs, and State Jumps. In 2019 IEEE 58th Conference on Decision and Control (CDC), 1538–1543.
24. Kleesattel, A. L., Mombaur, K. (2018). Inverse Optimal Control Based Enhancement of Sprinting Motion Analysis With and Without Running–Specific Prostheses. In 2018 7th IEEE International Conference on Biomechanical Robotics and Biomechatronics (Biorob), 556–562. IEEE.
25. Kleesattel, A. L., Clever, D., Funken, J., Potthast, W., Mombaur, K. (2017). Modeling and optimal control of able-bodied and unilateral amputee running. In ISBS Proceedings Archive, 35(1), Article 20.
26. Körkel, S., Kostina, E., Bock, H. G., Schlöder, J. P. (2004). Numerical Methods for Optimal Control Problems in Design of Robust Optimal Experiments for Nonlinear Dynamic Processes. Optimization Methods and Software, 19(3–4), 327–338.
27. Kostina, E. (2004). Robust Parameter Estimation in Dynamic Systems. Optimization and Engineering, 5(4), 461–484.
28. Kostina, E., Nattermann, M. (2015). Second-order sensitivity analysis of parameter estimation problems. International Journal for Uncertainty Quantification, 5(3), 209–231.
29. Leineweber, D., Schäfer, A., Bock, H., Schlöder, J. P. (2003). An efficient multiple shooting based reduced SQP strategy for large-scale dynamic process optimization. Part II: Software Aspects and Applications. Computers and Chemical Engineering, 27(2), 167–174.
30. Manns, P., Sreenivasa, M., Millard, M., Mombaur, K. (2017). Motion Optimization and Parameter Identification for a Human and Lower Back Exoskeleton Model. IEEE Robotics and Automation Letters, 2(3), 1564–1570.
31. Millard, M., Sreenivasa, M., Mombaur, K. (2017). Predicting the Motions and Forces of Wearable Robotic Systems Using Optimal Control. Frontiers in Robotics and AI, 4, Article 41.
32. Millard, M., Uchida, T., Seth, A., Delp, S. L. (2013). Flexing Computational Muscle: Modeling and Simulation of Musculotendon Dynamics. Journal of Biomechanical Engineering, 135(2), 021005.
33. Mombaur, K. (2001). Stability Optimization of Open-Loop Controlled Walking Robots. Doctoral dissertation, Heidelberg University.
34. Mombaur, K., Sharbafi, M., Seyfarth, A. (2017). Optimization as Guiding Principle of Locomotion. In Bioinspired Legged Locomotion: Models, Concepts, Control and Applications, 164–195. Elsevier Butterworth–Heinemann.
35. Ren, L., Howard, D., Ren, D., Nester, C. Tilan, L. (2010). A generic analytical foot rollover model for predicting translational ankle kinematics in gait simulation studies. Journal of Biomechanics, 43(2), 194–202.
36. Sager, S. (2005). Numerical methods for mixed–integer optimal control problems. Der andere Verlag, Tönning.

37. Sartori, M., Fernandez, J. W., Modenese, L., Carty, C. P., Barber, L. A., Oberhofer, K., Zhang, J., Handsfield, G. G., Stott, N. S., Besier, T. F., Farina, D., Lloyd, D. G. (2017). Toward modeling locomotion using electromyography-informed 3D models: application to cerebral palsy. Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 9(2), e1368.
38. Schultz, G., Mombaur, K. (2010). Modeling and Optimal Control of Human-Like Running. IEEE/ASME Transactions on Mechatronics, 15(5), 783–792.
39. Sichau, A., Ulbrich, S. (2012). A Second Order Approximation Technique for Robst Shape Optimization. Applied Mechanics and Materials, 104, 13–22.
40. Suleiman, W., Yoshida, E., Laumond, J.-P., Monin, A. (2007). On Humanoid Motion Optimization. In 2007 7th IEEE-RAS International Conference on Humanoid Robots, 180–187.
41. Wolf, S. (2019). Heidelberg MotionLab. Heidelberg University Hospital, Department of Orthopaedics and Trauma Surgery, www.heidel-motionlab.de

ROM-Based Multiobjective Optimization of Elliptic PDEs via Numerical Continuation

Stefan Banholzer, Bennet Gebken, Michael Dellnitz, Sebastian Peitz, and Stefan Volkwein

Abstract Multiobjective optimization plays an increasingly important role in modern applications, where several objectives are often of equal importance. The task in multiobjective optimization and multiobjective optimal control is therefore to compute the set of optimal compromises (the *Pareto set*) between the conflicting objectives. Since the Pareto set generally consists of an infinite number of solutions, the computational effort can quickly become challenging which is particularly problematic when the objectives are costly to evaluate as is the case for models governed by partial differential equations (PDEs). To decrease the numerical effort to an affordable amount, surrogate models can be used to replace the expensive PDE evaluations. Existing multiobjective optimization methods using model reduction are limited either to low parameter dimensions or to few (ideally two) objectives. In this chapter, we present a combination of the reduced basis model reduction method with a continuation approach using inexact gradients. The resulting approach can handle an arbitrary number of objectives while yielding a significant reduction in computing time.

Keywords optimal control · Multiobjective optimization · Pareto set · Reduced-order modelling · Reduced basis method · Parameter optimization · Elliptic PDE

Mathematics Subject Classification (2020) 90C29; 49J20

S. Banholzer · S. Volkwein (✉)
Department of Mathematics and Statistics, University of Konstanz, Konstanz, Germany
e-mail: stefan.banholzer@uni-konstanz.de

B. Gebken · M. Dellnitz · S. Peitz
Department of Mathematics, Paderborn University, Paderborn, Germany
e-mail: bgebken@math.upb.de

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_3

1 Introduction

The dilemma of deciding between multiple, equally important goals is present in almost all areas of engineering and economy. A prominent example comes from production, where we want to produce a product at minimal cost while simultaneously preserving a high quality. In the same manner, multiple goals are present in most technical applications, maximizing the velocity while minimizing the energy consumption of electric vehicles [24] being only one of many examples. These conflicting goals result in *multiobjective optimization problems* (MOPs) [9], where we want to optimize all objectives simultaneously. Since the objectives are in general contradictory, there exists an infinite number of *optimal compromises*. The set of these compromise solutions is called the *Pareto set*, and the goal in multiobjective optimization is to approximate this set in an efficient manner, which is significantly more expensive than solving a single-objective problem. Due to this, the development of efficient numerical approximation methods is an active area of research, and methods range from scalarization [9, 14] over set-oriented approaches [8] and continuation [14] to evolutionary algorithms [7]. Recent advances have paved the way to new challenging application areas for multiobjective optimization such as feedback control or problems constrained by partial differential equations (PDEs); cf. [22] for a survey.

In the presence of PDE constraints, the computational effort can quickly become infeasible such that special means have to be taken in order to accelerate the computation. To this end, computationally cheap approximations of the original problem, so-called *surrogate models*, form a promising approach for significantly reducing the computational effort. A widely used approach is to directly construct a mapping from the parameter to the objective space using as few function evaluations of the expensive model as possible, cf. [6, 30] for extensive reviews. In the case of PDE constraints, an alternative approach is via dimension reduction techniques such as Proper Orthogonal Decomposition (POD) [18, 29] or the reduced basis (RB) method [11]. In these methods, a small number of high-fidelity solutions is used to construct a low-dimensional surrogate model for the PDE which can be evaluated significantly faster while guaranteeing convergence using error estimates. In recent years, several methods have been proposed where model reduction is used in multiobjective optimization and optimal control. In [17] and [16], scalarization using the so-called weighted sum method was combined with RB and POD, respectively. In [1, 2], convex problems were solved using reference point scalarization and POD, and set-oriented approaches were used in [3, 4]. A comparison of both was performed in [23] for the Navier–Stokes equations.

In this chapter we combine an extension of the continuation methods presented in [14, 27] to inexact gradients (Sect. 2) with a reduced basis approach for elliptic PDEs (Sect. 3). To deal with the error introduced by the RB approach, we combine the KKT conditions for MOPs with error estimates for the RB method to obtain a tight superset of the Pareto set. For the example considered here, the proposed method yields a speed-up factor of approximately 63 compared to the direct solution of

the expensive problem (Sect. 4). Additionally, our approach allows us to control the quality of the result by controlling the errors for each objective function individually.

2 A Continuation Method for MOPs with Inexact Objective Gradients

In this section, we will begin by briefly introducing the basic concepts of multiobjective optimization upon which we will build in this chapter (see [9, 14] for detailed introductions). Afterwards, we will discuss the continuation method for MOPs and present two modifications of it that can deal with inexact gradient information.

2.1 Multiobjective Optimization

The goal of multiobjective optimization is to minimize several conflicting criteria at the same time. In other words, we want to minimize an objective $J = (J_1, \ldots, J_k) : \mathbb{R}^n \to \mathbb{R}^k$ that is vector valued. It maps the *variable space* $\mathbb{R}^n$ to the *image space* $\mathbb{R}^k$. In contrast to single-objective optimization (i.e., $k = 1$), there exists no natural total order of the image space $\mathbb{R}^k$ for $k > 1$. As a result, the classical concept of *optimality* has to be generalized:

Definition 2.1

(a) $\bar{u} \in \mathbb{R}^n$ is called *(globally) Pareto optimal* if there is no other point $u \in \mathbb{R}^n$ such that $J_i(u) \leq J_i(\bar{u})$ for all $i \in \{1, \ldots, k\}$ and $J_j(u) < J_j(\bar{u})$ for some $j \in \{1, \ldots, k\}$.

(b) The set P of all Pareto optimal points is called the *Pareto set*. Its image under J is the *Pareto front*.

The Pareto set is the solution of the *multiobjective optimization problem (MOP)*

$$\min_{u \in \mathbb{R}^n} J(u). \tag{MOP}$$

Constrained MOPs can be formulated analogously by restricting u in Definition 2.1 to a subset $U \subseteq \mathbb{R}^n$. Similar to the scalar-valued case, if J is differentiable, we can use the derivative of J to obtain necessary conditions for Pareto optimality, the *Karush–Kuhn–Tucker (KKT) conditions* [14]:

Theorem 2.2 *Let $\bar{u}$ be a Pareto optimal point of* (MOP)*. Then there exist multipliers*

$$\alpha \in \Delta_k := \left\{ \alpha \in (\mathbb{R}^{\geq 0})^k : \sum_{i=1}^{k} \alpha_i = 1 \right\}$$

such that

$$DJ(\bar{u})^\top \alpha = \sum_{i=1}^{k} \alpha_i \nabla J_i(\bar{u}) = 0. \tag{KKT}$$

For $k = 1$, this reduces to the well-known optimality condition $\nabla J(\bar{u}) = 0$. If J is non-convex, then the points satisfying (KKT) form a proper superset of the Pareto set P:

Definition 2.3 If $\bar{u} \in \mathbb{R}^n$ and $\bar{\alpha} \in \Delta_k$ satisfy (KKT), then $\bar{u}$ is called *Pareto critical* with corresponding *KKT vector* $\bar{\alpha}$, containing the *KKT-multipliers* $\bar{\alpha}_i$, $i \in \{1, \dots, k\}$. The set P_c of all Pareto critical points is called the *Pareto critical set*.

When solving an MOP, an initial step can be to compute the Pareto critical set. This set possesses additional structure which can be exploited in numerical schemes. Introducing the function

$$F : \mathbb{R}^n \times (\mathbb{R}^{>0})^k \to \mathbb{R}^{n+1}, (u, \alpha) \mapsto \begin{pmatrix} \sum_{i=1}^{k} \alpha_i \nabla J_i(u) \\ 1 - \sum_{i=1}^{k} \alpha_i \end{pmatrix},$$

we see that Pareto critical points and their corresponding KKT vectors can be described as the zero level set of F. As shown by Hillermeier [14], this has the following implication:

Theorem 2.4 *Let J be twice continuously differentiable.*

(a) *Let $\mathcal{M} := \{(u, \alpha) \in \mathbb{R}^n \times (\mathbb{R}^{>0})^k : F(u, \alpha) = 0\}$. If the Jacobian of F has full rank everywhere, i.e.,*

$$rk(DF(u, \alpha)) = n + 1 \quad \forall (u, \alpha) \in \mathcal{M}, \tag{2.1}$$

then $\mathcal{M}$ is a $(k-1)$-dimensional differentiable submanifold of $\mathbb{R}^{n+k}$. The tangent space of $\mathcal{M}$ at (u, α) is given by

$$T_{(u,\alpha)}\mathcal{M} = ker(DF(u, \alpha)).$$

(b) *Let $(u, \alpha) \in \mathcal{M}$ such that (2.1) holds in (u, α). Then there is an open set $U \subseteq \mathbb{R}^n \times \mathbb{R}^k$ with $(u, \alpha) \in U$ such that $\mathcal{M} \cap U$ is a manifold as in* (a). *In other words, $\mathcal{M}$ locally possesses a manifold structure in all points satisfying* (2.1).

Theorem 2.4 forms the basis for the continuation method we use in this chapter.

2.2 Continuation Method with Exact Gradients

We only give a brief description of the method here and refer to [27] and [14] for details. By Theorem 2.4, the Pareto critical set is—except for the boundary—the projection of the differentiable manifold $\mathcal{M} \subseteq \mathbb{R}^n \times \mathbb{R}^k$ onto its first n components. In [10] it has been shown that generically, this also holds for the first-order approximations, i.e., the projection of the tangent space of $\mathcal{M}$ yields the tangent cone of P_c. Given a Pareto critical point $\bar{u} \in P_c$, this means that we can find first-order candidates for new Pareto critical points in the vicinity of $\bar{u}$ by moving in the projected tangent space of $\mathcal{M}$. The idea of the continuation method is to do this iteratively to explore the entire Pareto critical set.

Instead of approximating P_c by a set of points, we use a set-oriented numerical approach; cf. [27] for details. This has the key advantage that it is easy to check whether a certain part of the set has already been computed, which is difficult when working with points. Additionally, a covering of P_c by boxes makes it easy to obtain (and exploit) its topological properties. In the approach, we evenly divide the variable space $\mathbb{R}^n$ into hypercubes or *boxes* B with radius $r > 0$:

$$\mathcal{B}(r) := \{[-r, r]^n + (2i_1 r, \ldots, 2i_n r)^\top : (i_1, \ldots, i_n) \in \mathbb{Z}^n\}. \tag{2.2}$$

Remark 2.5 For ease of notation and readability, we will only consider the case where points $u \in \mathbb{R}^n$ are contained in single boxes. In other words, we only consider the case where u is in the interior of a box and not in the intersection of multiple boxes. Since this is the generic case, this has no impact on the numerical methods we will propose later. ■

For $u \in \mathbb{R}^n$, let $B(u, r)$ be the box containing u. We want to compute the subset of $\mathcal{B}(r)$ covering the Pareto critical set for a given radius r, i.e.,

$$\mathcal{B}_c(r) := \{B \in \mathcal{B}(r) : P_c \cap B \neq \emptyset\}.$$

Since we are interested in a covering via boxes instead of an approximation via points, when moving in a tangent direction of the critical set, we will search for *tangent boxes* instead of single points. For $u \in \mathbb{R}^n$ let

$$N(u, r) := \{B \in \mathcal{B}(r) : B(u, r) \cap B \neq \emptyset\}$$

be the set of neighboring boxes of $B(u, r)$. Starting from a box $B(\bar{u}, r)$ containing a critical point $\bar{u}$ with KKT vector $\bar{\alpha}$, we want to explore the neighboring boxes covering the projected tangent space at $\bar{u}$, i.e.,

$$\mathcal{B}'(\bar{u}, r) = \{B' \in \mathcal{B}(r) : B' \in N(\bar{u}, r),\ B' \cap \bar{u} + pr_u(T_{(\bar{u},\bar{\alpha})}\mathcal{M}) \neq \emptyset\}. \tag{2.3}$$

Here, $pr_u : \mathbb{R}^{n+k} \to \mathbb{R}^n$ is the projection of the tangent space onto the first n components, i.e., the variable space. The typical situation is visualized in Fig. 1.

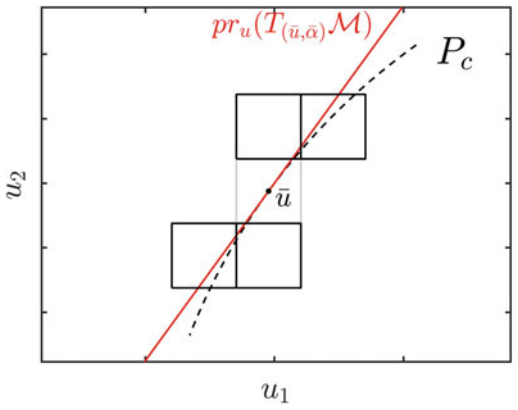

Fig. 1 Tangent boxes (black) of the initial box (grey) containing $\bar{u}$, which is contained in the Pareto critical set P_c (dashed). The red line indicates the projection of the tangent space of $\mathcal{M}$ onto the variable space

As the tangent space of the Pareto critical set is only a linear approximation, a corrector step is required to verify that a given tangent box actually contains part of the Pareto critical set. This means that there has to be at least one $u \in B$ satisfying (KKT). To this end, for a box B, we consider the problem

$$\min_{u \in B, \alpha \in \Delta_k} \|DJ(u)^\top \alpha\|_2^2 \tag{PC-Box}$$

Let $\theta(B)$ be the optimal value of this problem. Then obviously

$$\theta(B) = 0 \Leftrightarrow B \cap P_c \neq \emptyset.$$

In particular, if $\theta(B) = 0$ and $(\bar{u}, \bar{\alpha})$ is the solution of (PC-Box), then $\bar{u}$ is Pareto critical with corresponding KKT vector $\bar{\alpha}$. After solving (PC-Box) in each tangent box, all boxes with $\theta(B) = 0$ are added to a queue and a new iteration of the method is started with the first element in the queue. The method stops when the queue is empty, i.e., when there is no neighboring box of the current set of boxes that contains part of the Pareto critical set. For the remainder of this chapter, we will refer to this method as the *exact continuation method*.

2.3 Continuation Method with Inexact Gradients

Using ROM to solve the state equation of an MOP of an elliptic PDE will introduce an error in the objective functions and the corresponding gradients, which has to be taken into account in order to ensure Pareto criticality of the solution. We here present a method that calculates a tight superset of the Pareto critical set via numerical continuation, using upper bounds for the errors in the approximated gradients. Formally, we now assume that for each gradient ∇J_i, we only have an approximation ∇J_i^r such that

$$\sup_{u\in\mathbb{R}^n} \left\|\nabla J_i(u) - \nabla J_i^r(u)\right\|_2 \leq \epsilon_i,\ i \in \{1,\dots,k\}, \tag{2.4}$$

with upper bounds $\epsilon = (\epsilon_1,\dots,\epsilon_k)^\top \in \mathbb{R}^k$. Let P_c and P_c^r be the Pareto critical sets corresponding to $(\nabla J_i)_i$ and $(\nabla J_i^r)_i$, respectively. The following lemma shows how these error bounds translate to error bounds for the KKT conditions:

Lemma 2.6 *Let $\bar{u} \in \mathbb{R}^n$ be Pareto critical for J with KKT vector $\bar{\alpha} \in \Delta_k$. Then*

$$\|DJ^r(\bar{u})^\top \bar{\alpha}\|_2 \leq \sum_{i=1}^k \bar{\alpha}_i \epsilon_i \leq \|\epsilon\|_\infty.$$

Proof From the estimate

$$\left\|DJ^r(\bar{u})^\top\bar{\alpha}\right\|_2 = \left\|DJ^r(\bar{u})^\top\bar{\alpha} - DJ(\bar{u})^\top\bar{\alpha}\right\|_2 = \left\|\sum_{i=1}^k (\nabla J_i^r(\bar{u}) - \nabla J_i(\bar{u}))^\top \bar{\alpha}_i\right\|_2$$
$$\leq \sum_{i=1}^k \left\|\nabla J_i^r(\bar{u}) - \nabla J_i(\bar{u})\right\|_2 \bar{\alpha}_i \leq \sum_{i=1}^k \bar{\alpha}_i\epsilon_i \leq \|\epsilon\|_\infty$$

we derive the claim. □

Remark 2.7 Lemma 2.6 can be generalized to equality and inequality constrained MOPs using the constrained version of the optimality conditions from [14]. In this case, in the norm on the left-hand side of the inequality in Lemma 2.6, one additionally has to add a linear combination of the gradients of the equality and inequality constraints. ■

Lemma 2.6 shows that we have to weaken the conditions for Pareto criticality of the reduced objective function to obtain a superset of the actual Pareto critical set P_c. Formally, let

$$P_1^r := \left\{u \in \mathbb{R}^n : \min_{\alpha\in\Delta_k} \|DJ^r(u)^\top\alpha\|_2^2 \leq \|\epsilon\|_\infty^2\right\},$$
$$P_2^r := \left\{u \in \mathbb{R}^n : \min_{\alpha\in\Delta_k} \left(\|DJ^r(u)^\top\alpha\|_2^2 - (\alpha^\top\epsilon)^2\right) \leq 0\right\}.$$

P_1^r was also considered in [21] in the context of descent directions, where the solution of $\min_{\alpha\in\Delta_k} \|DJ^r(u)^\top\alpha\|_2^2$ is the squared length of the steepest descent direction in u. The condition for a point being in P_1^r only depends on the maximal error $\|\epsilon\|_\infty$ and can be seen as a relaxed version of the KKT conditions for the inexact objective function. In contrast to this, the condition in P_2^r actually considers the individual error bounds. By Lemma 2.6,

$$P_c \subseteq P_2^r \subseteq P_1^r \text{ and } P_c^r \subseteq P_2^r \subseteq P_1^r,$$

i.e., both P_1^r and P_2^r are supersets of P_c and P_c^r (the points $\bar{u}$ for which the inexact gradients satisfy (KKT)). In fact, P_2^r is a tight superset of P_c in the following sense:

Lemma 2.8 *Let $\tilde{u} \in P_2^r$. Then there is some continuously differentiable $\tilde{J} : \mathbb{R}^n \to \mathbb{R}^k$ with*

$$\sup_{u \in \mathbb{R}^n} \|\nabla \tilde{J}_i(u) - \nabla J_i^r(u)\|_2 \leq \epsilon_i \ \forall i \in \{1, \dots, k\}$$

such that $\tilde{u}$ is Pareto critical for $\tilde{J}$.

Proof Let

$$\begin{aligned}
\tilde{\alpha} &\in \operatorname{argmin}_{\alpha \in \Delta_k} \left(\|DJ^r(u)^\top \alpha\|_2^2 - (\alpha^\top \epsilon)^2 \right), \\
v &:= DJ^r(\tilde{u})^\top \tilde{\alpha}, \\
g(u) &:= -\left(\frac{1}{\tilde{\alpha}^\top \epsilon} \sum_{i=1}^n v_i u_i \right) \epsilon, \\
\tilde{J}(u) &:= J^r(u) + g(u).
\end{aligned}$$

Since $\tilde{u} \in P_2^r$ by assumption, we have $\|v\|_2 \leq \tilde{\alpha}^\top \epsilon$. Thus

$$\|\nabla \tilde{J}_i(u) - \nabla J_i^r(u)\|_2 = \|\nabla g_i(u)\|_2 = \frac{\epsilon_i}{\tilde{\alpha}^\top \epsilon} \|v\|_2 \leq \epsilon_i \quad \forall u \in \mathbb{R}^n \text{ and } \forall i \in \{1, \dots, k\},$$

and

$$D\tilde{J}(\tilde{u})^\top \tilde{\alpha} = v + \sum_{i=1}^k \tilde{\alpha}_i \nabla g_i(\tilde{u}) = v - \sum_{i=1}^k \tilde{\alpha}_i \frac{\epsilon_i}{\tilde{\alpha}^\top \epsilon} v = 0,$$

which proves the lemma. □

Lemma 2.8 shows that for each point $\tilde{u}$ in P_2^r, there is an objective function satisfying the error bounds (2.4) for which $\tilde{u}$ is Pareto critical. As a result, P_2^r is the tightest superset of P_c we can hope for if we only have the estimates in (2.4). The following example shows both supersets for a simple MOP (cf. [21]).

Example Let

$$J^r : \mathbb{R}^2 \to \mathbb{R}^2, \ u \mapsto \begin{pmatrix} (u_1 - 1)^2 + (u_2 - 1)^4 \\ (u_1 + 1)^2 + (u_2 + 1)^2 \end{pmatrix}.$$

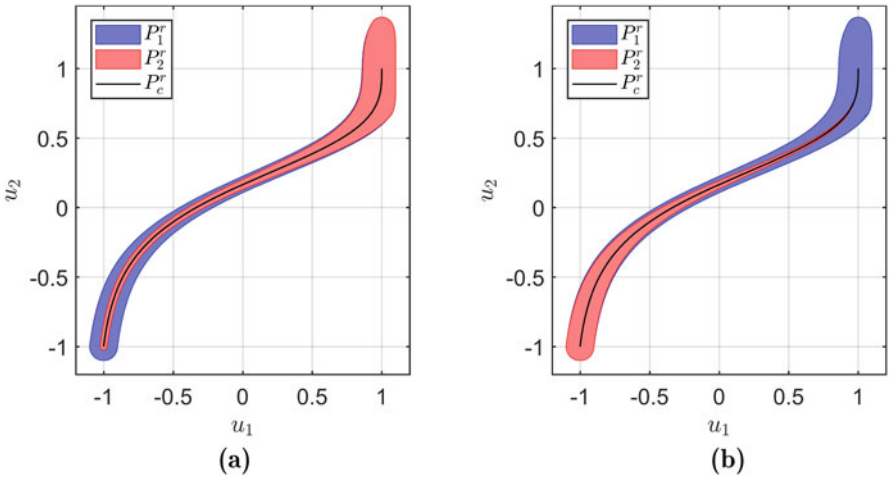

Fig. 2 P_1^r and P_2^r for different error bounds ϵ. (**a**) $\epsilon = \epsilon^1 = (0.2, 0.05)^\top$. (**b**) $\epsilon = \epsilon^2 = (0, 0.2)^\top$

We consider the two error bounds $\epsilon^1 = (0.2, 0.05)^\top$ and $\epsilon^2 = (0, 0.2)^\top$. The corresponding supersets P_1^r and P_2^r are shown in Fig. 2.

As $\|\epsilon_1\|_\infty = \|\epsilon_2\|_\infty = 0.2$, P_1^r is identical for both error bounds. Considering each component of J^r individually, the critical points of J_1^r and J_2^r are located at $u^1 = (1, 1)^\top$ and $u^2 = (-1, -1)^\top$, respectively. For P_2^r, we see that the difference between P_c^r and P_2^r becomes smaller the closer we get to the critical point of the objective function with the smaller error bound. This can be expected, as the influence (or weight) of $\nabla J_i^r(u)$ in the KKT conditions (KKT) becomes larger the closer u is to u^i. In particular, in Fig. 2b, the difference between P_2^r and P_c^r at $(1, 1)^\top$ becomes zero, as $\epsilon_1^2 = 0$. ◊

If we set $\epsilon_i = \|\epsilon\|_\infty$ for all $i \in \{1, \ldots, k\}$, then $P_1^r = P_2^r$. Thus, we will from now on only consider P_2^r. As shown in the previous example, the "dimension" of P_2^r is higher than the "dimension" of P_c^r. More precisely, P_2^r contains the closure of an open subset of $\mathbb{R}^n$, which is shown in the following lemma:

Lemma 2.9 *Let ∇J_i^r be continuous for all $i \in \{1, \ldots, k\}$. Let*

$$A := \left\{ u \in \mathbb{R}^n : \min_{\alpha \in \Delta_k} \left(\|DJ^r(u)^\top \alpha\|_2^2 - (\alpha^\top \epsilon)^2 \right) < 0 \right\}.$$

Then

(a) *P_2^r is closed. In particular, $\overline{A} \subseteq P_2^r$.*
(b) *A is open.*

Proof

(a) The case $P_2^r = \emptyset$ is trivial, so we assume that $P_2^r \neq \emptyset$. Let $\bar{u} \in \overline{P_2^r}$. Then there is a sequence $(u^i)_i \in P_2^r$ with $\lim_{i\to\infty} u^i = \bar{u}$. Consider the sequence $(\alpha^i)_i \in \Delta_k$ with

$$\alpha^i \in \operatorname{argmin}_{\alpha\in\Delta_k} \left(\|DJ^r(u^i)^\top \alpha\|_2^2 - (\alpha^\top \epsilon)^2 \right).$$

By compactness of Δ_k, we can assume w.l.o.g. that there is some $\bar{\alpha} \in \Delta_k$ with $\lim_{i\to\infty} \alpha^i = \bar{\alpha}$. Let

$$\Psi : \mathbb{R}^n \times \Delta_k \to \mathbb{R}, \quad (u, \alpha) \mapsto \|DJ^r(u)^\top \alpha\|_2^2 - (\alpha^\top \epsilon)^2.$$

By our assumption, Ψ is continuous. From $\Psi(u^i, \alpha^i) < 0$ for all $i \in \mathbb{N}$ it follows that $\Psi(\bar{u}, \bar{\alpha}) \leq 0$, which yields $\bar{u} \in P_2^r$.

(b) The case $A = \emptyset$ is again trivial such that we assume $A \neq \emptyset$. Let $\bar{u} \in A$ with

$$\bar{\alpha} \in \operatorname{argmin}_{\alpha\in\Delta_k} \left(\|DJ^r(\bar{u})^\top \alpha\|_2^2 - (\alpha^\top \epsilon)^2 \right).$$

Let $\psi : \mathbb{R}^n \to \mathbb{R}, u \mapsto \|DJ^r(u)^\top \bar{\alpha}\|_2^2 - (\bar{\alpha}^\top \epsilon)^2$. Then $\psi(\bar{u}) < 0$ and by our assumption, ψ is continuous. Therefore, there is some open set $U \subseteq \mathbb{R}^n$ with $\bar{u} \in U$ such that $\psi(u) < 0$ for all $u \in U$. Since

$$\min_{\alpha\in\Delta_k} \left(\|DJ^r(u)^\top \alpha\|_2^2 - (\alpha^\top \epsilon)^2 \right) \leq \psi(u) < 0 \quad \forall u \in U$$

we have $U \subseteq A$ such that A is open. □

We will now present two strategies for the numerical computation of P_2^r. Analogously to the case with exact gradients, we will approximate P_2^r via the box covering

$$\mathcal{B}_c^r(r) := \{B \in \mathcal{B}(r) : B \cap P_2^r \neq \emptyset\}.$$

2.3.1 Strategy 1

The idea of our first method is to mimic the exact continuation method to calculate $\mathcal{B}_c^r$. For this, there are mainly two modifications we have to make:

1. By Lemma 2.9, P_2^r is not a lower-dimensional object in $\mathbb{R}^n$, so it makes no sense to use tangent information to find first-order candidates as in (2.3). Instead, we have to consider all neighboring boxes.
2. The problem (PC-Box) has to be replaced by a problem that checks the defining inequality of P_2^r.

As a replacement for (PC-Box), we consider the following problem:

$$\min_{u\in B,\alpha\in\Delta_k} \|DJ(u)^\top\alpha\|_2^2 - (\alpha^\top\epsilon)^2. \tag{ϵPC-Box}$$

Let $\theta_\epsilon(B)$ be the optimal value of this problem. Note that $\theta_\epsilon(B) < 0$ is sufficient to verify that a box B contains part of P_2^r. As a result, we do not need to solve (ϵPC-Box) exactly. For example, when using an iterative method for the solution of (ϵPC-Box), we can stop when the function value is negative. The above mentioned changes yield Algorithm 1.

Algorithm 1 Strategy 1: box-continuation algorithm with inexact gradients

Given: Radius $r > 0$ of boxes.
Choose an initial point $u_0 \in P_2^r$ and initialize $\mathcal{B} = \{B(u_0, r)\}$ and a queue $Q = \{u_0\}$.
while $Q \neq \emptyset$ **do**
  Remove the first element $\bar{u}$ from Q.
  for $B' \in N(\bar{u}, r) \setminus \mathcal{B}$ **do**
    Solve (ϵPC-Box) for B'. Let $\theta_\epsilon(B')$ be the optimal value and (u', α') be the solution.
    if $\theta_\epsilon(B') \leq 0$ **then**
      Add u' to Q and B' to $\mathcal{B}$.
    end if
  end for
end while

Due to the loss of low-dimensionality of P_2^r, the formulation of the continuation method becomes much simpler. As a consequence, it is straightforward to show that Algorithm 1 yields the desired covering $\mathcal{B}_c^r(r)$.

When executing the exact continuation method directly using inexact gradients (i.e., forgetting about the inexactness) and comparing it to Algorithm 1 (with the same box radius), the former will generally be much faster than the latter. A suitable way to evaluate the run time is to compare the number of times Problems (PC-Box) and (ϵPC-Box) need to be solved, respectively, as they require the majority of the computing time and are equally difficult to solve. (Here, we assume that both problems are solved with equal precision.) For each box added to the collection $\mathcal{B}$ in either algorithm, one of these problems has to be solved. Consequently, the longer run time of Algorithm 1 is partly due to the fact that P_2^r is a superset of P_c^r, which means that more boxes are required to cover P_2^r than P_c^r. However, even if the error bounds ϵ are small such that P_2^r and P_c^r are almost equal, Algorithm 1 will be slower. This is due to the fact that instead of only the tangent boxes, all neighboring boxes have to be tested with (ϵPC-Box) in each loop of Algorithm 1. While this does not matter in the interior of P_2^r (as all neighboring boxes are in fact in P_2^r in that case), it is very inefficient at the boundary of P_2^r. This is the motivation for the second strategy.

2.3.2 Strategy 2

By Lemma 2.9, P_2^r has the same dimension as the space of variables $\mathbb{R}^n$. This means that it can be described much more efficiently by its topological boundary ∂P_2^r. To be more precise, $\mathbb{R}^n \setminus \partial P_2^r$ consists of different connected components that lie either completely inside or completely outside P_2^r. So if we know ∂P_2^r, we merely have to test one point of each connected component if it is contained in P_2^r or not to completely determine P_2^r. Therefore, the idea of our second strategy is to only compute ∂P_2^r.

Let

$$\varphi : \mathbb{R}^n \to \mathbb{R}, \quad u \mapsto \min_{\alpha \in \Delta_k} \left(\|DJ^r(u)^\top \alpha\|_2^2 - (\alpha^\top \epsilon)^2 \right). \tag{2.5}$$

This map is well-defined since Δ_k is compact, i.e., the minimum always exists. By Lemma 2.9, we have $\partial P_2^r \subseteq \varphi^{-1}(0)$. Our goal is to compute $\varphi^{-1}(0)$ via a continuation approach. To this end, we first have to show that φ is differentiable. We will do this by investigating the properties of the optimization problem in (2.5), i.e., of the problem

$$\begin{aligned} &\min_{\alpha \in \mathbb{R}^k} \omega(\alpha), \\ s.t. \quad &\sum_{i=1}^{k} \alpha_i = 1, \\ &\alpha_i \geq 0 \quad \forall i \in \{1, \dots, k\}, \end{aligned} \tag{2.6}$$

for

$$\omega(\alpha) := \|DJ^r(u)^\top \alpha\|_2^2 - (\alpha^\top \epsilon)^2 = \alpha^\top (DJ^r(u) DJ^r(u)^\top - \epsilon\epsilon^\top)\alpha.$$

This leads to the following result.

Theorem 2.10 *Let $\bar{u} \in \varphi^{-1}(0)$ such that (2.6) has a unique solution $\bar{\alpha} \in \Delta_k$ with $\bar{\alpha}_i > 0$ for all $i \in \{1, \dots, k\}$. Let (2.6) be uniquely solvable in a neighborhood of $\bar{u}$. Then there is an open set $U \subseteq \mathbb{R}^n$ with $\bar{u} \in U$ such that $\varphi|_U$ is continuously differentiable.*

Proof See Appendix A. □

For a standard continuation approach, we also have to show that $\varphi^{-1}(0)$ is a manifold. By the Level Set Theorem (cf. [19], Corollary 5.14), to properly show that $\varphi^{-1}(0)$ is a manifold in a neighborhood of some $\bar{u} \in \varphi^{-1}(0)$, we would have to show that $D\varphi|_U(\bar{u}) \neq 0$ (cf. (A.5)). From the theoretical point of view, this poses a problem as there is no obvious way to achieve this. In practice however, we can test this by checking if the norm of $D\varphi|_U(\bar{u})$ is below a certain threshold. If this

is the case, and if $\varphi^{-1}(0)$ is indeed not a manifold, we again have to consider all neighboring boxes as tangent boxes as in strategy 1. Otherwise, if $D\varphi|_U(\bar{u}) \neq 0$, we can compute the tangent space $T_{\bar{u}}$ of $\varphi^{-1}(0)$ at $\bar{u}$ via

$$T_{\bar{u}} = ker(D\varphi(\bar{u})).$$

Finally, in analogy to (PC-Box) and (ϵPC-Box), we will use the following problem to test if a box B contains part of ∂P_2^r:

$$\min_{u \in B} \varphi(u)^2. \qquad (\partial\epsilon\text{PC-Box})$$

The resulting continuation method is presented in Algorithm 2.

Remark 2.11

1. Since for every evaluation of φ the solution of the quadratic problem (2.6) has to be computed, (∂ϵPC-Box) is significantly more difficult to solve than (ϵPC-Box). Additionally, we are looking for the points u where $\varphi(u) = 0$, i.e., where the problem (2.6) is not positive definite. This increased difficulty of Strategy 2 is compensated by the fact that far fewer boxes have to be checked with (∂ϵPC-Box) than with (ϵPC-Box) in Strategy 1.
2. When all $\epsilon_i = \bar{\epsilon}$ are equal, $\varphi(u) = -\bar{\epsilon}^2$ for all $u \in P_c^r$, i.e., φ is constant on the Pareto critical set P_c^r. This means that local solvers may fail to find a minimum of φ when the box B in (∂ϵPC-Box) has a nonempty intersection with P^r. An obvious but expensive way to circumvent this problem is to start the local solver multiple times with different initial points. Alternatively, one can use sufficient conditions for a box B containing part of $\varphi^{-1}(0)$ before actually solving (∂ϵPC-Box). For example, by the intermediate value theorem, if there are two points in B where φ has different sign, we immediately know that $\varphi(u) = 0$ for some $u \in B$. (But note that for this method, we still need to find a point in $\varphi^{-1}(0) \cap B$ to be able to calculate the tangent space of $\varphi^{-1}(0)$).
3. In practice, error bounds which are zero can cause problems for the stability of Strategy 2. For example, in Fig. 2b, the width of P_2^r becomes arbitrarily small near $(1, 1)^\top$. As a result, Strategy 2 may jump between different parts of the boundary and thus miss certain parts. Additionally, since the boundary of P_2^r typically intersects the Pareto critical set P_c in this case, (∂ϵPC-Box) may be difficult to solve (as in 2.). Thus, in practice, one should use error bounds that are slightly larger than zero, even if the corresponding gradients are exact. ■

Algorithm 2 Strategy 2: boundary-continuation algorithm for inexact gradients

Given: Radius $r > 0$ of boxes.

1: Choose an initial point $u_0 \in \partial P_2^r$ and initialize $\mathcal{B} = \{B(u_0, r)\}$ and a queue $Q = \{u_0\}$.
2: **while** $Q \neq \emptyset$ **do**
3: Remove the first element $\bar{u}$ from Q.
4: If $\|D\varphi(\bar{u})\|_2$ is small set $T = \mathbb{R}^n$. Otherwise, compute the tangent space $T = ker(D\varphi(\bar{u}))$.

Predictor:

5: Find all neighboring boxes of $B(\bar{u}, r)$ that have a nonempty intersection with $\bar{u} + T$ and have not been considered before, i.e.,

$$\mathcal{B}'(\bar{u}, r) = \{B' \in \mathcal{B}(r) : B' \cap B(\bar{u}, r) \neq \emptyset,\ B' \cap \bar{u} + T \neq \emptyset\} \setminus \mathcal{B}.$$

Corrector:

6: **for** $B' \in \mathcal{B}'(\bar{u}, r)$ **do**
7: Solve ($\partial\epsilon$PC-Box) for B'. Let $\theta(B')$ be the optimal value and u' be the solution.
8: **if** $\theta(B') = 0$ **then**
9: Add u' to Q and B' to $\mathcal{B}$.
10: **end if**
11: **end for**
12: **end while**

2.4 Globalization Approach

Note that all algorithms presented in this section so far approximate either P_c, P_2^r, or ∂P_2^r by starting in an initial point u_0 and then locally exploring in all (tangent) directions. Thus, if the set we want to approximate is disconnected, we can only compute the connected component that contains u_0. In the following, we will describe how we can solve this problem, i.e., how our methods can be globalized.

As mentioned earlier, an advantage of using boxes in the continuation method instead of points is the fact that it is easy to detect whether a region has already been explored. In particular, this allows us to start the continuation in multiple initial points at the same time, by simply adding all of them to the queue Q in step 1 of Algorithms 1 or 2 (and initializing the covering $\mathcal{B}$ with the corresponding boxes). As a result, to globalize our methods, we merely have to find an initial set U_0 of points such that the intersection of U_0 with each connected component is nonempty.

For obtaining an initial set, we make use of the optimization problems that verify if a box contains part of the set we want to approximate, i.e., the problems (PC-Box), (ϵPC-Box), and ($\partial\epsilon$PC-Box). The idea is to consider a box covering as in (2.2) with large radius R and then simply test each box for relevant points using these problems. Let B_0 be a compact superset of the set that we want to approximate (i.e., of P_c, P_2^r or ∂P_2^r), e.g., a large outer box. For ease of notation, we assume that B_0 is a union of boxes in $\mathcal{B}(R)$. For the case of the Pareto critical set P_c, i.e., the globalization of the exact continuation method, the resulting method is presented in Algorithm 3. The corresponding globalization methods for Algorithm 1

and 2 are obtained by replacing (PC-Box) in step 3 by (ϵPC-Box) and (∂ϵPC-Box), respectively.

Algorithm 3 Global initialization

Given: Outer box B_0, Radius $R > 0$ of boxes.
1: Initialize $U_0 = \emptyset$.
2: **for** $B \in \mathcal{B}(R)$ with $B \cap B_0 \neq \emptyset$ **do**
3: Solve (PC-Box) for B. Let $\theta(B)$ be the optimal value and $\bar{u}$ be the solution.
4: **if** $\theta(B) = 0$ **then**
5: Add $\bar{u}$ to U_0.
6: **end if**
7: **end for**

The radius R has to be chosen such that for each connected component, there is at least one box in our covering that only has an intersection with the desired component. In theory, R can obviously become very small if two different connected components are very close to each other. In this case, Algorithm 3 becomes infeasible to use, as the number of boxes that have to be tested becomes too large. In practice however, the components are often sufficiently far apart such that a large radius is sufficient and only few boxes have to be considered.

For the globalization of the exact continuation method and Algorithm 1, we only have to take the non-connectivity of P_c and P_2^r into account. For Algorithm 2, an additional problem may arise since the boundary ∂P_2^r does not necessarily need to be smooth. Non-smoothness of ∂P_2^r is caused by points in which φ is not differentiable. (By Theorem 2.10, these are points where the solution of (∂ϵPC-Box) is not unique.) In these points, ∂P_2^r does not possess a tangent space, and our method will be unable to continue. As a result, we have to ensure in the initialization of Algorithm 2 that we choose an initial point in U_0 on each smooth component of ∂P_2^r. Visually, these can be thought of as the faces of P_2^r.

We conclude this section with some remarks on the practical use of Algorithm 3.

Remark 2.12

1. For MOPs with a high-dimensional variable space, Algorithm 3 quickly becomes infeasible due to the exponential growth of the number of boxes in $\mathcal{B}(R)$. For these cases, an initialization based on points instead of boxes should be used, for example by applying methods from global optimization to modified versions of (PC-Box), (ϵPC-Box), and (∂ϵPC-Box), where u is not constrained to a box B.
2. Instead of directly looping over all boxes in step 2 of Algorithm 3, in some cases it might be more beneficial to first execute a few steps of the subdivision algorithm (cf. [8]) to quickly discard boxes that are far away from the Pareto critical set. ■

3 Multiobjective Optimization of an Elliptic PDE Using the RB Method

In this section, we will present a multiobjective (parameter) optimization problem of an elliptic advection-diffusion-reaction equation and show how the reduced basis method can be applied in view of the continuation method for inexact gradients from Sect. 2.3 (see Algorithms 1 and 2).

3.1 Multiobjective Optimization of an Elliptic PDE

Given a domain $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, which is divided into m' pairwise disjoint subdomains $\Omega = \Omega_1 \dot\cup \ldots \dot\cup \Omega_{m'}$, we consider the problem

$$\min_{y,u} \ \mathcal{J}(y,u) := \begin{pmatrix} \frac{1}{2}\left\|y - y_d^1\right\|^2_{L^2(\Omega)} \\ \vdots \\ \frac{1}{2}\left\|y - y_d^{k-1}\right\|^2_{L^2(\Omega)} \\ \frac{1}{2}\|u\|^2_{\mathbb{R}^m} \end{pmatrix} \tag{MPOP}$$

s.t.

$$\begin{aligned} -\nabla \cdot \left[\left(\textstyle\sum_{i=1}^{m'} \kappa_i \chi_{\Omega_i}(x)\right) \nabla y(x)\right] + c\, b(x) \cdot \nabla y(x) + r\, y(x) &= f(x) \quad \text{for } x \in \Omega, \\ \tfrac{\partial y}{\partial \eta}(x) &= 0 \qquad \text{for } x \in \partial\Omega, \end{aligned} \tag{EPDE}$$

and the bilateral box constraints

$$u_a \le u \le u_b, \tag{BC}$$

where $u = (u_1, \ldots, u_m) = (\kappa_1, \ldots, \kappa_{m'}, c, r) \in \mathbb{R}^m$ is the parameter of dimension $m := m' + 2$, $U_{ad} := \{u \in \mathbb{R}^m \mid u_a \le u \le u_b\}$ is the admissible parameter set, and $y \in L^2(\Omega) =: H$ is the state variable.

For every $i \in \{1, \ldots, m'\}$ the parameter κ_i is the diffusion coefficient on the subdomain Ω_i. The vector field $b \in L^\infty(\Omega, \mathbb{R}^d)$ is the given advection, whose strength and orientation can be controlled by the parameter $c \in \mathbb{R}$. Moreover, the reaction coefficient is given by the parameter $r > 0$, and $f \in H$ is the inhomogeneity on the right-hand side of the equation. On the boundary, we impose homogeneous Neumann boundary conditions.

The cost functions $\mathcal{J}_1, \ldots, \mathcal{J}_{k-1} : H \times \mathbb{R}^m \to \mathbb{R}$ are of tracking type with respect to the desired states $y_d^1, \ldots, y_d^{k-1} \in H$, and the cost function $\mathcal{J}_k : H \times \mathbb{R}^m \to \mathbb{R}$ measures the parameter cost.

Setting $V := H^1(\Omega)$ and using the parameter-dependent bilinear form $a(u; \cdot, \cdot) : V \times V \to \mathbb{R}$ defined by

$$\begin{aligned} a(u; \varphi, \psi) &:= \sum_{i=1}^{m} u_i a_i(\varphi, \psi) \\ &:= \sum_{i=1}^{m'} \kappa_i \int_{\Omega_i} \nabla\varphi(x) \cdot \nabla\psi(x)\, dx + c \int_{\Omega} b(x) \cdot \nabla\varphi(x)\psi(x)\, dx \\ &\quad + r \int_{\Omega} \varphi(x)\psi(x)\, dx, \end{aligned}$$

for all $u \in \mathbb{R}^m$ and $\varphi, \psi \in V$, and the linear functional $F : V \to \mathbb{R}$ given by $F(\varphi) := \langle f, \varphi \rangle_H$ for all $\varphi \in V$, we can write (EPDE) in its weak formulation as: Find $y \in V$ such that

$$a(u; y, \varphi) = F(\varphi) \quad \text{for all } \varphi \in V \tag{3.1}$$

is satisfied. It is possible to show the unique solvability of (3.1) under some conditions on the parameter u.

Theorem 3.1 *There are $\kappa_{min} \in (0, \infty)^{m'}$, $c_{min}, c_{max} \in \mathbb{R}$ with $c_{min} < c_{max}$ and $r_{min} \in (0, \infty)$ such that (3.1) has a unique solution $y(u) \in V$ for every parameter $u = (\kappa, c, r) \in \mathbb{R}^m$ with $\kappa > \kappa_{min}$, $c_{min} < c < c_{max}$ and $r > r_{min}$.*

Proof It is straightforward to show that for all parameters $u \in \mathbb{R}^m$ the bilinear form $a(u; \cdot, \cdot)$ and the linear functional F are continuous, and that there are $\kappa_{min} \in (0, \infty)^{m'}$, $c_{min}, c_{max} \in \mathbb{R}$ with $c_{min} < c_{max}$ and $r_{min} \in (0, \infty)$ such that $a(u; \cdot, \cdot)$ is coercive for all $u = (\kappa, c, r) \in \mathbb{R}^m$ with $\kappa > \kappa_{min}$, $c_{min} < c < c_{max}$ and $r > r_{min}$. Now the Lax–Milgram Theorem can be applied to show the unique solvability of (3.1). □

With Theorem 3.1 in mind, we can introduce the solution operator of the elliptic PDE.

Definition 3.2 Define the set $U_{eq} := (\kappa_{min}, \infty) \times (c_{min}, c_{max}) \times (r_{min}, \infty)$ with the constants from Theorem 3.1. Let $\mathcal{S} : U_{eq} \to V \hookrightarrow H$ be defined as the solution operator of (3.1), i.e., the function $y := \mathcal{S}(u)$ solves the weak formulation (3.1) for any parameter $u \in U_{eq}$.

Remark 3.3 In the following, we suppose that it holds $U_{ad} \subset U_{eq}$. ■

Using the explicit dependence of the state y on the parameter u for all $u \in U_{eq}$, the essential cost functions $J_1, \ldots, J_k : U_{eq} \subset \mathbb{R}^m \to \mathbb{R}$ can be defined.

Definition 3.4 For any $i \in \{1, \ldots, k\}$ let the essential cost function $J_i : U_{eq} \subset \mathbb{R}^m \to \mathbb{R}$ be given by $J_i(u) := \mathcal{J}_i(\mathcal{S}(u), u)$ for all $u \in U_{eq}$.

For applying the continuation method from Sect. 2, which is based on Theorem 2.4, to solve this multiobjective parameter optimization problem, the cost functions $J_1, \ldots, J_k$ need to be twice continuously differentiable. This is the statement of the next lemma.

Lemma 3.5 *The cost functions $J_1, \ldots, J_k : U_{eq} \subset \mathbb{R}^m \to \mathbb{R}$ are twice continuously differentiable.*

Proof It is clear that the cost function J_k is twice continuously differentiable. Furthermore, it is possible to show that the solution operator $\mathcal{S}$ of (3.1) is twice continuously differentiable (this can be shown by rewriting (3.1) in the form $e(y, u) = 0$ and then using the implicit function theorem, cf. [15, Section 1.6]). From this, it immediately follows that the cost functions $J_1, \ldots, J_{k-1}$ are twice continuously differentiable as well. □

For later use, we need an explicit formula for the gradients $\nabla J_1, \ldots, \nabla J_k$. Therefore, we introduce the so-called adjoint equation for all $i \in \{1, \ldots, k-1\}$: Find $p \in V$ such that it holds

$$a(u; \varphi, p) = \langle y_d^i - \mathcal{S}(u), \varphi \rangle_H \quad \text{for all } \varphi \in V. \tag{3.2}$$

With the same arguments as in Theorem 3.1, it is possible to show that (3.2) has a unique solution for all $u \in U_{eq}$.

Definition 3.6 Denote by $\mathcal{A}_i : U_{eq} \to V \hookrightarrow H$ the solution operator of the adjoint Eq. (3.2) for all $i \in \{1, \ldots, k-1\}$.

Now a small computation shows that

$$J_i'(u)h = \langle \mathcal{S}(u) - y_d^i, \mathcal{S}'(u)h \rangle_H = \partial_u a(u; \mathcal{S}(u), \mathcal{A}_i(u))h,$$

which yields

$$\nabla J_i(u) = \begin{pmatrix} \partial_u a(u; \mathcal{S}(u), \mathcal{A}_i(u))e_1 \\ \vdots \\ \partial_u a(u; \mathcal{S}(u), \mathcal{A}_i(u))e_m \end{pmatrix} = \begin{pmatrix} a_1(\mathcal{S}(u), \mathcal{A}_i(u)) \\ \vdots \\ a_m(\mathcal{S}(u), \mathcal{A}_i(u)) \end{pmatrix} \tag{3.3}$$

for all $i \in \{1, \ldots, k-1\}$. Lastly, it is obvious that $\nabla J_k(u) = u$.

3.2 The Reduced Basis Method

For computing the Pareto critical set of the problem (MPOP) by the exact continuation method introduced in Sect. 2.2, the problem (PC-Box) has to be solved numerous times. However, already one gradient evaluation of all cost functions

$\nabla J_1(u), \ldots, \nabla J_k(u)$ involves the solution of one state and $k-1$ adjoint equations. Thus, using a finite element discretization for the weak formulations (3.1) and (3.2), which leads to large linear equation systems, is numerically very costly and time consuming. Therefore, the use of *reduced-order modelling* (ROM) is a common tool to lower the computational costs.

The idea of ROM is to use a low-dimensional subspace $V^r \subset V$ as a surrogate for the infinite-dimensional space V in the weak formulations (3.1) and (3.2). Given a finite-dimensional reduced-order space $V^r \subset V$, the reduced-order state equation reads: Find $y^r \in V^r$ such that

$$a(u; y^r, \varphi) = F(\varphi) \quad \text{for all } \varphi \in V^r \tag{3.4}$$

is satisfied.

With the same arguments as in Theorem 3.1 it can be shown that (3.4) has a unique solution for all $u \in U_{eq}$. Therefore, we can follow the procedure of Sect. 3.1 and introduce the solution operator $\mathcal{S}^r : U_{eq} \to V^r \subset V \hookrightarrow H$ of the ROM state equation (3.4) and consequently the ROM essential cost functions $J_1^r, \ldots, J_k^r$, which are defined by $J_i^r(u) := \mathcal{J}_i(\mathcal{S}^r(u), u)$ for all $u \in U_{eq}$ and all $i \in \{1, \ldots, k\}$. Again, it can be shown that the functions $J_1^r, \ldots, J_k^r$ are twice continuously differentiable so that they fit into the framework of Theorem 2.4. The gradient of the cost functions can also be displayed by the reduced-order adjoint equations

$$a(u; \varphi, p^r) = \langle y_d^i - \mathcal{S}^r(u), \varphi \rangle_H \quad \text{for all } \varphi \in V^r, \tag{3.5}$$

for all $i \in \{1, \ldots, k-1\}$, whose solution operator we denote by $\mathcal{A}_i^r : U_{eq} \to V^r \subset V \hookrightarrow H$. With this definition it holds

$$\nabla J_i^r(u) = \begin{pmatrix} \partial_u a(u; \mathcal{S}^r(u), \mathcal{A}_i^r(u))e_1 \\ \vdots \\ \partial_u a(u; \mathcal{S}^r(u), \mathcal{A}_i^r(u))e_m \end{pmatrix} = \begin{pmatrix} a_1(\mathcal{S}^r(u), \mathcal{A}_i^r(u)) \\ \vdots \\ a_m(\mathcal{S}^r(u), \mathcal{A}_i^r(u)) \end{pmatrix} \tag{3.6}$$

for all $i \in \{1, \ldots, k-1\}$. Moreover, we have $\nabla J_k^r(u) = u = \nabla J_k(u)$.

In this paper, we use a particular model-order reduction technique, namely the *reduced basis* (RB) method (see e.g. [13, 25, 26]). In the RB method, the snapshot space V^r is spanned by solutions of the state equation and the adjoint equations to different parameter values $u \in U_{ad}$. The reduced basis is then given by an orthonormal basis $(\Phi_1, \ldots, \Phi_N)$ of the space V^r.

By using the RB method we introduce an error in the state equation, which transfers to the cost functions, its gradients, and eventually to the Pareto critical set, which we want to compute. In Sect. 2.3, two strategies were presented to deal with the inflicted inexactness in the gradients of the multiobjective optimization problem. Both are based on the estimates (2.4) for the errors in the gradients of the cost functions. Thus, when applying the RB method we need to ensure these

estimates. This is done by using the well-known greedy algorithm (cf. [5]). Given a sufficiently fine finite parameter training set $\mathcal{P} \subset U_{ad}$, new solution snapshots are computed until the error in the gradients of all cost functions is smaller than the predefined error tolerance for all parameters in $\mathcal{P}$. The parameter for the new snapshots is thereby chosen as the one for which the error in the gradient is the largest. The procedure is summarized in Algorithm 4.

Algorithm 4 Greedy algorithm

Given: Parameter set $\mathcal{P} \subset U_{ad}$, greedy tolerances $\varepsilon_1, \dots, \varepsilon_k > 0$.
1: Choose $u \in \mathcal{P}$, compute $\mathcal{S}(u), \mathcal{A}_1(u), \dots, \mathcal{A}_{k-1}(u)$.
2: Set $V^r = span\{\mathcal{S}(u), \mathcal{A}_1(u), \dots, \mathcal{A}_{k-1}(u)\}$ and compute the reduced basis by orthonormalization.
3: **while** $\max_{u\in\mathcal{P}} \max_{i\in\{1,\dots,k-1\}} \left\|\nabla J_i(u) - \nabla J_i^r(u)\right\|_2 > \epsilon_i$ **do**
4: Choose $(\bar{u}, i) = \arg\max_{u\in\mathcal{P},\, i\in\{1,\dots,k-1\}} \left\|\nabla J_i(u) - \nabla J_i^r(u)\right\|_2$.
5: Compute $\mathcal{S}(\bar{u})$ and $\mathcal{A}_i(\bar{u})$.
6: Set $V^r = span\{V^r \cup \{\mathcal{S}(\bar{u}), \mathcal{A}_i(\bar{u})\}\}$ and compute the reduced basis by orthonormalization.
7: **end while**

3.3 Error Estimation for the Gradients

In the greedy procedure in Algorithm 4, the error between the full-order and the reduced-order gradients has to be evaluated. There are two strategies to do so.

1. The full-order gradients are computed and stored at the beginning of the greedy procedure. Therefore, in each greedy iteration, only the reduced-order gradients have to be computed and the error can be easily evaluated. Of course, this implies large computational costs at the beginning of the greedy procedure. This method is called strong greedy algorithm (cf. [5, 12]).
2. An a posteriori error estimator for the errors in the gradient is used, which can be efficiently evaluated. This results in computational costs for the greedy algorithm, which only depend on the reduced-order dimension N.

To be able to follow the second strategy, we introduce a rigorous a posteriori error estimator for the error in the gradient of the cost functions.

Using the gradient representations (3.3) and (3.6), we can write for $i \in \{1, \dots, k-1\}$

$$\left\|\nabla J_i(u) - \nabla J_i^r(u)\right\|_2^2 = \sum_{j=1}^{m} \left|a_j(\mathcal{S}(u), \mathcal{A}_i(u)) - a_j(\mathcal{S}^r(u), \mathcal{A}_i^r(u))\right|^2.$$

Due to the bilinearity and the continuity of $a_1, \dots, a_m$ and the triangle inequality, we can further write

$$
\begin{aligned}
&\left|a_j(\mathcal{S}(u), \mathcal{A}_i(u)) - a_j(\mathcal{S}^r(u), \mathcal{A}_i^r(u))\right| \\
\leq &\left|a_j(\mathcal{S}(u) - \mathcal{S}^r(u), \mathcal{A}_i^r(u))\right| + \left|a_j(\mathcal{S}(u) - \mathcal{S}^r(u), \mathcal{A}_i(u) - \mathcal{A}_i^r(u))\right| \\
&+ \left|a_j(\mathcal{S}^r(u), \mathcal{A}_i(u) - \mathcal{A}_i^r(u))\right| \qquad (3.7) \\
\leq &C_j \left(\left\|\mathcal{S}(u) - \mathcal{S}^r(u)\right\|_V \left\|\mathcal{A}_i^r(u)\right\|_V + \left\|\mathcal{S}(u) - \mathcal{S}^r(u)\right\|_V \left\|\mathcal{A}_i(u) - \mathcal{A}_i^r(u)\right\|_V \right. \\
&\left. + \left\|\mathcal{S}^r(u)\right\|_V \left\|\mathcal{A}_i(u) - \mathcal{A}_i^r(u)\right\|_V\right) \qquad (3.8)
\end{aligned}
$$

for all $j \in \{1, \ldots, m\}$.

Therefore, we need a posteriori error estimators for the state and the adjoint equations in order to be able to estimate the approximation error induced in the gradients. To this end, we use the following well-known estimators (cf. [26]):

$$
\begin{aligned}
\left\|\mathcal{S}(u) - \mathcal{S}^r(u)\right\|_V &\leq \frac{\|r_{\mathcal{S}}(u)\|_{V'}}{\alpha(u)} =: \Delta_{\mathcal{S}}(u), \\
\left\|\mathcal{A}_i(u) - \mathcal{A}_i^r(u)\right\|_V &\leq \frac{\left\|r_{\mathcal{A}_i}(u)\right\|_{V'}}{\alpha(u)} + \Delta_{\mathcal{S}}(u) =: \Delta_{\mathcal{A}_i}(u),
\end{aligned}
$$

where the residuals $r_{\mathcal{S}}(u)$ and $r_{\mathcal{A}_i}(u)$ are given by

$$
\begin{aligned}
\langle r_{\mathcal{S}}(u), \varphi\rangle_{V',V} &:= F(\varphi) - a(u; \mathcal{S}^r(u), \varphi) && \text{for all } \varphi \in V, \\
\langle r_{\mathcal{A}_i}(u), \varphi\rangle_{V',V} &:= \langle y_d^i - \mathcal{S}^r(u), \varphi\rangle_H - a(u; \varphi, \mathcal{A}_i^r(u)) && \text{for all } \varphi \in V.
\end{aligned}
$$

For methods on how to estimate $\alpha(u)$ and to evaluate the terms $\|r_{\mathcal{S}}(u)\|_{V'}$ and $\left\|r_{\mathcal{A}_i}(u)\right\|_{V'}$ efficiently, we refer for example to [26].

Remark 3.7 Since $J_k = J_k^r$, the gradients of the two functions also coincide, so that the ∇J_k is approximated exactly by ∇J_k^r. ■

4 Numerical Results

In this section, we will numerically investigate the application of the continuation method presented in Sect. 2 to the PDE-constrained multiobjective optimization problem using the reduced basis method in Sect. 3.

For the discretization of the state and adjoint equations, we used linear finite elements with 714 degrees of freedom.

4.1 Generation of the Reduced Basis

For investigating the generation of the reduced basis by the greedy algorithm in Algorithm 4, we consider the MPOP

$$\begin{pmatrix} J_1(u) \\ J_2(u) \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \left\| \mathcal{S}(u) - y_d^1 \right\|_H^2 \\ \frac{1}{2} \left\| u \right\|_{\mathbb{R}^4}^2 \end{pmatrix} \tag{4.1}$$

with $u = (\kappa_1, \kappa_2, c, r)$, $\Omega_1 = (0,1) \times (0, 0.5)$, $\Omega_2 = (0,1) \times (0.5, 1)$, and the admissible parameter set

$$U_{ad} = \{u = (\kappa_1, \kappa_2, c, r) \in \mathbb{R}^4 \mid 0.2 \leq \kappa_i \leq 5 \, (i = 1, 2), \ c = 0, \ r = 0.5\}.$$

The reason for setting $c = 0$ in this example is that the coercivity constant $\alpha(u)$ of the bilinear form $a(u; \cdot, \cdot)$ is explicitly given by $\alpha(u) = \min\{\kappa_1, \kappa_2, r\}$ for all $u \in U_{ad}$, so that we expect a good efficiency of the error estimator of both the state and adjoint equations. This is verified by the results shown in Fig. 3a, where the efficiency of the error estimator for both equations is shown for a given reduced basis for 1000 randomly chosen parameter values. However, the resulting efficiency of the error estimator for the error in the gradient is between 10^3 and 10^6 (see Fig. 3b) and thus not well-suited for a greedy procedure, which depends on a good error estimation. The huge overestimation of the error estimator is mainly due to the use of the triangle inequality (3.7) and the continuity estimates (3.8), as can be seen in Fig. 3b.

Compared to the strong greedy algorithm, we can see in Table 1 that this overestimation results in far more basis elements than actually needed to reach the given error bound. Since we want to investigate the influence of the error bounds in the estimate (2.4) on the problem, we want that the estimate (2.4) is satisfied sharply

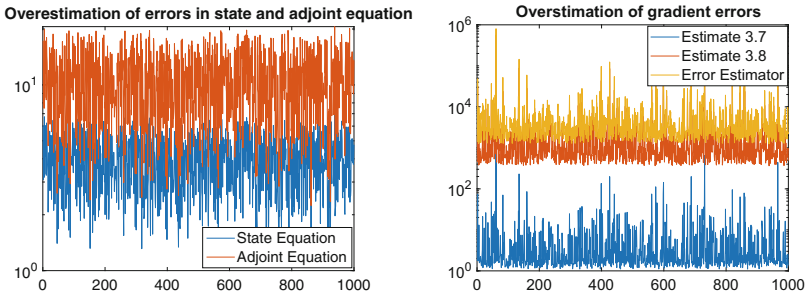

Fig. 3 Overestimations for 1000 randomly selected parameter values. (**a**) State and adjoint equation. (**b**) Gradient

Table 1 Number of basis functions for different error bounds

Error bound	Strong greedy	Error estimate
$\varepsilon = 1e-6$	24	56
$\varepsilon = 1e-5$	20	50
$\varepsilon = 1e-4$	16	40
$\varepsilon = 1e-3$	12	32
$\varepsilon = 1e-2$	12	26
$\varepsilon = 1e-1$	10	20

by the RB. Therefore, we will not use the error estimator to generate the basis, but instead use the strong greedy algorithm.

4.2 Application of the Continuation Methods to an MPOP

For the numerical investigation of the continuation method applied to a PDE-constrained multiobjective parameter optimization problem together with the use of the reduced basis method, we consider the MPOP

$$\begin{pmatrix} J_1(u) \\ J_2(u) \\ J_3(u) \\ J_4(u) \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\|\mathcal{S}(u) - \mathcal{S}((0.7, 0.8, 0.5))\|_H^2 \\ \frac{1}{2}\|\mathcal{S}(u) - \mathcal{S}((2, 0.5, 0.5))\|_H^2 \\ \frac{1}{2}\|\mathcal{S}(u) - \mathcal{S}((3, -0.5, 0.5))\|_H^2 \\ \frac{1}{2}\|u\|_{\mathbb{R}^3}^2 \end{pmatrix} \tag{4.2}$$

with $u = (\kappa, c, r)$ and

$$U_{ad} = \{u = (\kappa, c, r) \in \mathbb{R}^3 \mid 0.5 \leq \kappa \leq 3,\ -1 \leq c \leq 1,\ r = 0.5\},$$

i.e., the reaction parameter r is a constant so that we only optimize the diffusivity in the whole domain Ω and the strength and orientation of the advection field b. Thus, this can be seen as a problem with two parameters.

As described before, the reduced basis is generated by the strong greedy Algorithm 4, where the error bounds $\epsilon_1, \ldots, \epsilon_4$ are chosen in accordance with the estimate (2.4). As a reference, the exact solution of (4.2) (via exact continuation and FEM discretization of the weak formulations) is shown in Fig. 4.

Remark 4.1 Since (4.2) is constrained to a box, we have to use a constrained version of the exact continuation method (cf. [14]) to calculate Pareto critical points that lie on the boundaries of (4.2). But note that for this example, all Pareto critical points on the boundary are also Pareto critical if we ignore the constraints. In other words, for each Pareto critical point $\bar{u}$ on the boundary, there is a sequence of Pareto critical point in the interior that converges to $\bar{u}$. By continuity of DJ, the gradients of the (active) inequality constraints in the KKT conditions can be ignored. As a result, we can treat (4.2) as an unconstrained problem that we only solve in a certain area. ■

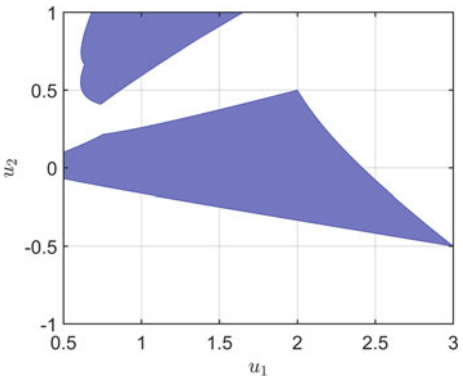

Fig. 4 The Pareto critical set of (4.2)

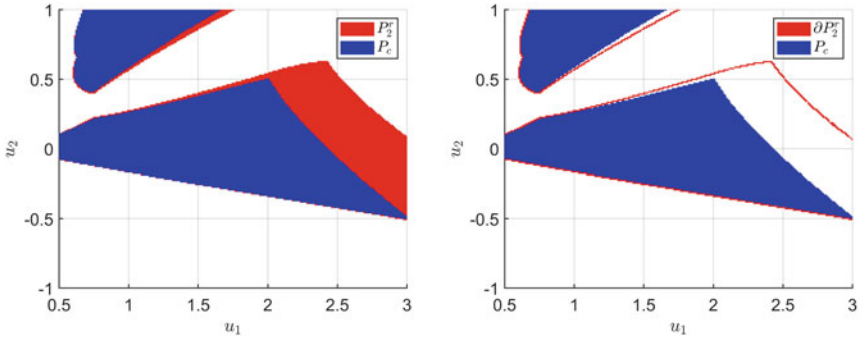

Fig. 5 Results of Strategy 1 (left) and 2 (right) for the MPOP (4.2) with $\epsilon = (0.03, 0.03, 0.01, 0.01)$

As a first test, we will compare the time needed to compute the exact solution of (4.2) with the time needed for Strategy 1 and 2. For the error bounds, we choose $\epsilon = (0.03, 0.03, 0.01, 0.01)$, and for the box radius we choose $r = \frac{3-0.5}{2^9} \approx 0.0049$. The results are shown in Fig. 5.

All three methods were implemented in Matlab. For the solution of the subproblems (PC-Box), (ϵPC-Box), and ($\partial\epsilon$PC-Box), the SQP-Algorithm of `fmincon` was used. (For increased stability during the continuation, each subproblem where the SQP-Algorithm found an optimal value larger than zero was restarted using the Interior-Point-Method and the Active-Set-Method of `fmincon`). The runtime, number of boxes, and number of subproblems needed are shown in Table 2. When comparing Strategy 1 and Strategy 2, we see that Strategy 2 needs about 20 times fewer boxes and solutions of subproblems than Strategy 1. This is to be expected, since Strategy 2 only computes a covering of the boundary of P_2^r, i.e., of a lower-dimensional set. When comparing the actual runtime, Strategy 2 is about 5 times faster than Strategy 1, since the subproblems in Strategy 2 are more expensive to solve than the ones in Strategy 1 (cf. Remark 2.11). Finally, Strategy 2 is about 63

Table 2 Comparison of the performance of the exact continuation method, Strategy 1, and Strategy 2 for Example (4.2). The number of subproblems is split up into subproblems for the continuation and initialization (cf. Sect. 2.4)

Algorithm	# Boxes	# Subproblems	Runtime (in seconds)
Exact cont.	15916	$18721 + 25$	$17501s$
Strategy 1	21750	$24490 + 25$	$1426s$
Strategy 2	899	$1027 + 225$	$276s$

times faster than the exact continuation method with FEM discretization, illustrating the large increase in efficiency we gain from our approach.

Although it is a lot quicker to use inexact gradients from ROM instead of the exact gradients via FEM, it is important to keep in mind that our methods are computing a superset of the actual Pareto critical set. For example, in Fig. 5, the right side of the lower connected component is only approximated poorly by P_2^r. Therefore, we will now investigate the influence of the error bounds $\epsilon = (\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$ on P_2^r, by applying Strategy 2 with reduced bases for different values of ϵ. Note that in all our tests we set $\epsilon_4 = 0.01$, although the error in the gradient of the fourth cost function is zero for all parameters. This is done to make the solution of ($\partial\epsilon$PC-Box) in line 7 of Algorithm 2 numerically stable (cf. Remark 2.11).

The results of our experiment can be seen in Fig. 6. Generally, as expected, the boundary ∂P_2^r encloses the Pareto critical set P_c sharper and sharper for decreasing ϵ. Moreover, we observe that it is crucial to choose an ϵ which is not too large: For the value $\epsilon = (0.1, 0.1, 0.1, 0.01)$ the shape of the boundary ∂P_2^r implies that the set P_2^r is connected, i.e., we lose the topological information that the Pareto critical set actually consists of two connected components. Decreasing ϵ to $\epsilon = (0.0885, 0.0885, 0.0885, 0.01)$ we are in the limit case in which the boundary ∂P_2^r touches the box constraints at around $(2.3, 1)$, so that this is the approximate ϵ for which we regain the basic topological information of a disconnected Pareto critical set.

If we compare the results for $\epsilon = (0.03, 0.03, 0.03, 0.01)$, $\epsilon = (0.03, 0.03, 0.01, 0.01)$, and $\epsilon = (0.03, 0.01, 0.01, 0.01)$, the influence of changing one component of ϵ becomes obvious. For $\epsilon = (0.03, 0.03, 0.03, 0.01)$ the set ∂P_2^r encloses the set P_c quite sharply at the upper connected component and at the left part of the lower connected component, where the second and third component of the corresponding KKT-multipliers α are small. On the other hand, in the right part of the lower connected component of P_c, where the second and third component of the corresponding KKT-multipliers are relatively large, the deviation of ∂P_2^r to P_c is still large. Consequently, first reducing ϵ_3 and then also ϵ_2 from 0.03 to 0.01 leads to a clearly visible sharper enclosing of this part of P_c.

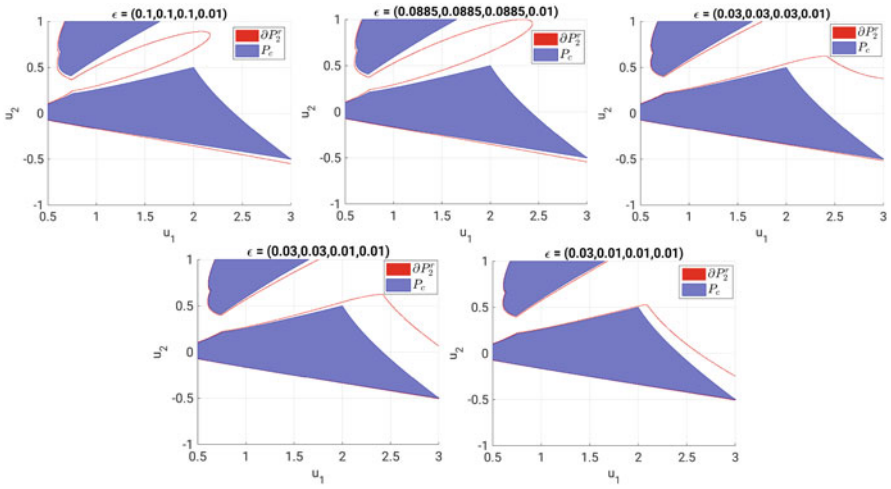

Fig. 6 Results of Strategy 2 for different values of ϵ

5 Conclusion and Outlook

In this chapter, we present a way to efficiently solve multiobjective parameter optimization problems of elliptic PDEs by combining the reduced basis method from PDE-constrained optimization with the continuation approach from multiobjective optimization, which computes a box covering of the Pareto critical set. Using the RB method in this setting introduces an error in the objective functions and their gradients that has to be considered when solving the MPOP. To this end, we require that the reduced basis guarantees error bounds for the gradients of the objective functions. These error bounds are then incorporated into the KKT optimality conditions for MOPs to derive a tight superset P_2^r of the actual Pareto (critical) set. This superset can be computed using a straightforward modification of the continuation method for MOPs (Strategy 1). Since P_2^r has the same dimensions as the variable space of the MOP, we afterwards present a second method that only computes the boundary ∂P_2^r of P_2^r (Strategy 2). We do this by showing that ∂P_2^r can be written as the level set of a differentiable mapping, which again enables the use of a continuation approach to compute it. For constructing the reduced basis, we use a greedy procedure which incorporates, and thus ensures, the error bounds for the gradients of the objective functions.

Our numerical tests show that the presented a posteriori error estimator for the error in the gradients is not well-suited for the application in a greedy procedure due to its bad efficiency. Therefore, a strong greedy algorithm is used to build the reduced basis. Concerning the solution of the MPOP we investigate two aspects: First, the runtimes of our methods are compared. In our case, Strategy 1 is about 13 times and Strategy 2 about 63 times faster than the exact solution of the MPOP (via

the classical continuation method with FEM discretization). Second, the influence of the error bound for the gradients of the objective functions is investigated. As expected, a smaller error bound leads to a tighter covering of the Pareto critical set. Moreover, we observe that single components of the error bound strongly influence the tightness of the covering in areas, in which the corresponding components of the KKT-multipliers are large. Thus, by individually adapting the single components of the error bound, we can nicely control the tightness of the covering.

For future work, there are some theoretical and practical aspects that should be investigated further:

- As mentioned in Remark 2.11, in certain situations there can be difficulties when solving the problem ($\partial\epsilon$PC-Box). In these situations, specialized methods that take these difficulties into account should be developed and used instead of standard methods for constrained optimization.
- If the number of objectives of the MPOP is larger than the number of variables, it may be possible to combine our approaches in this chapter with the hierarchical decomposition of the Pareto critical set presented in [10].
- The development of a more efficient a posteriori error estimator for the error in the gradients of the objective functions would allow to use it in the greedy procedure. In that way, the expensive strong greedy procedure would be avoided in the offline phase. One way to do so might be the application of localized RB methods, see e.g. [20].
- As explained in the globalization approach in Sect. 2.4, we have to use multiple initial points to ensure that we find all connected components of P_2^r (and faces of ∂P_2^r). Due to the local nature of the continuation method, this approach can potentially be parallelized, increasing the efficiency of our methods even more.
- If a decision maker is present with a certain preference, it may be worth to steer our continuation method in a direction that results from that preference instead of approximating the complete Pareto set. For the case with exact gradients, this was done in [28].

Acknowledgments This research was funded by the DFG Priority Programme 1962 "Non-smooth and Complementarity-based Distributed Parameter Systems".

Appendix A: Proof of Theorem 2.10

To prove Theorem 2.10, we first have to investigate some of the properties of the optimization problem (2.6). This problem is quadratic with linear equality and inequality constraints. We will first investigate the uniqueness of the solution in the following lemma.

Lemma A.1 *Let $u \in \varphi^{-1}(0)$ and let α^1 and α^2 be two solutions of (2.6) with $\alpha^1 \neq \alpha^2$. Then $\omega(\alpha) = 0$ for all $\alpha \in span(\{\alpha^1, \alpha^2\})$ and*

$$span(\{\alpha^1, \alpha^2\}) \cap ker(DJ(u)^\top) \neq \emptyset. \tag{A.1}$$

Proof For $c_1, c_2 \in \mathbb{R} \setminus \{0\}$ we have

$$\begin{aligned}
&\omega(c_1\alpha^1 + c_2\alpha^2) \\
&= (c_1\alpha^1 + c_2\alpha^2)^\top (DJ^r(u)DJ^r(u)^\top - \epsilon\epsilon^\top)(c_1\alpha^1 + c_2\alpha^2) \\
&= c_1^2\omega(\alpha^1) + 2({c_1\alpha^1}^\top DJ^r(u)DJ^r(u)^\top c_2\alpha^2 - {c_1\alpha^1}^\top \epsilon\epsilon^\top c_2\alpha^2) + c_2^2\omega(\alpha^2) \\
&= 2({c_1\alpha^1}^\top DJ^r(u)DJ^r(u)^\top c_2\alpha^2 - {c_1\alpha^1}^\top \epsilon\epsilon^\top c_2\alpha^2) \\
&= 2c_1c_2((DJ^r(u)^\top\alpha^1)^\top (DJ^r(u)^\top\alpha^2) - (\epsilon^\top\alpha^1)(\epsilon^\top\alpha^2)).
\end{aligned}$$

From $\omega(\alpha^1) = \omega(\alpha^2) = 0$ it follows that $\epsilon^\top\alpha^1 = \|DJ^r(u)^\top\alpha^1\|$ and $\epsilon^\top\alpha^2 = \|DJ^r(u)^\top\alpha^2\|$. Let $\sphericalangle$ be the angle between $DJ^r(u)^\top\alpha^1$ and $DJ^r(u)^\top\alpha^2$. Then

$$\begin{aligned}
&\omega(c_1\alpha^1 + c_2\alpha^2) \\
&= 2c_1c_2(cos(\sphericalangle)\|DJ^r(u)^\top\alpha^1\|\|DJ^r(u)^\top\alpha^2\| - \|DJ^r(u)^\top\alpha^1\|\|DJ^r(u)^\top\alpha^2\|) \\
&= 2c_1c_2(cos(\sphericalangle) - 1)\|DJ^r(u)^\top\alpha^1\|\|DJ^r(u)^\top\alpha^2\|.
\end{aligned} \tag{A.2}$$

Assume $cos(\sphericalangle) \neq 1$ (i.e., $cos(\sphericalangle) - 1 < 0$), $\|DJ^r(u)^\top\alpha^1\| \neq 0$, and $\|DJ^r(u)^\top\alpha^2\| \neq 0$. If we choose $c_1 = t$ and $c_2 = 1 - t$ for $t \in (0, 1)$, then $t\alpha^1 + (1-t)\alpha^2 \in \Delta_k$ and $\omega(t\alpha^1 + (1-t)\alpha^2) < 0$, which contradicts $u \in \varphi^{-1}(0)$. If $\|DJ^r(u)^\top\alpha^1\| = 0$ or $\|DJ^r(u)^\top\alpha^2\| = 0$, then (A.1) holds for $\bar{\alpha} = \alpha^1$ or $\bar{\alpha} = \alpha^2$, respectively. If $cos(\sphericalangle) - 1 = 0$, then $DJ^r(u)^\top\alpha^1$ and $DJ^r(u)^\top\alpha^2$ are linearly dependent, so there are $\bar{c}_1, \bar{c}_2 \in \mathbb{R} \setminus \{0\}$ such that $DJ^r(u)^\top\bar{\alpha} = 0$ for $\bar{\alpha} = \bar{c}_1\alpha^1 + \bar{c}_2\alpha^2$. In particular, in any case we must have $\omega(\alpha) = 0$ for all $\alpha \in span(\{\alpha^1, \alpha^2\})$. □

The previous lemma implies that for $k = 2$, the solution of (2.6) for $u \in \varphi^{-1}(0)$ is non-unique iff $DJ^r(u)DJ^r(u)^\top - \epsilon\epsilon^\top = 0$. For $k > 2$, we can only have non-uniqueness if (A.1) holds. If we consider the dimensions of the spaces in (A.1), we see that in the generic case, it can only hold if

$$\begin{aligned}
&dim(span(\{\alpha^1, \alpha^2\}) \cap ker(DJ(u)^\top)) \geq 1 \\
\Leftrightarrow\ &2 + k - rk(DJ(u)^\top) - k \geq 1 \\
\Leftrightarrow\ &rk(DJ(u)^\top) \leq 1,
\end{aligned}$$

i.e., if all gradients of the objectives are linearly dependent in u. This motivates us to assume that in general, the solution of (2.6) is unique for almost all $u \in \varphi^{-1}(0)$.

We will now investigate the differentiability of φ. Our strategy is to apply the implicit function theorem to the KKT conditions of (2.6) to obtain a differentiable

function ϕ that maps a point $u \in \mathbb{R}^n$ onto the solution of (2.6) in u. This would imply the differentiability of φ via concatenation with ω. An obvious problem here is the fact that (2.6) has inequality constraints which, when activated or deactivated under variation of u, lead to non-differentiabilities in ϕ. Note that an inequality constraint being active means that one component of α is zero, i.e., one of the objective functions has no impact on the current problem. Thus, for our theoretical purposes, if there is an active inequality constraint in (2.6) we will just ignore the corresponding objective function. This approach is strongly related to the hierarchical decomposition of the Pareto critical set (cf. [10]).

For the reasons mentioned above, we will now consider the case where the solution of (2.6) is strictly positive in each component. The following lemma shows a technical result that will be used in a later proof.

Lemma A.2 *Let $u \in \varphi^{-1}(0)$ and let $\bar{\alpha} \in \Delta_k$ be a solution of (2.6) with $\alpha_i > 0$ $\forall i \in \{1, \dots, k\}$. Then $\bar{\alpha}$ is unique if and only if there is no $\beta \in \mathbb{R}^k \setminus \{0\}$ with $\omega(\beta) = 0$ and $\sum_{i=1}^k \beta_i = 0$.*

Proof We will show that α is non-unique if and only if there is some $\beta \in \mathbb{R}^k$ with $\omega(\beta) = 0$ and $\sum_{i=1}^k \beta_i = 0$.

$\Rightarrow$: Let $\tilde{\alpha}$ be another solution of (2.6). Then, as in the proof of Lemma A.1, we must have $\omega(c_1\bar{\alpha} + c_2\tilde{\alpha}) = 0$ for all $c_1, c_2 \in \mathbb{R}$. This means we can choose $\beta = \bar{\alpha} - \tilde{\alpha}$.

$\Leftarrow$: Let $\beta \in \mathbb{R}^k$ with $\omega(\beta) = 0$ and $\sum_{i=1}^k \beta_i = 0$. Let $s > 0$ be small enough such that $\bar{\alpha} + s\beta \in \Delta_k$. Then, as in (A.2), we have

$$\omega(\bar{\alpha} + s\beta) = 2s(cos(\sphericalangle) - 1)\|DJ^r(u)^\top \bar{\alpha}\|\|DJ^r(u)^\top \beta\| \leq 0.$$

Since by assumption $\varphi(u) = 0$ we must have $\omega(\bar{\alpha} + s\beta) = 0$, so $\bar{\alpha} + s\beta$ is another solution of (2.6). □

To be able to use the KKT conditions of (2.6) to obtain its solution, we have to make sure that these conditions are sufficient. Since (2.6) is a quadratic problem, this means we have to show that the matrix in the objective ω is positive semidefinite.

Lemma A.3 *Let $u \in \varphi^{-1}(0)$ and let $\bar{\alpha} \in \Delta_k$ be the unique solution of (2.6) with $\bar{\alpha}_i > 0$ $\forall i \in \{1, \dots, k\}$. Then $\omega(\beta) \geq 0$ for all $\beta \in \mathbb{R}^k$. In particular, $DJ(u)DJ(u)^\top - \epsilon\epsilon^\top$ is positive semidefinite.*

Proof Assume there is some $\beta \in \mathbb{R}^k$ with $\omega(\beta) < 0$, i.e., $\epsilon^\top \beta > \|DJ^r(u)^\top \beta\|$. We distinguish between two cases:

Case 1: $\sum_{i=1}^k \beta_i = 0$: Similar to the proof of Lemma A.1 we get

$$\begin{aligned}\omega(\bar{\alpha} + s\beta) &< 2s((DJ^r(u)^\top \bar{\alpha})^\top (DJ^r(u)^\top \beta) - (\epsilon^\top \bar{\alpha})(\epsilon^\top \beta)) \\ &< 2s(cos(\sphericalangle) - 1)\|DJ^r(u)^\top \bar{\alpha}\|\|DJ^r(u)^\top \beta\| \leq 0\end{aligned}$$

for all $s > 0$. In particular, since $\bar{\alpha}$ is positive, there is some $\bar{s} > 0$ such that $\bar{\alpha} + \bar{s}\beta \in \Delta_k$ with $\omega(\bar{\alpha} + \bar{s}\beta) < 0$, which is a contradiction.

Case 2: $\sum_{i=1}^{k} \beta_i \neq 0$. W.l.o.g. assume that $\sum_{i=1}^{k} \beta_i = 1$. Consider

$$\bar{\omega} : \mathbb{R} \to \mathbb{R}, \quad s \mapsto \omega(\bar{\alpha} + s(\beta - \bar{\alpha})).$$

Then $\bar{\omega}(0) = 0$ and $\bar{\omega}(1) < 0$. By assumption, we must have $\bar{\omega}(s) > 0$ for all s such that $\bar{\alpha} + s(\beta - \bar{\alpha}) \in \Delta_k$. By continuity of $\bar{\omega}$ there must be some s^* with $\bar{\omega}(s^*) = 0$. Let $\bar{\beta} := \bar{\alpha} + s^*(\beta - \bar{\alpha})$. Using (A.2) we get

$$\begin{aligned}\omega(\bar{\alpha} + ts^*(\beta - \bar{\alpha})) &= \omega((1-t)\bar{\alpha} + t\bar{\beta}) \\ &= 2t(1-t)(cos(\sphericalangle) - 1)\|DJ^r(u)^\top \bar{\alpha}\|\|DJ^r(u)^\top \bar{\beta}\| \leq 0\end{aligned}$$

for all $t \in (0, 1)$, which is a contradiction. □

The previous results now allow us to prove Theorem 2.10.

Theorem 2.10 *Let $\bar{u} \in \varphi^{-1}(0)$ such that (2.6) has a unique solution $\bar{\alpha} \in \Delta_k$ with $\bar{\alpha}_i > 0$ for all $i \in \{1, \dots, k\}$. Let (2.6) be uniquely solvable in a neighborhood of $\bar{u}$. Then there is an open set $U \subseteq \mathbb{R}^n$ with $\bar{u} \in U$ such that $\varphi|_U$ is continuously differentiable.*

Proof The KKT conditions for (2.6) are

$$\begin{aligned}(DJ(u)DJ(u)^\top - \epsilon\epsilon^\top)\alpha - \begin{pmatrix}\lambda + \mu_1 \\ \vdots \\ \lambda + \mu_k\end{pmatrix} &= 0, \\ \sum_{i=1}^{k} \alpha_i - 1 &= 0, \\ \alpha_i \geq 0 \ \forall i \in \{1, \dots, k\}, \\ \mu_i \geq 0 \ \forall i \in \{1, \dots, k\}, \\ \mu_i \alpha_i = 0 \ \forall i \in \{1, \dots, k\}.\end{aligned} \tag{A.3}$$

for $\lambda \in \mathbb{R}$ and $\mu \in \mathbb{R}^k$. By Lemma A.3 these conditions are sufficient for optimality. By our assumption there is an open set U' with $\bar{u} \in U'$ such that the solution of (2.6) is unique and positive. Thus, on U', (A.3) is equivalent to

$$(DJ(u)DJ(u)^\top - \epsilon\epsilon^\top)\alpha - \begin{pmatrix}\lambda \\ \vdots \\ \lambda\end{pmatrix} = 0,$$

$$\sum_{i=1}^{k} \alpha_i - 1 = 0,$$

for some $\lambda \in \mathbb{R}$. This system can be rewritten as $G(u, (\alpha, \lambda)) = 0$ for

$$G : \mathbb{R}^n \times \mathbb{R}^{k+1} \to \mathbb{R}^{k+1}, \quad (u, (\alpha, \lambda)) \mapsto \begin{pmatrix} (DJ(u)DJ(u)^\top - \epsilon\epsilon^\top)\alpha - (\lambda, \dots, \lambda)^\top \\ \sum_{i=1}^{k} \alpha_i - 1 \end{pmatrix}.$$

Derivating G with respect to (α, λ) yields

$$D_{(\alpha,\lambda)}G(u, (\alpha, \lambda)) = \begin{pmatrix} (DJ(u)DJ(u)^\top - \epsilon\epsilon^\top) & (-1, \dots, -1)^\top \\ (1, \dots, 1) & 0 \end{pmatrix} \in \mathbb{R}^{(k+1)\times(k+1)}.$$

Let $\bar{\lambda} \in \mathbb{R}$ such that $G(\bar{u}, (\bar{\alpha}, \bar{\lambda})) = 0$. (Note that uniqueness of $\bar{\alpha}$ implies uniqueness of $\bar{\lambda}$ here.) For $D_{(\alpha,\lambda)}G(\bar{u}, (\bar{\alpha}, \bar{\lambda}))$ to be singular, there would have to be some $v = (v^1, v^2) \in \mathbb{R}^{k+1}$ with

$$0 = D_{(\alpha,\lambda)}G(\bar{u}, (\bar{\alpha}, \bar{\lambda}))v = \begin{pmatrix} (DJ(\bar{u})DJ(\bar{u})^\top - \epsilon\epsilon^\top)v^1 - (v^2, \dots, v^2)^\top \\ \sum_{i=1}^{k} v_i^1 \end{pmatrix}$$

and thus

$$\begin{aligned} 0 &= {v^1}^\top (DJ(\bar{u})DJ(\bar{u})^\top - \epsilon\epsilon^\top)v^1 - {v^1}^\top (v^2, \dots, v^2)^\top \\ &= {v^1}^\top (DJ(\bar{u})DJ(\bar{u})^\top - \epsilon\epsilon^\top)v^1 - v_2 \sum_{i=1}^{k} v_i^1 \\ &= w(v^1). \end{aligned}$$

By Lemma A.2, this is a contradiction to the assumption that $\bar{\alpha}$ is a unique solution of (2.6). So $D_{(\alpha,\lambda)}G(\bar{u}, (\bar{\alpha}, \bar{\lambda}))$ has to be regular. This means we can apply the implicit function theorem to obtain open sets $U \subseteq U' \subseteq \mathbb{R}^n$, $V \subseteq \mathbb{R}^{k+1}$ with $\bar{u} \in U$, $(\bar{\alpha}, \bar{\lambda}) \in V$ and a continuously differentiable function $\phi = (\phi_\alpha, \phi_\lambda) : U \to V$ with

$$G(u, (\alpha, \lambda)) = 0 \;\Leftrightarrow\; (\alpha, \lambda) = \phi(u) \quad \forall u \in U, (\alpha, \lambda) \in V.$$

In particular,

$$\varphi|_U(u) = \min_{\alpha \in \Delta_k} \left(\|DJ(u)^\top \alpha\|^2 - (\alpha^\top \epsilon)^2 \right) = \|DJ(u)^\top \phi_\alpha(u)\|^2 - (\phi_\alpha(u)^\top \epsilon)^2, \tag{A.4}$$

so $\varphi|_U$ is continuously differentiable. □

Remark A.4 From the proof of Theorem 2.10 we can even derive an explicit formula for the derivative of $\varphi|_U$ in $\bar{u}$: First of all, the derivative of the implicit function ϕ is given by

$$\begin{aligned}
&D\phi(\bar{u}) \\
&= -G_{(\alpha,\lambda)}(\bar{u},(\bar{\alpha},\bar{\lambda}))^{-1} G_u(\bar{u},(\bar{\alpha},\bar{\lambda})) \\
&= -\begin{pmatrix} DJ(\bar{u})DJ(\bar{u})^\top - \epsilon\epsilon^\top & -1_{k\times 1} \\ 1_{1\times k} & 0 \end{pmatrix}^{-1} \cdot \\
&\qquad \left(\begin{pmatrix} \bar{\alpha}^\top DJ(\bar{u})\nabla^2 J_1(\bar{u}) \\ \vdots \\ \bar{\alpha}^\top DJ(\bar{u})\nabla^2 J_k(\bar{u}) \\ 0_{1\times n} \end{pmatrix} + \begin{pmatrix} DJ(\bar{u}) \sum_{i=1}^k \bar{\alpha}_i \nabla^2 J_i(\bar{u}) \\ 0_{1\times n} \end{pmatrix} \right).
\end{aligned}$$

By applying the chain rule to (A.4), we obtain

$$\begin{aligned}
&D\varphi|_U(\bar{u}) \\
&= 2(DJ(\bar{u})^\top \bar{\alpha})^\top \sum_{i=1}^k \bar{\alpha}_i \nabla^2 J_i(\bar{u}) + \left(2(DJ(\bar{u})^\top \bar{\alpha})^\top DJ(\bar{u})^\top - 2(\bar{\alpha}^\top \epsilon)\epsilon^\top\right) D\phi_\alpha(\bar{u}) \\
&= 2(DJ(\bar{u})^\top \bar{\alpha})^\top \sum_{i=1}^k \bar{\alpha}_i \nabla^2 J_i(\bar{u}) + 2\bar{\alpha}^\top \left(DJ(\bar{u})DJ(\bar{u})^\top - \epsilon\epsilon^\top\right) D\phi_\alpha(\bar{u}) \\
&= 2(DJ(\bar{u})^\top \bar{\alpha})^\top \sum_{i=1}^k \bar{\alpha}_i \nabla^2 J_i(\bar{u}) + 2(\bar{\lambda},\ldots,\bar{\lambda}) D\phi_\alpha(\bar{u}). \qquad \text{(A.5)}
\end{aligned}$$

■

References

1. S. Banholzer, D. Beermann, and S. Volkwein. POD-Based Bicriterial Optimal Control by the Reference Point Method. *IFAC-PapersOnLine*, 49(8):210–215, 2016.
2. S. Banholzer, D. Beermann, and S. Volkwein. POD-Based Error Control for Reduced-Order Bicriterial PDE-Constrained Optimization. *Annual Reviews in Control*, 44:226–237, 2017.
3. D. Beermann, M. Dellnitz, S. Peitz, and S. Volkwein. POD-based multiobjective optimal control of PDEs with non-smooth objectives. In *Proceedings in Applied Mathematics and Mechanics (PAMM)*, pages 51–54, 2017.
4. D. Beermann, M. Dellnitz, S. Peitz, and S. Volkwein. Set-Oriented Multiobjective Optimal Control of PDEs using Proper Orthogonal Decomposition. In *Reduced-Order Modeling (ROM) for Simulation and Optimization*, pages 47–72. Springer, 2018.

5. A. Buffa, Y. Maday, A. T. Patera, C. Prudhomme, and G. Turinici. A priori convergence of the greedy algorithm for the parametrized reduced basis method. *ESAIM: Mathematical Modelling and Numerical Analysis*, 46(3):595–603, 2012.
6. T. Chugh, K. Sindhya, J. Hakanen, and K. Miettinen. A survey on handling computationally expensive multiobjective optimization problems with evolutionary algorithms. *Soft Computing*, pages 1–30, 2017.
7. C. A. Coello Coello, G. B. Lamont, and D. A. Van Veldhuizen. *Evolutionary Algorithms for Solving Multi-Objective Problems*, volume 2. Springer Science & Business Media, 2007.
8. M. Dellnitz, O. Schütze, and T. Hestermeyer. Covering Pareto Sets by Multilevel Subdivision Techniques. *Journal of Optimization Theory and Applications*, 124(1):113–136, Jan 2005.
9. M. Ehrgott. *Multicriteria optimization*. Springer Berlin Heidelberg New York, 2 edition, 2005.
10. B. Gebken, S. Peitz, and M. Dellnitz. On the hierarchical structure of Pareto critical sets. *Journal of Global Optimization*, 73(4):891–913, 2019.
11. M. A. Grepl and A. T. Patera. A posteriori error bounds for reduced-basis approximations of parametrized parabolic partial differential equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 39(1):157–181, 2005.
12. B. Haasdonk, J. Salomon, and B. Wohlmuth. A reduced basis method for the simulation of American options. In *Numerical Mathematics and Advanced Applications 2011*, pages 821–829, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.
13. J. Hesthaven, G. Rozza, and B. Stamm. Certified reduced basis methods for parametrized partial differential equations. *SpringerBriefs in Mathematics*, 2016.
14. C. Hillermeier. *Nonlinear Multiobjective Optimization: A Generalized Homotopy Approach*. Birkhäuser Basel, 2001.
15. M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints*. Springer, 2009.
16. L. Iapichino, S. Trenz, and S. Volkwein. Multiobjective optimal control of semilinear parabolic problems using POD. In B. Karasözen, M. Manguoglu, M. Tezer-Sezgin, S. Goktepe, and Ö. Ugur, editors, *Numerical Mathematics and Advanced Applications (ENUMATH 2015)*, pages 389–397. Springer, 2016.
17. L. Iapichino, S. Ulbrich, and S. Volkwein. Multiobjective PDE-Constrained Optimization Using the Reduced-Basis Method. *Advances in Computational Mathematics*, 43(5):945–972, 2017.
18. K. Kunisch and S. Volkwein. Galerkin proper orthogonal decomposition methods for parabolic problems. *Numerische Mathematik*, 90(1):117–148, 2001.
19. J. Lee. *Introduction to Smooth Manifolds*. Springer-Verlag New York, 2012.
20. M. Ohlberger and F. Schindler. Non-conforming localized model reduction with online enrichment: Towards optimal complexity in PDE constrained optimization. In *Finite Volumes for Complex Applications VIII - Hyperbolic, Elliptic and Parabolic Problems*, pages 357–365. Springer International Publishing, 2017.
21. S. Peitz and M. Dellnitz. *Gradient-Based Multiobjective Optimization with Uncertainties*, pages 159–182. Springer International Publishing, 2017.
22. S. Peitz and M. Dellnitz. A Survey of Recent Trends in Multiobjective Optimal Control Surrogate Models, Feedback Control and Objective Reduction. *Mathematical and Computational Applications*, 23(2), 2018.
23. S. Peitz, S. Ober-Blöbaum, and M. Dellnitz. Multiobjective Optimal Control Methods for the Navier-Stokes Equations Using Reduced Order Modeling. *Acta Applicandae Mathematicae*, 161(1):171–199, 2019.
24. S. Peitz, K. Schäfer, S. Ober-Blöbaum, J. Eckstein, U. Köhler, and M. Dellnitz. A Multiobjective MPC Approach for Autonomously Driven Electric Vehicles. *IFAC PapersOnLine*, 50(1):8674–8679, 2017.
25. A. Quarteroni, A. Manoni, and F. Negri. *Reduced Basis Methods for Partial Differential Equations*. Springer, 2016.

26. G. Rozza, D. B. P. Huynh, and A. T. Patera. Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. *Archives of Computational Methods in Engineering*, 15(3):1, 2007.
27. O. Schütze, A. Dell'Aere, and M. Dellnitz. On Continuation Methods for the Numerical Treatment of Multi-Objective Optimization Problems. In *Practical Approaches to Multi-Objective Optimization*, number 04461 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2005. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.
28. O. Schtze, O. Cuate, A. Martn, S. Peitz, and M. Dellnitz. Pareto explorer: a global/local exploration tool for many-objective optimization problems. *Engineering Optimization*, pages 1–24, 05 2019.
29. L. Sirovich. Turbulence and the dynamics of coherent structures part I: coherent structures. *Quarterly of Applied Mathematics*, XLV(3):561–571, 1987.
30. M. Tabatabaei, J. Hakanen, M. Hartikainen, K. Miettinen, and K. Sindhya. A survey on handling computationally expensive multiobjective optimization problems using surrogates: non-nature inspired methods. *Structural and Multidisciplinary Optimization*, 52(1):1–25, 2015.

Analysis and Solution Methods for Bilevel Optimal Control Problems

Stephan Dempe, Felix Harder, Patrick Mehlitz, and Gerd Wachsmuth

Abstract In this chapter, we first provide an overview of literature addressing the so-called bilevel optimal control problems which are hierarchical optimization problems with two decision makers where at least one of them has to solve an optimal control problem of ODEs or PDEs. By means of two examples from inverse PDE control, we demonstrate how problem-tailored regularization and relaxation approaches can be used to infer necessary optimality conditions in bilevel optimal control. Finally, we present an algorithm which can be used to solve a class of bilevel optimal control problems to global optimality.

Keywords Bilevel optimal control · Global optimization · Inverse optimal control · Optimality conditions · Solution algorithm

Mathematics Subject Classification (2020) Primary 49K20, 49M20; Secondary 90C26, 90C31

1 Introduction

A bilevel programming problem is an optimization problem (the so-called upper level problem) whose objective functional and feasible region depend implicitly on the solution set of a given parametric mathematical program (the so-called lower

S. Dempe
TU Bergakademie Freiberg, Fakultät für Mathematik und Informatik, Freiberg, Germany
e-mail: dempe@math.tu-freiberg.de

F. Harder · P. Mehlitz · G. Wachsmuth (✉)
BTU Cottbus-Senftenberg, Cottbus, Germany
e-mail: felix.harder@b-tu.de; patrick.mehlitz@b-tu.de; gerd.wachsmuth@b-tu.de

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_4

level problem). In abstract form, the upper level problem can be stated as

$$\begin{aligned} J(w,z) &\to \min_{w,z} \\ w &\in W \\ z &\in \Psi(w), \end{aligned}$$

where, for each w, $\Psi(w)$ denotes the solution set of the lower level problem

$$\begin{aligned} j(w,z) &\to \min_{z} \\ z &\in Z(w). \end{aligned}$$

Above, $J, j\colon \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ are given functionals on the product of two Banach spaces $\mathcal{W}$ and $\mathcal{Z}$, $W \subset \mathcal{W}$ is the so-called upper level feasible set, and the set-valued mapping $Z\colon \mathcal{W} \rightrightarrows \mathcal{Z}$ assigns to each $w \in \mathcal{W}$ the set of lower level feasible points $Z(w) \subset \mathcal{Z}$. The decision order in bilevel programming is given as follows. First, the upper level decision maker, the so-called leader, chooses an instance $w \in W$. Afterwards, the lower level decision maker, the so-called follower, is capable of computing the solution set $\Psi(w)$. Finally, the leader evaluates his objective functional. Noting that the leader is allowed to choose arbitrary elements of $\Psi(w)$, the above formulation of a bilevel programming problem reflects a cooperative behavior of both decision makers which is related to the so-called optimistic approach of bilevel optimization. Naturally, bilevel programming problems are not convex even if all the data is convex. Furthermore, transformations which convert the hierarchical model into a single-level program turn out to invoke nonsmoothness and irregularity which is why bilevel programs are generally challenging from the theoretical and numerical point of view. On the other hand, several real-world applications e.g. from chemical engineering, road pricing, gas shipment, and parameter reconstruction naturally result in mathematical models of bilevel structure. A detailed introduction to the topic of bilevel programming can be found in [2, 9, 13, 42] while a comprehensive overview of existing literature is presented in [10]. The latter comprises a list of more than 1350 published books, PhD-theses, and research articles concerned with bilevel optimization.

In bilevel optimal control (BOC), bilevel optimization problems are considered where at least one of the decision makers has to solve an optimal control problem of ordinary or partial differential equations (ODEs and PDEs) , see [23, 29, 43, 44] for an introduction to optimal control. Therefore, models from BOC generally unite the intrinsic difficulties of bilevel programming *and* optimal control, see [32]. In [3, 4], the authors discuss the situation where the upper level decision maker has to solve an optimal control problem of ODEs while certain penalty costs resulting from the associated terminal state are computed at the lower level stage. The investigation of such models is motivated by underlying applications from gas balancing in energy networks. The estimation or reconstruction of parameters in optimal control problems of ODEs or PDEs is considered theoretically in [12, 19, 20, 24, 46, 47].

These so-called *inverse* optimal control problems, where only the lower level decision maker has to solve an optimal control problem, arise e.g. in the context of human locomotion, multi-agent scheduling, or aircraft control, see [1, 14, 36, 40]. Finally, it is also possible that both decision makers need to face optimal control problems at their respective decision level, see [33, 39]. Problems of this type model e.g. the time-dependent coupling of container crane movements, see [27].

In order to tackle bilevel optimal control problems theoretically or numerically, one generally aims for the elimination of the hierarchical structure first. In the literature on bilevel programming, three corresponding approaches are suggested for that purpose. First, whenever the lower level solution is uniquely determined for each choice of the parameter, one could exploit the implicitly given associated solution operator in order to plug the lower level solution into the upper level problem, see [19, 30]. Second, one can replace the lower level problem by means of necessary and sufficient optimality conditions of Karush–Kuhn–Tucker (KKT) type, see [24, 33], which results in a *mathematical problem with complementarity constraints* (MPCC) in function spaces, see e.g. [5, 8, 16, 17, 33, 34], whenever the lower level problem comprises (generalized) inequality constraints. Third, one could exploit the implicitly given optimal value function of the lower level problem, which assigns to each parameter the associated lower level globally minimal function value, in order to formulate a single-level surrogate problem, see [3, 4, 12, 39].

Here, we want to demonstrate by means of two examples from inverse optimal PDE control how optimality conditions in bilevel optimal control can be derived via a regularization and relaxation approach, respectively. Furthermore, we present an algorithm which solves a specific class of inverse optimal control problems to global optimality. All these considerations are based on [12, 19].

2 Two Example Problems

In this section, we introduce two example problems from inverse optimal control where finitely many real parameters in the objective functional of an optimal control problem need to be reconstructed from measurements of optimal state and control.

For a bounded domain $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, with Lipschitz boundary $\Gamma \subset \mathbb{R}^d$, functions $u_a, u_b \in L^\infty(\Omega)$ satisfying $u_a(x) < u_b(x)$ for almost all $x \in \Omega$, a regularization parameter $\sigma > 0$, and a parameter vector $\alpha \in \mathbb{R}^n$, we consider the lower level parametric optimal control problem

$$
\begin{aligned}
f(\alpha, y, u) + \tfrac{\sigma}{2}\|u\|^2_{L^2(\Omega)} &\to \min_{y,u} \\
y \in H_0^1(\Omega), \quad u \in L^2(\Omega), & \\
-\Delta y &= u, \\
u_a \le u &\le u_b \qquad \text{a.e. on } \Omega,
\end{aligned}
\tag{$\mathrm{P}_f(\alpha)$}
$$

where $f\colon \mathbb{R}^n \times H_0^1(\Omega) \times L^2(\Omega) \to \mathbb{R}$ is a twice differentiable functional which is jointly convex in its second and third argument. Below, we will specify two typical choices for f. The equation $-\Delta y = u$ has to be understood in weak sense, i.e. in $H^{-1}(\Omega) := H_0^1(\Omega)^\star$. For later use, we introduce the *set of admissible controls* $U_{\text{ad}} \subset L^2(\Omega)$ via

$$U_{\text{ad}} := \{u \in L^2(\Omega) \,|\, u_a(x) \le u(x) \le u_b(x) \text{ f.a.a. } x \in \Omega\}.$$

Let $(y_{\text{o}}, u_{\text{o}}) \in H_0^1(\Omega) \times L^2(\Omega)$ be a pair of *observed* optimal state and control of ($\text{P}_f(\alpha)$) for an unknown parameter vector α. Then it is reasonable to consider the superordinate upper level problem

$$\begin{aligned} \tfrac{1}{2}\|y - y_{\text{o}}\|^2_{L^2(\Omega)} + \tfrac{\vartheta}{2}\|u - u_{\text{o}}\|^2_{L^2(\Omega)} \;&\to\; \min_{\alpha,y,u} \\ \alpha \in \mathbb{R}^n, \quad y \in H_0^1(\Omega), \quad u \in L^2(\Omega), \\ \alpha \in \Lambda, \\ (y,u) \in \Psi_f(\alpha). \end{aligned} \tag{IOC$_f$}$$

Noting that we aim for the identification of the unknown vector α in ($\text{P}_f(\alpha)$), (IOC$_f$) may be referred to as an inverse optimal control problem. Above, $\vartheta \ge 0$ is a weight parameter, $\Lambda \subset \mathbb{R}^n$ represents the standard simplex given by

$$\Lambda := \{\alpha \in \mathbb{R}^n \,|\, \alpha \ge 0,\ \textstyle\sum_{i=1}^n \alpha_i = 1\},$$

and $\Psi_f(\alpha) \subset H_0^1(\Omega) \times L^2(\Omega)$ denotes the solution set of ($\text{P}_f(\alpha)$) associated with the parameter α. We would like to mention that replacing Λ by any other polytope (a bounded intersection of finitely many halfspaces) $\tilde{\Lambda} \subset \mathbb{R}^n_+$ does not change the subsequently stated theory. Similarly, the objective function of (IOC$_f$) can be replaced by any continuously differentiable, convex function.

We focus our attention on two possible choices for the function f. First, we investigate the function $f_1\colon \mathbb{R}^n \times H_0^1(\Omega) \times L^2(\Omega) \to \mathbb{R}$ given by

$$f_1(\alpha, y, u) := \textstyle\sum_{i=1}^n \alpha_i\, h_i(y,u), \tag{2.1}$$

where $h_1, \ldots, h_n\colon H_0^1(\Omega) \times L^2(\Omega) \to \mathbb{R}$ are given convex functions satisfying additional assumptions that are specified in Sect. 3.2.1. In the associated problem (IOC$_{f_1}$), we aim to restore the precise form of the lower level objective function which is given as an unknown convex combination of given reference functionals. Second, let $f_2\colon \mathbb{R}^n \times H_0^1(\Omega) \times L^2(\Omega) \to \mathbb{R}$ be given by

$$f_2(\alpha, y, u) := \tfrac{1}{2}\left\| y - \textstyle\sum_{i=1}^n \alpha_i\, g_i \right\|^2_{L^2(\Omega)}, \tag{2.2}$$

where $g_1, \dots, g_n \in L^2(\Omega)$ are given shape functions. In this scenario, (IOC_{f_2}) is used to reconstruct the desired state within a classical tracking-type objective functional.

Remark 2.1 We note that for fixed parameter $\alpha \in \Lambda$, ($\mathrm{P}_{f_1}(\alpha)$) and ($\mathrm{P}_{f_2}(\alpha)$) are convex optimization problems which naturally means that the minimization is performed globally at the lower level stage. In classical bilevel programming, the lower level problem always has to be solved to global optimality even if the latter problem is *not* convex, see [9]. Particularly, locally optimal solutions of the lower level problem which are not globally optimal do not yield feasible points of the underlying bilevel programming problem. As a result, numerical bilevel programming with non-convex lower level problem is rather challenging since classical methods, which generally only can guarantee local optimality at the lower level stage, may turn out to compute infeasible points.

Remark 2.2 It is not difficult to show that the optimization problems ($\mathrm{P}_{f_1}(\alpha)$) and ($\mathrm{P}_{f_2}(\alpha)$) possess uniquely determined solutions for fixed $\alpha \in \Lambda$. Following arguments provided in [19] and [12], the associated single-valued solution operators are continuous as mappings from Λ to $H_0^1(\Omega) \times L^2(\Omega)$. As a consequence, both programs (IOC_{f_1}) and (IOC_{f_2}) possess a global minimizer.

Due to the above remark, it is reasonable to introduce mappings $\psi^y \colon \Lambda \to H_0^1(\Omega)$ and $\psi^u \colon \Lambda \to L^2(\Omega)$ such that $(\psi^y(\alpha), \psi^u(\alpha))$ is the uniquely determined solution of ($\mathrm{P}_f(\alpha)$) for each parameter $\alpha \in \Lambda$ and (depending on the context) $f \in \{f_1, f_2\}$.

3 Optimality Conditions

3.1 Definition of Optimality Systems for (IOC_f)

Since solutions of the lower level problem can be characterized by the KKT conditions of the lower level problem where the respective Lagrange multipliers are uniquely determined, (IOC_f) is equivalent to

$$
\begin{aligned}
&\tfrac{1}{2}\|y - y_\mathrm{o}\|^2_{L^2(\Omega)} + \tfrac{\vartheta}{2}\|u - u_\mathrm{o}\|^2_{L^2(\Omega)} \to \min_{\alpha, y, u, p, \lambda} \\
&\alpha \in \mathbb{R}^n, \quad y, p \in H_0^1(\Omega), \quad u, \lambda \in L^2(\Omega), \\
&\qquad\qquad \alpha \in \Lambda, \\
&\qquad\qquad D_y f(\alpha, y, u) - \Delta p = 0, \\
&\qquad\qquad D_u f(\alpha, y, u) + \sigma u - p + \lambda = 0, \\
&\qquad\qquad -\Delta y = u, \\
&\qquad\qquad (u, \lambda) \in \mathrm{gph}\, \mathcal{N}_{U_\mathrm{ad}}.
\end{aligned}
\tag{3.1}
$$

Here, $\operatorname{gph} \mathcal{N}_{U_{\mathrm{ad}}}$ denotes the graph of the set-valued mapping $\mathcal{N}_{U_{\mathrm{ad}}} \colon L^2(\Omega) \rightrightarrows L^2(\Omega)$ which assigns to each $u \in L^2(\Omega)$ the associated normal cone $\mathcal{N}_{U_{\mathrm{ad}}}(u)$ in the sense of convex analysis. Thus, this set possesses the representation

$$\operatorname{gph} \mathcal{N}_{U_{\mathrm{ad}}} = \left\{ (u, \lambda) \in U_{\mathrm{ad}} \times L^2(\Omega) \;\middle|\; \begin{array}{ll} \lambda \geq 0 & \text{a.e. on } \{u_a < u\} \\ \lambda \leq 0 & \text{a.e. on } \{u < u_b\} \end{array} \right\}.$$

Note that derivatives w.r.t. α do not appear in the constraints of (3.1), since α is not an optimization variable of the lower level problem. If we consider the MPCC-Lagrangian of (3.1) and the roots of its partial derivatives, we arrive (after some substitutions and using the pointwise structure of U_{ad} and $\mathcal{N}_{U_{\mathrm{ad}}}$) at the system in the following definition.

Definition 3.1 A feasible point $(\bar{\alpha}, \bar{y}, \bar{u}) \in \mathbb{R}^n \times H_0^1(\Omega) \times L^2(\Omega)$ of (IOC$_f$) is said to be weakly stationary (W-stationary) for (IOC$_f$) whenever there exist multipliers $\bar{p} \in H_0^1(\Omega)$, $\bar{\lambda} \in L^2(\Omega)$, $\bar{z} \in \mathbb{R}^n$, $\bar{\mu}, \bar{\rho} \in H_0^1(\Omega)$, and $\bar{w}, \bar{\xi} \in L^2(\Omega)$ which satisfy

$$0 = D^2_{y\alpha} f(\bar{\alpha}, \bar{y}, \bar{u})^\star(\bar{\mu}) + D^2_{u\alpha} f(\bar{\alpha}, \bar{y}, \bar{u})^\star(\bar{w}) + \bar{z}, \tag{3.2a}$$

$$0 = \bar{y} - y_{\mathrm{o}} + D^2_{yy} f(\bar{\alpha}, \bar{y}, \bar{u})^\star(\bar{\mu}) + D^2_{uy} f(\bar{\alpha}, \bar{y}, \bar{u})^\star(\bar{w}) - \Delta\bar{\rho}, \tag{3.2b}$$

$$0 = \vartheta(\bar{u} - u_{\mathrm{o}}) + D^2_{yu} f(\bar{\alpha}, \bar{y}, \bar{u})^\star(\bar{\mu}) + D^2_{uu} f(\bar{\alpha}, \bar{y}, \bar{u})^\star(\bar{w}) + \sigma\bar{w} - \bar{\rho} + \bar{\xi}, \tag{3.2c}$$

$$0 = -\Delta\bar{\mu} - \bar{w}, \tag{3.2d}$$

$$\bar{z} \in \mathcal{N}_\Lambda(\bar{\alpha}), \tag{3.2e}$$

$$0 = D_y f(\bar{\alpha}, \bar{y}, \bar{u}) - \Delta\bar{p}, \tag{3.2f}$$

$$0 = D_u f(\bar{\alpha}, \bar{y}, \bar{u}) + \sigma\bar{u} - \bar{p} + \bar{\lambda}, \tag{3.2g}$$

$$\bar{\lambda} \geq 0 \quad \text{a.e. on } I^{a+}(\bar{u}), \tag{3.2h}$$

$$\bar{\lambda} \leq 0 \quad \text{a.e. on } I^{b-}(\bar{u}), \tag{3.2i}$$

$$\bar{\xi} = 0 \quad \text{a.e. on } I^{a+}(\bar{u}) \cap I^{b-}(\bar{u}), \tag{3.2j}$$

$$\bar{w} = 0 \quad \text{a.e. on } \{\bar{\lambda} \neq 0\}. \tag{3.2k}$$

Above, we used

$$I^{a+}(\bar{u}) := \{u_a < \bar{u}\}, \qquad I^{b-}(\bar{u}) := \{\bar{u} < u_b\}.$$

If these multipliers additionally satisfy the condition

$$\bar{\xi}\bar{w} \geq 0 \quad \text{a.e. on } \Omega, \tag{3.3}$$

then $(\bar{\alpha}, \bar{y}, \bar{u})$ is referred to as Clarke-stationary (C-stationary). For pointwise Mordukhovich-stationarity (pM-stationarity) of $(\bar{\alpha}, \bar{y}, \bar{u})$, we require that the conditions

$$\begin{aligned} &\bar{\xi}\bar{w} = 0 \ \vee \ (\bar{\xi} < 0 \ \wedge \ \bar{w} < 0) \quad \text{a.e. on } \{\bar{\lambda} = 0\} \cap \{\bar{u} = u_a\}, \\ &\bar{\xi}\bar{w} = 0 \ \vee \ (\bar{\xi} > 0 \ \wedge \ \bar{w} > 0) \quad \text{a.e. on } \{\bar{\lambda} = 0\} \cap \{\bar{u} = u_b\} \end{aligned} \tag{3.4}$$

hold for the multipliers which solve the W-stationarity system. Finally, $(\bar{\alpha}, \bar{y}, \bar{u})$ is said to be strongly stationary (S-stationary) if it is W-stationary and the conditions

$$\begin{aligned} &\bar{\xi} \leq 0 \ \wedge \ \bar{w} \leq 0 \quad \text{a.e. on } \{\bar{\lambda} = 0\} \cap \{\bar{u} = u_a\}, \\ &\bar{\xi} \geq 0 \ \wedge \ \bar{w} \geq 0 \quad \text{a.e. on } \{\bar{\lambda} = 0\} \cap \{\bar{u} = u_b\} \end{aligned} \tag{3.5}$$

hold for the associated multipliers.

Note that (3.2f)–(3.2i) just provide the lower level optimality conditions which guarantee that $(\bar{\alpha}, \bar{y}, \bar{u}, \bar{p}, \bar{\lambda})$ is a feasible point of the complementarity-constrained optimization problem (3.1).

By definition, the subsequently stated relations hold between the stationarity notions introduced above for each feasible point of (IOC$_f$):

$$\text{S-stationary} \implies \text{pM-stationary} \implies \text{C-stationary} \implies \text{W-stationary}.$$

Clearly, this is not surprising since the additional conditions (3.3), (3.4), and (3.5) which characterize C-, pM-, and S-stationarity, respectively, were obtained by transferring the associated counterparts from finite-dimensional complementarity programming pointwise to the setting at hand. In Sect. 3.4, we will comment on the observation that this approach is not compatible with the underlying tools of variational analysis when Mordukhovich's stationarity concept is investigated. C-stationarity-type systems turn out to provide reliable first-order necessary optimality conditions for different classes of equilibrium problems in function spaces, see e.g. [12, 17, 19, 21, 22], while it is an open question whether this holds for associated Mordukhovich-stationarity-type systems as well, see e.g. [45]. Second-order sufficient optimality conditions for such problems are classically based on S-stationary points, see e.g. [7, 28].

In order to specify the system (3.2) to the settings in (IOC$_{f_1}$) and (IOC$_{f_2}$), the derivatives of f_1 and f_2 from (2.1) and (2.2) have to be computed, respectively. For f_1, we have

$$\begin{aligned} D^2_{y\alpha} f_1(\alpha, y, u)^\star(\mu) &= (D_y h_i(y, u)(\mu))_{i=1}^n, \\ D^2_{u\alpha} f_1(\alpha, y, u)^\star(w) &= (D_u h_i(y, u)(w))_{i=1}^n \end{aligned}$$

for each $\mu \in H_0^1(\Omega)$ and $w \in L^2(\Omega)$. The expressions for $D^2_{yy} f_1$, $D^2_{yu} f_1$, $D^2_{uy} f_1$, $D^2_{uu} f_1$, $D_y f_1$, and $D_u f_1$ can be obtained by simple linearity. Next, let us specify the appearing derivatives of the function f_2. First, we note that f_2 does not depend on u which is why the derivatives $D_u f_2$, $D^2_{u\alpha} f_2$, $D^2_{uy} f_2$, $D^2_{uu} f_2$, and $D^2_{yu} f_2$ vanish. For the remaining derivatives, we obtain

$$\begin{aligned} D_y f_2(\alpha, y, u) &= y - \textstyle\sum_{i=1}^n \alpha_i g_i, \\ D^2_{y\alpha} f_2(\alpha, y, u)^\star(\mu) &= (-\langle g_i, \mu \rangle_{L^2(\Omega)})_{i=1}^n, \\ D^2_{yy} f_2(\alpha, y, u)^\star(\mu) &= \mu \end{aligned}$$

for each $\mu \in H_0^1(\Omega)$.

3.2 Regularization Approach

In this subsection, we discuss how optimality conditions for the problem (IOC$_{f_1}$) can be achieved by penalizing the lower level control constraints in the objective function of the lower level problem. Our goal is to derive (under reasonable assumptions) C-stationarity of all local minimizers associated with (IOC$_{f_1}$).

The results in this subsection are based on [19].

3.2.1 Assumptions and Properties of the Lower Level Problem

In order to proceed with our analysis, some assumptions on the functions h_i are necessary. For each $i \in \{1, \dots, n\}$, we require that there exists a function $\theta^{(i)} \colon \Omega \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that

- $h_i(y, u) = \int_\Omega \theta^{(i)}(x, y(x), u(x)) \mathrm{d}x$ for all $y \in H_0^1(\Omega)$ and $u \in L^2(\Omega)$,
- $\theta^{(i)}(x, \cdot, \cdot) \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is convex for all $x \in \Omega$,
- $\theta^{(i)}(x, \cdot, \cdot) \in C^2(\mathbb{R}^2)$ for all $x \in \Omega$,
- $D^2_{yy}\theta^{(i)}(\cdot, \cdot, \cdot)$, $D^2_{yu}\theta^{(i)}(\cdot, \cdot, \cdot)$, $D^2_{uu}\theta^{(i)}(\cdot, \cdot, \cdot) \colon \Omega \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ are uniformly bounded,
- $\theta^{(i)}(\cdot, y, u) \colon \Omega \to \mathbb{R}$ is measurable for all fixed $y, u \in \mathbb{R}$, and
- $\theta^{(i)}(\cdot, 0, 0)$, $D_y\theta^{(i)}(\cdot, 0, 0)$, $D_u\theta^{(i)}(\cdot, 0, 0) \in L^\infty(\Omega)$.

These conditions impose a pointwise structure on the functions h_i. However, due to the appearance of the Laplace operator, we are faced with non-pointwise effects for the optimal control problem (P$_{f_1}(\alpha)$). We mention that weaker but more complicated conditions on the functions h_i are available in [19].

In the subsequent lemma, we discuss the continuity and smoothness properties of the lower level solution maps $\psi^y \colon \Lambda \to H_0^1(\Omega)$ and $\psi^u \colon \Lambda \to L^2(\Omega)$ associated with (P$_{f_1}(\alpha)$). The proofs for these results can be found in [19].

Lemma 3.2

1. *The solution maps* $\psi^u\colon \Lambda \to L^2(\Omega)$ *and* $\psi^y\colon \Lambda \to H_0^1(\Omega)$ *are Lipschitz continuous.*
2. *The solution maps* $\psi^u\colon \Lambda \to L^2(\Omega)$ *and* $\psi^y\colon \Lambda \to H_0^1(\Omega)$ *are directionally differentiable.*

3.2.2 C-Stationarity for Local Minimizers

The upcoming result can be shown by exploiting the second statement of Lemma 3.2. We refer to [19, Theorem 3.2] for details and a proof.

Theorem 3.3 *Consider the unconstrained case* $U_{\mathrm{ad}} := L^2(\Omega)$. *Then each local minimizer of* (IOC_{f_1}) *is S-stationary.*

This result is useful for obtaining C-stationarity of local solutions in the presence of control constraints. In order to do that, for a penalty parameter $k > 0$, we consider the regularization

$$\begin{aligned} f_1(\alpha, y, u) + \tfrac{\sigma}{2}\|u\|^2_{L^2(\Omega)} + k\Pi(u) &\to \min_{y,u} \\ y \in H_0^1(\Omega), \quad u \in L^2(\Omega), & \qquad (\mathrm{P}_{f_1}(k,\alpha)) \\ -\Delta y &= u \end{aligned}$$

of ($\mathrm{P}_{f_1}(\alpha)$). Here, the function $\Pi\colon L^2(\Omega) \to \mathbb{R}$ is a function that penalizes the control constraints $u \in U_{\mathrm{ad}}$ and is defined as

$$\forall u \in L^2(\Omega)\colon \quad \Pi(u) := \int_\Omega \big[\pi(u_a(x) - u(x)) + \pi(u(x) - u_b(x))\big]\mathrm{d}x,$$

where $\pi : \mathbb{R} \to \mathbb{R}$ is given by

$$\forall s \in \mathbb{R}\colon \quad \pi(s) := \begin{cases} 0 & s \le 0, \\ 2s^3 - s^4 & 0 < s < 1, \\ 2s - 1 & s \ge 1. \end{cases}$$

Note that π is twice continuously differentiable and that its second-order derivative is bounded. Thus, $k\Pi$ satisfies the assumptions in Sect. 3.2.1. Due to the absence of control constraints, the solution operators $\psi_k^y\colon \Lambda \to H_0^1(\Omega)$ and $\psi_k^u\colon \Lambda \to L^2(\Omega)$ of the problem ($\mathrm{P}_{f_1}(k,\alpha)$) are differentiable. Thus, they can be considered as a regularization of the (in general non-differentiable) solution operators ψ^y and ψ^u.

The regularized bilevel problem can be written as

$$\begin{aligned} \tfrac{1}{2}\|\psi_k^y(\alpha) - y_0\|^2_{L^2(\Omega)} + \tfrac{\vartheta}{2}\|\psi_k^u(\alpha) - u_0\|^2_{L^2(\Omega)} &\to \min_{\alpha} \\ \alpha &\in \Lambda. \end{aligned} \tag{IOC$_{f_1}(k)$}$$

Let us assume that $(\bar\alpha, \bar y, \bar u)$ is a unique global minimizer of (IOC$_{f_1}$). We denote the global solution of (IOC$_{f_1}(k)$) by α_k. Since there are no control constraints in (P$_{f_1}(k, \alpha)$), we can apply Theorem 3.3 to (IOC$_{f_1}(k)$). Then we get multipliers $p_k, \mu_k, \rho_k \in H_0^1(\Omega)$, $z_k \in \mathbb{R}^n$, $\lambda_k, \xi_k, w_k \in L^2(\Omega)$ that satisfy (3.2). Note that we have $\lambda_k = \xi_k = 0$ since $I^{a+}(\psi_k^u(\alpha_k)) = I^{b-}(\psi_k^u(\alpha_k)) = \Omega$ holds. The idea for obtaining C-stationarity of $(\bar\alpha, \bar y, \bar u)$ is to observe the behavior of this stationarity system for (IOC$_{f_1}(k)$) as $k \to \infty$.

The result is given in the following theorem.

Theorem 3.4 *There exist $\bar p \in H_0^1(\Omega)$, $\bar\lambda \in L^2(\Omega)$ such that the convergences*

$$\alpha_k \to \bar\alpha \quad \text{in } \mathbb{R}^n, \tag{3.6a}$$

$$\psi_k^y(\alpha_k) \to \bar y \quad \text{in } H_0^1(\Omega), \tag{3.6b}$$

$$\psi_k^u(\alpha_k) \to \bar u \quad \text{in } L^2(\Omega), \tag{3.6c}$$

$$p_k \to \bar p \quad \text{in } H_0^1(\Omega), \tag{3.6d}$$

$$k\, D\Pi(\psi_k^u(\alpha_k)) \to \bar\lambda \quad \text{in } L^2(\Omega) \tag{3.6e}$$

hold for $k \to \infty$. Additionally, there exist $\bar\mu, \bar\rho \in H_0^1(\Omega)$, $\bar\xi, \bar w \in L^2(\Omega)$, $\bar z \in \mathbb{R}^n$ such that we obtain the convergences

$$z_k \to \bar z \quad \text{in } \mathbb{R}^n, \tag{3.6f}$$

$$\mu_k \to \bar\mu \quad \text{in } H_0^1(\Omega), \tag{3.6g}$$

$$\rho_k \to \bar\rho \quad \text{in } H_0^1(\Omega), \tag{3.6h}$$

$$w_k \rightharpoonup \bar w \quad \text{in } L^2(\Omega), \tag{3.6i}$$

$$k\, D^2\Pi(\psi_k^u(\alpha_k))w_k \rightharpoonup \bar\xi \quad \text{in } L^2(\Omega) \tag{3.6j}$$

along a subsequence. The limits satisfy (3.2) *and* (3.3)*.*

Let us give a sketch of the proof. As a first step, one can show the convergences (3.6a), (3.6b), and (3.6c), mostly using standard methods, see [19, Lemmas 4.1 and 4.2]. Next, we can use continuity properties of $D_{(y,u)}f(\alpha, y, u)$ to obtain (3.6d), (3.6e), with the limits satisfying (3.2f) and (3.2g). The complementarities (3.2h) and (3.2i) follow from pointwise arguments, see [19, Lemma 4.3]. Next, one can show that $\{w_k\}_{k\in\mathbb{N}}$ is bounded in $L^2(\Omega)$ so that we can conclude (3.6i).

Then the system of S-stationarity for ($\mathrm{IOC}_{f_1}(k)$) can be used to obtain the convergences (3.6f), (3.6g), (3.6h), and (3.6j) as well as the conditions (3.2a)–(3.2e). We refer to [19, Lemma 4.4] for details. To complete the system of W-stationarity, the conditions (3.2j), (3.2k) need to be shown. This can be done with the help of Egorov's theorem and (3.6e), (3.6j), see [19, Lemma 4.5, Lemma 4.6]. Finally, after some calculations and using the nonnegativity of $D^2_{uu} f(\alpha_k, \psi^u_k(\alpha_k), \psi^y_k(\alpha_k))$, it turns out that (3.3) also holds, see [19, Lemma 4.7].

In order to generalize Theorem 3.4 from unique global solutions to local minimizers of (IOC_{f_1}), we can utilize a standard localization argument whose proof is sketched below.

Theorem 3.5 *Let* $(\bar\alpha, \bar y, \bar u) \in \mathbb{R}^n \times H^1_0(\Omega) \times L^2(\Omega)$ *be a local minimizer of* (IOC_f). *Then it is a C-stationary point of this program.*

We just present the idea of a proof here. By assumption, $\bar\alpha$ is a local minimizer of

$$\min_\alpha \left\{ \tfrac{1}{2}\|\psi^y(\alpha) - y_\mathrm{o}\|^2_{L^2(\Omega)} + \tfrac{\vartheta}{2}\|\psi^u(\alpha) - u_\mathrm{o}\|^2_{L^2(\Omega)} \,\middle|\, \alpha \in \Lambda \right\}.$$

Hence, we find some neighborhood $U \subset \mathbb{R}^n$ of $\bar\alpha$ such that $(\bar\alpha, \bar y, \bar u)$ is the unique global minimizer of the inverse optimal control problem

$$\begin{aligned} \tfrac{1}{2}\|\alpha - \bar\alpha\|^2_{\mathbb{R}^n} + \tfrac{1}{2}\|y - y_\mathrm{o}\|^2_{L^2(\Omega)} + \tfrac{\vartheta}{2}\|u - u_\mathrm{o}\|^2_{L^2(\Omega)} &\to \min_{\alpha,y,u} \\ \alpha \in \mathbb{R}^n, \quad y \in H^1_0(\Omega), \quad u &\in L^2(\Omega), \\ \alpha &\in \Lambda \cap U, \\ (y,u) &\in \Psi_{f_1}(\alpha). \end{aligned}$$

Noting that the derivative of $\alpha \mapsto \frac{1}{2}\|\alpha - \bar\alpha\|^2_{\mathbb{R}^n}$ vanishes at $\bar\alpha$ while we have $\mathcal{N}_{\Lambda\cap U}(\bar\alpha) = \mathcal{N}_\Lambda(\bar\alpha)$, the above considerations can be used to show that $(\bar\alpha, \bar y, \bar u)$ is indeed C-stationary for (IOC_{f_1}).

3.3 Relaxation Approach

Here, we illustrate how necessary optimality conditions for (IOC_{f_2}) can be derived. These considerations are based on [12]. We are going to strike a path which is essentially different from the one which was used in Sect. 3.2 and which exploits the so-called *optimal value function* $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ of ($\mathrm{P}_{f_2}(\alpha)$). This function is defined via

$$\forall \alpha \in \mathbb{R}^n\colon \quad \varphi(\alpha) := \min_{y,u} \left\{ f_2(\alpha, y, u) + \tfrac{\sigma}{2}\|u\|^2_{L^2(\Omega)} \,\middle|\, -\Delta y = u,\ u \in U_\mathrm{ad} \right\},$$

i.e. φ assigns to each parameter vector $\alpha \in \mathbb{R}^n$ the globally minimal objective value of ($P_{f_2}(\alpha)$). Due to Remark 2.2, there are continuous mappings $\psi^y \colon \Lambda \to H_0^1(\Omega)$ and $\psi^u \colon \Lambda \to L^2(\Omega)$ such that $(\psi^y(\alpha), \psi^u(\alpha))$ is the unique solution of ($P_{f_2}(\alpha)$) for each $\alpha \in \Lambda$. Thus, we have $\varphi(\alpha) = f_2(\alpha, \psi^y(\alpha), \psi^u(\alpha)) + \frac{\sigma}{2}\|\psi^u(\alpha)\|^2_{L^2(\Omega)}$ for each $\alpha \in \Lambda$ which shows that φ is continuous on Λ. It is easy to see that the domains of ψ^y and ψ^u can be extended to the whole space $\mathbb{R}^n$ without losing continuity, i.e. φ is continuous everywhere. Observing that the functional f_2 is jointly convex w.r.t. *all* variables, one can easily check that φ is a convex function as well. Additionally, [12, Lemma 4.3] guarantees the continuous differentiability of φ.

3.3.1 The Optimal Value Reformulation and Its Relaxation

By definition of φ, it is obvious that

$$\begin{aligned} \tfrac{1}{2}\|y - y_\Omega\|^2_{L^2(\Omega)} + \tfrac{\vartheta}{2}\|u - u_\Omega\|^2_{L^2(\Omega)} &\to \min_{\alpha,y,u} \\ \alpha \in \mathbb{R}^n, \quad y \in H_0^1(\Omega), \quad u &\in L^2(\Omega), \\ \alpha &\in \Lambda, \\ f_2(\alpha, y, u) + \tfrac{\sigma}{2}\|u\|^2_{L^2(\Omega)} - \varphi(\alpha) &\le 0, \\ -\Delta y &= u, \\ u &\in U_{\mathrm{ad}} \end{aligned} \tag{OVR}$$

is equivalent to (IOC_{f_2}). We refer to (OVR) as the *optimal value reformulation* of (IOC_{f_2}). Although (OVR) is a smooth single-level optimization problem, we cannot simply tackle it with the aid of suitable KKT-type optimality conditions since reasonable constraint qualifications which apply to (OVR) fail to hold at all of its feasible points, see [12, Section 5.1].

In order to deal with this issue, we *relax* the feasible set of (OVR). Therefore, we choose a sequence $\{\varepsilon_k\}_{k\in\mathbb{N}}$ of positive relaxation parameters tending to zero as $k \to \infty$ and consider

$$\begin{aligned} \tfrac{1}{2}\|y - y_\Omega\|^2_{L^2(\Omega)} + \tfrac{\vartheta}{2}\|u - u_\Omega\|^2_{L^2(\Omega)} &\to \min_{\alpha,y,u} \\ \alpha \in \mathbb{R}^n, \quad y \in H_0^1(\Omega), \quad u &\in L^2(\Omega), \\ \alpha &\in \Lambda, \\ f_2(\alpha, y, u) + \tfrac{\sigma}{2}\|u\|^2_{L^2(\Omega)} - \varphi(\alpha) &\le \varepsilon_k, \\ -\Delta y &= u, \\ u &\in U_{\mathrm{ad}}. \end{aligned} \tag{OVR(ε_k)}$$

The following lemma parallels [12, Lemma 5.2].

Lemma 3.6 *Fix $k \in \mathbb{N}$. Then Robinson's constraint qualification is valid at each feasible point of* (OVR(ε_k)).

For details regarding Robinson's constraint qualification, we refer the interested reader to the monograph [6]. It is not difficult to check that (OVR(ε_k)) possesses an optimal solution for each $k \in \mathbb{N}$, see [12, Lemma 5.3]. In the upcoming theorem, whose proof can be found in [12] as well, we study the behavior of a sequence of global minimizers associated with (OVR(ε_k)).

Theorem 3.7 *For each $k \in \mathbb{N}$, let $(\bar{\alpha}_k, \bar{y}_k, \bar{u}_k) \in \mathbb{R}^n \times H_0^1(\Omega) \times L^2(\Omega)$ be a global minimizer of* (OVR(ε_k)). *Then the sequence $\{(\bar{\alpha}_k, \bar{y}_k, \bar{u}_k)\}_{k\in\mathbb{N}}$ possesses a convergent subsequence whose limit point is a global minimizer of* (OVR) *and, thus, of* (IOC$_{f_2}$).

Noting that φ is only implicitly given while (OVR(ε_k)) is a non-convex program for each $k \in \mathbb{N}$, Theorem 3.7 seems to be of limited practical use at the first glance. However, as we will see later, this result plays a crucial role for the successful derivation of stationarity conditions for (IOC$_{f_2}$). For that purpose, let us state the KKT conditions of (OVR(ε_k)) for fixed $k \in \mathbb{N}$ at one of its global minimizers $(\bar{\alpha}_k, \bar{y}_k, \bar{u}_k) \in \mathbb{R}^n \times H_0^1(\Omega) \times L^2(\Omega)$. Due to Lemma 3.6, this system, which is given by

$$\begin{aligned}
z_k + \gamma_k \left(\left(\left\langle \textstyle\sum_{i=1}^n (\bar{\alpha}_k)_i g_i - \bar{y}_k, g_j \right\rangle_{L^2(\Omega)} \right)_{j=1}^n - \varphi'(\bar{\alpha}_k) \right) &= 0, \\
\bar{y}_k - y_\mathrm{o} + \gamma_k \left(\bar{y}_k - \textstyle\sum_{i=1}^n (\bar{\alpha}_k)_i g_i \right) - \Delta p_k &= 0, \\
\vartheta(\bar{u}_k - u_\mathrm{o}) + \gamma_k \sigma \bar{u}_k - p_k + \lambda_k &= 0, \\
z_k &\in \mathcal{N}_\Lambda(\bar{\alpha}_k), \\
0 \le \gamma_k \perp \tfrac{1}{2} \| \bar{y}_k - \textstyle\sum_{i=1}^n (\bar{\alpha}_k)_i g_i \|^2_{L^2(\Omega)} + \tfrac{\sigma}{2} \| \bar{u}_k \|^2_{L^2(\Omega)} - \varphi(\bar{\alpha}_k) - \varepsilon_k &\le 0, \\
\lambda_k &\in \mathcal{N}_{U_\mathrm{ad}}(\bar{u}_k)
\end{aligned} \tag{3.7}$$

for multipliers $z_k \in \mathbb{R}^n$, $\gamma_k \in \mathbb{R}$, $p_k \in H_0^1(\Omega)$, and $\lambda_k \in L^2(\Omega)$, indeed provides a necessary optimality condition.

3.3.2 C-Stationarity for Local Minimizers

Our goal is now to show that each local minimizer of (IOC$_{f_2}$) is C-stationary in the sense of Definition 3.1. In order to do so, we exploit the relaxation approach described above. Therefore, we pick a sequence $\{\varepsilon_k\}_{k\in\mathbb{N}}$ of positive relaxation parameters tending to zero as $k \to \infty$ and fix a sequence $\{(\bar{\alpha}_k, \bar{y}_k, \bar{u}_k)\}_{k\in\mathbb{N}} \subset \mathbb{R}^n \times$

$H_0^1(\Omega) \times L^2(\Omega)$ of global minimizers associated with (OVR(ε_k)) that converges in norm to a global minimizer $(\bar{\alpha}, \bar{y}, \bar{u}) \in \mathbb{R}^n \times H_0^1(\Omega) \times L^2(\Omega)$ of (IOC$_{f_2}$). Due to Sect. 3.3.1, there are multipliers $z_k \in \mathbb{R}^n$, $\gamma_k \in \mathbb{R}$, $p_k \in H_0^1(\Omega)$, and $\lambda_k \in L^2(\Omega)$ which satisfy (3.7). Furthermore, let us note that the conditions

$$\begin{aligned} \psi^y(\bar{\alpha}_k) - \textstyle\sum_{i=1}^n (\bar{\alpha}_k)_i g_i - \Delta\phi^p(\bar{\alpha}_k) &= 0, \\ \sigma\psi^u(\bar{\alpha}_k) - \phi^p(\bar{\alpha}_k) + \phi^\lambda(\bar{\alpha}_k) &= 0, \\ \phi^\lambda(\bar{\alpha}_k) &\in \mathcal{N}_{U_{\mathrm{ad}}}(\psi^u(\bar{\alpha}_k)) \end{aligned}$$

hold for each $k \in \mathbb{N}$ where $\phi^p \colon \Lambda \to H_0^1(\Omega)$ and $\phi^\lambda \colon \Lambda \to L^2(\Omega)$ denote the Lagrange multiplier mappings of (P$_{f_2}(\alpha)$) which are continuous since the solution mappings $\psi^y \colon \Lambda \to H_0^1(\Omega)$ and $\psi^u \colon \Lambda \to L^2(\Omega)$ possess this property. As the upcoming lemma shows, we can combine all these multipliers in a deft way such that by taking the limit $k \to \infty$, a reasonable stationarity condition is obtained for $(\bar{\alpha}, \bar{y}, \bar{u})$, see [12, Lemma 5.5].

Lemma 3.8 *Under the assumptions made above, there exist multipliers $\bar{z} \in \mathbb{R}^n$, $\bar{\mu}, \bar{\rho} \in H_0^1(\Omega)$, and $\bar{w}, \bar{\xi} \in L^2(\Omega)$ such that the convergences*

$$\begin{aligned} z_k &\to \bar{z} && \text{in } \mathbb{R}^n, \\ \gamma_k(\bar{y}_k - \psi^y(\bar{\alpha}_k)) &\to \bar{\mu} && \text{in } H_0^1(\Omega), \\ p_k - \gamma_k\phi^p(\bar{\alpha}_k) &\to \bar{\rho} && \text{in } H_0^1(\Omega), \\ \gamma_k(\bar{u}_k - \psi^u(\bar{\alpha}_k)) &\rightharpoonup \bar{w} && \text{in } L^2(\Omega), \\ \lambda_k - \gamma_k\phi^\lambda(\bar{\alpha}_k) &\rightharpoonup \bar{\xi} && \text{in } L^2(\Omega) \end{aligned}$$

hold along a subsequence. Furthermore, these multipliers satisfy (3.2a)–(3.2e).

Furthermore, it is possible to show that the multipliers from Lemma 3.8 already satisfy the remaining C-stationarity conditions. This follows exploiting the convergences from Lemma 3.8, the definition of the normal cone, and some pointwise arguments, see [12, Lemmas 5.6, 5.7 and Remark 5.1].

Lemma 3.9 *The multipliers $\bar{w}, \bar{\xi} \in L^2(\Omega)$ from Lemma 3.8 additionally satisfy the conditions* (3.2j), (3.2k), *and* (3.3).

Noting that the conditions (3.2f)–(3.2i) just characterize lower level optimality of $(\bar{y}, \bar{u})$ for the fixed parameter $\bar{\alpha}$, the above lemmas show that the global minimizer $(\bar{\alpha}, \bar{y}, \bar{u})$ is C-stationary for (IOC$_{f_2}$).

Finally, we are in position to derive C-stationarity of *all* local minimizers associated with (IOC$_{f_2}$) by performing a localization as sketched after Theorem 3.5.

Theorem 3.10 *Let $(\bar{\alpha}, \bar{y}, \bar{u}) \in \mathbb{R}^n \times H_0^1(\Omega) \times L^2(\Omega)$ be a local minimizer of (IOC$_{f_2}$). Then it is a C-stationary point of this program.*

3.4 Variational Analysis Approach and Mordukhovich-Stationarity

Noting that the derivatives $D_y f_2$ and $D_u f_2$ are linear mappings, the feasible set of the complementarity program (3.1) is characterized via linear maps (apart from the challenging variational structure of the set $\operatorname{gph} \mathcal{N}_{U_{\mathrm{ad}}}$). That is why results from finite-dimensional programming, see e.g. [15, Theorem 3.9], suggest that local minimizers of (IOC_{f_2}) could be already pM-stationary which is a stronger condition than C-stationarity, see Definition 3.1.

In order to come up with this enhanced stationarity system, one may think of the following variational analysis approach promoted in [37]: Local minimizers of (IOC_{f_2}) satisfy a first-order necessary optimality condition comprising the derivative of the objective as well as the limiting normal cone to the feasible set. The latter can be expressed in terms of the constraining data function's derivatives and the limiting normal cone to the set $\operatorname{gph} \mathcal{N}_{U_{\mathrm{ad}}}$ under validity of suitable constraint qualifications comprising the so-called sequential normal compactness of $\operatorname{gph} \mathcal{N}_{U_{\mathrm{ad}}}$ at the reference point. However, noting that this set is a non-trivial decomposable set, see [41, Section 6.4] for an introduction, one obtains from [34, 35] that the limiting normal cone to $\operatorname{gph} \mathcal{N}_{U_{\mathrm{ad}}}$ at one of its points (u, λ) is given by

$$\left\{ (\xi, w) \in L^2(\Omega)^2 \;\middle|\; \begin{array}{ll} \xi = 0 & \text{a.e. on } I^{a+}(u) \cap I^{b-}(u) \\ w = 0 & \text{a.e. on } \{\lambda \neq 0\} \end{array} \right\}$$

while $\operatorname{gph} \mathcal{N}_{U_{\mathrm{ad}}}$ is nowhere sequentially normally compact, see [31]. Thus, the variational analysis approach yields at most W-stationarity of the local minimizers associated with (IOC_{f_2}). This is a very weak result, since we already know C-stationarity of local minimizers due to Theorem 3.10. In view of the outstanding success of variational analysis in finite dimensions, this deficiency in the infinite-dimensional case is quite surprising. It remains an open question whether local minimizers of (IOC_{f_2}) satisfy the aforementioned pM-stationarity conditions from Definition 3.1. Similar arguments apply to (IOC_{f_1}).

Another example for the limited use of the variational analysis approach in function space optimization has been reported in [18]. In this paper, the authors show that the limiting normal cone to the complementarity set

$$\left\{ (y, \eta) \in H_0^1(\Omega) \times H^{-1}(\Omega) \;\middle|\; y \geq 0,\ \eta \leq 0,\ \langle y, \eta \rangle_{H_0^1(\Omega)} = 0 \right\}$$

is uncomfortably large whenever the dimension of the underlying domain Ω is at least 2. This negative result addresses e.g. the optimal control of the obstacle problem which is a hierarchical optimization problem in function spaces, see [17] for an overview. Exploiting the variational analysis approach, satisfactory optimality conditions can only be obtained in case $d = 1$ since $H_0^1(\Omega)$ is continuously embedded into $C(\overline{\Omega})$ in this situation, see [26] for details.

3.5 Comments on Biactivity and S-Stationarity

We check that the biactive set has measure zero unless some rather strict assumptions on the data are fulfilled. To simplify the considerations, we suppose that the control bounds u_a and u_b are constants, and we restrict the discussion to f_2 defined in (2.2). Further, we require $g_1, \ldots, g_n \in H^2(\Omega)$ for the data. Now, let $(\bar{\alpha}, \bar{y}, \bar{u})$ be a W-stationary point satisfying (3.2). We shall need the regularity $\bar{y}, \bar{p} \in H^2(\Omega)$ which is satisfied under moderate assumptions on Ω.

Now, we consider the biactive set $B_a := \{\bar{\lambda} = 0\} \cap \{\bar{u} = u_a\}$. On this set, (3.2g) implies

$$0 = \bar{\lambda} = \bar{p} + \sigma\, \bar{u} = \bar{p} + \sigma\, u_a.$$

Hence, Stampacchia's lemma implies $-\Delta \bar{p} = \sigma\, \Delta u_a = 0$ a.e. on B_a. Thus, the adjoint equation (3.2f) gives

$$0 = \bar{y} - \textstyle\sum_{i=1}^{n} \bar{\alpha}_i\, g_i$$

a.e. on B_a. Applying Stampacchia's lemma again yields

$$u_a = \bar{u} = -\Delta \bar{y} = -\textstyle\sum_{i=1}^{n} \bar{\alpha}_i\, \Delta g_i \qquad \text{a.e. on } B_a.$$

For most choices of the data g_i, the set on which such a condition can be satisfied is a null set and this implies that B_a is a null set. A similar argumentation applies to $B_b := \{\bar{\lambda} = 0\} \cap \{\bar{u} = u_b\}$. Moreover, it is clear that every W-stationary point is already S-stationary if these biactive sets have measure zero.

However, it is possible to construct examples possessing a biactive set and for which local minimizers may fail to be S-stationary, see [19, Example 3.4].

4 Numerical Solution

A typical route for the solution of bilevel optimization problems is to solve regularized or relaxed problems and pass to the limit with the regularization or relaxation parameter. However, this approach delivers only very weak convergence results. In particular, if only local solutions of the regularized or relaxed problems are computed, then the limit point may fail to be a local solution to the bilevel problem.

4.1 Global Solution Algorithm for (IOC$_{f_2}$)

A first idea about how to solve (IOC$_{f_2}$) is presented in Theorem 3.7: For a sequence $\{\varepsilon_k\}_{k\in\mathbb{N}}$ of positive relaxation parameters tending to zero as $k \to \infty$, we could aim for the global solution of the associated sequence of surrogate problems (OVR(ε_k)) since the sequence of computed minimizers converges along a subsequence to a global minimizer of (IOC$_{f_2}$). However, noting that the optimal value function φ is only implicitly known while (OVR(ε_k)) is a non-convex program for each $k \in \mathbb{N}$, this approach does not seem to be too promising although it motivates to take a closer look at the potential of the optimal value reformulation (OVR) concerning the numerical solution of (IOC$_{f_2}$).

On the other hand, we need to observe that due to the convexity of φ, (OVR) is a so-called DC-problem where DC abbreviates *difference of convex functions*, see [25] for an overview of DC-programming. Clearly, in (OVR), concavity is hidden only in $-\varphi$. In order to exploit this observation for the construction of a solution algorithm, we aim for the construction of a piecewise affine approximation of the function φ from above which is refined during each iteration of the method. The resulting relaxed surrogate problems then can be decomposed into finitely many convex subproblems which can be solved to global optimality with ease. This idea is inspired by techniques from finite-dimensional bilevel programming, see [11, Section 4].

Let $A := \{\alpha^1, \ldots, \alpha^m\} \subset \mathbb{R}^n$ be a nonempty set satisfying $\Lambda \subset \operatorname{int}\operatorname{conv} A$ and consider the function $\xi_A \colon \operatorname{conv} A \to \mathbb{R}$ given by

$$\xi_A(\alpha) := \min_{\nu} \left\{ \sum_{j=1}^m \nu_j \, \varphi(\alpha^j) \;\middle|\; \nu \geq 0,\ \sum_{j=1}^m \nu_j = 1,\ \sum_{j=1}^m \nu_j \alpha^j = \alpha \right\}$$

for each $\alpha \in \operatorname{conv} A$. By convexity of φ, we easily see that ξ_A approximates φ from above on $\operatorname{conv} A$. Furthermore, we have $\varphi(\alpha^j) = \xi_A(\alpha^j)$ for all $j = 1, \ldots, m$. Noting that ξ_A is the optimal value function of a linear parametric optimization problem where the parameter only appears at the right hand side of the constraints, ξ_A is piecewise affine and convex, see [38, Section 6]. Particularly, $\operatorname{conv} A$ can be decomposed into finitely many so-called regions of stability where ξ_A is affine, respectively. All these properties motivate the formulation of Algorithm 1.

Under the postulated assumptions, it is not difficult to show that (OVR(A_k)) possesses a global minimizer for each $k \in \mathbb{N}$ which can be computed by decomposition of (OVR(A_k)) into finitely many convex subproblems exploiting the regions of stability associated with ξ_{A_k}. By construction of the algorithm, we have

$$\forall \alpha \in \Lambda: \quad \varphi(\alpha) \leq \xi_{A_{k'}}(\alpha) \leq \xi_{A_k}(\alpha)$$

for any two natural numbers $k, k' \in \mathbb{N}$ satisfying $k \leq k'$. Thus, with increasing iteration counter k, the relaxation (OVR(A_k)) gets tighter. Finally, let us mention that due to the choice of $A_1 \subset \mathbb{R}^n$, there exists a constant $L > 0$ such that all the functions ξ_{A_k}, $k \in \mathbb{N}$, are Lipschitz continuous on Λ with modulus L, see [12,

Algorithm 1 Computation of global minimizers to (IOC$_{f_2}$)

S1. Let $A_1 \subset \mathbb{R}^n$ be a finite set such that $\Lambda \subset \operatorname{int}\operatorname{conv} A_1$ is valid and set $k := 1$.

S2. Compute a global minimizer $(\bar{\alpha}_k, \bar{y}_k, \bar{u}_k)$ of the optimization problem

$$\begin{aligned} \tfrac{1}{2}\|y - y_o\|^2_{L^2(\Omega)} + \tfrac{\vartheta}{2}\|u - u_o\|^2_{L^2(\Omega)} &\to \min_{\alpha,y,u} \\ \alpha \in \mathbb{R}^n, \quad y \in H^1_0(\Omega), \quad u &\in L^2(\Omega), \\ \alpha &\in \Lambda, \\ f_2(\alpha, y, u) + \tfrac{\sigma}{2}\|u\|^2_{L^2(\Omega)} - \xi_{A_k}(\alpha) &\le 0, \\ -\Delta y &= u, \\ u &\in U_{\mathrm{ad}}. \end{aligned} \tag{OVR(A_k))}$$

S3. Compute $\varphi(\bar{\alpha}_k)$. If $f_2(\bar{\alpha}_k, \bar{y}_k, \bar{u}_k) + \frac{\sigma}{2}\|\bar{u}_k\|^2_{L^2(\Omega)} = \varphi(\bar{\alpha}_k)$ holds, then the point $(\bar{\alpha}_k, \bar{y}_k, \bar{u}_k)$ is a global minimizer of (OVR) (and, thus, of (IOC$_{f_2}$)) and the algorithm terminates. Otherwise, set $A_{k+1} := A_k \cup \{\bar{\alpha}_k\}$ as well as $k \mapsto k+1$ and go to **S2**.

Section 6.1] for details. This property is crucial in order to prove the subsequently stated theorem, see [12, Theorem 6.1].

Theorem 4.1 *Either, Algorithm 1 terminates after finitely many steps having computed a global minimizer of* (OVR) *and, thus, of* (IOC$_{f_2}$)*, or it produces a sequence* $\{(\bar{\alpha}_k, \bar{y}_k, \bar{u}_k)\}_{k\in\mathbb{N}} \subset \mathbb{R}^n \times H^1_0(\Omega) \times L^2(\Omega)$ *of global minimizers of* (OVR(A_k))*. This sequence possesses a convergent subsequence and all accumulation points are global minimizers of* (OVR) *and, thus, of* (IOC$_{f_2}$)*.*

4.2 Numerical Example

A numerical example for the solution of (IOC$_{f_2}$) is given in [12, Section 6.2]. In this section, we present a different numerical example. We use the data

$$\begin{aligned} &\Omega = (0,1)^2, && g_1(x) = 4\,\sin(\pi\, x_1), && u_a \equiv -10, \\ &n = 2, && g_2(x) = 3\,\sin(\pi\, x_2), && u_b \equiv 10, \\ &\sigma = 10^{-2}, && y_o = 0.1\, g_1 + 0.5\, g_2 && \vartheta = 0. \end{aligned}$$

The upper level objective in (IOC$_{f_2}$) is replaced by

$$(\alpha, y, u) \mapsto \tfrac{1}{2}\|y - y_o\|^2_{L^2(\Omega)} + \tfrac{\vartheta}{2}\|u - u_o\|^2_{L^2(\Omega)} + \tau^\top \alpha,$$

i.e. we have added a linear term w.r.t. α with $\tau = (-2, 0)^\top$. Furthermore, we replaced the standard simplex Λ by the unit simplex $\tilde{\Lambda} \subset \mathbb{R}^2$ given by

$$\tilde{\Lambda} := \operatorname{conv}\left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\right\}.$$

In Algorithm 1, we use the starting set

$$A_1 := \left\{\begin{pmatrix}-1/2\\-1/2\end{pmatrix}, \begin{pmatrix}3/2\\0\end{pmatrix}, \begin{pmatrix}0\\3/2\end{pmatrix}, \begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\right\}.$$

In Fig. 1, we show the reduced objective function $\gamma\colon \tilde{\Lambda} \to \mathbb{R}$, i.e. the map given by

$$\forall \alpha \in \tilde{\Lambda}\colon \quad \gamma(\alpha) := \tfrac{1}{2}\|\psi^y(\alpha) - y_\mathrm{o}\|^2_{L^2(\Omega)} + \tfrac{\vartheta}{2}\|\psi^u(\alpha) - u_\mathrm{o}\|^2_{L^2(\Omega)} + \tau^\top \alpha.$$

It can be seen that this function is not convex and possesses at least two local minimizers. In order to reduce the curvature of the optimal value function, which also increases its approximability, we perform a curvature reduction. That is, we replace the lower level objective f_2 by the function

$$(\alpha, y, u) \mapsto f_2(\alpha, y, u) - \tfrac{1}{2}\,\alpha^\top H\,\alpha,$$

where $H \in \mathbb{R}^{2\times 2}$ is a suitably chosen matrix depending on the initial data of the lower level problem. We note that this transformation does not change the lower level solution set since f_2 is only shifted by a quadratic term in α while we minimize only w.r.t. y and u at the lower level stage. Although the adjusted lower level objective functional is not jointly convex anymore, it is still jointly convex on

$$\left\{(\alpha, y, u) \in \mathbb{R}^2 \times H_0^1(\Omega) \times L^2(\Omega) \;\middle|\; -\Delta y = u\right\}$$

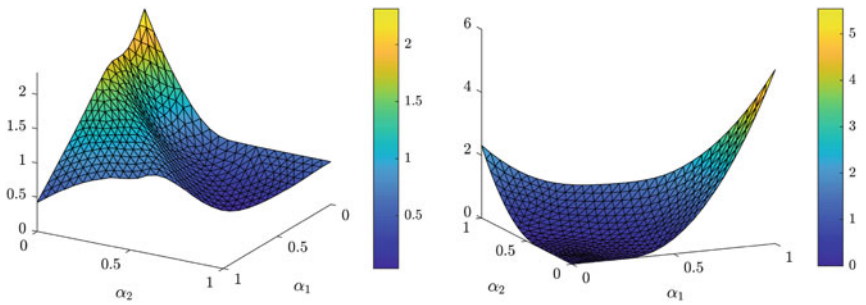

Fig. 1 Reduced objective function (left) and optimal value function with curvature reduction (right) of the problem given in Sect. 4.2

which is enough for the above theory to work. In practice, it turned out that performing a curvature reduction on φ significantly reduces the total number of iterations in Algorithm 1. Regarding the example discussed here, using curvature reduction saves about 25% of the total number of iterations. Even more convincing results were obtained for the example from [12, Section 6.2] where the curvature reduction saves more than 80% of the total number of iterations. The details of this approach, particularly the precise choice of the matrix H, can be found in [12, Section 6.2]. The optimal value function with reduced curvature is shown in Fig. 1.

Moreover, we exploit the following modifications in Algorithm 1 in order to increase its performance. We remark that the convergence result Theorem 4.1 carries over to this modified version.

- In Step 3 of the algorithm, we do not only add $\bar{\alpha}_k$ to A_k, but also all three midpoints of the edges of the region of stability in which $\bar{\alpha}_k$ lies. This increases the approximation quality of $\xi_{A_{k+1}}$.
- In order to speed-up the computations, the finitely many convex subproblems of (OVR(A_k)) are solved in parallel.

Let us describe the results of Algorithm 1. Since we are solving a relaxed optimization problem in each iteration, we obtain lower bounds on the optimal value of the bilevel problem (IOC$_{f_2}$). On the other hand, calculating $\gamma(\bar{\alpha}_k)$ yields upper bounds. Since $\bar{\alpha}_k$ is the solution of the relaxed problem, the true value $\gamma(\bar{\alpha}_k)$ can be quite large. Therefore, we denote by $\hat{\alpha}_k$ the best known point of γ in iteration k, i.e.

$$\hat{\alpha}_k := \arg\min_{\alpha \in \{\bar{\alpha}_1, \ldots, \bar{\alpha}_k\}} \gamma(\alpha).$$

This yields a decreasing upper bound $\gamma(\hat{\alpha}_k)$. We record the Euclidean distances $\|\bar{\alpha} - \bar{\alpha}_k\|_{\mathbb{R}^2}$ and $\|\bar{\alpha} - \hat{\alpha}_k\|_{\mathbb{R}^2}$ in Fig. 2, and one can believe that $\bar{\alpha}_k \to \bar{\alpha}$ holds for $\bar{\alpha} \approx (0.3306, 0.6694)$ as predicted by Theorem 4.1.

5 Future Perspectives

Let us comment about future research directions. First of all, it is important to understand the role of pM-stationarity, in which the additional sign conditions (3.4) are required. In fact, we are going to investigate if this is a necessary optimality condition or if it is possible to construct an example for which (3.4) is violated at the minimizer. It is also necessary to study second-order optimality conditions. Particularly, sufficient second-order optimality conditions are important for checking stability of solutions.

The most important feature of Algorithm 1 is its guaranteed convergence toward global minimizers of the non-convex problem (IOC$_{f_2}$). The main ingredient in the construction is the convexity of the optimal value function φ. Here, it should be checked whether this requirement can be relaxed and how this algorithm can be applied to different instances of (IOC$_f$).

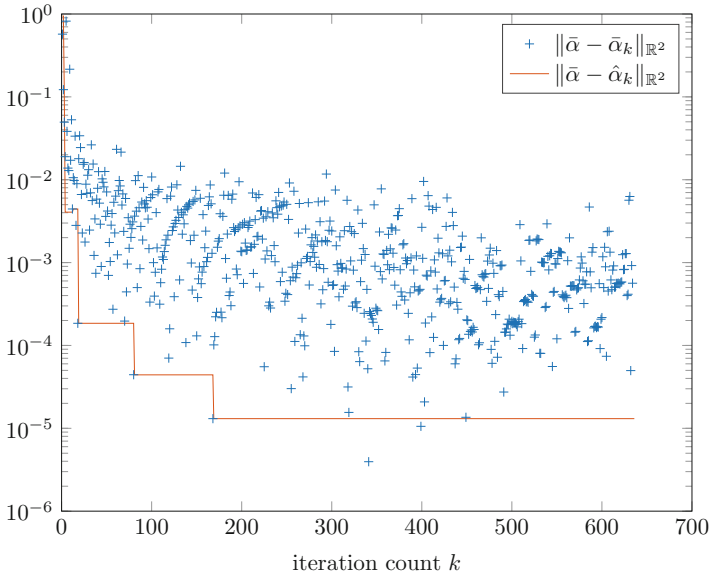

Fig. 2 Errors of $\bar{\alpha}_k$ (blue crosses) and $\hat{\alpha}_k$ (red line)

A challenging topic for future research is the numerical analysis of Algorithm 1. In particular, it would be nice to couple the iterations of Algorithm 1 with an adaptive refinement strategy which guarantees the convergence toward the global solution of the continuous problem.

Acknowledgments This work is supported by the DFG grant *Analysis and Solution Methods for Bilevel Optimal Control Problems* (grant numbers DE 650/10-1 and WA3636/4-1) within the Priority Program SPP 1962 (Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization).

References

1. S. Albrecht and M. Ulbrich, *Mathematical programs with complementarity constraints in the context of inverse optimal control for locomotion*, Optimization Methods and Software **32** (2017), no. 4, 670–698.
2. J. F. Bard, *Practical Bilevel Optimization: Algorithms and Applications*, Kluwer Academic, Dordrecht, 1998.
3. F. Benita, S. Dempe, and P. Mehlitz, *Bilevel Optimal Control Problems with Pure State Constraints and Finite-dimensional Lower Level*, SIAM Journal on Optimization **26** (2016), no. 1, 564–588.
4. F. Benita and P. Mehlitz, *Bilevel Optimal Control With Final-State-Dependent Finite-Dimensional Lower Level*, SIAM Journal on Optimization **26** (2016), no. 1, 718–752.
5. ———, *Optimal Control Problems with Terminal Complementarity Constraints*, SIAM Journal on Optimization **28** (2018), no. 4, 3079–3104.

6. J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer, New York, Berlin, Heidelberg, 2000.
7. C. Christof and G. Wachsmuth, *On Second-Order Optimality Conditions for Optimal Control Problems Governed by the Obstacle Problem*, Optimization (2020), 1–41.
8. C. Clason, Y. Deng, P. Mehlitz, and U. Prüfert, *Optimal control problems with control complementarity constraints: existence results, optimality conditions, and a penalty method*, Optimization Methods and Software 35 (2020), no. 1, 142–170.
9. S. Dempe, *Foundations of Bilevel Programming*, Kluwer, Dordrecht, 2002.
10. S. Dempe, *Bilevel optimization: theory, algorithms and applications*, Bilevel Optimization: Advances and Next Challenges (S. Dempe and A. B. Zemkoho, eds.), Springer, Cham, 2020, pp. 581–672.
11. S. Dempe and S. Franke, *On the solution of convex bilevel optimization problems*, Computational Optimization and Applications **63** (2016), no. 3, 685–703.
12. S. Dempe, F. Harder, P. Mehlitz, and G. Wachsmuth, *Solving inverse optimal control problems via value functions to global optimality*, Journal of Global Optimization **74** (2019), no. 2, 297–325.
13. S. Dempe, V. Kalashnikov, G. Pérez-Valdéz, and N. Kalashnykova, *Bilevel Programming Problems - Theory, Algorithms and Applications to Energy Networks*, Springer, Berlin, 2015.
14. F. Fisch, J. Lenz, F. Holzapfel, and G. Sachs, *On the Solution of Bilevel Optimal Control Problems to Increase the Fairness in Air Races*, Journal of Guidance, Control, and Dynamics **35** (2012), no. 4, 1292–1298.
15. M. L. Flegel and C. Kanzow, *On M-stationary points for mathematical programs with equilibrium constraints*, Journal of Mathematical Analysis and Applications **310** (2005), no. 1, 286–302.
16. L. Guo and J. J. Ye, *Necessary optimality conditions for optimal control problems with equilibrium constraints*, SIAM Journal on Control and Optimization **54** (2016), no. 5, 2710–2733.
17. F. Harder and G. Wachsmuth, *Comparison of optimality systems for the optimal control of the obstacle problem*, GAMM-Mitteilungen **40** (2018), no. 4, 312–338.
18. ———, *The limiting normal cone of a complementarity set in Sobolev spaces*, Optimization **67** (2018), no. 10, 1579–1603.
19. ———, *Optimality conditions for a class of inverse optimal control problems with partial differential equations*, Optimization **68** (2018), no. 2–3, 615–643.
20. K. Hatz, J. P. Schlöder, and H. G. Bock, *Estimating Parameters in Optimal Control Problems*, SIAM Journal on Scientific Computing **34** (2012), no. 3, A1707–A1728.
21. R. Herzog, C. Meyer, and G. Wachsmuth, *C-Stationarity for Optimal Control of Static Plasticity with Linear Kinematic Hardening*, SIAM Journal on Control and Optimization **50** (2012), no. 5, 3052–3082.
22. M. Hintermüller and D. Wegner, *Optimal Control of a Semidiscrete Cahn–Hilliard–Navier–Stokes System*, SIAM Journal on Control and Optimization **52** (2014), no. 1, 747–772.
23. M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE constraints*, Springer, 2009.
24. G. Holler, K. Kunisch, and R. C. Barnard, *A bilevel approach for parameter learning in inverse problems*, Inverse Problems **34** (2018), no. 11, 1–28.
25. R. Horst and N. V. Thoai, *DC Programming: Overview*, Journal of Optimization Theory and Applications **103** (1999), no. 1, 1–43.
26. J. Jarušek and J. V. Outrata, *On sharp necessary optimality conditions in control of contact problems with strings*, Nonlinear Analysis: Theory, Methods & Applications **67** (2007), no. 4, 1117–1128.
27. M. Knauer, *Fast and save container cranes as bilevel optimal control problems*, Mathematical and Computer Modelling of Dynamical Systems **18** (2012), no. 4, 465–486.
28. K. Kunisch and D. Wachsmuth, *Sufficient optimality conditions and semi-smooth Newton methods for optimal control of stationary variational inequalities*, ESAIM: Control, Optimisation and Calculus of Variations **18** (2012), no. 2, 520–547.

29. F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal Control*, John Wiley & Sons, Hoboken, 2012.
30. P. Mehlitz, *Necessary optimality conditions for a special class of bilevel programming problems with unique lower level solution*, Optimization **66** (2017), no. 10, 1533–1562.
31. ———, *On the Sequential Normal Compactness Condition and its Restrictiveness in Selected Function Spaces*, Set-Valued and Variational Analysis **27** (2019), no. 3, 763–782.
32. P. Mehlitz and G. Wachsmuth, *Bilevel optimal control: existence results and stationarity conditions*, Bilevel Optimization: Advances and Next Challenges (S. Dempe and A. B. Zemkoho, eds.), Springer, Cham, 2020, pp. 451–484.
33. P. Mehlitz and G. Wachsmuth, *Weak and strong stationarity in generalized bilevel programming and bilevel optimal control*, Optimization **65** (2016), no. 5, 907–935.
34. ———, *The Limiting Normal Cone to Pointwise Defined Sets in Lebesgue Spaces*, Set-Valued and Variational Analysis **26** (2018), no. 3, 449–467.
35. ———, *The weak sequential closure of decomposable sets in Lebesgue spaces and its application to variational geometry*, Set-Valued and Variational Analysis **27** (2019), no. 1, 265–294.
36. K. Mombaur, A. Truong, and J.-P. Laumond, *From human to humanoid locomotion—an inverse optimal control approach*, Autonomous Robots **28** (2010), no. 3, 369–383.
37. B. Mordukhovich, *Variational Analysis and Generalized Differentiation*, Springer, Berlin, 2006.
38. F. Nožička, J. Guddat, H. Hollatz, and B. Bank, *Theorie der linearen parametrischen Optimierung*, Akademie-Verlag, Berlin, 1974.
39. K. D. Palagachev and M. Gerdts, *Exploitation of the Value Function in a Bilevel Optimal Control Problem*, System Modeling and Optimization (L. Bociu, J.-A. Désidéri, and A Habbal, eds.), Springer, Cham, 2016, pp. 410–419.
40. ———, *Numerical Approaches Towards Bilevel Optimal Control Problems with Scheduling Tasks*, Math for the Digital Factory (L. Ghezzi, D. Hömberg, and C. Landry, eds.), Springer, Cham, 2017, pp. 205–228.
41. N. S. Papageorgiou and S. T. Kyritsi-Yiallourou, *Handbook of applied analysis*, Advances in Mechanics and Mathematics, vol. 19, Springer, New York, 2009.
42. K. Shimizu, Y. Ishizuka, and J. F. Bard, *Nondifferentiable and two-level mathematical programming*, Kluwer Academic, Dordrecht, 1997.
43. F. Tröltzsch, *Optimal Control of Partial Differential Equations*, Vieweg, Wiesbaden, 2009.
44. J. L. Troutman, *Variational Calculus and Optimal Control*, Springer, New York, 1996.
45. G. Wachsmuth, *Towards M-stationarity for Optimal Control of the Obstacle Problem with Control Constraints*, SIAM Journal on Control and Optimization **54** (2016), no. 2, 964–986.
46. J. J. Ye, *Necessary Conditions for Bilevel Dynamic Optimization Problems*, SIAM Journal on Control and Optimization **33** (1995), no. 4, 1208–1223.
47. ———, *Optimal Strategies For Bilevel Dynamic Problems*, SIAM Journal on Control and Optimization **35** (1997), no. 2, 512–531.

A Calculus for Non-smooth Shape Optimization with Applications to Geometric Inverse Problems

Marc Herrmann, Roland Herzog, Stephan Schmidt, and José Vidal-Núñez

Abstract We are concerned with a class of non-smooth shape optimization problems involving the total variation of the normal vector field along the shape's boundary in their objective. The discrete version of the total variation functional on triangulated surfaces promotes a separation into flat and non-flat regions. Applications include mesh denoising problems as well as geometric inverse problems.

Keywords Non-smooth shape optimization · Total variation · Non-smooth geometries · Geometric inverse problems · Consistent discretization

1 Introduction

This survey article is concerned with a class of non-smooth shape optimization problems. Specifically, we consider applications of the total variation (TV) functional, applied to the normal vector field along the shape's boundary. As a preliminary step,

M. Herrmann · S. Schmidt
Faculty of Mathematics and Computer Science, Lehrstuhl für Mathematik VI, Julius-Maximilians-Universität Würzburg, Würzburg, Germany
e-mail: marc.herrmann@mathematik.uni-wuerzburg.de; stephan.schmidt@mathematik.uni-wuerzburg.de
https://www.mathematik.uni-wuerzburg.de/~herrmann
https://www.mathematik.uni-wuerzburg.de/~schmidt

R. Herzog (✉)
Institute for Applied Mathematics, University of Heidelberg, Heidelberg, Germany
e-mail: roland.herzog@mathematik.tu-chemnitz.de
https://www.tu-chemnitz.de/herzog

J. Vidal-Núñez
Faculty of Mathematics, Technische Universität Chemnitz, Chemnitz, Germany
e-mail: jose.vidal-nunez@mathematik.tu-chemnitz.de
https://www.tu-chemnitz.de/mathematik/part_dgl/people/vidal

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_5

we also address image denoising and inpainting problems on stationary surfaces which are not subject to optimization.

The total variation (TV) functional is popular as a regularizer in imaging and inverse problems; see for instance Rudin et al. [33], Chan et al. [12], Bachmayr and Burger [2], Langer [27], and Vogel [34, Chapter 8]. For a real-valued function $u \in W^{1,1}(\Omega)$ on a bounded domain $\Omega \subset \mathbb{R}^d$, the TV-seminorm is defined as

$$|u|_{TV(\Omega)} \coloneqq \int_\Omega |\nabla u|_s \, \mathrm{d}\boldsymbol{x}, \tag{1.1}$$

where $|\cdot|_s$ denotes the s-norm of vectors in $\mathbb{R}^d$ for some $s \in [1, \infty]$. The most frequent choices are $s = 2$ (isotropic case) and $s = 1$ (anisotropic case). Definition (1.1) extends to less regular functions whose distributional gradient exists only in the sense of measures. In this more general case, we have the representation

$$|u|_{TV(\Omega)} = \sup \left\{ \int_\Omega u \operatorname{div} \boldsymbol{p} \, \mathrm{d}\boldsymbol{x} : \boldsymbol{p} \in C_c^\infty(\Omega; \mathbb{R}^d), \ |\boldsymbol{p}|_{s^*} \le 1 \right\}, \tag{1.2}$$

where $s^* = s/(s-1)$ denotes the conjugate of s. Equation (1.2) is known as a dual representation of $|u|_{TV(\Omega)}$.

In order to further discuss issues of duality, we introduce the space

$$\boldsymbol{H}(\operatorname{div}; \Omega) \coloneqq \left\{ \boldsymbol{v} \in \boldsymbol{L}^2(\Omega; \mathbb{R}^d) : \operatorname{div} \boldsymbol{v} \in L^2(\Omega) \right\}, \tag{1.3}$$

equipped with the norm

$$\|\boldsymbol{v}\|_{\boldsymbol{H}(\operatorname{div};\Omega)} \coloneqq \left(\|\boldsymbol{v}\|^2_{\boldsymbol{L}^2(\Omega;\mathbb{R}^d)} + \|\operatorname{div} \boldsymbol{v}\|^2_{L^2(\Omega)} \right)^{1/2}. \tag{1.4}$$

It is well known that $\boldsymbol{H}(\operatorname{div}; \Omega)$ agrees with the closure of $C^\infty(\Omega; \mathbb{R}^d)$ w.r.t. (1.4). We refer the reader to [16, Chapter I] or [30, Chapter 3] for a thorough background.

A class of classical TV-L^2 image reconstruction problems can be cast as

$$\left\{ \begin{array}{ll} \text{Minimize} & \dfrac{1}{2} \displaystyle\int_\Omega |Ku - f|^2 \, \mathrm{d}\boldsymbol{x} + \dfrac{\alpha}{2} \displaystyle\int_\Omega |u|^2 \, \mathrm{d}\boldsymbol{x} + \beta \, |u|_{TV(\Omega)} \\ \text{over} & u \in BV(\Omega). \end{array} \right. \tag{1.5}$$

The space $BV(\Omega)$ contains functions of bounded variation and the operator $K \in \mathcal{L}(L^2(\Omega))$ appearing in (1.5) expresses available a-priori knowledge about the relation between the image u to be reconstructed and the observed data f. Common examples include $K = \mathrm{id}$ for classical image denoising [33], $K =$ masking for inpainting problems [13, Chapter 6.5], $K =$ blur for deblurring problems [8, 14], and $K =$ coarsen for un-zooming problems [29]. We assume that $\alpha > 0$ holds, or else that K is injective and has closed range. Consequently, $B \coloneqq \alpha \, \mathrm{id} + K^* K$ is a coercive operator in $\mathcal{L}(L^2(\Omega))$. Here and throughout, id denotes the identity mapping.

It was shown in Hintermüller and Kunisch [24] that the following problem serves as a Fenchel predual for (1.5):

$$\begin{cases} \text{Minimize} & \dfrac{1}{2}\|\operatorname{div} \boldsymbol{p} + K^* f\|^2_{B^{-1}} \\ \text{over} & \boldsymbol{p} \in \boldsymbol{H}(\operatorname{div}; \Omega) \\ \text{subject to} & |\boldsymbol{p}|_1 \leq \beta \quad \text{a.e. on } \Omega. \end{cases} \tag{1.6}$$

Here we utilize the abbreviation $\|w\|^2_{B^{-1}} = (w, B^{-1}w)_{L^2(\Omega)} = (w, w)_{B^{-1}}$. Solving the predual problem has a number of advantages compared to solving the primal problem directly. First, we do not have to employ a non-smooth algorithm since, in contrast to (1.5), (1.6) is smooth. The second issue concerns the discretization of the regularization term in (1.5), which is not immediate and can be avoided through the treatment of the predual problem (1.6). Third, as was pointed out in Bartels [4], the finite element solution of minimization problems in BV spaces may suffer from low convergence rates.

The work summarized in this survey generalizes (1.5) and (1.6) in various directions. In Sect. 2 we consider image denoising and inpainting problems on stationary surfaces and prove the equivalence of (1.5) and (1.6) in the proper differential geometric setting. We also formulate a function space interior point approach for the solution of the predual problem. Numerical results utilizing a conforming discretization by Raviart–Thomas (RT) surface finite elements of various polynomial degrees are presented. Unfortunately, the straightforward discretization does not allow for a convenient primal-dual pairing on the level of finite element coefficients, except for the case of lowest-order elements. We therefore discuss in Sect. 2.4 a discrete version of the TV-seminorm tailored to discretizations of (1.5) by piecewise polynomials (DG finite elements) and of (1.6) by RT finite elements. Our formulation admits a convenient primal-dual pairing also for higher-order elements, which is important for efficient numerical algorithms.

In Sect. 3 we consider shape optimization problems featuring the total variation of the normal vector field $\boldsymbol{n}$ along the shape's boundary as part of the objective functional. We pay particular attention to the case of triangulated surfaces, in which the normal vector is piecewise constant and its total variation is concentrated in jumps across inter-element edges. For the solution of this non-smooth problem, we formulate a variant of the split Bregman algorithm [18], which falls into the category of ADMM methods (alternating direction method of multipliers). Each iteration consists of a shape optimization, a variation minimization, and a Lagrange multiplier update step. Since each normal vector is an element of the sphere $\mathcal{S}^2$ due to its unit length, we formulate the method in a Riemannian framework. Interestingly, the variation minimization step can be solved explicitly by shrinkage operations in the appropriate tangent spaces to $\mathcal{S}^2$. We present numerical results for problems whose objective does or does not depend on the state of a partial differential equation (PDE) defined on the volume Ω enclosed by its boundary Γ.

2 Image Reconstruction on Surfaces

In this section, we mainly focus on Herrmann et al [23] where we analyzed the total variation and its application to image reconstruction problems on smooth surfaces. We consider the image reconstruction problem

$$\begin{cases} \text{Minimize} & \frac{1}{2}\int_S |Ku - f|^2 \,\mathrm{d}\boldsymbol{s} + \frac{\alpha}{2}\int_S |u|^2 \,\mathrm{d}\boldsymbol{s} + \beta |u|_{TV(S)} \\ \text{over} & u \in BV(S), \end{cases} \tag{2.1}$$

where $S \subset \mathbb{R}^3$ is a smooth, compact, orientable, and connected surface without boundary. Moreover, S is endowed with a Riemannian metric $\mathfrak{g}$. This problem was proposed in Lai and Chan [26] as an analogue of the TV–L^2 reconstruction model (1.5) for images on smooth surfaces. In the following Sect. 2.1 we define the space of functions of bounded variation $BV(S)$ and derive in Sect. 2.2 a predual representation of (2.1). This leads to a convex quadratic problem with pointwise constraints in $\boldsymbol{H}(\mathrm{div}; S)$. The predual of problem (2.1) can be solved by a variety of methods, including for instance primal-dual Chambolle–Pock [11] and split Bregman iterations [18]. We analyze, however, its solution by means of a function space interior point method. Finally, we present in Sect. 2.4 a discrete version of the total variation functional which has been proposed in Herrmann et al. [22] specifically in the context of higher-order finite element discretizations by piecewise polynomial, globally discontinuous (DG) functions. This part is presented in the flat domain setting (1.5) but can be extended to (2.1) on surfaces as well.

2.1 Functions of Bounded Variation on Surfaces

Let $C^\infty(S)$ denote the space of smooth, real-valued functions on the surface S and let $C_c^\infty(S)$ be the subspace of functions of compact support. Moreover, $\boldsymbol{C}^\infty(S; T(S))$ denotes the space of smooth vector fields, i.e., sections of the tangent bundle over S. As usual, the support of a function f is defined as

$$\operatorname{supp} f \coloneqq \operatorname{cl}\{p \in S : f(p) \neq 0\}$$

with $\operatorname{cl} C$ denoting the closure of a set $C \subset S$.

For $1 \leq p < \infty$ the Lebesgue space $L^p(S)$ is defined as the completion of $C^\infty(S)$ w.r.t. the norm

$$\|f\|_{L^p(S)} \coloneqq \left(\int_S |f|^p \,\mathrm{d}\boldsymbol{s}\right)^{1/p}. \tag{2.2}$$

We refer the reader, e.g., to Hebey [20, Ch. 1.2] for more details. Naturally, this definition extends to vector fields $\boldsymbol{f} \in \boldsymbol{L}^p(S; T(S))$, which are endowed with the norm

$$\|\boldsymbol{f}\|_{\boldsymbol{L}^p(S;T(S))} \coloneqq \left(\int_S |\boldsymbol{f}|_{\mathfrak{g}}^p \, \mathrm{d}\boldsymbol{s} \right)^{1/p}.$$

The spaces $L^2(S)$ and $\boldsymbol{L}^2(S; T(S))$ are Hilbert spaces w.r.t. the usual inner products $(\cdot,\cdot)_{L^2(S)}$ and $(\cdot,\cdot)_{\boldsymbol{L}^2(S;T(S))}$.

We are now in the position to recall the definition of functions of bounded variation on smooth, compact, connected surfaces. Background material on BV functions on flat domains can be found, for instance, in Giusti [17], Ziemer [35, Ch. 5] or Attouch et al. [1, Ch. 10]. The following definition can be found in Lai and Chan [26, Sect. 3.1] or Ben-Artzi and LeFloch [5, Sect. 4].

Definition 2.1 A function $u \in L^1(S)$ belongs to $BV(S)$ if the TV-seminorm defined by

$$|u|_{TV(S)} \coloneqq \sup \left\{ \int_S u \operatorname{div} \boldsymbol{\eta} \, \mathrm{d}\boldsymbol{s} : \boldsymbol{\eta} \in \boldsymbol{C}^\infty(S; T(S)) : |\boldsymbol{\eta}(\boldsymbol{s})|_{\mathfrak{g}} \leq 1 \right\} \tag{2.3}$$

is finite. The space $BV(S)$ is endowed with the norm

$$\|u\|_{BV(S)} = \|u\|_{L^1(S)} + |u|_{TV(S)}, \quad u \in BV(S). \tag{2.4}$$

Notice that in (2.3), $\boldsymbol{\eta}$ is a smooth vector field with pointwise norm (induced by the Riemannian metric $\mathfrak{g}$) bounded by one. It can be shown that $BV(S) \hookrightarrow L^2(S)$ holds.

2.2 Dual Representation

The derivation of a pair of primal/dual problems for TV–L^2 is not straightforward due to the lack of reflexivity of BV spaces. For the flat case, it was shown in Hintermüller and Kunisch [24] that the *predual* problem (1.6), posed in $\boldsymbol{H}(\mathrm{div})$, is the appropriate concept. A formal version of this problem on smooth surfaces was proposed in Lai and Chan [26]. A rigorous analysis was provided in Herrmann et al. [23], where the following result can be found. As before, we set

$$B \coloneqq \alpha \, \mathrm{id} + K^* K \in \mathcal{L}(L^2(S)). \tag{2.5}$$

Theorem 2.2 *The Fenchel dual problem of*

$$\begin{cases} \text{Minimize} & \frac{1}{2}\|\operatorname{div} \boldsymbol{p} + K^* f\|_{B^{-1}}^2 \\ \text{over} & \boldsymbol{p} \in \boldsymbol{H}(\operatorname{div}; S) \\ \text{subject to} & |\boldsymbol{p}|_{\mathfrak{g}} \leq \beta \quad \text{a.e. on } S \end{cases} \tag{2.6}$$

is equivalent to (2.1). *Moreover, if* $\overline{\boldsymbol{p}}$ *is an optimal solution to* (2.6) *and* $\overline{u}$ *is optimal to* (2.1)*, then*

$$B\,\overline{u} = \operatorname{div} \overline{\boldsymbol{p}} + K^* f. \tag{2.7}$$

A natural discretization of the surface S is given in terms of a triangulation into flat triangles. Consequently, it is natural to discretize u in terms of piecewise polynomials, i.e., discontinuous finite elements on this triangulation. Then the discretization of the non-smooth term $|u|_{TV(S)}$ is not straightforward, at least not for linear or higher-order finite element functions. We come back to this issue in Sect. 2.4. Finally, as observed previously in Carter [9], Chambolle [10], Chan et al. [12], Hintermüller and Kunisch [24], we mention that the predual variable $\boldsymbol{p}$ serves as an edge detector and thus is a quantity of interest. The image u can be recovered from $\boldsymbol{p}$ using (2.7).

Problem (2.6) features nonlinear inequality constraints $|\boldsymbol{p}|_{\mathfrak{g}} \leq \beta$. It can be solved, e.g., by a function space logarithmic barrier method as proposed in Herrmann et al. [23]. To this end, we consider the following family of convex problems for a decreasing sequence of barrier parameters $\mu \searrow 0$:

$$\begin{cases} \text{Minimize} & \frac{1}{2}\|\operatorname{div} \boldsymbol{p} + K^* f\|_{B^{-1}}^2 - \mu \int_S \ln\left(\beta^2 - |\boldsymbol{p}|_{\mathfrak{g}}^2\right) \mathrm{d}\boldsymbol{s} \\ \text{over} & \boldsymbol{p} \in \boldsymbol{H}(\operatorname{div}; S) \\ \text{subject to} & |\boldsymbol{p}|_{\mathfrak{g}} \leq \beta \text{ a.e. on } S. \end{cases} \tag{2.8}$$

For any fixed barrier parameter $\mu > 0$, problem (2.8) can be solved using Newton's method. The following result characterizes optimality for (2.8).

Theorem 2.3 *For every* $\mu > 0$*, problem* (2.8) *possesses a unique solution* $\boldsymbol{p} \in \boldsymbol{H}(\operatorname{div}; S)$*. The vector field* $\boldsymbol{p} \in \boldsymbol{H}(\operatorname{div}; S)$ *is the unique solution for* (2.8) *if and only if* $|\boldsymbol{p}|_{\mathfrak{g}} \leq \beta$ *holds a.e. on* S *and*

$$\left(\operatorname{div} \boldsymbol{p} + K^* f,\ \operatorname{div} \delta\boldsymbol{p}\right)_{B^{-1}} + \mu \int_S \frac{2\,(\boldsymbol{p}, \delta\boldsymbol{p})_{\mathfrak{g}}}{\beta^2 - |\boldsymbol{p}|_{\mathfrak{g}}^2}\,\mathrm{d}\boldsymbol{s} = 0 \tag{2.9}$$

for all $\delta\boldsymbol{p} \in \boldsymbol{H}(\operatorname{div}; S)$.

2.3 Implementation Details and Numerical Results

We implemented a discrete version of the barrier method in FEniCS [28]. We discretized the data f in the finite element space $\mathcal{DG}_r$ on the triangulated surface, consisting of piecewise polynomials of degree up to r. The dual variable is discretized using $\boldsymbol{H}(\mathrm{div}; S)$-conforming discretization by surface Raviart–Thomas (RT) finite elements of degree $r + 1$, see Raviart and Thomas [32], Ern and Guermond [15, Chapter 1.4.7] or Logg et al. [28, Ch. 3.4.1]. To recover the image u from $\boldsymbol{p}$ by (2.7) we choose matching polynomial degrees, i.e., $u \in \mathcal{DG}_r$.

We reproduce here three results from Herrmann et al. [23]. The first is a classical gray-scale denoising problem with $K = \mathrm{id}$ on the geometry depicted in Fig. 1, consisting of 354,330 triangles and 177,167 vertices. We choose the polynomial degree $r = 2$ for this case. The image data f is scaled to $[0, 1]$. We add artificial noise based on a normal distribution with standard deviation $\sigma = 0.1$ and zero mean to each entry in the coefficient vector representing f. The denoising results for parameters $\alpha = 0$ and $\beta = 0.3$ are shown in Fig. 1.

The second example is a color denoising problem. The geometry consists of 100,000 triangles and 50,002 vertices. Due to the fine features on the sole and a leathery texture on the outside, we utilize a polynomial degree $r = 3$ here. Noise is added in the same way to each of the RGB channels as described for the gray-scale test case above. The denoising results for parameters $\alpha = 0$ and $\beta = 0.5$ are shown in Fig. 2. The denoising procedure was conducted individually per RGB channel.

We conclude this subsection mentioning a third numerical experiment based on color inpainting in the shoe geometry. We simulate a 3% loss of data, as shown in Fig. 3. Contrary to the denoising situation, K^*K is no longer invertible and $\alpha > 0$ is required. Results for $\alpha = 0.1$ are given in Fig. 3.

2.4 Discrete Total Variation

Primal-dual algorithms constitute a popular class of methods for the solution of TV–L^2 and related problems. Their efficient implementation relies on a convenient

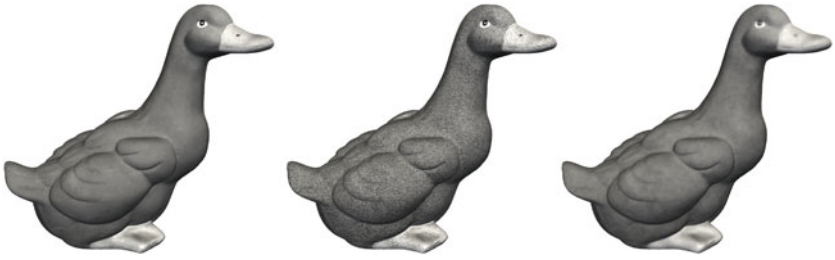

Fig. 1 Duck test case: noise free (left) and noisy (middle) texture and denoising result (right) for $\alpha = 0$ and $\beta = 0.3$. The object was kindly scanned by the Rechenzentrum of Würzburg University

Fig. 2 Shoe test case: noise free (left) and noisy (middle) originals and denoising result (right) for $\alpha = 0$ and $\beta = 0.5$

Fig. 3 Shoe with missing texture (left) and TV-inpainting solutions for $\alpha = 0.1$, $\beta = 0.7$ (middle), and $\beta = 1.0$ (right)

dual representation of the TV-seminorm in an appropriate discretized setting. As we mentioned previously, this is not obvious for discretizations by higher-order finite element functions.

For simplicity of the presentation, we restrict the discussion now to TV–L^2 on flat domains Ω rather than surfaces. We consider discretizations of problem (1.5) where $u \in \mathcal{DG}_r(\Omega)$, i.e., images are represented as piecewise polynomials of degree at most r on a given triangulation of Ω. We assume a geometrically conforming mesh throughout, i.e., hanging nodes are excluded.

In the particular case where $u \in \mathcal{DG}_0(\Omega)$, the gradient ∇u is a collection of line measures concentrated on the interior edges E of the triangulation. Indeed, $\nabla u_{|E} = [\![u]\!]_E \, \boldsymbol{n}_E$ holds, where $[\![u]\!]_E$ is the scalar jump of u across E, and $\boldsymbol{n}_E$ is the edge normal. The latter can have either of the two orientations.

The TV-seminorm of $u \in \mathcal{DG}_0(\Omega)$ in the isotropic case ($s = 2$) is easily seen to be

$$|u|_{TV(\Omega)} = \sum_E \left| [\![u]\!]_E \right| \, |E|, \tag{2.10}$$

where $|E|$ is the length of the edge E. It has been proved for the lowest-order case $r = 0$ that (2.10) can also be characterized as

$$|u|_{TV(\Omega)} = \sup \left\{ \int_\Omega u \operatorname{div} \boldsymbol{p} \, \mathrm{d}\boldsymbol{x} : \boldsymbol{p} \in \mathcal{RT}_1^0, \ |\boldsymbol{p} \cdot \boldsymbol{n}_E| \le 1 \text{ for all interior edges } E \right\}, \tag{2.11}$$

see for instance Bartels [3, 4]. Here $\mathcal{RT}_1(\Omega)$ is the lowest-order Raviart–Thomas finite element space on the same triangulation which underlies $\mathcal{DG}_0(\Omega)$. Moreover, $\mathcal{RT}_1^0(\Omega)$ is the subspace of functions satisfying $\boldsymbol{p} \cdot \boldsymbol{n} = 0$ along the boundary $\partial\Omega$.

The observation that the dual representation (2.11) utilizes only a finite dimensional subspace of $\boldsymbol{H}(\operatorname{div}; \Omega)$ is important, particularly for primal/dual or purely dual numerical approaches for problem (1.5). It motivated us to consider in Herrmann et al. [22] primal/dual representations for discretizations by higher-order finite elements, and specifically for $u \in \mathcal{DG}_r(\Omega)$ with $0 \le r \le 4$, and for general $s \in [1, \infty]$. In this case, it is not hard to see that the TV-seminorm (1.2) can be evaluated as

$$|u|_{TV(\Omega)} = \sum_T \int_T |\nabla u|_s \, \mathrm{d}\boldsymbol{x} + \sum_E \int_E |[\![u]\!]|_s \, \mathrm{d}\boldsymbol{s}, \tag{2.12}$$

where T is a triangle and $[\![u]\!]$ denotes the vector-valued jump of a function in normal direction across an interior edge E of the triangulation. Unfortunately, as soon as the polynomial degree exceeds $r = 0$, (2.12) does not possess a dual representation such as (2.11), for which it suffices to let $\boldsymbol{p}$ range over a finite subspace of $\boldsymbol{H}(\operatorname{div}; \Omega)$. We therefore propose to replace (2.12) by the approximate analogue

$$|u|_{DTV(\Omega)} \coloneqq \sum_T \int_T \mathcal{I}_T \{ |\nabla u|_s \} \, \mathrm{d}\boldsymbol{x} + \sum_E \int_E \mathcal{I}_E \{ |[\![u]\!]|_s \} \, \mathrm{d}\boldsymbol{s}, \tag{2.13}$$

which we term the *discrete TV-seminorm*. Here $\mathcal{I}_T$ and $\mathcal{I}_E$ are local interpolation operators into the polynomial spaces $\mathcal{P}_{r-1}(T)$ and $\mathcal{P}_r(E)$, respectively. In terms of the standard Lagrangian bases $\{\varphi_{T,i}\}$ and $\{\varphi_{E,j}\}$ of these spaces, they are defined as

$$\begin{aligned} \int_T \mathcal{I}_T \{ |\nabla u|_s \} \, \mathrm{d}\boldsymbol{x} &= \sum_{i=1}^{r(r+1)/2} |\nabla u(x_{T,i})|_s \, c_{T,i}, \\ \int_E \mathcal{I}_E \{ |[\![u]\!]|_s \} \, \mathrm{d}\boldsymbol{s} &= \sum_{j=1}^{r+1} |[\![u]\!](x_{E,j})| \, |\boldsymbol{n}_E|_s \, c_{E,j}, \end{aligned} \tag{2.14}$$

where the weights are given by

$$c_{T,i} \coloneqq \int_T \varphi_{T,i} \, \mathrm{d}\boldsymbol{x} \quad \text{and} \quad c_{E,j} \coloneqq \int_E \varphi_{E,j} \, \mathrm{d}\boldsymbol{s}. \tag{2.15}$$

In virtue of the fact that $\nabla u_{|T} \in \mathcal{P}_{r-1}(T)^2$ and $[\![u]\!] \in \mathcal{P}_r(E)$, it is clear that $|\cdot|_{DTV(\Omega)}$ is indeed a seminorm on $\mathcal{DG}_r(\Omega)$, provided that all weights $c_{T,i}$ and $c_{E,j}$ are non-negative. As shown in Herrmann et al. [22, Lemma 3.1], this is the case for the polynomial degrees $0 \le r \le 4$ under consideration. We thus obtain the following dual representation of $|u|_{DTV(\Omega)}$, whose proof can be found in Herrmann et al. [22, Theorem 3.2].

Theorem 2.4 *Suppose $0 \le r \le 4$. Then for any $u \in \mathcal{DG}_r(\Omega)$, the discrete TV-seminorm (2.13) satisfies*

$$\begin{aligned} |u|_{DTV(\Omega)} = \sup \Bigg\{ \int_\Omega u \operatorname{div} \boldsymbol{p} \,\mathrm{d}\boldsymbol{x} \,:\, & \boldsymbol{p} \in \mathcal{RT}^0_{r+1}(\Omega), \\ & |\boldsymbol{\sigma}_{T,i}(\boldsymbol{p})|_{s^*} \le c_{T,i} \text{ for all } T, i = 1, \dots, r\,(r+1)/2, \\ & |\sigma_{E,j}(\boldsymbol{p})| \le |\boldsymbol{n}_E|_s \, c_{E,j} \text{ for all } E,\ j = 1, \dots, r+1 \Bigg\}. \end{aligned} \tag{2.16}$$

In (2.16), $\boldsymbol{\sigma}_{T,i}$ and $\sigma_{E,j}$ are a standard set of degrees of freedom for the Raviart–Thomas space $\mathcal{RT}_{r+1}(\Omega)$, namely

$$\boldsymbol{\sigma}_{T,i}(\boldsymbol{p}) \coloneqq \int_T \varphi_{T,i}\, \boldsymbol{p} \,\mathrm{d}\boldsymbol{x}, \quad i = 1, \dots, r\,(r+1)/2, \tag{2.17a}$$

$$\sigma_{E,j}(\boldsymbol{p}) \coloneqq \int_E \varphi_{E,j}\, (\boldsymbol{p} \cdot \boldsymbol{n}_E) \,\mathrm{d}s, \quad j = 1, \dots, r+1. \tag{2.17b}$$

Observe that for $r = 0$, Eq. (2.16) boils down to (2.11).

As an application for the discrete total variation and its dual formulation, we focus on the resolution of the following discrete TV–L^2 problem:

$$\text{Minimize} \quad \frac{1}{2}\|u - f\|^2_{L^2(\Omega_0)} + \beta\, |u|_{DTV(\Omega)}. \tag{DTV-L2}$$

This problem is a pure denoising problem when $\Omega_0 = \Omega$ and a combined inpainting and denoising problem when $\Omega_0 \subsetneq \Omega$. Using Theorem 2.4, then Fenchel dual of (DTV-L2) can be derived in a straightforward way.

Theorem 2.5 *Let $0 \le r \le 4$. Then the dual problem of* (DTV-L2) *is*

$$\textit{Minimize} \quad \frac{1}{2}\|\operatorname{div} \boldsymbol{p} + f\|^2_{L^2(\Omega_0)} \textit{ s.t. } \boldsymbol{p} \in \beta \boldsymbol{P}, \tag{DTV-L2-D}$$

where the admissible set is

$$\begin{aligned} \boldsymbol{P} \coloneqq \Big\{ \boldsymbol{p} \in \mathcal{RT}^0_{r+1}(\Omega) : & |\boldsymbol{\sigma}_{T,i}(\boldsymbol{p})|_{s^*} \le c_{T,i} \text{ for all } T \text{ and all } i, \\ & |\sigma_{E,j}(\boldsymbol{p})| \le |\boldsymbol{n}_E|_s \, c_{E,j} \text{ for all } E \text{ and all } j \Big\}. \end{aligned} \tag{2.18}$$

Notice $\boldsymbol{p} \in \beta \boldsymbol{P}$ means that $\boldsymbol{p}$ satisfies constraints as in (2.18) but with $c_{T,i}$ and $c_{E,j}$ replaced by $\beta\, c_{T,i}$ and $\beta\, c_{E,j}$, respectively. As in Sect. 2.2, we can recover the (unique) solution of the primal problem from the solution of the dual problem in case $\Omega_0 = \Omega$ as follows:

$$u = \operatorname{div} \boldsymbol{p} + f \in \mathcal{DG}_r(\Omega). \tag{2.19}$$

In case $\Omega_0 \subsetneq \Omega$, the primal and dual solutions are, in general, not unique.

Having established the dual problem gives us the opportunity of applying purely dual methods [23, 24], or primal-dual algorithms, including Chambolle–Pock [11] and split Bregman [18]. In what follows, we focus on the latter two and refer the reader to the extended preprint Herrmann et al. [21] for a full account, including TV–L^1 problems. Our goal is to show that these methods can be applied to solve (DTV-L2) for polynomial degrees $1 \le r \le 4$ in an equally efficient way as for the standard setting $r = 0$. We also aim to exhibit the benefits of polynomial orders $r \ge 1$ for image quality, both for denoising and inpainting applications. We restrict the discussion to the isotropic case $s = 2$ and $r \in \{0, 1, 2\}$ and refer the reader to Herrmann et al. [22] for more results.

In our tests, we use the two images displayed in Fig. 4, and we measure image quality according to the common peak signal-to-noise ratio, shortened by PSNR. Our first numerical example is the denoising images in $\mathcal{DG}_r(\Omega)$. We represent (interpolate) the non-discrete image displayed in Fig. 4 (middle) in the space $\mathcal{DG}_r(\Omega)$ for $r = 0, 1, 2$. We show the convergence results for the split Bregman and Chambolle–Pock methods in Table 1. We employed the primal-dual gap in combination with a measure of dual infeasibility as a stopping criterion for both algorithms, see Herrmann et al. [22] for details. As we can see from the PSNR value, the quality of the reconstructed image is best for $r = 1$ and drops again for $r = 2$. These results, however, have to be interpreted in the light of the fact that we added noise to each coefficient of the image, i.e., the image data is more corrupted in case of $r = 2$.

We continue with the denoising of low-resolution images. Here, we work with the cameraman image presented in Fig. 4 (left) and apply the split Bregman method

Fig. 4 Left: Cameraman pixel test image. Middle: Non-discrete test image. Right: Mesh used to represent the image in the middle

Table 1 Comparison of the performance of split Bregman (SB) *vs* Chambolle–Pock (CP) for the denoising problem in various discretizations

Space	Algorithm	Iterations	Time [s]	PSNR
$\mathcal{DG}_0$	SB	37	1.6	32.031
	CP	128	3.4	31.987
$\mathcal{DG}_1$	SB	57	5.8	36.092
	CP	91	6.7	33.480
$\mathcal{DG}_2$	SB	41	9.3	31.896
	CP	223	35.1	31.066

Table 2 Performance of split Bregman (SB) for the low-resolution denoising problem in various discretizations

Space	Algorithm	Iterations	Time [s]	PSNR
$\mathcal{DG}_0$	SB	20	6.3	19.333
$\mathcal{DG}_2$	SB	101	84.3	20.855

Table 3 Performance of Chambolle–Pock for the (DTV-L2) inpainting problem in various discretizations

Space	Algorithm	Iterations	Time [s]	PSNR
$\mathcal{DG}_0$	CP	2031	47.7	23.617
$\mathcal{DG}_1$	CP	697	49.0	26.788
$\mathcal{DG}_2$	CP	2286	354.0	26.385

because it performed slightly better than Chambolle–Pock in the previous denoising example. More precisely, we consider a low resolution of the cameraman image, which was obtained by interpolating the 256×256 pixel image onto a 64×64 square pixel grid with crossed diagonals. As shown in Table 2 and [22, Figure 7.4], the use of higher-order polynomial functions can partially compensate the loss of geometric resolution.

Finally, we demonstrate the utility of *discrete* algorithms for the purpose of denoising and inpainting. To this end, we consider the non-discrete "ball" image and randomly delete two thirds of all cells, which subsequently serves as the inpainting region $\Omega \setminus \Omega_0$. Noise is added to the remaining data and we solve problem (DTV-L2) in $\mathcal{DG}_r(\Omega)$ for $r \in \{0, 1, 2\}$. In this case, Chambolle–Pock performs better than split Bregman; see Table 3. Clearly, the higher-order results produce images closer to the original than the recovery in $\mathcal{DG}_0$, which is reflected in the PSNR values.

3 Shape Optimization Using Total Variation of the Normal Vector Field

In this section, we discuss the total variation of the normal vector field of a discrete surface Γ as a regularizer for shape optimization problems. This regularizer was first introduced in Bergmann et al. [6]. We demonstrate its utility in mesh denoising and geometric inverse problems. To solve these problems, we propose a Riemannian ADMM iteration, which generalizes the split Bregman algorithm for total variation problems from Goldstein and Osher [18].

3.1 Total Variation of Normal

Our total variation approach is tailored to the most common representation of discrete surfaces, by meshes consisting of flat triangles T and straight sided edges E. We assume geometric conformity, i.e., no hanging nodes are allowed. These surface representations give rise to piecewise constant normal fields $\boldsymbol{n}$, whose variation is concentrated in spontaneous changes across edges between triangles. We assume that each edge E has an arbitrary but fixed orientation, so that the normal vectors of its two neighboring triangles can be referenced as $\boldsymbol{n}_E^+, \boldsymbol{n}_E^-$. All normals are elements of the unit sphere $\mathcal{S}^2 = \{\boldsymbol{v} \in \mathbb{R}^3 : |\boldsymbol{v}|_2 = 1\}$. The latter is a smooth manifold equipped with a Riemannian metric $\mathfrak{g}$, which we take here to be the pullback of the Euclidean metric from $\mathbb{R}^3$.

An important detail in our approach is us measuring the distance between neighboring normals geodesically. On $\mathcal{S}^2$, the logarithmic map $\log_{\boldsymbol{n}_E^+} \boldsymbol{n}_E^- \in \mathcal{T}_{\boldsymbol{n}_E^+}\mathcal{S}^2$, which specifies the unique tangent vector at $\boldsymbol{n}_E^+$ such that the shortest geodesic departing from $\boldsymbol{n}_E^+$ in that direction will reach $\boldsymbol{n}_E^-$ at unit time, is given explicitly by

$$\log_{\boldsymbol{n}_E^+} \boldsymbol{n}_E^- = \arccos\left((\boldsymbol{n}_E^+)^\top \boldsymbol{n}_E^-\right) \frac{\boldsymbol{n}_E^- - (\boldsymbol{n}_E^{+\top} \boldsymbol{n}_E^-)\boldsymbol{n}_E^+}{\left|\boldsymbol{n}_E^- - (\boldsymbol{n}_E^{+\top} \boldsymbol{n}_E^-)\boldsymbol{n}_E^+\right|_{\mathfrak{g}}}.$$

The logarithmic map is well-defined whenever $\boldsymbol{n}_E^+ \neq -\boldsymbol{n}_E^-$. The distance between these two normals amounts to

$$d(\boldsymbol{n}_E^+, \boldsymbol{n}_E^-) := \left|\log_{\boldsymbol{n}_E^+} \boldsymbol{n}_E^-\right|_{\mathfrak{g}} = \arccos\left((\boldsymbol{n}_E^+)^\top \boldsymbol{n}_E^-\right), \tag{3.1}$$

which agrees with the angle between them. Similar to (2.10), we define the total variation of the normal as

$$|\boldsymbol{n}|_{TV(\Gamma)} := \sum_E d(\boldsymbol{n}_E^+, \boldsymbol{n}_E^-)|E| = \sum_E |\log_{\boldsymbol{n}_E^+} \boldsymbol{n}_E^-|_{\mathfrak{g}}|E|, \tag{3.2}$$

where $|E|$ is the Euclidean length of the straight edge E. Note that we are again facing a non-differentiable regularization term. The non-differentiability occurs for surfaces having at least one pair of neighboring normals that are equal, i.e., $\boldsymbol{n}_E^+ = \boldsymbol{n}_E^-$, indicating to a flat patch of two or more neighboring triangles.

3.2 Mesh Denoising

Meshes are widely employed in computer graphics and computer vision. There they are used to capture digital surface data, and they can approximate general surfaces with arbitrary geometry. Moreover, meshes are easy to obtain via 3D scanning devices and they are easy to store and manipulate by software. Even when using

high-fidelity scanners, measurement errors are bound to be present in the data, which motivates the use of preprocessing steps in order to clean up the mesh before further computations. The process of removing such errors while preserving relevant features is known as *mesh denoising*.

The main difficulty in removing undesired noise from a mesh is that both, noise and sharp features, are considered high frequency signals from the point of view of imaging processing. This issue makes it difficult to distinguish between them. The aforementioned problem has been of interest in the community of image processing since late in the 1980s, and many algorithms and approaches for mesh denoising have been developed so far. We refer to Botsch et al. [7] for a survey in denoising algorithms.

We consider the problem of denoising a noisy mesh $\tilde{\Gamma}$, which is topologically identical to Γ. That is, a triangle $\tilde{T}$ of $\tilde{\Gamma}$ can be written as the image $\tilde{T} = \{(\mathrm{id} + \boldsymbol{V}_\Gamma)(\boldsymbol{s}), \boldsymbol{s} \in T\}$, where T is the triangle of the unknown surface mesh Γ corresponding to $\tilde{T}$ and $\boldsymbol{V}_\Gamma$ is an affine transformation. Similar to (DTV-L2), we arrive at

$$\text{Minimize} \quad \frac{1}{2}\|\boldsymbol{V}_\Gamma\|^2_{\boldsymbol{L}^2(\Gamma)} + \beta \sum_E |\log_{\boldsymbol{n}_E^+} \boldsymbol{n}_E^-|_{\mathfrak{g}} |E|, \tag{3.3}$$

where the vertex positions of Γ serve as optimization variables.

Riemannian Split Bregman Method To address the non-smoothness of (3.2) and due to the similarity with (DTV-L2), we propose a Riemannian split Bregman method. To this end, we introduce the following new variable $\boldsymbol{d}_E$ per edge:

$$\boldsymbol{d}_E = \log_{\boldsymbol{n}_E^+} \boldsymbol{n}_E^- \in \mathcal{T}_{\boldsymbol{n}_E^+}\mathcal{S}^2.$$

To couple the new variable into the problem, we propose the following augmented Lagrangian formulation:

$$\mathcal{L}(\Gamma, \boldsymbol{d}, \boldsymbol{b}) := \frac{1}{2}\|\boldsymbol{V}_\Gamma\|^2_{\boldsymbol{L}^2(\Gamma)} + \beta \sum_E |\boldsymbol{d}_E|_{\mathfrak{g}} |E| + \frac{\gamma}{2} \sum_E \left|\boldsymbol{d}_E - \log_{\boldsymbol{n}_E^+} \boldsymbol{n}_E^- - \boldsymbol{b}_E\right|^2_{\mathfrak{g}} |E|, \tag{3.4}$$

where each $\boldsymbol{b}_E \in \mathcal{T}_{\boldsymbol{n}_E^+}\mathcal{S}^2$ is a Lagrange multiplier and the vectors $\boldsymbol{d}$ and $\boldsymbol{b}$ are simply the collections of their entries $\boldsymbol{d}_E, \boldsymbol{b}_E \in \mathcal{T}_{\boldsymbol{n}_E^+}\mathcal{S}^2$, one per edge E. The main difference to an ADMM method in Hilbert spaces is that the vectors $\boldsymbol{d}_E$ and $\boldsymbol{b}_E$ have values in the tangent space $\mathcal{T}_{\boldsymbol{n}_E^+}\mathcal{S}^2$. This necessitates an update of the Lagrange multiplier estimates $\boldsymbol{b}_E$ whenever the vertex positions of Γ and thus the normal vectors are changing. These updates are realized via parallel transport of a tangent vector from one tangent space to another, along the unique shortest geodesic connecting the base points. Specifically, the parallel transport $P_{\boldsymbol{n}\to\boldsymbol{n}'}\colon \mathcal{T}_{\boldsymbol{n}}\mathcal{S}^2 \to$

$\mathcal{T}_{\boldsymbol{n}'}\mathcal{S}^2$ is given by

$$P_{\boldsymbol{n}\to\boldsymbol{n}'}(\boldsymbol{\xi}) = \boldsymbol{\xi} - \frac{\boldsymbol{\xi}^\top(\log_{\boldsymbol{n}}\boldsymbol{n}')}{\arccos(\boldsymbol{n}^\top\boldsymbol{n}')}(\log_{\boldsymbol{n}}\boldsymbol{n}' + \log_{\boldsymbol{n}'}\boldsymbol{n}), \tag{3.5}$$

see for instance Hosseini and Uschmajew [25] and Persch [31, Section 2.3.1]. Apart from these necessary adaptations, the main idea of the split Bregman method remains to successively minimize (3.4) w.r.t. each variable, Γ and $\boldsymbol{d}$, independently in addition to using a simple update formula for the vector of Lagrange multipliers $\boldsymbol{b}$.

Instead of minimizing (3.4) w.r.t. the vertex positions defining Γ to a certain accuracy, in practice we only perform one gradient step per iteration. This is in line with Goldstein and Osher [18], where a Gauss–Seidel sweep is proposed. We obtain the sensitivity information for this gradient step via automatic differentiation w.r.t. the vertex positions available in the FEniCS framework, see Ham et al. [19].

Afterwards, the minimization of (3.4) w.r.t. $\boldsymbol{d}$ can be done explicitly by one vectorial shrinkage operation per edge E in the respective tangent space $\mathcal{T}_{\boldsymbol{n}_E^+}\mathcal{S}^2$. We refer the reader to Bergmann et al. [6] for more details.

We present numerical experiments confirming the performance of our approach. For this purpose, we use as a benchmark the well-known "fandisk" geometry, where noise was added in normal direction with standard deviation $\sigma = 0.2$. The results obtained can be seen in Fig. 5.

3.3 Inverse Problem

Finally, we demonstrate the utility of the total variation of the normal as a regularizer in an inverse problem of electrical impedance tomography (EIT) type.

EIT Model Problem Traditionally, EIT problems are modeled with Neumann (current) boundary conditions, and the internal conductivity is an unknown function across the entire domain. In order to focus on the demonstration of the utility of (3.2) as a regularizer in geometric inverse problems, we consider a simplified situation in which we seek to reconstruct a perfect conductor inside a domain of otherwise homogeneous electrical properties. Hence, we arrive at the following problem:

$$\begin{aligned} &\text{Minimize} \quad \frac{1}{2}\sum_{i=1}^{r}\int_{\Gamma_2}|u_i - z_i|^2 \,\mathrm{d}\boldsymbol{s} + \beta\ |\boldsymbol{n}|_{TV(\Gamma_1)} \\ &\text{s.t.} \quad \begin{cases} -\Delta u_i = 0 & \text{in } \Omega, \\ \dfrac{\partial u_i}{\partial \boldsymbol{n}} = 0 & \text{on } \Gamma_1, \\ \dfrac{\partial u_i}{\partial \boldsymbol{n}} + \alpha\, u_i = f_i & \text{on } \Gamma_2. \end{cases} \end{aligned} \tag{3.6}$$

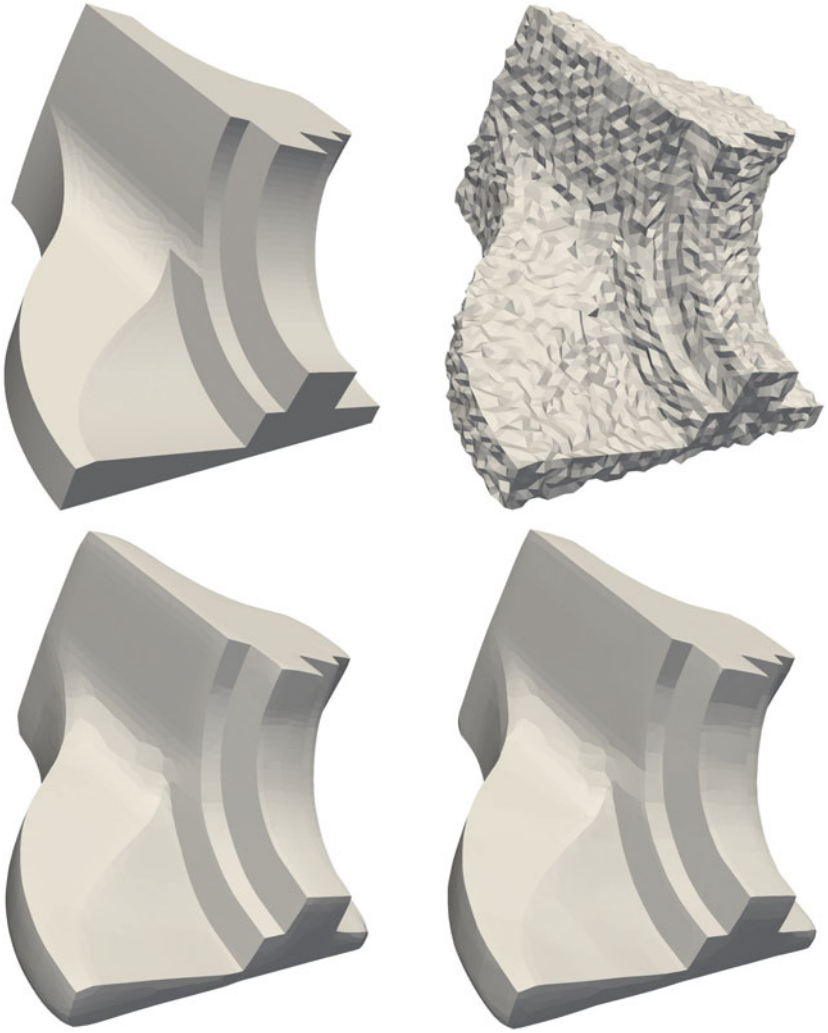

Fig. 5 Top left: original geometry, top right: noisy geometry, bottom left: result for $\beta = 10^{-4}$ and $\gamma = 10^{-1}$, bottom right: result for $\beta = 10^{-6}$ and $\gamma = 10^{-3}$

The unknowns of the problem are $(u_1, \ldots, u_r, \Gamma_1)$. Here, Γ_1 denotes the unknown inclusion in the interior of Ω, and u_i is the electric potential generated by the boundary forcing f_i for $i = 1, \ldots, r$. The tracking data are surface measurements z_i on the known outer boundary Γ_2. We treat the PDE state u_i as directly dependent on Γ_1 and the sensitivities of $u_i(\Gamma_1)$ are eliminated via an adjoint approach.

We present numerical results for the impedance tomography model problem described in the previous section. The state u and adjoint state were discretized using piecewise linear, globally continuous finite elements on a tetrahedral grid of Ω

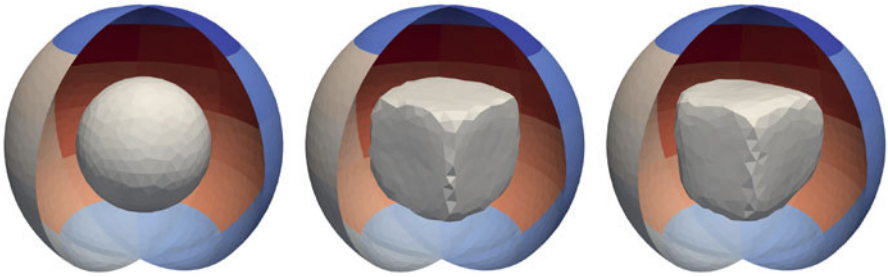

Fig. 6 Initial shape, TV-regularized reconstruction without noise and TV-regularized reconstruction with noise. The color on the outer shell denotes the support of each source term for $i = 1, \ldots, 48$ in (3.6)

minus the volume enclosed by Γ_1. The mesh has 4429 vertices and 41272 tetrahedra. The algorithmic parameters are $\beta = 10^{-6}$, $\gamma = 10^{-1}$, and $\alpha = 10^{-5}$. Each f_i is constant with value 1.0 on its support, which is shown color-coded in Fig. 6.

We also show in Fig. 6 the reconstruction results for Γ_1 obtained in the noise-free setting and with noise. In the latter case, we added normally distributed random noise with zero mean and standard deviation $\sigma = 10^{-2}$ per degree of freedom on the fixed measurement surface Γ_2 for each of the $r = 48$ simulations of the forward model (3.6). The amount of noise has to be interpreted in relation to the range of values for the simulated state, which is

$$\max_{s \in \Gamma_2} u_i(s) - \min_{s \in \Gamma_2} u_i(s) \approx 0.3, \quad i = 1, \ldots, r.$$

In each case, the initial guess for Γ_1 was the surface of the ball $B_{0.5}(0)$ while the true solution is a cube. As can be seen from Fig. 6, the total variation functional helps to reconstruct this shape quite nicely.

4 Conclusion and Outlook

The main aspect of this work is to survey TV-based regularizers fostering non-smoothness in a range of reconstruction problems. To this end, both image and surface problems are considered. With respect to images, we discussed a (pre-)dual denoising scheme for images on surfaces. This approach is subsequently adapted to a discrete setting with higher-order finite elements by introducing a discrete total variation concept, which allows for a convenient dual problem. Numerical results are provided both for flat images as well as on surfaces.

Concerning three dimensional geometric reconstruction problems, we defined the total variation of the normal vector field as a regularizer. This adds the additional difficulty that the unknown is now a geometric object. Furthermore, the input to

the regularizer, i.e., the surface normal, has values in a manifold, the unit sphere. Hence, previously introduced algorithms need to be adapted to operate in the appropriate Riemannian setting, which includes parallel transport to compensate shifting tangent spaces when the unknown surface is updated. The survey concludes with an application in electrical impedance tomography.

As an outlook into the future, the methodologies reviewed above can all be used as a starting point to extend a variety of approaches originating from mathematical imaging to the reconstruction of surfaces, when the corresponding algorithms are adapted to work both on and with manifolds. Such examples could include directional effects into the TV concept or the inpainting of meshes, where subregions of the mesh are missing, a problem naturally and frequently arising in 3D scanning of objects with non-convex parts.

References

1. Attouch, H., G. Buttazzo, and G. Michaille (2006). *Variational Analysis in Sobolev and BV Spaces*. Vol. 6. MPS/SIAM Series on Optimization. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), pp. xii+634.
2. Bachmayr, M. and M. Burger (2009). "Iterative total variation schemes for non-linear inverse problems". In: *Inverse Problems* 25.10, pp. 105004, 26. https://doi.org/10.1088/0266-5611/25/10/105004.
3. Bartels, S. (2012). "Total variation minimization with finite elements: convergence and iterative solution". In: *SIAM Journal on Numerical Analysis* 50.3, pp. 1162–1180. https://doi.org/10.1137/11083277X.
4. Bartels, S. (2015). "Error control and adaptivity for a variational model problem defined on functions of bounded variation". In: *Mathematics of Computation* 84.293, pp. 1217–1240. https://doi.org/10.1090/S0025-5718-2014-02893-7.
5. Ben-Artzi, M. and P. G. LeFloch (2007). "Well-posedness theory for geometry-compatible hyperbolic conservation laws on manifolds". In: *Annales de l'Institut Henri Poincaré. Analyse Non Linéaire* 24.6, pp. 989–1008. https://doi.org/10.1016/j.anihpc.2006.10.004.
6. Bergmann, R., M. Herrmann, R. Herzog, S. Schmidt, and J. Vidal-Núñez (2019). *Total Variation of the Normal Vector Field as Shape Prior*. arXiv: 1902.07240.
7. Botsch, M., M. Pauly, L. Kobbelt, P. Alliez, B. Lévy, S. Bischoff, and C. Rössl (2007). "Geometric Modeling Based on Polygonal Meshes". In: *ACM SIG-GRAPH 2007 Courses*. SIGGRAPH '07. ACM. https://doi.org/10.1145/1281500.1281640.
8. Cai, J.-F., S. Osher, and Z. Shen (2009). "Linearized Bregman iterations for frame-based image deblurring". In: *SIAM Journal on Imaging Sciences* 2.1, pp. 226–252. https://doi.org/10.1137/080733371.
9. Carter, J. L. (2001). "Dual Methods for Total Variation-Based Image Restoration". PhD thesis. UCLA.
10. Chambolle, A. (2004). "An algorithm for total variation minimization and applications". In: *Journal of Mathematical Imaging and Vision* 20.1-2. Special issue on mathematics and image analysis, pp. 89–97. https://doi.org/10.1023/B:JMIV.0000011325.36760.1e.
11. Chambolle, A. and T. Pock (2011). "A first-order primal-dual algorithm for convex problems with applications to imaging". In: *Journal of Mathematical Imaging and Vision* 40.1, pp. 120–145. https://doi.org/10.1007/s10851-010-0251-1.
12. Chan, T. F., G. H. Golub, and P. Mulet (1999). "A nonlinear primal-dual method for total variation-based image restoration". In: *SIAM Journal on Scientific Computing* 20.6, pp. 1964–1977. https://doi.org/10.1137/S1064827596299767.

13. Chan, T. F. and J. Shen (2005). *Image processing and analysis*. Variational, PDE, wavelet, and stochastic methods. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, pp. xxii+400. https://doi.org/10.1137/1.9780898717877.
14. Dobson, D. C. and F. Santosa (1996). "Recovery of blocky images from noisy and blurred data". In: *SIAM Journal on Applied Mathematics* 56.4, pp. 1181–1198. https://doi.org/10.1137/S003613999427560X.
15. Ern, A. and J.-L. Guermond (2004). *Theory and Practice of Finite Elements*. Berlin: Springer.
16. Girault, V. and P.-A. Raviart (1986). *Finite Element Methods for Navier-Stokes Equations*. Springer.
17. Giusti, E. (1984). *Minimal Surfaces and Functions of Bounded Variation*. Vol. 80. Monographs in Mathematics. Birkhäuser Verlag, Basel, pp. xii+240. https://doi.org/10.1007/978-1-4684-9486-0.
18. Goldstein, T. and S. Osher (2009). "The split Bregman method for $L1$-regularized problems". In: *SIAM Journal on Imaging Sciences* 2.2, pp. 323–343. https://doi.org/10.1137/080725891.
19. Ham, D. A., L. Mitchell, A. Paganini, and F. Wechsung (2018). *Automated shape differentiation in the Unified Form Language*. Tech. rep. arXiv: 1808.08083
20. Hebey, E. (1996). *Sobolev spaces on Riemannian manifolds*. Vol. 1635. Lecture Notes in Mathematics. Springer-Verlag, Berlin, pp. x+116. https://doi.org/10.1007/BFb0092907.
21. Herrmann, M., R. Herzog, S. Schmidt, J. Vidal-Núñez, and G. Wachsmuth (2018). *Discrete total variation with finite elements and applications to imaging*. Extended preprint. arXiv: 1804.07477.
22. Herrmann, M., R. Herzog, S. Schmidt, J. Vidal-Núñez, and G. Wachsmuth (2019). "Discrete total variation with finite elements and applications to imaging". In: *Journal of Mathematical Imaging and Vision* 61.4, pp. 411–431. https://doi.org/10.1007/s10851-018-0852-7.
23. Herrmann, M., R. Herzog, H. Kröner, S. Schmidt, and J. Vidal-Núñez (2018). "Analysis and an interior point approach for TV image reconstruction problems on smooth surfaces". In: *SIAM Journal on Imaging Sciences* 11.2, pp. 889–922. https://doi.org/10.1137/17M1128022.
24. Hintermüller, M. and K. Kunisch (2004). "Total bounded variation regularization as a bilaterally constrained optimization problem". In: *SIAM Journal on Applied Mathematics* 64.4, pp. 1311–1333. https://doi.org/10.1137/S0036139903422784.
25. Hosseini, S. and A. Uschmajew (2017). "A Riemannian gradient sampling algorithm for nonsmooth optimization on manifolds". In: *SIAM Journal on Op-timization* 27.1, pp. 173–189. https://doi.org/10.1137/16M1069298.
26. Lai, R. and T. F. Chan (Dec. 2011). "A framework for intrinsic image processing on surfaces". In: *Computer Vision and Image Understanding* 115.12, pp. 1647–1661. https://doi.org/10.1016/j.cviu.2011.05.011.
27. Langer, A. (2017). "Automated parameter selection in the $L^1 - L^2$-TV model for removing Gaussian plus impulse noise". In: *Inverse Problems. An International Journal on the Theory and Practice of Inverse Problems, Inverse Methods and Computerized Inversion of Data* 33.7, pp. 074002, 41. https://doi.org/10.1088/1361-6420/33/7/074002.
28. Logg, A., K.-A. Mardal, G. N. Wells, et al. (2012). *Automated Solution of Differential Equations by the Finite Element Method*. Springer. https://doi.org/10.1007/978-3-642-23099-8.
29. Malgouyres, F. and F. Guichard (2001). "Edge direction preserving image zooming: a mathematical and numerical analysis". In: *SIAM Journal on Numerical Analysis* 39.1, 1–37 (electronic). https://doi.org/10.1137/S0036142999362286.
30. Monk, P. (2003). *Finite Element Methods for Maxwell's Equations*. Numerical Mathematics and Scientific Computation. New York: Oxford University Press.
31. Persch, J. (2018). "Optimization Methods in Manifold-Valued Image Processing". Dissertation. Technische Universität Kaiserslautern.
32. Raviart, P.-A. and J. M. Thomas (1977). "A mixed finite element method for 2nd order elliptic problems". In: *Mathematical aspects of finite element methods (Proc. Conf., Consiglio Naz. delle Ricerche (C.N.R.), Rome, 1975)*. Vol. 606. Lecture Notes in Mathematics. Berlin: Springer, pp. 292–315.

33. Rudin, L. I., S. Osher, and E. Fatemi (1992). “Nonlinear total variation based noise removal algorithms”. In: *Physica D* 60.1–4, pp. 259–268. https://doi.org/10.1016/0167-2789(92)90242-F.
34. Vogel, C. R. (2002). *Computational Methods for Inverse Problems* Philadelphia: SIAM.
35. Ziemer, W. P. (1989). *Weakly differentiable functions* Vol. 120. Graduate Texts in Mathematics. Sobolev spaces and functions of bounded variation. New York: Springer-Verlag, pp. xvi+308. https://doi.org/10.1007/978-1-4612-1015-3.

Rate-Independent Systems and Their Viscous Regularizations: Analysis, Simulation, and Optimal Control

Roland Herzog, Dorothee Knees, Christian Meyer, Michael Sievers, Ailyn Stötzner, and Stephanie Thomas

Abstract This chapter provides a survey on the analysis, simulation, and optimal control of a class of non-smooth evolution systems that appears in the modeling of dissipative solids. Our focus is on models that include internal constraints, such as a flow rule in plasticity, and that account for the temperature dependence of the respective materials. We discuss here two cases, namely purely rate-independent models and viscously regularized models coupled to the temperature equation.

Keywords Rate-independent system · Parametrized BV solution · Approximation schemes · Optimal control · Thermo-viscoplasticity

Mathematics Subject Classification (2020) 49J20 49J27 49J40 74C05 74H15

1 Introduction

This chapter provides a survey on the analysis, simulation, and optimal control of a class of non-smooth evolution systems that appears in the modeling of dissipative solids. Our focus is on models that include internal constraints, such as a flow

R. Herzog
Institute for Applied Mathematics, University of Heidelberg, Heidelberg, Germany
e-mail: roland.herzog@mathematik.tu-chemnitz.de

D. Knees (✉) · S. Thomas
Institute for Mathematics, University of Kassel, Kassel, Germany
e-mail: dknees@mathematik.uni-kassel.de; sthomas@mathematik.uni-kassel.de

C. Meyer · M. Sievers
TU Dortmund, Faculty for Mathematics, Dortmund, Germany
e-mail: christian.meyer@math.uni-dortmund.de; michael.sievers@math.uni-dortmund.de; michael.sievers@mathematik.tu-dortmund.de

A. Stötzner
TU Chemnitz, Faculty of Mathematics, Chemnitz, Germany
e-mail: ailyn.stoetzner@mathematik.tu-chemnitz.de

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_6

rule in plasticity, and that account for the temperature dependence of the respective materials. We discuss here two cases, namely purely rate-independent models and viscously regularized models coupled to the temperature equation.

Due to the internal constraints, the evolution of dissipative solids typically shows a non-smooth behavior. In the case of rate-independent models, one even has to deal with solutions that are discontinuous in time. All this is caused by a complex interplay of a non-smooth dissipation and a typically non-convex energy. For these reasons, the simulation of such systems is still a challenging topic, and the analytic frame and the numerical tools for optimal control thereof are not yet fully developed.

This chapter is structured as follows: We first give a short introduction to rate-independent systems and discuss different solution concepts (Sect. 2) for such systems. The class of solutions that we are focusing on in this chapter is the class of (parametrized) balanced viscosity solutions (BV solutions). In order to shorten the notation, we restrict ourselves to the semilinear case, [38]. Section 3 is devoted to the discussion of different discretization strategies for such systems. Here, we present results for the abstract semilinear case as well as results for concrete systems formulated in terms of partial differential equations and variational inequalities. In Sect. 4, we then describe first results on the optimal control of problems with rate-independent systems as constraints. Finally, in Sect. 5, we switch to rate-dependent systems that are coupled with the temperature equation.

2 Rate-Independent Systems and Solution Concepts

Within certain regimes, the behavior of many dissipative solids can be assumed to be rate-independent. Rate independence means that rescaling both the applied forces and the solutions in time in the same way implies that the rescaled solutions solve the rescaled model equations. Prominent examples are quasi-static elasto-plasticity, brittle damage or fracture, and shape memory alloys. Setting up the models in the framework of generalized standard materials [16], one assumes that the actual state of a mechanical structure is completely described by the displacement field belonging to some state space $\mathcal{U}$ and some internal variables belonging to a state space $\mathcal{Z}$. These internal variables describe on a macroscopic level the inner structure of the material and reflect the loading history. Constraints like friction laws or flow rules imply that the resulting model typically consists of a balance of linear momentum that is coupled with a doubly nonlinear differential inclusion characterizing the evolution of the internal variable z. Given a stored energy functional $\mathcal{E} : [0, T] \times \mathcal{U} \times \mathcal{Z} \to \mathbb{R}$ that depends on the time via time-dependent loads and a dissipation potential $\mathcal{R} : \mathcal{Z} \to [0, \infty]$, the evolution in the quasi-static case (i.e., inertia terms and viscoelastic behavior are neglected) is characterized by the following system:

$$u(t) = \operatorname*{arg\,min}_{v \in \mathcal{U}} \mathcal{E}(t, v, z(t)), \tag{BM}$$

$$0 \in \partial \mathcal{R}(\dot{z}(t)) + D_z \mathcal{E}(t, u(t), z(t)) \quad \text{in } \mathcal{Z}^*. \tag{RIS}$$

In this section, we assume that $\mathcal{R}$ is convex and positively homogeneous of degree one, and hence, the system (BM)–(RIS) is rate-independent. Generalizations do exist, where for instance the state space $\mathcal{Z}$ is replaced by a metric space and where the potential $\mathcal{R}$ is replaced by a dissipation distance, see [29].

The analytic properties of the system (BM)–(RIS) depend significantly on the convexity properties of the energy potential $\mathcal{E}$. If $\mathcal{E}$ is quadratic and uniformly convex in the pair (u, z), then classical results guarantee the existence and uniqueness of solutions to (BM)–(RIS) that are Lipschitz continuous in time. We refer the reader to [6] for the general theory on evolution systems involving maximal monotone operators and to [14, 17], where similar results were derived for models for elasto-plasticity. We subsequently denote all curves $(u, z) \in W^{1,\infty}(0, T; \mathcal{U} \times \mathcal{Z})$ such that (BM)–(RIS) hold for almost all $t \in [0, T]$ as *differential solutions*.

The picture changes completely when the energy $\mathcal{E}$ is not convex in (u, z). This is for instance the case for damage models of Ambrosio–Tortorelli type, for finite strain elasto-plasticity, or for rate-independent versions of ferroelectric models [12, 27, 45]. In the non-convex case, global solutions that are continuous with respect to time do not exist even if the loading path is smooth in time. Hence, suitable notions of weak solutions are necessary, which allow for discontinuities and at the same time are enhanced by jump criteria selecting physically reasonable discontinuities. In addition, there is a need for numerical algorithms that reliably approximate the type of solution one is interested in.

In the seminal paper [37], the authors introduced the notion of *Global Energetic Solutions* (GES) for rate-independent systems. In our context, a curve $t \mapsto (u(t), z(t))$ is a GES if the following global stability condition (S) and energy dissipation balance (E) are satisfied for all $t \in [0, T]$:

$$\mathcal{E}(t, u(t), z(t)) \le \mathcal{E}(t, v, \zeta) + \mathcal{R}(\zeta - z(t)) \quad \text{for all } v \in \mathcal{U}, \zeta \in \mathcal{Z}, \tag{S}$$

$$\mathcal{E}(t, u(t), z(t)) + \mathrm{diss}_{\mathcal{R}}(z; [0, t]) = \mathcal{E}(0, z(0)) + \int_0^t \partial_t \mathcal{E}(r, u(r), z(r))\, \mathrm{d}r. \tag{E}$$

Here, $\mathrm{diss}_{\mathcal{R}}(z; [0, t])$ corresponds to the total variation of the trajectory of z measured in terms of the dissipation functional $\mathcal{R}$. This concept was successfully applied to various models from continuum mechanics, and the existence of GES is nowadays well established for a variety of models involving non-convex energies. This concept also proves to be very flexible when studying limits of families of rate-independent systems depending on parameters. In particular, tools from the calculus of variations such as Γ-convergence can be applied to investigate such systems, see [30].

However, due to the global minimality condition (S), global energetic solutions tend to develop discontinuities that could be considered as nonphysical since solutions might jump across energy barriers. As an alternative, the authors of [11] proposed to start from a viscously regularized version of (RIS) of the type

$$0 \in \partial \mathcal{R}(\dot{z}_\gamma(t)) + \gamma \dot{z}_\gamma(t) + D_z \mathcal{E}(t, u_\gamma(t), z_\gamma(t)) \tag{2.1}$$

and to study the limit as the viscosity parameter γ tends to zero. This approach was intensively investigated in the last 12 years leading to the notion of *balanced viscosity solutions* (BV solutions). We refer the reader to [31, 32] for results in an abstract setting, to [8] for an application in plasticity and to [20, 23], where a corresponding analysis was carried out for fracture and damage models. In comparison with GES, BV solutions tend to jump as late as possible and hence are substantially different from GES in the non-convex case.

There are different ways of characterizing BV solutions. For this survey article, let us deal with the following parametrized version, where we introduce a further space $\mathcal{V}$ such that $\mathcal{Z} \subset \mathcal{V}$:

Definition 2.1 A tuple $(S, \hat{t}, \hat{u}, \hat{z})$ with $\hat{z} \in W^{1,\infty}(0, S; \mathcal{V}) \cap L^\infty(0, S; \mathcal{Z})$, $S > 0$ and $\hat{t} \in W^{1,\infty}(0, S; \mathbb{R})$ is a $\mathcal{V}$-parametrized BV solution if $\hat{t}(0) = 0$, $\hat{t}(S) = T$, $\hat{t}'(s) \geq 0, \hat{t}'(s) + \|\hat{z}(s)\|_{\mathcal{V}} \leq 1$ for a.a. $s \in (0, S)$, $\hat{u}(s) \in \arg\min_{v \in \mathcal{U}} \mathcal{E}(\hat{t}(s), v, \hat{z}(s))$ for all s, and if there exists a measurable function $\lambda : (0, S) \to [0, \infty)$ such that the complementarity condition and the inclusion here below are satisfied for almost all s

$$\lambda(s) \geq 0, \quad \lambda(s)\hat{t}'(s) = 0, \tag{2.2}$$

$$0 \in \partial\mathcal{R}(\hat{z}'(s)) + \lambda(s)\hat{z}'(s) + D_z\mathcal{E}(\hat{t}(s), \hat{u}(s), \hat{z}(s)). \tag{2.3}$$

The solution is $\mathcal{V}$-normalized if in addition $\hat{t}'(s) + \|\hat{z}'(s)\|_{\mathcal{V}} = 1$ for a.a. s.

Let us stress that there is a certain flexibility in representing parametrized solutions. In order to arrive at (2.2)–(2.3), one introduces the arc-length-type parameter $s_\gamma(t) := t + \int_0^t \|\dot{z}_\gamma(r)\|_{\mathcal{V}} \mathrm{d}r$ and reformulates (2.1) in terms of this new variable. The limit $\gamma \to 0$ then leads to $\mathcal{V}$-parametrized BV solutions. Alternatively, one could for instance use the stronger $\mathcal{Z}$-reparametrization where everywhere in the previous definitions $\mathcal{V}$ has to be replaced with $\mathcal{Z}$, or parametrizations relying on the so-called vanishing viscosity contact potential

$$p(v, \xi) := \mathcal{R}(v) + \|v\|_{\mathcal{V}} \operatorname{dist}_{\mathcal{V}^*}(-\xi, \partial\mathcal{R}(0)). \tag{2.4}$$

In this case, $s_\gamma(t) := t + \int_0^t p(\dot{z}_\gamma(r), D_z\mathcal{E}(r, u_\gamma(r), z_\gamma(r)))\mathrm{d}r$, and the normalization condition in the limit reads $\hat{t}'(s) + p(\hat{z}'(s), D_z\mathcal{E}(\hat{t}(s), \hat{u}(s), \hat{z}(s))) = 1$. Each of these reparametrizations has its (analytical) advantages and disadvantages. The stronger the norms in the parametrization the more regular the limit functions are in the parametrized picture. However, only a reparametrization by the vanishing viscosity contact potential implies that limits of vanishing viscosity sequences are already normalized [32], which up to now is an essential ingredient for the study of optimal control problems with BV solutions as constraints [22], see also Sect. 4. This cannot be guaranteed a priori for the other choices.

Alternative notions of solutions that in some sense lie between GES and BV solutions were discussed in [24], where jumps across small energy barriers are admissible, and in [33], where special scalings in the vanishing viscosity procedure are prescribed. Let us finally mention that GES, BV solutions and the solutions

introduced in [33] belong to the class of local solutions, the most general definition of solutions for rate-independent systems. Adapted to the notation here, a pair $(u, z) : [0, T] \to \mathcal{U} \times \mathcal{Z}$ is a local solution if for almost all t we have (BM), the local stability condition $-D_z\mathcal{E}(t, u(t), z(t)) \in \partial\mathcal{R}(0)$ and the energy dissipation estimate (E) with $\leq$ instead of $=$ (for all t). We refer the reader to [29] for a detailed overview on different solution concepts for rate-independent systems.

3 Discretization Schemes for Rate-Independent Systems and Their Convergence

3.1 The Semilinear Setting

In this section, we will simplify the model by assuming that the energy potential $\mathcal{E}$ depends on the internal variable z alone, and not on the displacement field. In order to set up the model, let the state space $\mathcal{Z}$ be a separable Hilbert space that fulfills the embeddings

$$\mathcal{Z} \subset\subset \mathcal{V} \subset \mathcal{X} \tag{3.1}$$

for another separable Hilbert space $\mathcal{V}$ and a Banach space $\mathcal{X}$, that is, $\mathcal{Z}$ is compactly embedded into $\mathcal{V}$ and continuously embedded into $\mathcal{X}$. We are working with a semilinear model, [38], meaning that the energy potential consists of a quadratic leading term and a lower order non-convexity, as well as a linear term that is given by the external load ℓ. To be more precise, let $A \in \mathrm{Lin}(\mathcal{Z}, \mathcal{Z}^*)$ be a linear, self-adjoint, and bounded operator and $\mathcal{F} : \mathcal{Z} \to [0, \infty)$ a nonlinearity such that $\mathcal{E} : [0, T] \times \mathcal{Z} \to \mathbb{R}$ is given by

$$\mathcal{E}(t, z) := \tfrac{1}{2}\langle Az, z\rangle_{\mathcal{Z}^*,\mathcal{Z}} + \mathcal{F}(z) - \langle \ell(t), z\rangle_{\mathcal{V}^*,\mathcal{V}} = \mathcal{I}(z) - \langle \ell(t), z\rangle_{\mathcal{V}^*,\mathcal{V}}, \tag{3.2}$$

where

$$\mathcal{I} : \mathcal{Z} \to \mathbb{R}, \quad \mathcal{I}(z) = \tfrac{1}{2}\langle Az, z\rangle_{\mathcal{Z}^*,\mathcal{Z}} + \mathcal{F}(z)$$

depends on the state z alone. Here, we assume that A is $\mathcal{Z}$-elliptic, i.e., there exists a constant $\alpha > 0$ such that

$$\forall z \in \mathcal{Z} : \quad \langle Az, z\rangle \geq \alpha \|z\|^2_{\mathcal{Z}}.$$

The non-convexity $\mathcal{F}$ is assumed to fulfill the following:

$$\mathcal{F} \in C^2(\mathcal{Z}, \mathbb{R}) \text{ with } \mathcal{F} \geq 0, \tag{3.3}$$

$$\mathrm{D}_z\mathcal{F} \in C^1(\mathcal{Z}, \mathcal{V}^*), \tag{3.4}$$

$$\exists q \geq 1 : \|\mathrm{D}_z \mathcal{F}(z)v\|_{\mathcal{V}^*} \leq C(1 + \|z\|_{\mathcal{Z}}^q)\|v\|_{\mathbb{Z}}, \tag{3.5}$$

$$\mathcal{F} : \mathcal{Z} \to \mathbb{R} \text{ and } \mathrm{D}_z \mathcal{F} : \mathcal{Z} \to \mathcal{Z}^* \text{ are weakly continuous.} \tag{3.6}$$

Furthermore, if not otherwise stated, we assume in this section that the initial state $z_0 \in \mathcal{Z}$ and the external load ℓ are compatible in the sense that

$$z_0 \in \mathcal{Z}, \ell \in H^1((0,T);\mathcal{V}^*) \text{ and } \mathrm{D}_z \mathcal{E}(0,z_0) = \mathrm{D}_z \mathcal{I}(z_0) - \ell(0) \in \mathcal{V}^*. \tag{3.7}$$

Finally, let the dissipation potential $\mathcal{R} : \mathcal{X} \to [0,\infty)$ be convex, continuous, positively homogeneous of degree one, and bounded by $\|\cdot\|_{\mathcal{X}}$, i.e.,

$$\exists c, C > 0 \forall x \in \mathcal{X} : \quad c\|x\|_{\mathcal{X}} \leq \mathcal{R}(x) \leq C\|x\|_{\mathcal{X}}. \tag{3.8}$$

As a consequence of the assumptions (3.1), (3.3)–(3.5), and (3.8), $\mathcal{I}$ is λ-convex on sublevels, meaning that the following holds true (cf. [19, Lemma 1.1]): For every $\rho > 0$, there exists $\lambda = \lambda(\rho) > 0$ such that for all $z_1, z_2 \in \mathcal{Z}$ with $\|z_i\|_{\mathcal{Z}} \leq \rho$ we have the estimate

$$\mathcal{I}(z_2) - \mathcal{I}(z_1) \geq \langle \mathrm{D}_z \mathcal{I}(z_1), z_2 - z_1 \rangle_{\mathcal{Z}^*,\mathcal{Z}} + \tfrac{\alpha}{2}\|z_1 - z_2\|_{\mathcal{Z}}^2 - \lambda \mathcal{R}(z_2 - z_1)\|z_2 - z_1\|_{\mathcal{V}}. \tag{3.9}$$

In this setting, (BM)–(RIS) reduce to the problem: Find $z : [0,T] \to \mathcal{Z}$ with $z(0) = z_0$ such that for almost all $t \in (0,T)$, we have

$$0 \in \partial \mathcal{R}(\dot{z}(t)) + \mathrm{D}_z \mathcal{E}(t, z(t)) \,. \tag{3.10}$$

The existence of $\mathcal{V}$-parametrized BV solutions is guaranteed by [32, Theorem 3.12].

3.2 Discretization Schemes, Abstract Semilinear Setting

Several different discretization schemes are proposed in the literature for the approximation of (3.10). In order to simplify the notation, we assume an equidistant partition of $[0,T]$, i.e., for $N \in \mathbb{N}$ we define $\tau^N := T/N$ and $t_k^N = k\tau$ for $0 \leq k \leq N$.

Global energetic solutions can be approximated by the following time incremental global minimization scheme:

$$z_0^N := z_0, \quad z_k^N \in \operatorname{Arg\,min}\{\, \mathcal{E}(t_k^N, z) + \mathcal{R}(z - z_{k-1}^N) \,;\, z \in \mathcal{Z} \,\}, \quad 1 \leq k \leq N \,, \tag{3.11}$$

and we refer the reader to [29] for details.

In order to approximate BV solutions, the most natural approach is to introduce a time-discrete version of the viscous system (2.1) and to pass to the limit $\tau^N \to 0$, $\gamma \to 0$ simultaneously. This results in the following (implicit Euler) scheme: Let $z_0^{N,\gamma} := z_0$. Then for $1 \le k \le N$

$$z_k^{N,\gamma} \in \operatorname{Arg\,min}\{\, \mathcal{E}(t_k^N, z) + \mathcal{R}(z - z_{k-1}^{N,\gamma}) + \tfrac{\gamma}{2\tau^N} \|z - z_{k-1}^{N,\gamma}\|_{\mathcal{V}}^2 ;\ z \in \mathcal{Z} \,\}\,. \tag{3.12}$$

If $\tau^N \to 0$, $\gamma \to 0$, and $\tau^N/\gamma \to 0$, then (subsequences of) suitable interpolants of the points $(z_k^{N,\tau})_k$ converge to BV solutions, see for instance [32, Theorem 3.12]. However, in practice, it is often difficult to find a good relation between the parameters τ^N and γ such that the right jump behavior is already visible for rather coarse discretizations. See, e.g., [21], where a crack propagation model was analyzed.

As a first alternative to the previously discussed method, a local minimization scheme was proposed and analyzed in [11]. There, the idea is to discretize the system in the parametrized picture that results in a scheme that has a time-adaptive character: Let $\tau > 0$ and $z_0^\tau = z_0$, $t_0^\tau = 0$. For $k \ge 1$, determine $z_k^\tau \in \mathcal{Z}$ and t_k^τ by

$$z_k^\tau \in \operatorname{Arg\,min}\{\, \mathcal{E}(t_{k-1}^\tau, v) + \mathcal{R}(v - z_{k-1}^\tau) ;\ v \in \mathcal{Z},\ \|v - z_{k-1}^\tau\|_{\mathcal{V}} \le \tau \,\} \tag{3.13a}$$

$$t_k^\tau = \min\left\{ t_{k-1}^\tau + \tau - \|z_k^\tau - z_{k-1}^\tau\|_{\mathcal{V}},\ T \right\}. \tag{3.13b}$$

While this question was not addressed in [11], it was shown in [19] that for each $\tau > 0$ the final time T is reached after a finite number of incremental minimization steps $N(\tau)$. Moreover, the $\mathcal{Z}$-length of the polygonal path defined by the points $(z_k^\tau)_{0 \le k \le N(\tau)}$ is uniformly bounded w.r.t. τ, i.e., $\sup_{\tau>0} \sum_{k=1}^{N(\tau)} \|z_k^\tau - z_{k-1}^\tau\|_{\mathcal{Z}} < \infty$. Let us parametrize the polygonal path as follows: With $S_\tau := T + \sum_{k=1}^{N(\tau)} \|z_k^\tau - z_{k-1}^\tau\|_{\mathcal{V}}$ and $s_k^\tau := k\tau$, let $\hat{z}_\tau : [0, S_\tau] \to \mathcal{Z}$ and $\hat{t}_\tau : [0, S_\tau] \to [0, T]$ denote the affine interpolants related to the points $(z_k^\tau)_k$ and $(t_k^\tau)_k$ with $\hat{z}_\tau(s_k^\tau) = z_k^\tau$ and similar for $\hat{t}_\tau$.

Theorem 3.1 *For every vanishing sequence $(\tau_n)_{n\in\mathbb{N}}$, there exists a subsequence and a $\mathcal{V}$-parametrized BV solution $(S, \hat{t}, \hat{z})$ in the sense of Definition 2.1 such that $S_\tau \to S$, $\hat{t}_\tau \overset{*}{\rightharpoonup} \hat{t}$ in $W^{1,\infty}((0,S);\mathbb{R})$ and $\hat{z}_\tau \overset{*}{\rightharpoonup} \hat{z}$ in $W^{1,\infty}((0,S);\mathcal{V}) \cap L^\infty(0,S;\mathcal{Z})$.*

We refer the reader to [11] for a proof in the finite-dimensional case and to [19, Theorem 2.5] for the semilinear case introduced above.

If one wants to avoid the local minimization in (3.13a), an alternative is given by the following ansatz: Let $\eta > 0$. Determine $(z_k^N)_{0 \le k \le N}$ with $z_0^N := z_0$ from the following incremental minimization scheme, where $t_k^N = k\tau^N$, $z_{k,0} := z_{k-1}^N$, and

$i \geq 1$:

$$z_{k,i} \in \operatorname{Arg\,min}\{\, \mathcal{I}(t_k^N, v) + \frac{\eta}{2}\|v - z_{k,i-1}\|_{\mathcal{V}}^2 + \mathcal{R}(v - z_{k,i-1})\,;\ v \in \mathcal{Z}\,\}, \tag{3.14a}$$

$$z_k^N := \lim_{i\to\infty} z_{k,i} \quad \text{(weak limit in } \mathcal{Z}\text{)}. \tag{3.14b}$$

Again it is shown in [19] that for $\tau \to 0$ and $\eta \to \infty$ suitable interpolants of $(z_k^N)_{k\in\mathbb{N}}$ converge to parametrized BV solutions. In order to construct the interpolating curves, also the intermediate points $(z_{k,i})_{i\in\mathbb{N}}$ are taken into account in the proof. Observe that thanks to estimate (3.9) the minimization problem (3.14a) is uniformly convex provided that η is large enough. Combining Theorem 3.7 and Theorem 4.5 from [19], incremental solutions of (3.14a)–(3.14b) still converge to BV solutions if one replaces (3.14b) by a stopping criterion of the type "stop if $\|z_{k,i} - z_{k,i-1}\|_{\mathcal{V}} \leq \delta$" and defines $z_k^N := z_{k,i}$. It is shown that for each k after a finite number of steps the stopping criterion is active and that for $\tau^N, \delta \to 0$ and $\eta \to \infty$ interpolating curves converge to BV solutions. Scheme (3.14a)–(3.14b) is inspired by an ansatz discussed in [3], where instead of (3.14a) the authors work with $z_{k,i} \in \operatorname{Arg\,min}\{\, \mathcal{I}(t_k^N, v) + \frac{\eta}{2}\|v - z_{k,i-1}\|_{\mathcal{V}}^2 + \mathcal{R}(v - z_k^N)\,;\ v \in \mathcal{Z}\,\}$. For $\tau^N \to 0$ but $\eta > 0$ fixed, they show the convergence of (subsequences of) interpolants to local solutions.

Apart from standard a priori estimates that follow from discrete versions of an energy dissipation balance, the essential ingredient for the convergence proofs is a uniform (with respect to the discretization parameters) estimate of the arc length of the interpolating curves in $\mathcal{Z}$, i.e., an estimate of the type $\sum_{k=1}^N \sum_i \|z_{k,i} - z_{k,i-1}\|_{\mathcal{Z}} \leq C$ for (3.14a)–(3.14b) or $\sum_{k=1}^N \|z_k^\tau - z_{k-1}^\tau\|_{\mathcal{Z}} \leq C$ for (3.13a). The derivation of such estimates involves similar arguments as for the vanishing viscosity analysis of general rate-independent systems. The incremental system is then reformulated in a parametrized picture, and limits are identified via compactness arguments and weak convergence principles.

Just like for the choice of the viscous regularization term in the vanishing viscosity ansatz in (2.1), there is a certain flexibility in choosing the norm (e.g., $\mathcal{V}$-norm or $\mathcal{Z}$-norm) in (3.13a) as well as in the quadratic terms in (3.12) and (3.14a). Clearly, different choices will lead to different limit models. However, one could consider the choice of the norm as a further degree of freedom in order to set up a physically reasonable model.

Let us finally address alternate minimization approaches (staggered schemes). Such schemes are widely used for coupled systems in the framework of generalized standard materials when two or more variables are involved. The basic idea is to freeze the variables alternatingly and to calculate the other variables from the resulting subsystem. However, for the case when the whole system or a subsystem shows rate-independent behavior and when the energy is not convex, a convergence analysis was initiated only recently. In [42, 43], the convergence of discrete solutions generated by an alternate minimization scheme to semistable local

solutions was shown, a slightly stronger notion than the notion of local solution. For the purely rate-independent case, a first convergence proof was given in [18] for the Ambrosio–Tortorelli damage model and in [19] for an abstract semilinear setting. In the Ambrosio–Tortorelli model, the underlying energy is non-convex but separately uniformly convex in the displacement and the damage variable. The main observation in [18] is that even without introducing any viscosity terms into the minimization scheme the discrete solutions converge to BV-type solutions.

Let us finally illustrate the abovementioned schemes with the following finite-dimensional example ($\mathcal{Z} = \mathcal{V} = \mathcal{X} = \mathbb{R}$) from [19]:

$$\mathcal{I}(t, z) := 5z^2 - \frac{t^2}{2(0.1 + z^2)}, \quad \mathcal{R}(v) := 10|v|, \quad z_0 = 1 \quad \text{and } T = 1.5\,. \tag{3.15}$$

In Fig. 1 (left), the set $\{\,(t, z)\,;\; -\mathrm{D}\mathcal{I}(t, z) \in \partial\mathcal{R}(0)\,\}$ is marked in gray. The figure shows the global energetic solution (blue, calculated with (3.11), $N = 100$), the BV solution (purple, calculated with (3.13), $\tau = 90$ and $N(\tau) \approx 150$), and a viscous approximation (green, calculated with (3.12), $N = 100$, $\gamma = \sqrt{T/N}$). The right graph in Fig. 1 shows the interpolants $(\hat{z}_\tau, \hat{t}_\tau)$ corresponding to (3.13) as functions of the arc- length parameter s. One clearly sees the region, where $\hat{t}_\tau$ is constant, while $\hat{z}_\tau$ is decreasing. This corresponds to the vertical transition of the purple curve in the left picture and to a jump discontinuity of z with respect to the true physical time.

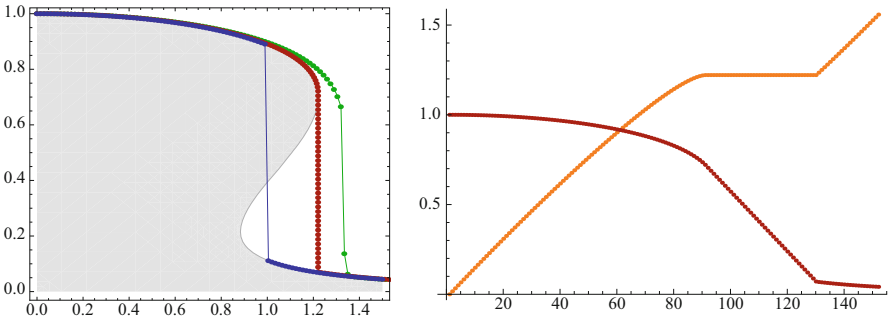

Fig. 1 Left: Global energetic solution (blue), BV solution (purple), and viscous approximation (green) as functions of time t; Right: BV solution $\hat{z}$ (purple, decreasing) and $\hat{t}$ (orange, increasing) as functions of the arc-length parameter s

3.3 A Priori Estimates, Abstract Semilinear Setting

While Theorem 3.1 shows that the local minimization scheme (3.13) is able to approximate parametrized solutions, in the sense that every accumulation point of approximations with increasingly finer step size τ is an element of this class of solutions, this section is concerned with a priori estimates for approximations using (3.13). For the notion of energetic solutions, this has been worked out in [25, 36] using the global incremental minimization (3.11) instead of the local one in (3.13) and assuming that the energy $\mathcal{E}$ is uniformly convex. The authors show that the error of the approximation compared to the unique global energetic solution is of order $\mathcal{O}(\sqrt{\tau})$, where τ denotes the fineness of the discretization in time. For quadratic and coercive energies, this result has been extended to $\mathcal{O}(\tau)$ in [26] and more general in [4], see also [1] for the special case of elasto-plasticity. In contrast, to the best of our knowledge, there exist no such results for the concept of parametrized solutions. One major difficulty here is due to the fact that the parametrized solutions are defined in the extended state space, each with their own artificial time. In the convex case, one can handle this problem by retransforming the solution back into the physical time and comparing it to the unique differential solution to obtain an a priori estimate. Hence, again, the main assumption is the (at least local) uniform convexity of the energy functional. Without this assumption, solutions are, in general, not unique, so that it is not clear if any of the solutions is preferred by the algorithm (see Figure 1 in [34]). However, one cannot expect a convergence result going beyond Theorem 3.1 without further assumptions. Thus, in addition to the assumptions in Sect. 3.1, we require the energy functional to fulfill $\mathcal{E}(t,\cdot) \in C^{2,1}_{loc}(\mathcal{Z};\mathbb{R})$, that is to say, for all $r > 0$ there exists $C(r) \geq 0$ such that for all $z_1, z_2 \in B_{\mathcal{Z}}(0,r)$ it holds

$$\langle \big[D_z^2\mathcal{E}(t,z_1) - D_z^2\mathcal{E}(t,z_2)\big]v, v\rangle_{\mathcal{Z}^*,\mathcal{Z}} \leq C(r)\|z_1 - z_2\|_{\mathcal{Z}}\|v\|^2_{\mathcal{Z}}.$$

Note that, due to the structure of the energy functional $\mathcal{I}$, the constant $C(r)$ does not depend on the time t and, moreover, this assumption holds iff $\mathcal{F} \in C^{2,1}_{loc}(\mathcal{Z};\mathbb{R})$. At first, we also assume that $\mathcal{E}$ is (globally) κ-uniformly convex, i.e., there exists a $\kappa > 0$ such that for all $t \in [0,T]$ it holds

$$\langle D_z^2\mathcal{E}(t,z)v, v\rangle_{\mathcal{Z}^*,\mathcal{Z}} \geq \kappa\|v\|^2_{\mathcal{Z}} \quad \forall z, v \in \mathcal{Z}.$$

On the one hand, this condition implies that there exists a unique differential solution to the rate-independent system, and on the other hand, it allows us to prove that $t_{k+1} > t_k$ for all iterations k, provided that the Lipschitz constant of the external load is sufficiently small, see Theorem 3.2. This is crucial in order to define the following affine interpolant for $t \in [t_{k-1}, t_k)$:

$$z_\tau(t) := z_{k-1} + \frac{t - t_{k-1}}{t_k - t_{k-1}}(z_k - z_{k-1}) \tag{3.16}$$

for which we obtain the following:

Theorem 3.2 *Let $\mathcal{E}(t,\cdot) \in C^{2,1}_{loc}(\mathcal{Z};\mathbb{R})$ be κ-uniformly convex. Moreover, let $\ell \in W^{2,1}([0,T];\mathcal{V}^*)$ with $|\ell|_{Lip} < \kappa$. Then, the sequence $\{z_\tau\}_{\tau>0}$ of retransformed discrete parametrized solutions converges to the unique (differential) solution z and satisfies the a priori error estimate*

$$\|z_\tau(t) - z(t)\|_{\mathcal{Z}} \le K\,\tau \qquad \forall t \in [0,T], \tag{3.17}$$

where $K = K(\alpha,\kappa,\ell,z_0,T,\mathcal{F},\|A\|_{\mathcal{L}(\mathcal{Z},\mathcal{Z}^*)}) > 0$ is independent of τ.

Note that due to the $1-$homogeneity of $\mathcal{R}$, the time can always be rescaled in such a way that the condition on the Lipschitz constant of ℓ is fulfilled. Under the specified assumptions, we thus obtain an optimal rate of convergence for the iterated local minimization scheme. However, if the energy is globally uniformly convex, then a local minimum of (3.13) is also a global one, so that the localization in (3.13) is obsolete. This, however, changes if the energy is not globally but locally uniformly convex w.r.t. some evolution z. By this, we mean that for $z : [0,T] \to \mathcal{Z}$ with $z \in W^{1,\infty}([0,T];\mathcal{Z})$, there exist $\kappa, \Delta > 0$, independent of t, such that $\mathcal{E}(t,\cdot)$ is κ-uniformly convex on $\overline{B_\Delta(z(t))}$ for all $t \in [0,T]$, i.e.,

$$\langle D_z^2\mathcal{E}(t,\tilde{z})v, v\rangle_{\mathcal{Z}^*,\mathcal{Z}} \ge \kappa\|v\|^2_{\mathcal{Z}} \qquad \forall \tilde{z} \in \overline{B_{\mathcal{Z}}(z(t),\Delta)},\ v \in \mathcal{Z}. \tag{3.18}$$

In this case, an a priori estimate like (3.17), in general, no longer holds for the global minimization scheme, see Fig. 2 (a global energetic and a differential solution must not even coincide, see [47]). In contrast, the affine interpolant of the local minimization iterates defined as in (3.16) still fulfills the following a priori estimate:

Theorem 3.3 *Let $z \in C^{0,1}([0,T];\mathcal{Z})$ be a (differential) solution with $-D_z\mathcal{E}(0,z_0) \in \partial\mathcal{R}(0)$. Furthermore, let $\mathcal{E}$ be locally κ-uniformly convex around z with radius $\Delta > 0$ and assume that $\ell \in W^{2,1}([0,T];\mathcal{V})$ with $|\ell|_{Lip} < \kappa$. Then there exists a constant $K_{loc} > 0$, independent of τ, such that, for the back-transformed discrete parametrized solution $z_\tau : [0,T] \to \mathcal{Z}$ and all $\tau \le \bar{\tau}$ with $\bar{\tau}$ sufficiently small, it holds:*

$$\|z_\tau(t) - z(t)\|_{\mathcal{Z}} \le K_{loc}\,\tau \quad \forall t \in [0,T]. \tag{3.19}$$

Figure 2, which is based on the example in [34, Section 4.2], shows exactly this situation. Here, the approximation using the global minimization scheme produces a jump, whereas the discrete solution obtained by (3.13) nicely approximates the differential solution with an error of order $\mathcal{O}(\tau)$. The gray regions have the same meaning as in Fig. 1 (left). The example was calculated with $\mathcal{R}(v) = |v|$, $\mathcal{I}(t,z) =$

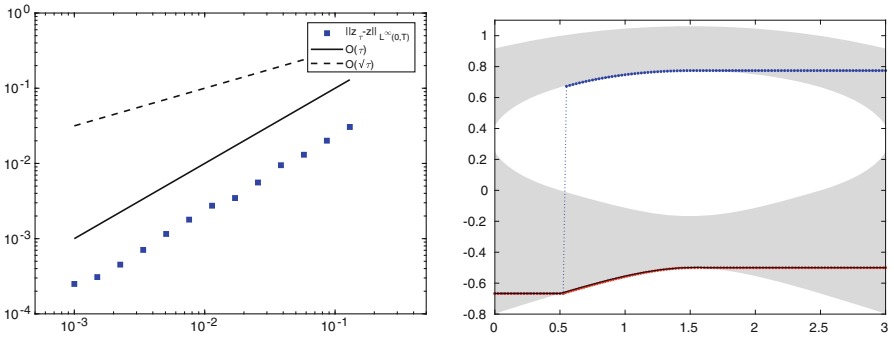

Fig. 2 Left: Errors for the approximation of a parametrized solution using the local minimization scheme depending on the step size τ; Right: Corresponding differential solution (black) as well as the numerical approximations of a global energetic solution (blue) and BV solution (red) as functions of the time t

$\frac{1}{2}z^2 + \mathcal{F}(z) - \ell(t)z$ with $\ell(t) = -1/2(t - 3/2)^2 + 3/2$ and

$$\mathcal{F}(z) = \begin{cases} 2z^3 - 5/2z^2 + 1 & z \geq 0 \\ -2z^3 - 5/2z^2 + 1 & z < 0 \end{cases}.$$

3.4 Finite-Element Discretization and Numerical Realization

This section is devoted to the numerical realization of the local minimization scheme (3.13). Additionally to the time discretization, we also employ a spatial discretization $\mathcal{Z}_h \subset \mathcal{Z}$ using finite elements. Moreover, we allow for a further approximation $\mathcal{R}_h$ of the dissipation functional $\mathcal{R}$. The overall algorithm reads as follows:

Algorithm 1 (Fully Discrete Local Minimization)

1: Set $z_0^\tau = P_h(z_0)$, $t_0 = 0$, and $k = 1$
2: **while** $t_k < T$ **do**
3: Compute z_k^τ as solution of

$$z_k^\tau \in \arg\min\{\mathcal{E}(t_{k-1}^\tau, z) + \mathcal{R}_h(z - z_{k-1}^\tau) \,:\, z \in \mathcal{Z}_h,\ \|z - z_{k-1}^\tau\|_{\mathcal{V}} \leq \tau\} \tag{3.20}$$

4: Time update:

$$t_k^\tau = \min\{t_{k-1}^\tau + \tau - \|z_k^\tau - z_{k-1}^\tau\|_{\mathcal{V}}, T\} \tag{3.21}$$

5: Set $k = k + 1$.
6: **end while**

Herein, P_h denotes the Ritz projection, i.e.,

$$P_h(u) \in \mathcal{Z}_h\,, \quad a(P_h(u), v) = a(u, v) \qquad \forall v \in \mathcal{Z}_h\,,$$

where a is the bilinear form induced by A. Before we turn to the numerical results obtained by Algorithm 1, we note that under suitable assumptions on the spatial discretization and the approximation $\mathcal{R}_h$, see [35], we obtain the following result, which is based on the analysis for Theorem 3.1.

Theorem 3.4 (Convergence Toward Parametrized Solutions) *There exists a sequence $\{\tau_n, h_n\}_{n\in\mathbb{N}} \subset \mathbb{R}_+ \times \mathbb{R}_+$ converging to zero such that the affine interpolants generated by the fully discrete local minimization Algorithm 1 and the artificial end time S_{τ_n,h_n} satisfy*

$$S_{\tau_n,h_n} \to S, \tag{3.22}$$

$$\hat{t}_{\tau_n,h_n} \overset{*}{\rightharpoonup} \hat{t} \qquad \text{in } W^{1,\infty}((0,S);\mathbb{R}), \tag{3.23}$$

$$\hat{z}_{\tau_n,h_n} \overset{*}{\rightharpoonup} \hat{z} \qquad \text{in } W^{1,\infty}((0,S);\mathcal{V}) \cap L^{\infty}((0,S);\mathcal{Z}), \tag{3.24}$$

$$\hat{z}_{\tau_n,h_n}(s) \rightharpoonup \hat{z}(s) \qquad \text{in } \mathcal{Z} \text{ for every } s \in [0,S] \tag{3.25}$$

and the limit $(\hat{t}, \hat{z})$ is a $\mathcal{V}$-parametrized solution.

Moreover, every accumulation point $(\hat{t}, \hat{z})$ of time incremental sequences in the sense of (3.22)–(3.25) is a $\mathcal{V}$-parametrized solution.

The results shown below are generated by Algorithm 1 for the following problem data (see [35]):

- $\Omega = [0,1]^2$ and $\mathcal{Z} = H_0^1(\Omega)$, $\mathcal{V} = L^2(\Omega)$, as well as $\mathcal{X} = L^1(\Omega)$.
- The operator A within the energy functional is set to $A = -\triangle : H_0^1(\Omega) \to H^{-1}(\Omega)$, so that the coercivity constant α equals Poincaré's constant.
- The nonlinearity $\mathcal{F}$ in the energy is defined as the well-known double-well potential

$$\mathcal{F}(z) := 48 \int_\Omega \left(1 - z(x)^2\right)^2 \mathrm{d}x.$$

- The external loads are only depending on t and given by

$$\ell(t,x) = \ell(t) := -48\,\sin(2\pi t), \quad (t,x) \in [0,T] \times \Omega.$$

- The dissipation functional is given by the L^1-norm, i.e., $\mathcal{R}(v) = \|v\|_{L^1(\Omega)}$.

While we choose linear finite elements for the spatial discretization, the discrete dissipation potential $\mathcal{R}_h : \mathcal{Z}_h \to \mathbb{R}$ is defined by

$$\mathcal{R}_h(z_h) := \int_\Omega \sum_{i=1}^{N_h} |z_i| \varphi_i(x) \mathrm{d}x, \tag{3.26}$$

wherein z_i denotes the coefficient vector of z_h w.r.t. the nodal basis φ_i, i.e., $z_h(x) = \sum_{i=1}^{N_h} z_i \varphi_i(x)$. This has the advantage that the convex subdifferential of $\mathcal{R}_h$ can be expressed componentwise. In addition, the first-order optimality system for the discretized optimization problem, see [35, Lemma 4.2], can be solved numerically by a semismooth Newton method. The corresponding numerical results (with $z_0 \equiv 0$ and $T = 1$) are shown in Fig. 3. For a detailed explanation hereof, see [35]. Note that the viscous behavior of the system during the jump time $t \approx 0.6724$ can be nicely observed (see Fig. 3c–i and Fig. 4).

4 Optimal Control of Rate-Independent Systems

Concerning the optimization and optimal control of rate-independent systems, the literature is rather scarce, in particular with regard to non-convex energies. If the energy is convex, then several results are known, concerning existence of optimal solutions as well as optimality conditions. In the uniformly convex case, all notions of solutions introduced above are basically equivalent and the rate-independent system holds in its strong form (BM)–(RIS). Moreover, in this case, (BM)–(RIS) admit a unique solution that gives rise to the definition of an associated solution operator, the control-to-state mapping $\ell \mapsto (u, z)$. In this way, optimal control problems governed by (BM)–(RIS) with convex energy can be transformed into a problem in the control variable only, frequently termed the reduced problem. This sometimes called implicit programming approach is widely used.

The situation changes completely if one relaxes the convexity assumptions or even turns to non-convex energies, and at this point, the literature becomes rather scarce. When it comes to non-convex energies, an entirely new challenge comes into play, namely the many different notions of solutions for the state system and the lack of uniqueness of the corresponding solutions. Existence results for optimal control problems in the GES setting are proven in [9, 10, 41, 48]. A first result in the setting of BV solutions was derived in [22], and we shortly summarize here the challenges.

Let us study an optimal control problem governed by (3.10), where the external load ℓ is the control variable. We remain within the semilinear setting introduced in Sect. 3.1 and restrain the problem to an admissible set consisting of all normalized parametrized BV solutions of (3.10). Independently of the solution concept under consideration, solutions of (3.10) are not unique in general, which is due to the non-convexity of the energy. In order to show existence of an optimal control, it is

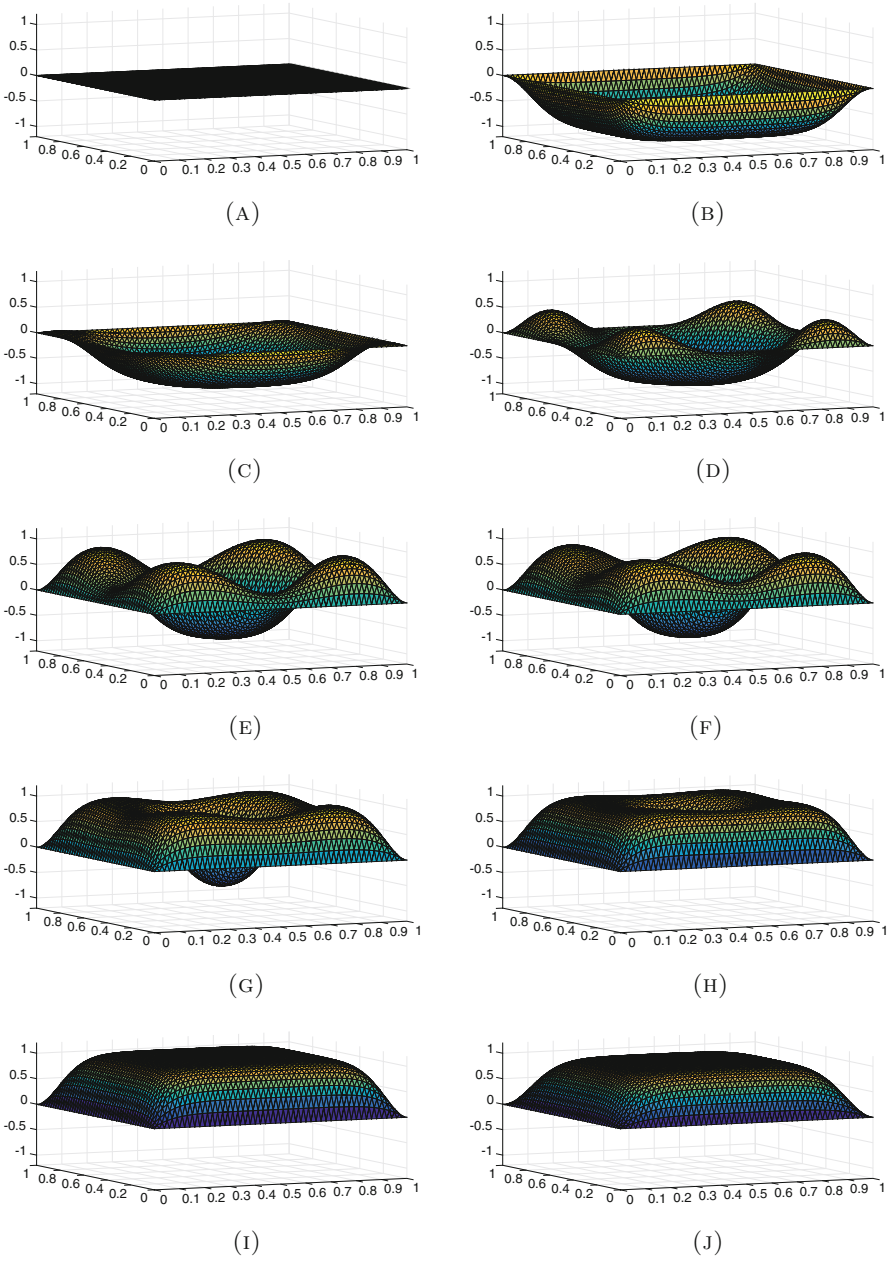

Fig. 3 Computed parametrized solution to the problem described in Sect. 3.4. Figure 3–3i shows the viscous transition corresponding to the discontinuity at time $t \approx 0.6724$. (**a**) $t = 0.00$. (**b**) $t = 0.01$. (**c**) $t = 0.6724$. (**d**) $t = 0.6724$. (**e**) $t = 0.6724$. (**f**) $t = 0.6724$. (**g**) t =0.6724. (**h**) $t = 0.6724$. (**i**) t =0.6729. (**j**) t =1.00

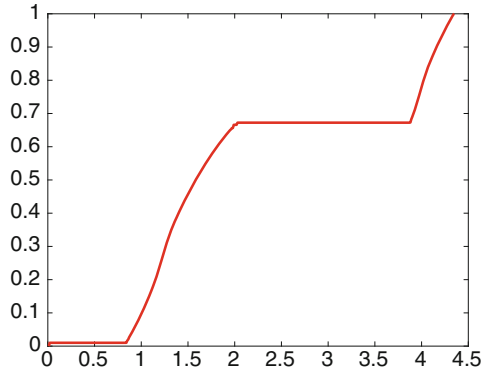

Fig. 4 Evolution of the physical time as a function of the artificial time. The physical time stands still during the viscous transitions at time $t \approx 0.01$ and $t \approx 0.6724$

therefore necessary to show sequential closedness of the set-valued control-to-state map as well as a compactness result.

To be more precise, for z_0 and ℓ that are compatible in the sense of (3.7), we define the admissible set

$$M_{ad} := \{\, (S, \hat{t}, \hat{z}, \ell) \,|\, (S, \hat{t}, \hat{z}) \text{ is a normalized } p\text{-parametrized BV solution for } (z_0, \ell) \,\}$$

and consider the optimal control problem

$$\left.\begin{array}{ll} \min & J(S, \hat{z}, \ell) := j(\hat{z}(S)) + \alpha \|\ell\|_{H^1(0,T;\mathcal{V}^*)} \\ \text{s.t.} & (S, \hat{t}, \hat{z}, \ell) \in M_{ad}. \end{array}\right\}. \tag{4.1}$$

Here, $\alpha > 0$ is a fixed Tikhonov parameter and $j : \mathcal{V} \to \mathbb{R}$ is continuous and bounded from below, e.g., $j(z) := \|z - z_{\text{des}}\|_{\mathcal{V}}$ for a desired end state $z_{\text{des}} \in \mathcal{V}$. As already mentioned in Sect. 2, there are different options for the representation of parametrized solutions, and only the reparametrization based on the vanishing viscosity contact potential p defined in (2.4) guarantees that the accumulation points of vanishing viscosity sequences are normalized. Since the optimal control problem calls for normalized solutions, we rely on the p-parametrization for the rest of this section. This leads to the following adjustments in the definition of BV solutions:

Definition 4.1 A tuple $(S, \hat{t}, \hat{z})$ with $\hat{z} \in AC^\infty([0, S]; \mathcal{X}) \cap L^\infty(0, S; \mathcal{Z})$, $S > 0$ and $\hat{t} \in W^{1,\infty}(0, S; \mathbb{R})$ is a p-parametrized BV solution if the following holds: The set

$$G := \{s \in [0, S] \,|\, \operatorname{dist}_{\mathcal{V}}(-D_z\mathcal{E}(\hat{t}(s), \hat{z}(s)), \partial\mathcal{R}(0)) > 0\} \tag{4.2}$$

is relatively open and $\hat{z} \in W^{1,1}_{\text{loc}}(G; \mathcal{V})$, $D_z\mathcal{E}(\hat{t}(\cdot), \hat{z}(\cdot)) \in L^\infty_{\text{loc}}(G; \mathcal{V}^*)$. Furthermore, $\hat{t}'(s) + \mathcal{R}[\hat{z}'](s) + \|\hat{z}'(s)\|_{\mathcal{V}} \operatorname{dist}_{\mathcal{V}}(-D_z\mathcal{E}(\hat{t}(s), \hat{z}(s)), \partial\mathcal{R}(0)) \leq 1$ for a.a.

$s \in (0, S)$, $\hat{t}(0) = 0$, $\hat{t}(S) = T$, $\hat{t}'(s) \geq 0$, and there exists a measurable function $\lambda : (0, S) \to [0, \infty)$ with $\lambda(s) = 0$ on $(0, S) \setminus G$ such that the complementarity condition and the inclusion here below are satisfied

$$\text{f.a.a. } s \in (0, S): \quad \hat{t}'(s)\text{dist}_{\mathcal{V}}(-D_z\mathcal{E}(\hat{t}(s), \hat{z}(s)), \partial\mathcal{R}(0)) = 0 \tag{4.3}$$

$$\text{f.a.a. } s \in G: \quad 0 \in \partial\mathcal{R}(\hat{z}'(s)) + \lambda(s)\hat{z}'(s) + D_z\mathcal{E}(\hat{t}(s), \hat{z}(s)). \tag{4.4}$$

The solution is p-normalized if in addition

$$\hat{t}'(s) + \mathcal{R}[\hat{z}'](s) + \|\hat{z}'(s)\|_{\mathcal{V}}\text{dist}_{\mathcal{V}}(-D_z\mathcal{E}(\hat{t}(s), \hat{z}(s)), \partial\mathcal{R}(0)) = 1 \text{ for a.a. } s.$$

It is important to note here that, since the norms involved in the reparametrization are relatively weak, the resulting limit cannot be shown to be differentiable w.r.t. the norm on $\mathcal{V}$ on the entire interval $[0, S]$. Instead, we are dealing with the so-called *generalized metric derivative*, which for $z \in AC^{\infty}([0, S], \mathcal{X})$ is defined by

$$\mathcal{R}[z'](s) := \lim_{h \searrow 0} \mathcal{R}((z(s + h) - z(s))/h) \in \mathbb{R}.$$

It is shown in [2, Thm 1.1.2] that $\mathcal{R}[z'](s)$ exists almost everywhere and that $\mathcal{R}[z'] \in L^{\infty}(0, S)$.

Theorem 4.2 *Let $\alpha > 0$ be a fixed Tikhonov parameter, $z_0 \in \mathcal{Z}$ be chosen such that there exists $\ell \in H^1(0, T; \mathcal{V}^*)$ such that (z_0, ℓ) complies with (3.7), and let $j : \mathcal{V} \to \mathbb{R}$ be bounded from below and continuous. Then, the optimal control problem (4.1) has a globally optimal solution.*

The proof of this existence theorem relies on the (weak) sequential compactness of solution sets, to be more precise of sets of the type

$$M_\rho := \big\{(S, \hat{t}, \hat{z}) \,;\, \exists\, z_0, \ell \text{ such that (3.7) and } \|z_0\|_{\mathcal{Z}} + \|\ell\|_{H^1((0,T);\mathcal{V}^*)} \leq \rho \text{ hold}$$
$$\text{and } (S, \hat{t}, \hat{z}) \text{ is a normalized } p\text{-parametrized BV solution for } (z_0, \ell)\big\}$$

for $\rho > 0$.

The key ingredient for proving compactness of the solution sets M_ρ is an a priori estimate for the driving forces $D_z\mathcal{E}(\hat{t}, \hat{z})$ in the space $\mathcal{V}^*$, which is based on the inclusion (4.4). This estimate is trivial on the complement of the set G from (4.2), since $\partial\mathcal{R}(0)$ is a bounded set in $\mathcal{V}^*$, but poses a major challenge on the set G. This can be remedied in the following way: thanks to (4.3), we know that the external load $\ell \circ \hat{t}$ is a constant on each connected component of G, which we denote by $\ell_* \in \mathcal{V}^*$. For each such component, we can therefore find a reparametrization such that the transformed functions $\widetilde{z}$ are solutions of the autonomous system

$$0 \in \partial\mathcal{R}(\dot{\widetilde{z}}(t)) + \dot{\widetilde{z}}(t) + D\mathcal{I}(\widetilde{z}(t)) - \langle \ell_*, \widetilde{z}(t)\rangle_{\mathcal{V}^*,\mathcal{V}} \quad \text{for } t > 0. \tag{4.5}$$

The essential estimates are then derived for the viscous model (4.5) and subsequently transferred to the original one. We refer the reader to [22] for the details.

5 Optimal Control of Thermo-Viscoplasticity

In this section, we elaborate on a system of thermo-visco(elasto)plasticity at small strains with linear kinematic hardening and von Mises yield condition. This topic is motivated by a multitude of important applications, where thermo-plastic material behavior leads to severe damage and material fatigue. We exemplarily mention thermally induced creeping, which occurs for instance in the operation of power plants, whose turbine blades deform permanently under the influence of stresses and heat. This process may eventually cause thermo-mechanical fatigue, cf. e.g., [5, 28, 44]. Another instance is the behavior of steel columns exposed to fire, cf. [46]. The high temperatures lead to a substantial damage of the material such that its yield strength is significantly decreased, a phenomenon known as creep buckling of the columns. This is regarded one of the main reasons for the collapse of the World Trade Center in New York on 2001/9/11, see [7].

The temperature part of these models is naturally rate-dependent, whereas many plasticity models are rate-independent. However, as a first step toward these coupled rate-dependent/rate-independent models, we consider a viscous regularization of the elasto-plastic system. This regularization is two-fold, and it applies to both the viscoplastic flow rule and the balance of momentum. The associated regularization parameters are $\nu > 0$ and $\gamma > 0$, respectively.

The overall system we consider is as follows:

$$\text{stress–strain relation:} \quad \boldsymbol{\sigma} = \mathbb{C}\big(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \boldsymbol{p} - \boldsymbol{t}(\theta)\big), \tag{5.1}$$

$$\text{conjugate forces:} \quad \boldsymbol{\chi} = -\mathbb{H}\,\boldsymbol{p}, \tag{5.2}$$

$$\text{viscoplastic flow rule:} \quad \gamma\,\dot{\boldsymbol{p}} + \partial_{\dot{\boldsymbol{p}}} D(\dot{\boldsymbol{p}}, \theta) \ni [\boldsymbol{\sigma} + \boldsymbol{\chi}]\,, \tag{5.3}$$

$$\text{balance of momentum:} \quad -\operatorname{div}\big(\boldsymbol{\sigma} + \nu\,\boldsymbol{\varepsilon}(\dot{\boldsymbol{u}})\big) = \boldsymbol{\ell}, \tag{5.4}$$

$$\text{heat equation:} \quad \varrho\, c_p\, \dot{\theta} - \operatorname{div}(\kappa\,\nabla\theta) = r + \nu\,\boldsymbol{\varepsilon}(\dot{\boldsymbol{u}}) : \boldsymbol{\varepsilon}(\dot{\boldsymbol{u}}) + (\boldsymbol{\sigma} + \boldsymbol{\chi}) : \dot{\boldsymbol{p}} - \theta\,\boldsymbol{t}'(\theta) : \mathbb{C}(\boldsymbol{\varepsilon}(\dot{\boldsymbol{u}}) - \dot{\boldsymbol{p}}). \tag{5.5}$$

The unknowns are the stress $\boldsymbol{\sigma}$, back-stress $\boldsymbol{\chi}$, plastic strain $\boldsymbol{p}$, displacement $\boldsymbol{u}$, and temperature θ. Further, $\mathbb{C}$ and $\mathbb{H}$ denote the elastic and hardening moduli, respectively, and $\boldsymbol{\varepsilon}(\boldsymbol{u})$ denotes the symmetrized gradient or linearized strain associated with $\boldsymbol{u}$. The temperature-dependent term $\boldsymbol{t}(\theta)$ expresses thermally induced strains. D denotes the dissipation function and is assumed to be convex and positively homogeneous of degree one. The right-hand sides $\boldsymbol{\ell}$ and r represent mechanical and thermal volume and boundary loads, respectively. ϱ, c_p, and κ describe the density,

specific heat capacity, and thermal conductivity of the material. For the derivation of the system (5.1)–(5.5) and more on its physical background, we refer the reader to [39, Chapter 22 and 23]. Observe that (5.2) and (5.3) play the role of (BM), while (5.1) and (5.4) correspond to a viscoelastic version of (RIS).

We analyzed the well-posedness of (5.1)–(5.5) in the presence of mixed essential and natural boundary conditions for the temperature and the elastic displacement equations. The details along with the full set of assumptions are given in [13, Theorem 10]. The proof utilizes maximal parabolic regularity results and Banach's fixed-point theorem, applied to a reduced problem formulated in the temperature variable θ alone. We also established the weak sequential continuity of the control-to-state map $(\boldsymbol{\ell}, r) \mapsto (\boldsymbol{u}, \boldsymbol{p}, \theta)$ in appropriate spaces. This allowed us to deduce the existence of optimal controls for a range of problems; see [13, Section 4].

In the follow-up paper [15], differentiability properties of the control-to-state map $(\boldsymbol{\ell}, r) \mapsto (\boldsymbol{u}, \boldsymbol{p}, \theta)$ were investigated. Under similar assumptions as in [13], we established its local Lipschitz continuity as well as directional differentiability. Together, this implies the Hadamard differentiability of the control-to-state map. Moreover, owing to a result by [40], the map is Fréchet differentiable on a dense set.

Based on these findings, a nonlinear conjugate gradient method for the solution of optimal control problems associated with (5.1)–(5.5) was devised and implemented in the dissertation [49]. The forward system was solved using a semi-implicit Euler scheme in time and an appropriate finite-element discretization in space. Under the assumption that the iterates occurring in an optimization method are points of differentiability, an adjoint representation of the gradient of the discretized problem was derived and, subsequently, a nonlinear conjugate gradient scheme with line search implemented.

We report here on one particular setting and associated numerical results. We refer the reader to [49, Chapter 7] for full details. The two-dimensional geometry is a square disc with a circular hole depicted in Fig. 5. Due to symmetry, the actual computational domain is only a quarter of the disc.

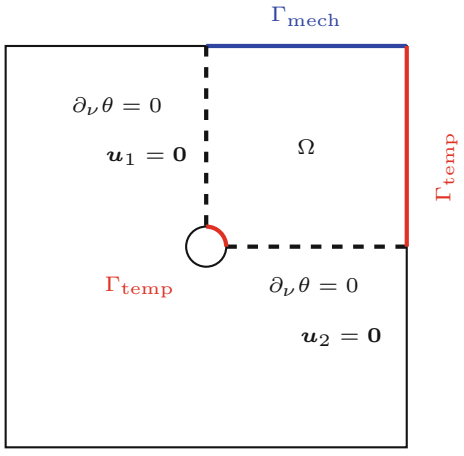

Fig. 5 Geometry of a square disc with a hole. Symmetry boundary conditions are applied, and only a quarter of the disc is used as the computational domain

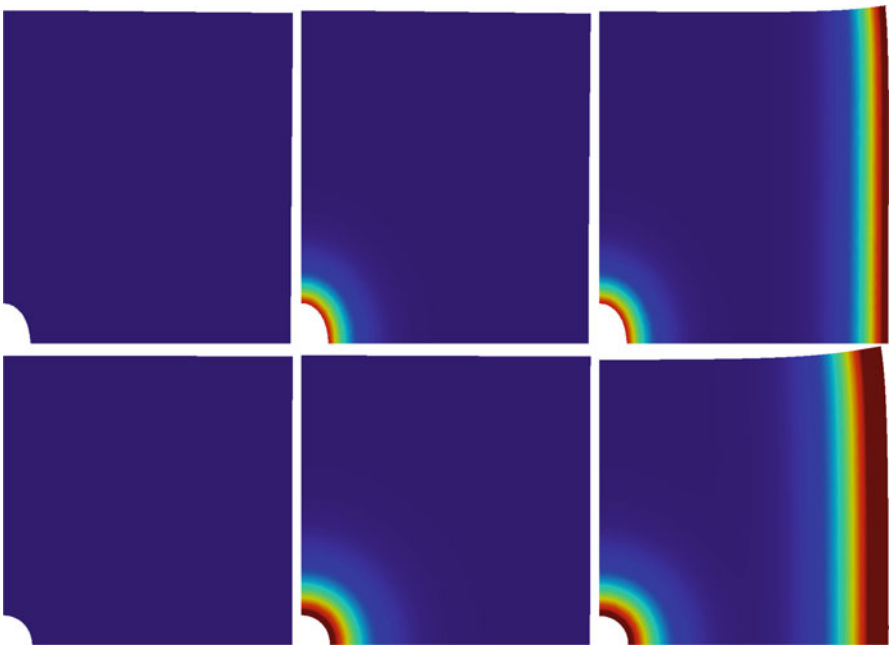

Fig. 6 Temperatures associated with computed controls in the three settings (left: no heating, middle: heating at the hole, right: heating at the hole and the right boundary) at two different points in time during the heating phase

The goal in this chapter was to obtain a uniform final displacement of the upper boundary after the process and subsequent cooling. To this end, mechanical traction forces are applied at the upper boundary, whose evolution in time and spatial distribution are subject to optimization. We compare the achievement of this goal in three settings. The first setting does not apply heating, while the second and third settings allow for additional heating at the hole, and at the hole and on the right boundary, respectively. In these cases, the thermal control acts as a second control variable. The computed mechanical controls, as well as the associated temperature and plastic strain in the three settings, are shown in Figs. 6 and 7 for two different points in time.

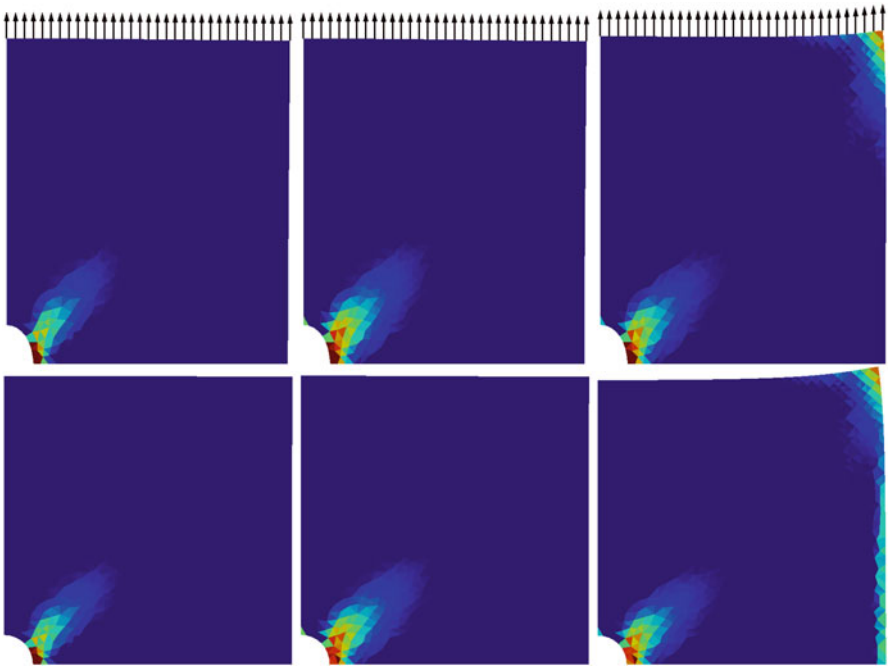

Fig. 7 Frobenius norm of the plastic strain $|\boldsymbol{p}|$ and computed mechanical controls (boundary tractions) in the three settings (left: no heating, middle: heating at the hole, right: heating at the hole and the right boundary) at the same points in time as in Fig. 6

Acknowledgments The research of this work was carried out in Project P09 (Optimal Control of Dissipative Solids: Viscosity Limits and Non-Smooth Algorithms) within the DFG Priority Program SPP 1962 (Non-Smooth and Complementarity-Based Distributed Parameter Systems: Simulation and Hierarchical Optimization). The support by the DFG is gratefully acknowledged.

References

1. Jochen Alberty and Carsten Carstensen. "Numerical Analysis of Time-Depending Primal Elastoplasticity with Hardening". In: *SIAM Journal on Numerical Analysis* 37.4 (2000), pp. 1271–1294.
2. Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2005.
3. Marco Artina, Filippo Cagnetti, Massimo Fornasier, and Francesco Solombrino. "Linearly constrained evolutions of critical points and an application to cohesive fractures." In: *Math. Models Methods Appl. Sci.* 27.2 (2017), pp. 231–290. https://doi.org/10.1142/S0218202517500014.
4. Sören Bartels. "Quasi-optimal Error Estimates for Implicit Discretizations of Rate-Independent Evolutions". In: *SIAM Journal on Numerical Analysis* 52.2 (2014), pp. 708–716.

5. Herbert F. Bahls and Hans-Thomas Bolms. “Turbinenbeschaufelung”. In: *Stationäre Gasturbinen*. Heidelberg: Springer, 2010, pp. 567–594.
6. Haïm Brézis. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*. North-Holland Publishing Co., 1973.
7. Zdeněk P. Bažant and Yong Zhou. “Why Did the World Trade Center Collapse? – Simple Analysis”. In: *Journal of Engineering Mechanics* 128.1 (2002), pp. 2–6.
8. Gianni Dal Maso, Antonio DeSimone, Maria Giovanna Mora, and Massimiliano Morini. “A vanishing viscosity approach to quasistatic evolution in plasticity with softening”. In: *Archive for Rational Mechanics and Analysis* 189.3 (2008), pp. 469–544. https://doi.org/10.1007/s00205-008-0117-5.
9. Michela Eleuteri and Luca Lussardi. “Thermal control of a rate-independent model for permanent inelastic effects in shape memory materials”. In: *Evol. Equ. Control Theory* 3.3 (2014), pp. 411–427. https://doi.org/10.3934/eect.2014.3.411.
10. Michela Eleuteri, Luca Lussardi, and Ulisse Stefanelli. “Thermal control of the Souza-Auricchio model for shape memory alloys”. In: *Discrete Contin. Dyn. Syst. Ser. S* 6.2 (2013), pp. 369–386. https://doi.org/10.3934/dcdss.2013.6.369.
11. Messoud A. Efendiev and Alexander Mielke. “On the rate-independent limit of systems with dry friction and small viscosity”. In: *Journal of Convex Analysis* 13.1 (2006), pp. 151–167.
12. Alessandro Giacomini. “Ambrosio-Tortorelli approximation of quasistatic evolution of brittle fractures”. In: *Calculus of Variations and Partial Differential Equations* 22.2 (2005), pp. 129–172. https://doi.org/10.1007/s00526-004-0269-6.
13. Roland Herzog, Christian Meyer, and Ailyn Stötzner. “Existence of solutions of a thermoviscoplastic model and associated optimal control problems”. In: *Nonlinear Analysis: Real World Applications* 35 (2017), pp. 75–101. https://doi.org/10.1016/j.nonrwa.2016.10.008.
14. Weimin Han and B. Daya Reddy. *Plasticity*. Second. Vol. 9. Interdisciplinary Applied Mathematics. Mathematical theory and numerical analysis. Springer, New York, 2013, pp. xvi+421. https://doi.org/10.1007/978-1-4614-5940-8.
15. Roland Herzog and Ailyn Stötzner. “Hadamard Differentiability of the Solution Map in Thermoviscoplasticity”. In: *Pure and Applied Functional Analysis* 4.2 (2019), pp. 271–295.
16. Bernard Halphen and Nguyen Quoc Son. “Sur les matériaux standards généralisés”. In: *Journal de Mécanique Théorique et Appliquée. Journal of Theoretical and Applied Mechanics* 14 (1975), pp. 39–63.
17. Claes Johnson. “On Plasticity with Hardening”. In: *Journal of Mathematical Analysis and Applications* 62.2 (1978), pp. 325–336. https://doi.org/10.1016/0022-247X(78)90129-4.
18. Dorothee Knees and Matteo Negri. “Convergence of alternate minimization schemes for phase-field fracture and damage.” In: *Math. Models Methods Appl. Sci.* 27.9 (2017), pp. 1743–1794. https://doi.org/10.1142/S0218202517500312.
19. Dorothee Knees. “Convergence analysis of time-discretisation schemes for rate-independent systems”. In: *ESAIM, Control Optim. Calc. Var.* 25 (2019). https://doi.org/10.1051/cocv/2018048.
20. Dorothee Knees, Riccarda Rossi, and Chiara Zanini. “Balanced viscosity solutions to a rate-independent system for damage”. In: *European Journal of Applied Mathematics* 30.1 (2019), pp. 117–175. https://doi.org/10.1017/S0956792517000407.
21. Dorothee Knees and Andreas Schröder. “Computational aspects of quasi-static crack propagation”. In: *Discrete and Continuous Dynamical Systems. Series S* 6.1 (2013), pp. 63–99. https://doi.org/10.3934/dcdss.2013.6.63.
22. Dorothee Knees and Stephanie Thomas. *Optimal control for rate independent systems constrained to BV solutions*. Tech. rep. arXiv:1810.12572. University of Kassel, 2018.
23. Dorothee Knees, Chiara Zanini, and Alexander Mielke. “Crack growth in polyconvex materials”. In: *Physica D. Nonlinear Phenomena* 239.15 (2010), pp. 1470–1484. https://doi.org/10.1016/j.physd.2009.02.008.
24. Christopher J. Larsen. “Epsilon-stable quasi-static brittle fracture evolution.” In: *Commun. Pure Appl. Math.* 63.5 (2010), pp. 630–654. https://doi.org/10.1002/cpa.20300.

25. Alexander Mielke, Laetitia Paoli, Adrien Petrov, and Ulisse Stefanelli. "Error estimates for space-time discretizations of a rate-independent variational inequality". In: *SIAM Journal on Numerical Analysis* 48.5 (2010), pp. 1625–1646.
26. Alexander Mielke. "Chapter 6 Evolution of Rate-Independent Systems". In: *Handbook of Differential Equations, Evolutionary Equations* 2 (Dec. 2006).
27. Andreas Mainik and Alexander Mielke. "Global existence for rate-independent gradient plasticity at finite strain." In: *J. Nonlinear Sci.* 19.3 (2009), pp. 221–248. https://doi.org/10.1007/s00332-008-9033-y.
28. Ekkehard Maldfeld and Michael Müller. "Statische und dynamische Auslegung des Turbinenläufers". In: *Stationäre Gasturbinen*. Heidelberg: Springer, 2010, pp. 649–682.
29. Alexander Mielke and Tomáš Roubíček. *Rate-independent systems. Theory and application.* Vol. 193. New York, NY: Springer, 2015. https://doi.org/10.1007/978-1-4939-2706-7.
30. Alexander Mielke, Tomáš Roubíček, and Ulisse Stefanelli. "Γ-limits and relaxations for rate-independent evolutionary problems." In: *Calc. Var. Partial Differ. Equ.* 31.3 (2008), pp. 387–416. https://doi.org/10.1007/s00526-007-0119-4.
31. Alexander Mielke, Riccarda Rossi, and Giuseppe Savaré. "BV solutions and viscosity approximations of rate-independent systems". In: *ESAIM. Control, Optimisation and Calculus of Variations* 18.1 (2012), pp. 36–80. https://doi.org/10.1051/cocv/2010054.
32. Alexander Mielke, Riccarda Rossi, and Giuseppe Savaré. "Balanced viscosity (BV) solutions to infinite-dimensional rate-independent systems." In: *J. Eur. Math. Soc. (JEMS)* 18.9 (2016), pp. 2107–2165. https://doi.org/10.4171/JEMS/639.
33. Luca Minotti and Giuseppe Savaré. "Viscous corrections of the time incremental minimization scheme and visco-energetic solutions to rate-independent evolution problems." In: *Arch. Ration. Mech. Anal.* 227.2 (2018), pp. 477–543. https://doi.org/10.1007/s00205-017-1165-5.
34. Christian Meyer and Michael Sievers. *A-priori error analysis of local incremental minimization schemes for rate-independent evolutions*. Tech. rep. SPP1962-115. 2019.
35. Christian Meyer and Michael Sievers. "Finite element discretization of local minimization schemes for rate-independent evolutions". In: *Calcolo. A Quarterly on Numerical Analysis and Theory of Computation* 56.1 (2019), Art. 6, 38. https://doi.org/10.1007/s10092-018-0301-4.
36. Alexander Mielke and Florian Theil. "On rate-independent hysteresis models". In: *NoDEA: Nonlinear Differential Equations and Applications* 11.2 (2004), pp. 151–189.
37. Alexander Mielke, Florian Theil, and Valery I. Levitas. "A variational formulation of rate-independent phase transformations using an extremum principle". In: *Archive for Rational Mechanics and Analysis* 162.2 (2002), pp. 137–177. https://doi.org/10.1007/s002050200194.
38. Alexander Mielke and Sergey Zelik. "On the vanishing-viscosity limit in parabolic systems with rate-independent dissipation terms". In: *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)* 13.1 (2014), pp. 67–135.
39. Niels S. Ottosen and Matti Ristinmaa. *The Mechanics of Constitutive Modeling*. Elsevier, 2005.
40. David Preiss. "Differentiability of Lipschitz functions on Banach spaces". In: *Journal of Functional Analysis* 91.2 (1990), pp. 312–345. https://doi.org/10.1016/0022-1236(90)90147-D.
41. Filip Rindler. "Optimal control for nonconvex rate-independent evolution processes". In: *SIAM J. Control Optim*. 47.6 (2008), pp. 2773– 2794. https://doi.org/10.1137/080718711.
42. Tomáš Roubíček. "Rate-independent processes in viscous solids at small strains." In: *Math. Methods Appl. Sci.* 32.7 (2009), pp. 825–862.
43. Tomáš Roubíček, Marita Thomas, and Christos G. Panagiotopoulos. "Stress-driven local-solution approach to quasistatic brittle delamination". In: *Nonlinear Analysis. Real World Applications. An International Multidisciplinary Journal* 22 (2015), pp. 645–663. https://doi.org/10.1016/j.nonrwa.2014.09.011.
44. Abdullah Aziz Saad. "Cyclic Plasticity and Creep of Power Plant Materials". PhD thesis. University of Nottingham, UK, 2012.

45. David Schrade, Marc-André Keip, Huy Ngoc Minh Thai, Jörg Schröder, Bob Svendsen, Ralf Müller, and Dietmar Gross. “Coordinate-invariant phase-field modeling of ferroelectrics, part I: Model formulation and single-crystal simulations”. In: *GAMM-Mitt.* 38 (2015), pp. 102–114.
46. Diego Somaini, Markus Knobloch, and Mario Fontana. “Buckling of steel columns in fire: non-linear behaviour and design proposal”. In: *Steel Construction - Design and Research* 5.3 (2012), pp. 175–182. https://doi.org/10.1002/stco.201210022.
47. Ulisse Stefanelli. “A variational characterization of rate-independent evolution”. In: *Mathematische Nachrichten* 282.11 (2009), pp. 1492–1512. https://doi.org/10.1002/mana.200810803.
48. Ulisse Stefanelli. “Magnetic control of magnetic shape-memory single crystals”. In: *Physica B* 407 (2012), pp. 1316–1321.
49. Ailyn Stötzner. “Optimal Control of Thermoviscoplasticity”. PhD thesis. Technische Universität Chemnitz, Germany, 2018. urn: urn:nbn: de:bsz:ch1-qucosa2-318874.

Generalized Nash Equilibrium Problems with Partial Differential Operators: Theory, Algorithms, and Risk Aversion

Deborah Gahururu, Michael Hintermüller, Steven-Marian Stengl, and Thomas M. Surowiec

Abstract PDE-constrained (generalized) Nash equilibrium problems (GNEPs) are considered in a deterministic setting as well as under uncertainty. This includes a study of deterministic GNEPs with nonlinear and/or multi-valued operator equations as forward problems and PDE-constrained GNEPs with uncertain data. The deterministic nonlinear problems are analyzed using the theory of generalized convexity for set-valued operators, and a variational approximation approach is proposed. The stochastic setting includes a detailed overview of the recently developed theory and algorithms for risk-averse PDE-constrained optimization problems. These new results open the way to a rigorous study of stochastic PDE-constrained GNEPs.

Keywords Generalized Nash equilibrium problems · PDE-constrained optimization · L-convexity · Set-valued analysis · Fixed-point theory · Risk-averse optimization · Coherent risk measures · Stochastic optimization · Method of multipliers

Mathematics Subject Classification (2020) 49J20, 49J55, 49K20, 49K45, 49M99, 65K10, 65K15, 90C15, 91A10

D. Gahururu · T. M. Surowiec
Philipps-Universität Marburg, Faculty of Mathematics and Computer Science, Marburg, Germany
e-mail: gahururu@mathematik.uni-marburg.de; surowiec@mathematik.uni-marburg.de

M. Hintermüller (✉)
Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany
e-mail: michael.hintermueller@wias-berlin.de

S.-M. Stengl
Weierstrass Institute, Berlin, Germany
e-mail: steven-marian.stengl@wias-berlin.de

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_7

1 Introduction

Many applications and areas in science study phenomena sharing the common requirement of minimizing more than one objective simultaneously. In general, the solution of these problems has to address conflicting interests of the involved agents. Hence, we turn our attention to modeling a degree of competition and noncooperative behavior leading to *Nash games*. This concept has been successfully applied to a variety of applications in economics and in the context of networks, see [9, 11] and additionally [35] for the combinatorial branch of optimization. In many practical cases, the actions of the players in these games are restricted by equilibrium constraints establishing a reinforced linkage between the diverging interests. As we know from the mathematical treatment of optimal control and design problems, this coupling is usually resolved as an operator equation. However, in the context of partial differential equation (PDE)-constrained optimization, this type of concept has not yet been frequently studied.

We start by motivating N agent *games*. In this context, mathematically speaking, a set of N agents (or players) solve each an individual minimization problem to find their respective optimal strategy. For player i, this reads as

$$\text{minimize}_{u_i \in U_{\text{ad}}^i} \; \mathcal{J}_i(u_i, u_{-i}) \text{ over } u_i \in U_i,$$

where $U_{\text{ad}}^i \subset U_i$, with U_i a Banach space, is the set of *feasible strategies*. The functional $\mathcal{J}_i$ is specific for the player and involves his strategy u_i as well as the (given) strategies of all other players denoted as u_{-i}. Here and in the following, the combined vector of *all* strategies is usually denoted as $u = (u_i, u_{-i})$ without any permutation of components. A vector $u \in U$ with $U = U_1 \times \cdots \times U_N$ is called a *Nash equilibrium* if every strategy chosen by an agent is his optimal choice given the strategies of the other agents. This yields

$$u_i \in \text{argmin}_{u_i' \in U_{\text{ad}}^i} \left\{ \mathcal{J}_i(u_i', u_{-i}) \text{ over } u_i' \in U_i \right\} \text{ for all } i = 1, \ldots, N. \tag{1.1}$$

The problem of finding such a strategy vector is then called a *Nash equilibrium problem* (NEP). In this setting, the influence of the other players' actions is limited to the objectives, whereas the strategy sets remain unchanged. Allowing the other players to also influence the set of feasible strategies leads to a set-valued strategy mapping $C_i : U_{\text{ad}}^{-i} \rightrightarrows U_{\text{ad}}^i$ in the underlying optimization problems. A Nash equilibrium is then a point $u \in U_{\text{ad}}$ with $U_{\text{ad}} = U_{\text{ad}}^1 \times \ldots \times U_{\text{ad}}^N$ satisfying

$$u_i \in \text{argmin}_{u_i' \in C_i(u_{-i})} \left\{ \mathcal{J}_i(u_i', u_{-i}) \text{ over } u_i' \in U_i \right\} \text{ for all } i = 1, \ldots, N.$$

Finding a solution for the latter type of problem is also known as *Generalized Nash equilibrium problem* (GNEP). Correspondingly, we assume the strategy mapping to be structured as

$$C_i(u_{-i}) = \left\{ u_i' \in U_{\text{ad}}^i : g(u_i', u_{-i}) \in K \right\}$$

with $g : U \to X$ and $K \subseteq X$ a nonempty, closed, convex subset of some Banach space X. In principle, it is possible to incorporate several mappings $g_i : U \to X_i$, but we want to keep our presentation concise. Therefore, in our context, a (GNEP) is given by

$$u_i \in \operatorname{argmin}\{\mathcal{J}_i(u_i', u_{-i}) \text{ subject to } u_i' \in U_{\text{ad}}^i \text{ and } g(u_i', u_{-i}) \in K\} \tag{1.2}$$

for all $i = 1, \ldots, N$. Concerning the general constraint, we are particularly interested in constraints on the *state* variable y that is generated through a continuous *solution mapping* $S : U \to Y$ involving the entirety of the players strategies via $y = S(u)$. Here, the set Y is again a Banach space. The origin of this operator might be a PDE or the minimization of an underlying parametrized optimization problem. Moreover, we assume in our setting that the players' objectives are separable of the type

$$\mathcal{J}_i(u_i, u_{-i}) = J_i^1(S(u_i, u_{-i})) + J_i^2(u_i).$$

Here J_i^1 only depends on the state, e.g., by a data-fitting, respectively, tracking-type term, and J_i^2 only on the control, e.g., in the form of a regularization or control cost. Note that by this setting a coupling between the players is established via the objectives. The dependence of the feasible sets occurs through the presence of a state constraint $\mathcal{G}(y) \in K$, which might stem from a physical or technical consideration. Hence, a (GNEP) in our setting has the general form

$$\begin{aligned} &\text{minimize}_{u_i, y} \quad J_i^1(y) + J_i^2(u_i) \text{ over } u_i \in U_i,\, y \in Y \\ &\text{subject to } u_i \in U_{\text{ad}}^i \text{ and } \mathcal{G}(y) \in K \text{ with } y = S(u_i, u_{-i}). \end{aligned} \tag{1.3}$$

Here, the continuous mapping $\mathcal{G} : Y \to X$, together with the set K, models the state constraint, leading to the relation $g = \mathcal{G} \circ S$. This model is flexible enough to allow for a wide variety of different mathematical and practical applications. However, some aspects discussed hereafter are more conveniently described using the more abstract setting of (1.2) rather than (1.3). We will, hence, switch between these formulations keeping their formal relation in mind.

As previously mentioned, the operator S may originate from a broad variety of problems including (possibly nonlinear) PDEs, Vis, or complementarity problems. Throughout, we assume the solution mapping to be a singleton, meaning that given u the state $y = y(u)$ is unique. This does not need to be the case in general. Our model may thus be seen as closely related to multi-leader-follower games (MLFG) that are investigated within the scope of this report, as well.

Mathematical games involve a broad variety of challenges, including existence, characterization of equilibria via first-order systems, as well as numerical analysis and solvers. Moreover, in many applications, problem data are uncertain, occurring, e.g., as random parameters. This gives rise to risk-related formulations of the involved PDE-constrained minimization as well as (G)NEP. In this chapter, we study in particular risk-averse agents by modeling appropriate individual objectives.

2 Nash Games Involving Nonlinear Operator Equations

We study the following Nash game with a linear operator equations and compare [21]:

$$\begin{aligned} &\text{minimize} \quad J_i^1(y) + J_i^2(u_i) \text{ over } u_i \in U_i,\, y \in Y, \\ &\text{subject to } u_i \in U_{\text{ad}}^i \text{ and } Ay = b + Bu \text{ in } W. \end{aligned} \tag{2.1}$$

Here, Y is as before, W a Banach space, $b \in W$ fixed, $A \in \mathcal{L}(Y, W)$ an invertible, bounded linear operator, and $B \in \mathcal{L}(U, W)$ a bounded, linear operator involving the strategies of all players at once. This motivates the solution operator $S(u) = A^{-1}(b + Bu)$ of the state equation $Ay = b + Bu$.

First, we study existence of an equilibrium of (2.1). Here, the coupling of the minimization problems of the individual agents prevents using a technique associated with a single minimization problem. Rather we need to invoke fixed-point theory for set-valued operators. For this, we reformulate (1.1) as

$$u \in \mathcal{B}(u), \tag{2.2}$$

with $\mathcal{B}(u) := \Pi_{i=1}^N \mathcal{B}_i(u_{-i})$, and $\mathcal{B}_i(u_{-i}) = \operatorname{argmin}\left\{\mathcal{J}_i(u_i', u_{-i}) : u_i' \in U_{\text{ad}}^i\right\}$. Here, the *best response mapping* $\mathcal{B} : U_{\text{ad}} \rightrightarrows U_{\text{ad}}$ assigns to every given strategy the Cartesian product of all players' feasible strategies yielding the optimal value. The existence proof of a solution to (2.2) uses a result of Kakutani, Fan and Glicksberg:

Theorem 2.1 (cf. [14]) *Given a closed point-to-(nonvoid)-convex-set mapping $\Phi : Q \rightrightarrows Q$ of a convex compact subset Q of a convex Hausdorff linear topological space into itself, then there exists a fixed point $x \in \Phi(x)$.*

Two assumptions are crucial in the above theorem: (i) the convexity assumption on the values of the mapping and (ii) the compactness of the underlying set. In our situation, (i) becomes a topological condition regarding the set of minimizers for the players' optimization problems. This property is guaranteed when the (reduced) objective functional is convex. Concerning (ii), in finite dimensions, the compactness is guaranteed by closedness and boundedness. In our infinite-dimensional setting, however, this condition is usually not fulfilled with respect to the strong topology. Hence, we require a transition to the weak topology leading to a strengthened condition on the closedness of the graph of the operator.

In order to apply Theorem 2.1, let J_i^1, J_i^2 be convex, continuous, functionals. Moreover, let J_i^2 or S be completely continuous on their respective domains. Additionally, let U_i be a reflexive, separable Banach space and U_{ad}^i a nonempty, closed, and bounded subset of U_i. Then, the latter is also compact with respect to the weak topology. These conditions guarantee the existence of an equilibrium by applying the theorem.

We next come to the (GNEP) in [21], which reads

$$\begin{aligned} &\text{minimize} \quad J_i^1(y) + J_i^2(u_i) \text{ over } u_i \in U_i,\, y \in Y, \\ &\text{subject to } u_i \in U_{\text{ad}}^i \text{ and } y \in K \text{ with} \\ &\qquad Ay = b + Bu \text{ in } W, \end{aligned} \tag{2.3}$$

with a continuous embedding $Y \hookrightarrow X$. Let

$$C_i(u_{-i}) := \left\{ u_i' \in U_{\text{ad}}^i : S(u_i', u_{-i}) \in K \right\}$$

denote the associated set-valued strategy map, with C again the Cartesian product. This setting adds another difficulty to the existence proof, as we are now confronted with moving sets of feasible strategies. Hence, the selection of sequences in the range of the operator to prove the closedness property of the best response map becomes an issue. To address this challenge, we notice that the condition restricting the players' feasible strategies is the same for all players. Hence, one is able to formulate the overall set of feasible strategies as

$$\mathcal{F} = \{u \in U_{\text{ad}} : S(u) \in K\}.$$

It is worth noting that the set $\mathcal{F}$ characterizes the whole strategy mapping via

$$u_i' \in C_i(u_{-i}) \Leftrightarrow (u_i', u_{-i}) \in \mathcal{F}$$

for all $i = 1, \ldots, N$, which implies in particular $\text{Fix}(C) = \mathcal{F}$, where $\text{Fix}(\cdot)$ denotes the set of fixed points of a map. In fact, this observation applies already to the more general setting of (1.2) and allows us to introduce the strengthened solution concept of *variational equilibria*. It relates to a strategy vector $u \in \mathcal{F}$ solving the fixed-point problem

$$u \in \widehat{\mathcal{B}}(u), \tag{2.4}$$

with $\widehat{\mathcal{B}} : \mathcal{F} \to \mathcal{F}$, and $\widehat{\mathcal{B}}(u) = \text{argmin}\left\{ \sum_{i=1}^N \mathcal{J}_i(u_i', u_{-i}) \text{ over } u' \in \mathcal{F} \right\}$. In this formulation, only a *single* minimization process occurs. It is straightforward to prove that every variational equilibrium is also a Nash equilibrium. Consequently, providing existence for the operator $\widehat{\mathcal{B}}$ is sufficient. To apply Theorem 2.1, we note that due to the linearity of S the joint set of feasible strategies is convex as well. If a (GNEP) has in addition only convex objectives, then it is referred to as a *jointly convex Nash game*.

Nonlinear PDEs lead to an underlying operator equation of the type

$$A(y) = b + B(u) \text{ in } W,$$

with a nonlinear operator $A : Y \to W$ and again a bounded linear $B : U \to W$. Now the solution mapping $S : U \to Y$ is nonlinear. In contrast to the previously discussed case, convexity of the reduced objectives is not necessarily fulfilled. Of course, the same holds in the generalized case for values of the strategy set C as well as for the joint set of strategy vectors $\mathcal{F}$. Hence, the existence proof becomes a very delicate task. One option to proceed is the identification of combinations of objectives and operator equations that still guarantee the required convexity conditions. In this context, it is interesting to discuss the necessary structure first for mere optimization problems and then for Nash games. If not otherwise stated, the subsequent results of the following subsection will be made available in [19] together with their proofs.

2.1 On the Convexity of Optimal Control Problems Involving Nonlinear Operator Equations

In the following, we investigate *generalized* operator equations of the type

$$w \in A(y) \text{ in } W.$$

This setting allows us to treat also variational inequalities (VIs). Here, $w \in W$ is a given control and $y \in Y$ the associated state. To ensure well-posedness, we assume that the set-valued operator $A : Y \rightrightarrows W$ has a single-valued inverse $A^{-1} : W \to Y$ with the entire space W as its domain. Moreover, associated with Y and W, let $K \subseteq Y$, respectively, $K_W \subseteq W$ denote nonempty, closed, and convex cones. These cones induce preorder relations $\leq_K$ and $\leq_{K_W}$ on their respective spaces by $y_0 \leq_K y_1 :\Leftrightarrow y_1 - y_0 \in K$ for $y_0, y_1 \in K$ (and analogously for W). Using these relations, it is possible to generalize the convexity notion from functionals to operators, and further even to set-valued operators between Banach spaces, cf. [5, Subsection 2.3.5].

Definition 2.2 Let X_1, X_2 be topological vector spaces with $L \subseteq X_2$ a nonempty closed, convex cone inducing a preorder relation as described above. A set-valued mapping $\Phi : X_1 \rightrightarrows X_2$ is called *L-convex*, if for all $t \in (0, 1)$ and $x_0, x_1 \in X_1$ the relation

$$t\Phi(x_1) + (1-t)\Phi(x_0) \subseteq \Phi(tx_1 + (1-t)x_0) + L$$

holds. Additionally, Φ is called L-concave if it is $(-L)$-convex.

Our next aim is to identify conditions on the operator A that guarantee that the solution operator $A^{-1} : W \to Y$ is L-convex.

Theorem 2.3 *Let Y, W be Banach spaces, both equipped with closed and convex cones $L \subseteq Y$ and $L_W \subseteq W$, respectively. Let $A : Y \rightrightarrows W$ be a set-valued operator fulfilling the following assumptions:*

(i) *The operator A is L_W-concave in the sense of Definition 2.2.*
(ii) *The mapping $A^{-1} : W \to Y$ is single-valued with domain* dom $A = W$*, and it is L_W-L-isotone (compare also to* [4, Section 1.2]*), i.e.,*

$$\text{for } w_1, w_0 \in W \text{ with } w_2 \geq_{L_W} w_1 \text{ it holds that } A^{-1}(w_2) \geq_L A^{-1}(w_1).$$

Then, the mapping $A^{-1} : W \to Y$ is L-convex.

We illustrate the previous Theorem 2.3 by two examples.

Example Let $d \in \mathbb{N}\backslash\{0\}$ and $D \subseteq \mathbb{R}^d$ be an open, bounded domain with Lipschitz boundary. Consider the operator

$$A(y) := -\Delta y + N(y) \tag{2.5}$$

on the Sobolev space $Y = H_0^1(D)$ with $W = H^{-1}(D)$. Let N be a superposition operator $N : L^2(D) \to L^2(D)$ induced by a concave, nondecreasing function on $\mathbb{R}$. We set $L := \left\{\varphi \in H_0^1(D) : \varphi \geq 0 \text{ a.e. on } D\right\}$ together with $L_W := L^+$ with

$$L^+ = \left\{\xi \in H^{-1}(D) : \langle \xi, \varphi\rangle_{H^{-1}, H_0^1} \geq 0 \text{ for all } \varphi \in H_0^1(D) \text{ with } \varphi \geq 0 \text{ a.e. on } D\right\}.$$

Then, A is L_W-concave: Indeed, let $t \in (0, 1)$ and $y_0, y_1 \in H_0^1(D)$ and $\varphi \in L$ be arbitrarily chosen; then we have

$$\begin{aligned}&\langle tA(y_1) + (1-t)A(y_0) - A(ty_1 + (1-t)y_0), \varphi\rangle_{H^{-1}, H_0^1}\\ &\qquad = (tN(y_1) + (1-t)N(y_0) - N(ty_1 + (1-t)y_0), \varphi)_{L^2(D)} \leq 0,\end{aligned}$$

showing the concavity of A. Moreover, the operator A is invertible and isotone in the L_W-L-sense. The first property can be deduced from the monotonicity of the operator N together with the coercivity of the Laplacian. To see the latter, choose $w_0, w_1 \in W$ with $w_0 \leq_{L_W} w_1$, and let $y_0, y_1 \in Y$ be the solution of $w_j = A(y_j)$ for $j = 0, 1$. Testing the difference of the equations by $(y_0 - y_1)^+$ yields

$$\begin{aligned}0 &\geq -\|\nabla(y_0 - y_1)^+\|^2_{L^2(D)} - (N(y_0) - N(y_1), (y_0 - y_1)^+)_{L^2(D)}\\ &= \langle A(y_1) - A(y_0), (y_0 - y_1)^+\rangle_{H_0^1, H^{-1}} = \langle w_1 - w_0, (y_0 - y_1)^+\rangle_{H_0^1, H^{-1}} \geq 0,\end{aligned}$$

which implies $y_1 \geq y_0$ a.e. and hence the isotonicity of A^{-1}, which gives us finally the L-convexity of the solution operator A^{-1}.

In the previous example, (2.5) relates to semilinear elliptic PDEs and hence addresses a constraint that has been widely discussed in the optimal control literature (cf. [42] for a general overview and [7, 30] for more recent research activities). An extension to semilinear parabolic equations is possible; see, e.g., [31, Chapter 3, Section 2]. Theorem 2.3 can be applied to VIs as well; see [32, Lemma 4.1] for a first result. In contrast, here we provide a more general result.

Example Let Y be a reflexive vector lattice with order cone L, i.e., Y is a reflexive Banach space and L a nonempty, closed, and convex cone with $L \cap (-L) = \{0\}$, and consider an L^+-concave, semicontinuous, and strongly monotone operator $A : Y \rightrightarrows Y^*$. Moreover, assume A to be strictly T-monotone, i.e., $\langle A(y+z) - A(y), (-z)^+ \rangle < 0$ for z with $(-z)^+ \neq 0$. Let $M \subseteq Y$ be a nonempty, closed, convex set, and lower bounded, i.e., $M + L \subseteq M$ and for all $y_0, y_1 \in M$ and $\min(y_0, y_1) \in M$. Moreover, let $w \in Y^*$ be given. We consider the following VI:

$$\text{Find } y \in M \ : \ w \in A(y) + N_M(y).$$

Then, one can show that the associated solution operator $S : Y^* \to Y$ is L-convex.

These examples illustrate the power of the proposed concept, which allows us to next consider optimization problems of the type

$$\begin{aligned} &\text{minimize} \quad J^1(y) + J^2(u) \text{ over } u \in U,\, y \in Y, \\ &\text{subject to } u \in U_{\text{ad}} \text{ and } y \in K \text{ with} \\ &\qquad b + Bu \in A(y) \text{ in } W, \end{aligned} \tag{2.6}$$

which may represent a model for a single agent's decision process. In order to guarantee the convexity of (2.6), we assume the convexity of both parts J^1 and J^2, respectively. Additionally, we assume the isotonicity of J^1 on Y, i.e., $y_0 \leq_L y_1 \Rightarrow J^1(y_0) \leq J^1(y_1)$. Considering single-valuedness, the L-convexity of the solution operator $S(u) := A^{-1}(b + B(u))$ reads $S(tu_1 + (1-t)u_0) \leq_L tS(u_1) + (1-t)S(u_0)$. Hence, $J^1 \circ S$ is convex and so is the entire objective as well. For a nonempty, closed, convex set $K \subseteq Y$ with $K - L \subseteq K$, the indicator functional $i_K : Y \to [0, +\infty]$ is isotone and convex. Thus, the convexity of the set of feasible controls in (2.6) can be stated as the following intersection of closed, convex sets:

$$\{u \in U_{\text{ad}} : S(u) \in K\} = U_{\text{ad}} \cap \{u \in U : i_K(S(u)) \leq 0\}.$$

Under these conditions, the convexity of the optimization problem (2.6) is guaranteed. We illustrate this by the following optimization of doping profiles; cf. [28].

Example Let $D \subseteq \mathbb{R}^2$ be a given, bounded, open domain with Lipschitz boundary and $D_o \subseteq D$ an open subset. For a function $z \in L^2(\Omega)$, we denote $z^{2+} :=$

$\max(0, z)^2$. Consider

$$\min_{u \in U_{\text{ad}}} \frac{1}{2} \int_{D_o} (S(u) + 1)^{2+} dx + \frac{\alpha}{2} \int_D u^2 dx, \tag{2.7}$$

where $S : L^2(D) \to H^1(D)$ is the solution operator of the following PDE:

$$-\kappa \Delta y + \sinh(y) = -Bu - b \text{ in } D, \ \kappa \frac{\partial y}{\partial n} = 0 \text{ on } \partial D,$$

with B the (linear) solution operator of the PDE

$$-r \Delta d + d = u \text{ in } D, \ r \frac{\partial d}{\partial n} = 0 \text{ on } \partial D,$$

and $U_{\text{ad}} := \{u \in L^2(D) : 0 \leq u \leq 1 \text{ a.e. on } D\}$. Note that by the use of the Trudinger–Moser inequality (cf. [34]), the function $\sinh(y)$ lies in $L^2(D)$ for $y \in H_0^1(D)$. Assume further that $b \geq 0$ a.e. on D. Then the solution operator is L-convex. To see this, define the auxiliary operator $A : H^1(D) \to H^{-1}(D)$, $\langle A(y), w \rangle_{H^{-1}, H^1} := (\nabla y, \nabla w)_{L^2} + (N(y), w)_{L^2}$, with

$$N(y) = \begin{cases} y, & \text{if } y \geq 0 \\ \sinh(y), & \text{else} \end{cases}$$

as a superposition operator. Recalling the result corresponding to (2.5), we see that the operator N is induced by a monotone and concave function on $\mathbb{R}$. Hence, the solution map is L-convex. The solution operator of the auxiliary problem and S coincide, because both operators are sign preserving. Since $u \geq 0$ a.e. by feasibility, we get $Bu \geq 0$ a.e. and together with $b \geq 0$ a.e. on D the nonnegativity of the solutions. Hence, the operators sinh and N coincide. Thus, we see that S is indeed L-convex on U_{ad}. Moreover, the objective is convex and isotone yielding the convexity of (2.7).

We would now like to derive first-order optimality conditions for (2.6). For this purpose, we extend the subdifferential concept from convex and nonsmooth analysis to vector-valued operators. For an element $y^* \in L^+$ with

$$L^+ := \{z^* \in Y^* : \langle z^*, y \rangle \geq 0 \text{ for all } y \in L\},$$

we define the *subdifferential* of the solution operator $S : U \to Y$ in direction y^* as

$$\partial S(u)(y^*) := \partial \langle y^*, S(\cdot) \rangle (u). \tag{2.8}$$

Due to the L-convexity of S also the functional $u \mapsto \langle y^*, S(u) \rangle$ is convex. Hence, the above expression (2.8) is well defined and reads as a scalarizing formulation;

compare [33, Theorem 1.90]. Note that this object is closely linked to the (Fréchet) coderivative (cf. [33, Definition 1.32], which is defined for a set-valued operator $F : X_1 \rightrightarrows X_2$ as

$$D^*F(x_1, x_2)(x_2^*) := \left\{x_1^* \in X_1^* : (x_1^*, -x_2^*) \in N_{\mathrm{gph}(F)}(x_1, x_2)\right\},$$

where $N_{\mathrm{gph}(F)}(x_1, x_2)$ denotes the (Fréchet) normal cone of $\mathrm{gph}(F)$ in $(x_1, x_2) \in \mathrm{gph}(F)$, the graph of F; see [5] for more details. In the case of a nonempty, closed, convex set, the Fréchet normal cone and its corresponding notion from convex analysis coincide. Using the mapping $S_L : U \rightrightarrows Y$ defined by $S_L(u) := S(u) + L$, we obtain for our notation in (2.8) the equivalent formulation

$$\partial S(u)(y^*) = \{u^* \in U^* : (u^*, -y^*) \in N_{\mathrm{gph}(S_L)}(u, S(u))\},$$

where we use $y^* \in K^+$. This concept allows for the following type of chain rule. In its formulation, $\mathcal{D}$ denotes the set of arguments of a set-valued map with nonempty image, and core the core of a set; see, e.g., [5, Definition 2.72] and [6, Subsection 4.1.3] for definitions and details.

Theorem 2.4 *Let U, Y be Banach spaces, the latter one equipped with a closed, convex cone L. Let $f_2 : U \to \mathbb{R} \cup \{+\infty\}$ and $f_1 : Y \to \mathbb{R} \cup \{+\infty\}$ be convex, proper, lower semicontinuous functionals, and moreover let f_1 be L-isotone. Let the operator $S : U \to Y$ be L-convex. Then, the functional $f_1 \circ S + f_2 : U \to \mathbb{R} \cup \{+\infty\}$ is convex. Furthermore, consider $u \in \mathcal{D}(\partial f_2)$ with $S(u) \in \mathcal{D}(\partial f_1)$ and let one of the following two conditions hold:*

(i) *Let S be locally bounded and the following constraint qualification hold*

$$0 \in \mathrm{core}\,(\mathrm{dom}\, f_2 \times \mathrm{dom}\, f_1 - \mathrm{gph}(S))\,.$$

(ii) *Let S be semicontinuous and the following constraint qualification hold*

$$0 \in \mathrm{core}\,(S\,(\mathrm{dom}\, f_2) - \mathrm{dom}\, f_1)\,.$$

Then, the following chain rule holds for the subdifferential of the composed objective:

$$\partial(f_1 \circ S + f_2)(u) = \partial S(u)\left(\partial f_1(S(u))\right) + \partial f_2(u).$$

The proposed chain rule in Theorem 2.4 as well as the proof and the other results of Sect. 2 will be made available in [19]. Using the functionals $f_2 = J^2 + i_{U_{\mathrm{ad}}}$ and

$f_1 = J^1 + i_C$, we obtain the first-order system

$$\begin{aligned} -q &\in \partial J^2(u) + N_{U_{\text{ad}}}(u), \\ y^* &\in \partial J^1(y) + N_K(y), \\ q &\in \partial S(u)(y^*). \end{aligned} \tag{2.9}$$

Theorem 2.4 enables one to derive necessary and sufficient optimality conditions even for constraints involving PDEs, VIs, or complementarity problems admitting a nonsmooth solution operator. Of course, not all optimal control problems will fit into the above framework and might not meet the assumptions required in Theorem 2.1. Hence, it might be worthwhile investigating the use of more general fixed-point results. One possibility in this direction is the Eilenberg–Montgomery Theorem (cf. [8]) where a weaker topological assumption replaces convexity. The application of this result still requires a characterization of the solution set for the players' optimization problems. This, however, is ongoing research.

3 Nash Games Using Penalization Techniques

The direct application of the nonsmooth approach in the previous section may be delicate for many Nash games. We therefore draw our attention to a characterization of first-order conditions for (1.3) involving a continuously differentiable solution operator. Indeed, let $A : Y \to W$ be an invertible, continuously differentiable operator with an everywhere invertible derivative. In the following, let K denote a nonempty, closed convex *cone*, and $\mathcal{G}$ a constraint map. The first-order system for a Nash equilibrium of the game associated with

$$\begin{aligned} &\text{minimize } J_i^1(y) + J_i^2(u_i) \text{ over } u_i \in U_i,\ y \in Y \text{ subject to} \\ &\quad u_i \in U_{\text{ad}}^i \text{ and } \mathcal{G}(y) \in K \text{ with} \\ &\quad A(y) = b + Bu \end{aligned} \tag{3.1}$$

for $i = 1, \ldots, N$ can be derived by the proposition of a constraint qualification of Robinson–Zowe–Kurcyusz type (RZK) (see [45]). In this setting, it reads

$$\left(D\mathcal{G}(y) \circ DA(y)^{-1} \circ B_i\right) U_{\text{ad}}^i - K(\mathcal{G}(y)) = X \text{ for all } i = 1, \ldots, N. \tag{3.2}$$

The first-order system then becomes

$$\begin{aligned} 0 &= \partial_i J_i^2(u_i) + B_i^* p_i + \lambda_i && \text{in } U_i^*, \\ A(y) &= b + Bu && \text{in } W, \\ DA(y)^* p_i &= \partial_y J_i^1(y) - D\mathcal{G}(y)^* \mu_i && \text{in } Y^*, \\ \lambda_i &\in N_{U_{\text{ad}}^i}(u_i) && \text{in } U_i^*, \\ X^* \supseteq K^+ \ni \mu_i \perp \mathcal{G}(y) &\in K \subseteq X && \text{for all } i = 1, \dots, N. \end{aligned} \tag{3.3}$$

In the case of a variational equilibrium, the single non-decoupling optimization process leads to a (possibly weaker) constraint qualification formulated as

$$\Big(D\mathcal{G}(y) \circ DA(y)^{-1} \circ B \Big) U_{\text{ad}}(u) - K(\mathcal{G}(y)) = X. \tag{3.4}$$

This leads to a special instance of (3.3) where all multipliers $\mu_i \in X^*, i = 1, \dots, N$, coincide, i.e., $\mu_i = \mu$ for all $i \in \{1, \dots, N\}$ in (3.3). In many situations involving function spaces, higher regularity of the state is needed to guarantee the constraint qualification. This on the other hand leads to a reduced regularity of the multiplier(s) $\mu_{(i)}$ and subsequently also of the adjoint states p_i in practice. The above results of the subsequent ones in this section can be found in [18], if not stated otherwise.

3.1 Γ*-Convergence*

Next we use the notion of Γ*-convergence* to approximate our state-constrained Nash game by a sequence of simpler Nash games with a weakened form of the state constraint.

First we introduce a unified view on the different notions of equilibria discussed here.

Definition 3.1 Let a Banach space U and a functional $\mathcal{E} : U \times U \to \overline{\mathbb{R}}$ be given. A point $u \in U$ is called equilibrium, if

$$\mathcal{E}(u, u) \leq \mathcal{E}(u', u) \text{ holds for all } u' \in U.$$

The first component in the functional fulfills the task of a control variable, whereas the second one acts as a parameter and hence establishes a feedback mechanism. Note that the dependence of the domain of the reduced functional $\mathcal{E}(\cdot, u)$ on u is possible. Recalling the definition of the strategy mapping C as $C(u) = \prod_{i=1}^N C_i(u_{-i})$ with $C_i(u_{-i}) = \{u_i' \in U_{\text{ad}}^i : g(u_i', u_{-i}) \in K\}$ and $g = \mathcal{G} \circ S$ as the composition of state constraint and solution operator, we reobtain by the choice

of functionals

$$\begin{aligned}\mathcal{E}(u', u) &= \sum_{i=1}^{N} \mathcal{J}_i(u_i', u_{-i}) + i_{C(u)}(u') = \sum_{i=1}^{N} \left(\mathcal{J}_i(u_i', u_{-i}) + i_{C_i(u_{-i})}(u_i')\right) \\ &= \sum_{i=1}^{N} \left(\mathcal{J}_i(u_i', u_{-i}) + i_{U_{\text{ad}}^i}(u_i) + i_K(g(u_i', u_{-i}))\right)\end{aligned} \tag{3.5}$$

and

$$\begin{aligned}\widehat{\mathcal{E}}(u', u) &= \sum_{i=1}^{N} \mathcal{J}_i(u_i', u_{-i}) + i_{\mathcal{F}}(u') \\ &= \sum_{i=1}^{N} \left(\mathcal{J}(u_i', u_{-i}) + i_{U_{\text{ad}}^i}(u_i')\right) + i_K(g(u_i', u_{-i}))\end{aligned} \tag{3.6}$$

the notion of Nash, respectively, variational equilibria. Our aim now is to generalize Γ-convergence to equilibrium problems of the above form.

Definition 3.2 Let U be a Banach space and let $\mathcal{T}$ denote either the strong or weak topology on U. A sequence of functionals $\mathcal{E}_n : U \times U \to \overline{\mathbb{R}}$ is called Γ-*convergent* to a functional $\mathcal{E} : U \times U \to \overline{\mathbb{R}}$ if the following two conditions hold:

(i) For all sequences $u_n \overset{\mathcal{T}}{\to} u$, it holds $\mathcal{E}(u, u) \leq \liminf_{n\to\infty} \mathcal{E}_n(u_n, u_n)$.

(ii) For all $u' \in U$ and all sequences $u_n \overset{\mathcal{T}}{\to} u$, there exists a sequence $u_n' \overset{\mathcal{T}}{\to} u'$ such that $\mathcal{E}(u', u) \geq \limsup_{n\to\infty} \mathcal{E}_n(u_n', u_n)$.

Of course, it is as well possible to combine the strong and weak topology in Definition 3.2. Note that the classical notion of Γ-convergence for a minimization problem is a special case of the above. The following convergence result holds true.

Proposition 3.3 *Let $\mathcal{E}_n$ be a Γ-convergent sequence of functionals as in Definition 3.1 with limit $\mathcal{E}$. Then, every accumulation point of a sequence of corresponding equilibria $(u_n)_{n\in\mathbb{N}}$ is an equilibrium of the limit.*

Our intention is to address the state constraint by applying a *penalization technique*. Therefore, the constraint $g(u) \in K$ encoded in the indicator function is substituted by a continuously differentiable *penalty function* $\beta : X \to [0, +\infty)$,

$$\beta(x) = 0 \text{ if and only if } x \in K,$$

scaled by a penalty parameter $\gamma > 0$. This leads to the formulation of the penalized functionals corresponding to the (GNEP) as

$$\mathcal{E}_\gamma(v,u) = \sum_{i=1}^{N} \Big(\mathcal{J}_i(v_i, u_{-i}) + \gamma\beta(g(v_i, u_{-i})) \Big) + i_{U_{\mathrm{ad}}}(v),$$

as well as to the variational equilibrium problem

$$\widehat{\mathcal{E}}_\gamma(v,u) = \sum_{i=1}^{N} \mathcal{J}_i(v_i, u_{-i}) + \gamma\beta(g(v)) + i_{U_{\mathrm{ad}}}(v).$$

Using the definition of the state as well as the composition $g = \mathcal{G} \circ S$, this leads to the *penalized Nash game*

$$\begin{aligned} &\text{minimize} \quad J_i^1(u_i) + J_i^2(y) + \gamma\beta(\mathcal{G}(y)) \text{ over } u_i \in U_i,\, y \in Y \\ &\text{subject to } u_i \in U_{\mathrm{ad}}^i \text{ with } A(y) = b + Bu, \end{aligned} \tag{3.7}$$

and in a similar fashion to the *penalized variational equilibrium problem*

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^{N} \Big(J_i^1(y_i) + J_i^2(u_i') \Big) + \gamma\beta(\mathcal{G}(y)) \text{ over } u_i' \in U_i,\, y_i \in Y \text{ and } y \in Y \\ &\text{subject to } u_i' \in U_{\mathrm{ad}}^i \text{ and } A(y_i) = b + B(u_i', u_{-i}) \text{ as well as} \\ &\qquad A(y) = b + Bu'. \end{aligned} \tag{3.8}$$

The definition of the states y_i and y comes from the presence of the terms $S(u_i', u_{-i})$ in the state-related functionals J_i^1 and of the expression $S(u')$ occurring in $\beta \circ \mathcal{G}$ for the penalization of the constraint $u' \in \mathcal{F}$. Moreover, we assume in the terms of the abstract setting (1.2) that the functionals $u \mapsto \mathcal{J}_i(u_i, u_{-i})$ are continuous with respect to the strong topology on U_i and the weak one on U_{-i}, i.e., for all sequences $u_i^n \to u_i$ and $u_{-i}^n \rightharpoonup u_{-i}$ it holds that $\mathcal{J}_i(u_i^n, u_{-i}^n) \to \mathcal{J}_i(u_i, u_{-i})$. This condition can usually be guaranteed for a wide variety of applications as in the setting of (1.3) by complete continuity of the solution map S together with continuity of the mappings J_1^i on Y and J_2^i on U_i. With these conditions at hand, it is possible to derive the Γ-convergence of (3.6) and by proposing $\mathrm{dom}(C) = U_{\mathrm{ad}}$ also the Γ-convergence of (3.5).

Turning to the derivation of a first-order system for the penalized problems, we assume for convenience that J_i^1, J_i^2, $i = 1, \ldots, N$, are all continuously

differentiable. In *both equilibrium cases*, this leads to the following system:

$$\begin{aligned} 0 &= \partial_i J_i^2(u_i) + B_i^* p_i + \lambda_i && \text{in } U_i^*, \\ A(y) &= b + Bu && \text{in } W, \\ DA(y)^* p_i &= DJ_i^1(y) - D\mathcal{G}(y)^* \mu && \text{in } Y^*, \\ \lambda_i &\in N_{U_{\text{ad}}^i}(u_i) && \text{in } U_i^*, \\ \mu &= -\gamma D\beta(\mathcal{G}(y)) && \text{in } X^*. \end{aligned} \tag{3.9}$$

In fact, for a jointly convex game, the first-order system would not only be necessary, but also sufficient implying the equivalence of the two penalized equilibrium problems. Assuming for the moment that at least the functionals J_i^2 are strongly convex, we find the strong monotonicity of the first derivative $\partial_i J_i^2 : U_i \to U_i^*$ and hence the unique solvability of the VI

$$\text{Find } u_i \in U_i \ : \ u_i^* \in \partial J_i^2(u_i) + N_{U_{\text{ad}}}(u_i),$$

given an arbitrary $u_i^* \in U_i^*$. This problem admits a Lipschitz-continuous solution operator denoted by $P_i : U_i^* \to U_i$. In the simplest case of $J_i^2(u_i) = \frac{1}{2}\|u_i\|_{U_i}^2$ for a separable Hilbert space U_i, this map reads as a composition with the projection mapping on U_{ad}. Often, the system can be rewritten as a fixed-point problem

$$u = T(u)$$

with $T : U_{\text{ad}} \to U_{\text{ad}}$ defined by $T(u) = (T_1(u), \dots, T_N(u))$ and

$$T_i(u) = P_i(-B_i^* p_i) \text{ with } p_i = p_i(y) = DA(y)^{-*}\left(\partial_y J_i^1(y) + \gamma D\mathcal{G}(y)^* D\beta(\mathcal{G}(y))\right),$$

and $y = S(u) = A^{-1}(b + Bu)$. Since this is a fixed-point problem involving only a single-valued operator—in contrast to the formulation for Nash and variational equilibria—the existence question does not suffer from a lack of topological characterization of its values and can thus be treated with classical Schauder-type results, cf. [44, Theorem IV.7.18]. Using the described penalization technique, one is hence able to propose a generalized solution concept that is also suitable for a numerical treatment of the state constraint by motivating a *path-following technique*. The idea is to observe the solution(s) of the above first-order system for a range of penalty parameters $\gamma \in [\gamma_{\min}, +\infty)$ leading to the *path*

$$\begin{aligned} \mathcal{P} = \Big\{ &(\gamma, u^\gamma, y^\gamma, p^\gamma, \mu^\gamma, \lambda^\gamma) \in [\gamma_{\min}, +\infty) \times U \times Y \times (W^*)^N \times X \times U^* \\ &\text{such that } (u^\gamma, y^\gamma, p^\gamma, \mu^\gamma, \lambda^\gamma) \text{ solves (3.9)} \Big\}. \end{aligned}$$

From the numerical viewpoint, it is interesting to study the behavior of the solutions of (3.9) for $\gamma \to +\infty$. As a first step toward a path analysis, we study the boundedness of the path. This is next done in the fully abstract setting only.

Lemma 3.4 *Let the mappings $v \mapsto \partial_i \mathcal{J}_i(v_i, v_{-i})$ (in the fully abstract setting) be bounded for all $i = 1, \dots, N$ (i.e., images of bounded sets are bounded). If additionally the RZK condition (3.4) holds, then the path $\mathcal{P}$ is bounded.*

Using this result, it is straightforward to utilize reflexivity and the Banach–Alaoglu theorem to obtain the existence of weakly and weakly* converging subsequences. The next result guarantees that the corresponding limits are the desired solutions.

Theorem 3.5 *Let the condition (3.4) as well as the boundedness condition of Lemma 3.4 be fulfilled, and let moreover the following additional assumptions hold:*

(i) *The first derivatives of the objectives $\mathcal{J}_i$ with respect to the players' strategy satisfy for every weakly convergent sequence $u_i^n \rightharpoonup u_i^*$ in U the property*

$$\langle \partial_i \mathcal{J}_i(u_i^*, u_{-i}^*), u_i^* \rangle_{U_i, U_i^*} \leq \limsup_{n \to +\infty} \langle \partial_i \mathcal{J}_i(u_i^n, u_{-i}^n), u_i^n \rangle_{U_i, U_i^*}.$$

(ii) *The mapping $g : U \to X$ is strongly continuous and* uniformly Fréchet differentiable on every bounded set*, i.e., on very bounded subset $M \subseteq U$ holds that*

$$\lim_{\|h\|_X \to 0} \sup_{u \in M} \frac{\|g(u+h) - g(u) - Dg(u)h\|_X}{\|h\|_U} = 0.$$

Then, every path has a limiting point $(u^*, q^*, \lambda^*, \mu^*)$ along a subsequence, and every limiting point fulfills the necessary first-order condition for a Nash equilibrium (resp. variational equilibrium).

Together with the existence for solutions to the first-order system for the penalized system (3.9), the combined fulfillment of the conditions guarantees the existence of a point fulfilling the first-order system for (VEP) and hence especially for (GNEP).

This procedure sketches the numerical treatment of the (GNEP) problem (2.4). Besides identifying a suitable algorithm to solve the system (3.9), also an adaptive parameter update technique is needed; compare [21] for the latter. Here take a highly related approach leading to the definition of the *value functions*

$$\begin{aligned} \mathcal{W}_\gamma(u^\gamma) &= \inf_{u' \in U_{\mathrm{ad}}} \mathcal{E}_\gamma(u', u^\gamma) = \inf_{u' \in U_{\mathrm{ad}}} \sum_{i=1}^{N} \left(\mathcal{J}_i(u_i', u_{-i}^\gamma) + \gamma \beta(g(u_i', u_{-i}^\gamma)) \right) \\ &= \sum_{i=1}^{N} \inf_{u_i' \in U_{\mathrm{ad}}^i} \left(\mathcal{J}_i(u_i', u_{-i}^\gamma) + \gamma \beta(g(u_i', u_{-i}^\gamma)) \right) \end{aligned} \tag{3.10}$$

and analogously for the penalized (VEP)

$$\widehat{\mathcal{W}}_\gamma(u^\gamma) = \inf_{u' \in U_{\mathrm{ad}}} \left(\sum_{i=1}^{N} \mathcal{J}_i(u_i', u_{-i}^\gamma) + \gamma \beta(g(u')) \right). \tag{3.11}$$

One observes that $\mathcal{E}_\gamma(u^\gamma, u^\gamma) - \mathcal{W}_\gamma(u^\gamma) \geq 0$ and $\widehat{\mathcal{E}}_\gamma(u^\gamma, u^\gamma) - \widehat{\mathcal{W}}_\gamma(u^\gamma) \geq 0$, with equality only if u^γ is a solution of the penalized Nash game, respectively, (VEP). Using the defined value functionals, we seek to evaluate the effect of an increase of γ on the behavior of our solution. Therefore, we consider the functional $\tilde{\gamma} \mapsto \mathcal{W}_{\tilde{\gamma}}(u_\gamma)$ respectively $\tilde{\gamma} \mapsto \widehat{\mathcal{W}}_{\tilde{\gamma}}(u_\gamma)$. For a local description of the behavior, we extract first-order information by providing bounds for the upper and lower limits for the directional derivative of the proposed functionals.

Lemma 3.6 *Let J_i^1, J_i^2 be continuous functionals, and let the best response mapping with respect to the penalty parameter $\tilde{\gamma}$, i.e.,*

$$\tilde{\gamma} \mapsto \mathcal{B}^{\tilde{\gamma}}(u^\gamma) = \operatorname{argmin} \left\{ \sum_{i=1}^{N} \left(\mathcal{J}_i(u_i', u_{-i}^\gamma) + \tilde{\gamma} \beta(g(u_i', u_{-i}^\gamma)) \right) \text{ over } u' \in U_{\mathrm{ad}} \right\} \text{ and}$$

$$\tilde{\gamma} \mapsto \widehat{\mathcal{B}}^{\tilde{\gamma}}(u^\gamma) = \operatorname{argmin} \left\{ \sum_{i=1}^{N} \mathcal{J}_i(u_i', u_{-i}^\gamma) + \tilde{\gamma} \beta(g(u')) \text{ over } u' \in U_{\mathrm{ad}} \right\}$$

be nonempty-valued. Let $u^\gamma \in U_{\mathrm{ad}}$ be an equilibrium for the penalized (GNEP) *in* (3.7)*, respectively,* (VEP) *in* (3.8)*. Then, the difference quotients satisfy*

$$0 \leq \liminf_{\eta \searrow 0} \frac{\mathcal{W}(\gamma+\eta) - \mathcal{W}(\gamma)}{\eta} \leq \limsup_{\eta \searrow 0} \frac{\mathcal{W}(\gamma+\eta) - \mathcal{W}(\gamma)}{\eta} \leq N\beta(g(u^\gamma)) \text{ and}$$

$$0 \leq \liminf_{\eta \searrow 0} \frac{\widehat{\mathcal{W}}(\gamma+\eta) - \widehat{\mathcal{W}}(\gamma)}{\eta} \leq \limsup_{\eta \searrow 0} \frac{\widehat{\mathcal{W}}(\gamma+\eta) - \widehat{\mathcal{W}}(\gamma)}{\eta} \leq \beta(g(u^\gamma)).$$

If, moreover, the best response map $\tilde{\gamma} \mapsto \mathcal{B}^{\tilde{\gamma}}(u^\gamma)$, respectively, $\tilde{\gamma} \mapsto \widehat{\mathcal{B}}^{\tilde{\gamma}}(u^\gamma)$, is single-valued and continuous, then the functional $\mathcal{W}$, respectively $\widehat{\mathcal{W}}$, is even differentiable with $\mathcal{W}'(\gamma) = N\beta(g(u^\gamma))$, respectively $\widehat{\mathcal{W}}'(\gamma) = \beta(g(u^\gamma))$.

Hence, the composition of the penalty and the state constraint serves as a way to adjust the penalty parameter for each step of the path-following procedure by

$$\gamma \mapsto \gamma + \max\left(\frac{\pi_{\mathrm{path}}}{\beta(g(u^\gamma))}, \varepsilon \right)$$

with a fixed parameter $\pi_{\mathrm{path}} > 0$. Using this technique, strong violations of the state constraint resulting in a big penalty term induce a more timid update, whereas low values cause a more aggressive behavior. The update is safeguarded with a fixed

upper bound $\varepsilon > 0$ for the case of very low values of the penalty functional. If the value is zero, then the algorithm terminates since it has found a solution of the original (GNEP), respectively, (VEP). The results of Sect. 3 together with the corresponding proofs and details will be made available in [18].

With this outline of an algorithm, we end the discussion of deterministic Nash equilibria and turn our attention to the case involving uncertainties.

4 PDE-Constrained GNEPs Under Uncertainty

4.1 Motivation

Most real-world problems in the natural sciences, engineering, economics, and finance are subject to uncertainty. This inherent stochasticity arises from a number of unavoidable factors, which range from noisy measurements and data acquisition to ambiguity in the choice of model and its underlying exogenous parameters. Consequently, we must incorporate random parameter into our mathematical models. Within the framework of PDE-constrained decision problems, we are then confronted with the task of optimizing systems of random partial differential equations.

In order to ensure these new infinite-dimensional stochastic decision problems yield robust solutions to outliers or potentially catastrophic events, we appeal to the theory of risk-averse optimization, which has been widely developed over the last several decades within the (finite dimensional) stochastic programming community, see, e.g., [41] and many references therein. Furthermore, using risk models in the context of Nash equilibrium problems allows us to model the preferences of the agents more accurately by assuming they have well-defined risk preferences.

Nevertheless, the literature on risk-averse PDE-constrained optimization was extremely scarce until recently [13, 24–28]. Therefore, in order to tackle risk-averse PDE-constrained GNEPs, it has been necessary to first develop the theory, approximation, and algorithms for the optimization setting. These results can now be leveraged for the NEP and ultimately GNEP setting.

In what follows, we will first present the recent theory of risk-averse PDE-constrained optimization in which the risk preferences of the individual agents are modeled by convex risk measures. Following this, we will apply the theory to a model risk-averse PDE-constrained Nash equilibrium problem. This will more clearly delineate the differences between the optimization and game-theoretic frameworks. We then present the recent approach in [25] for smoothing nonsmooth risk measures that is interesting from a theoretical perspective, but also useful for gradient-based optimization algorithms. In particular, we will see that epi-regularization of risk measures is an essential component of the primal–dual risk minimization algorithm recently developed in [27].

4.2 Additional Notation and Preliminary Results

In addition to the notation introduced above, we recall several further concepts necessary for the coming discussions. Unless otherwise stated, these are considered standing assumptions in the text below.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space where Ω is an arbitrary set of outcomes, $\mathcal{F} \subseteq 2^{\Omega}$ is the associated σ-algebra of events, and the set function $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a probability measure. We employ the standard abbreviations "a.e." and "a.a." for "almost everywhere" and "almost all" with respect to $\mathbb{P}$, respectively. If necessary, we will append these by $\mathbb{P}$ and write $\mathbb{P}$-a.e. or $\mathbb{P}$-a.a. As $\mathcal{F}$ is fixed, we write "$\mathcal{F}$-measurable" simply as "measurable" if clear in context. Since we will often deal with Banach-space-valued random terms, we recall that a random element X in a Banach space $\mathcal{X}$ is a measurable mapping $X : \Omega \to \mathcal{X}$, where $\mathcal{X}$ is endowed with the Borel σ-algebra. We denote expectation by $\mathbb{E}[X]$.

We assume that the control space U is a real reflexive Banach space and denote the set of admissible decisions by $U_{\text{ad}} \subset U$. The latter is assumed to be a nonempty, closed, and convex set. In the context of Nash equilibrium problems, U_{ad} is assumed to be bounded as well. The physical domain for the deterministic PDE solutions will be denoted by $D \subset \mathbb{R}^d$. We assume that D is an open and bounded set with Lipschitz boundary ∂D. The associated state space for the deterministic solutions will be denoted by $V := H^1(D)$ (or $H_0^1(D)$), where $H^1(D)$ is the usual Sobolev space of $L^2(D)$-functions with weak derivatives in $L^2(D)$ [1].

The natural function-space setting for solutions of random PDEs is in classical Bochner spaces, cf. [17]. We recall that the Bochner space $L^p(\Omega, \mathcal{F}, \mathbb{P}; W)$ comprises all measurable functions that map Ω into some Banach space W with p finite moments for $p = [1, \infty)$. When $p = \infty$, $L^\infty(\Omega, \mathcal{F}, \mathbb{P}; W)$ is the space of all essentially bounded W-valued measurable functions. The norms are given by

$$\|v\|_{L^p(\Omega,\mathcal{F},\mathbb{P};W)} = \mathbb{E}\left[\|v\|_W^p\right]^{1/p} \quad \text{for } p \in [1, \infty)$$
$$\|v\|_{L^\infty(\Omega,\mathcal{F},\mathbb{P};W)} = \operatorname*{ess\,sup}_{\omega\in\Omega} \|v(\omega)\|_W.$$

When $W = \mathbb{R}$, we set $L^p(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}) = L^p(\Omega, \mathcal{F}, \mathbb{P})$. In our optimization and equilibrium settings, the random objective maps U into $\mathcal{X} := L^p(\Omega, \mathcal{F}, \mathbb{P})$ for some $p \in [1, \infty)$. Whenever it is clear, we simply write $\mathcal{X}$.

As discussed in Sect. 4.1, we model risk-averse behavior by means of risk measures. There is a vast literature on the subject of risk measures and their usage in optimization. In our models, the individual agents' problems are assumed to take the form:

$$\min_{u\in U_{\text{ad}}} \mathcal{R}[\mathcal{J}(S(u))] + \wp(u),$$

where $\mathcal{R}$ is a nonlinear, typically nonsmooth, functional on $\mathcal{X}$. We refer the interested reader to [41, Chap. 6.] and the references therein as a starting point. For our purposes, it will suffice to introduce two general classes of risk measures here, each of which follows the standard axiomatic approach as in [3, 12, 39]. We start by recalling the definition of a regular measure of risk as suggested by Rockafellar and Uryasev in [39]. The conditions below were postulated as minimal regularity properties for risk measures in the context of optimization. A functional $\mathcal{R} : \mathcal{X} \to \overline{\mathbb{R}}$ where $\overline{\mathbb{R}} := (-\infty, \infty]$ is a regular measure of risk provided is proper, closed, convex and satisfies $\mathcal{R}[C] = C$ for all constant random variables $C \in \mathbb{R}$, and $\mathcal{R}$ is risk averse: $\mathcal{R}[X] > \mathbb{E}[X]$ for all nonconstant $X \in \mathcal{X}$. Therefore, the expected value is not a regular measure of risk in this setting. This is reasonable from the perspective that setting $\mathcal{R} = \mathbb{E}$ would indicate neutrality to risk and not yield a robust solution.

Perhaps the most well-known risk measures are the coherent risk measures. These were introduced in a systematic way in [3] as a means of axiomatizing the behavior of risk-averse decision makers. The risk measure $\mathcal{R}$ is coherent provided:

(C1) *Subadditivity*: If $X, X' \in \mathcal{X}$, then $\mathcal{R}[X + X'] \leq \mathcal{R}[X] + \mathcal{R}[X']$.
(C2) *Monotonicity*: If $X, X' \in \mathcal{X}$ and $X \geq X'$ almost surely, then $\mathcal{R}[X] \geq \mathcal{R}[X']$.
(C3) *Translation equivariance*: If $C \in \mathbb{R}$ and $X \in \mathcal{X}$, then $\mathcal{R}[X + C] = \mathcal{R}[X] + C$.
(C4) *Positive homogeneity*: If $C \in [0, \infty)$ and $X \in \mathcal{X}$, then $\mathcal{R}[CX] = C\mathcal{R}[X]$.

A rather popular coherent risk measure is the conditional or average value at risk (CVaR or AVaR). Given a risk or confidence level $\beta \in (0, 1)$, the average value at risk of a random variable X is the average of the associated quantiles $F_\alpha^{-1}(X)$ over $\alpha \in (\beta, 1)$. Here, we have

$$F_\alpha^{-1}(X) = \mathrm{VaR}_\beta(X) := \inf\{x \in \mathbb{R} \,:\, F_X(x) \geq \beta\},$$

i.e., the value at risk of X at confidence level β, and

$$\mathrm{AVaR}_\beta(X) := \frac{1}{1-\beta} \int_\beta^1 \mathrm{VaR}_\alpha(X)\, \mathrm{d}\alpha.$$

This gives a measure of the tail of the distribution of X. It is particularly well suited in the context of risk-averse optimization as a means of accounting for tail events. CVaR can be written in several ways; for optimization, we use

$$\mathrm{AVaR}_\beta(X) = \inf_{t \in \mathbb{R}} \left\{ t + \frac{1}{1-\beta} \mathbb{E}[(X - t)_+] \right\}, \tag{4.1}$$

where $(x)_+ := \max\{0, x\}$ [38]; the (smallest) minimizer in (4.1) is $\mathrm{VaR}_\beta(X)$.

As shown in [25, Thm 1], the only coherent risk measures that are continuously Fréchet differentiable are expectations. Therefore, regardless of how smooth the objective or control state mappings are, any risk-averse PDE-constrained opti-

mization problem using coherent regular risk measures is an infinite-dimensional nonsmooth optimization problem.

4.3 Risk-Averse PDE-Constrained Optimization: Theory

We now focus on developing the theory for the "single-player" setting. We start by considering the following abstract optimization problem:

$$\min_{u \in U_{\text{ad}}} \mathcal{R}[\mathcal{J}(S(u))] + \wp(u). \tag{4.2}$$

Here, $u \in U$ represents the decision variable (controls, parameters, designs, etc.), U_{ad} is the associated feasible set, $\wp$ is a deterministic cost function, $\mathcal{R}$ is a risk measure as in Sect. 4.2, $\mathcal{J}$ is a random objective in the form of a general superposition operator, and $S(u)$ is the solution mapping for the random PDE.

As motivation for the chosen setting, we recall the class of random PDEs considered in [28] (in strong form): For $u \in U$ and $\mathbb{P}-$a.e. $\omega \in \Omega$, $y = S(u)$ solves

$$\begin{aligned} -\nabla \cdot (\kappa(\omega)\nabla y(\omega)) + c(\omega)y(\omega) + N(y(\omega), \omega) &= [B(\omega)u] + b(\omega), && \text{in } D \\ \kappa(\omega)\frac{\partial y}{\partial n}(\omega) &= 0, && \text{on } \partial D. \end{aligned} \tag{4.3}$$

Here, we assume κ, c, b are random elements in an appropriate Bochner space and the operator N is a potentially nonlinear maximal monotone operator. $B(\omega)$ maps u into the image space of the differential operator.

Returning to the abstract setting, it was shown in [26] that a number of basic regularity assumptions need to be imposed on $\mathcal{R}$, $\mathcal{J}$, S, $\wp$, and U_{ad} in order to prove the existence of a solution and derive optimality conditions for (4.2). The inclusion of stochasticity and the nonlinearity and nonsmoothness of $\mathcal{R}$ add a further level of complexity not seen in deterministic problems. We impose the following conditions on S and $\mathcal{J}$ throughout.

Assumption 4.1 (Properties of the Solution Map) It holds that

1. $S(u) : \Omega \to V$ is strongly $\mathcal{F}$-measurable for all $u \in U_{\text{ad}}$.
2. There exists an increasing function $\rho : [0, \infty) \to [0, \infty)$ and $C \in L^q(\Omega, \mathcal{F}, \mathbb{P})$ with $C \geq 0$, $q \in [1, \infty]$ such that

$$\|S(u)\|_V \leq C\rho(\|u\|_U) \quad \mathbb{P}-\text{a.e.} \quad \forall u \in U_{\text{ad}}.$$

3. If $u_n \rightharpoonup u$ in U_{ad}, then $S(u_n) \rightharpoonup S(u)$ in V $\mathbb{P}-$a.e.

Each of these assumptions is minimal. For example, if $S(u)$ is not measurable, then $\mathcal{R} \circ \mathcal{J} \circ S$ is meaningless. The second assumption can be seen as an integrability requirement. Since $\mathcal{J}$ is typically a nonlinear operator, it is essential for S to possess such properties. The latter condition appears to be the weakest condition needed (along with the assumption on $\mathcal{R}$, $\mathcal{J}$, etc. below) to prove the existence of a solution. As shown in [24, Sec. 2.2], Assumption 4.1 implies:

1. $S(u) \in L^q(\Omega, \mathcal{F}, \mathbb{P}; V)$ for all $u \in U_{\text{ad}}$.
2. By letting

$$\mathcal{V} := L^q(\Omega, \mathcal{F}, \mathbb{P}; V),$$

we have $S(u_n) \rightharpoonup S(u)$ in $\mathcal{V}$ for any $\{u_n\} \subset U_{\text{ad}}$ such that $u_n \rightharpoonup u$.

Furthermore, in order to derive optimality conditions, S needs to be continuously differentiable.

Assumption 4.2 There exists an open set $W \subseteq U$ with $U_{\text{ad}} \subseteq W$ such that the solution map $u \mapsto S(u) : W \to \mathcal{V}$ is continuously Fréchet differentiable.

The results in [24] indicate that we could slightly weaken this to Hadamard directional differentiability, which would allow us to consider risk-averse control of random elliptic variational inequalities in the future.

Continuing, we will assume that the random objective $\mathcal{J}$ is the result of a superposition of some possibly random integral functional J and an element $y \in \mathcal{V}$. The necessary, and in part sufficient, conditions needed for J are given below.

Assumption 4.3 (Properties of $J : V \times \Omega \to \mathbb{R}$) It holds that

1. J is a Carathéodory function, i.e., $J(\cdot, \omega)$ is continuous for $\mathbb{P}-$a.e. $\omega \in \Omega$ and $J(u, \cdot)$ is measurable for all $v \in V$.
2. If $1 \leq p, q < \infty$, then there exists $a \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ with $a \geq 0$ $\mathbb{P}-$a.e. and $c > 0$ such that

$$|J(v, \omega)| \leq a(\omega) + c\|v\|_U^{q/p}. \tag{4.4}$$

If $1 \leq p < \infty$ and $q = \infty$, then the uniform boundedness condition holds: for all $c > 0$, there exists $\gamma = \gamma(c) \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ such that

$$|J(v, \omega)| \leq \gamma(\omega) \;\; \mathbb{P}-\text{a.e.} \quad \forall\, v \in V, \;\; \|v\|_V \leq c. \tag{4.5}$$

3. $J(\cdot, \omega)$ is convex for $\mathbb{P}-$a.e. $\omega \in \Omega$.

It follows from a well-known result due to Krasnosel'skii, see, e.g., [29], [43, Thm 19.1], see also Theorem 4 in [15], that Assumption 4.3.1–2 guarantees $\mathcal{J} : \mathcal{V} \to L^p(\Omega, \mathcal{F}, \mathbb{P})$ continuously. These are necessary and sufficient and cannot be weakened. For several examples of objectives that satisfy Assumption 4.3, we refer to [26, Sec. 3.1]. Finally, the convexity assumption guarantees Gâteaux directional

differentiability. If this is not available, then additional assumptions must be made on the partial derivatives of J with respect to u. We gather the related main statements on $\mathcal{J}$ from [26] here for the reader's convenience.

Theorem 4.4 (Continuity and Gâteaux Differentiability of $\mathcal{J}$) *Let Assumption 4.3.1–2 hold. Then $\mathcal{J} : \mathcal{V} \to L^p(\Omega, \mathcal{F}, \mathbb{P})$ is continuous. Furthermore, if Assumption 4.3.1–3 holds, then $\mathcal{J}$ is Gâteaux directionally differentiable.*

Since the objective functional in (4.2) is of the form $\mathcal{R} \circ \mathcal{J} \circ S$, Theorem 4.4 is not strong enough to guarantee the necessary smoothness properties of $\mathcal{J}$ as a nonlinear operator from $\mathcal{V}$ into $L^p(\Omega, \mathcal{F}, \mathbb{P})$ that would provide us with first-order optimality conditions. This requires further regularity conditions. The weakest type of directional differentiability that allows a chain rule is Hadamard directional differentiability, cf. [40]. In the current setting, this can be demonstrated if $\mathcal{J}$ is locally Lipschitz, see [26, Cor. 3.10]. For the development of function-space-based optimization algorithms, in particular the convergence analysis, we generally need continuous Fréchet differentiability. This can be proven provided the partial derivatives of $\partial_u J(\cdot, \omega)$ satisfy a Hölder continuity condition, see [26, Thm. 3.11].

We now have a sufficient amount of structure to prove existence of optimal solutions to (4.2). The following lemma is essential.

Lemma 4.5 (Weak Lower-Semicontinuity of the Composite Objective) *Let Assumptions 4.1 and 4.3 hold. If $\mathcal{R} : L^1(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ is proper, closed, monotonic, convex, and subdifferentiable at $\mathcal{J}(S(u))$ for some $u \in U_{\mathrm{ad}}$, then the composite functional $(\mathcal{R} \circ \mathcal{J} \circ S) : U_{\mathrm{ad}} \to \mathbb{R}$ is weakly lower semicontinuous at $u \in U_{\mathrm{ad}}$.*

Using Lemma 4.5, we can now prove existence of solutions.

Theorem 4.6 (Existence of Optimal Solutions) *Let Assumptions 4.1, 4.2, and 4.3 hold. Let $\mathcal{R} : L^1(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ be a proper, closed, convex, and monotonic risk measure, and let $\wp : U \to \overline{\mathbb{R}}$ be proper, closed, and convex. Finally, suppose either U_{ad} is bounded or $u \mapsto \mathcal{R}(\mathcal{J}(S(u))) + \wp(u)$ is coercive. Then, (4.2) has a solution.*

Next, we can also derive a general first-order optimality condition. The essential point here is the regularity condition on $\mathcal{R}$, which guarantees the composite reduced objective function $\mathcal{R} \circ \mathcal{J} \circ S$ is Hadamard directionally differentiable. The standard regularity assumptions: finiteness or $\operatorname{int} \operatorname{dom} \mathcal{R} \neq \emptyset$ are considerably mild given the types of risk measures used in practice.

Theorem 4.7 (A General Optimality Condition) *Suppose that in addition to the assumptions of Theorem 4.6, the risk measure $\mathcal{R}$ is either finite on $L^1(\Omega, \mathcal{F}, \mathbb{P})$ or $\operatorname{int} \operatorname{dom} \mathcal{R} \neq \emptyset$. Moreover, assume that $\mathcal{J} : \mathcal{V} \to L^p(\Omega, \mathcal{F}, \mathbb{P})$ is locally Lipschitz and $\wp$ is Gâteaux directionally differentiable. Then for any optimal solution $u^\star$*

to (4.2)*, the following first-order optimality condition holds:*

$$\sup_{\vartheta \in \partial \mathcal{R}(\mathcal{J}(S(u^\star)))} \mathbb{E}[\mathcal{J}'(S(u^\star); S(u^\star)'\delta u)\,\vartheta] + \wp'(u^\star; \delta u) \geq 0, \ \forall \delta u \in T_{U_{\text{ad}}}(u^\star), \tag{4.6}$$

where $T_{U_{\text{ad}}}(u^\star)$ *is the contingent cone to* U_{ad} *at* $u^\star$*, which is defined by*

$$T_{U_{\text{ad}}}(u^\star) := \left\{ d \in U \,\middle|\, \exists\, \tau_k \downarrow 0,\ \exists\, d_k \to d \text{ in } U \ : z^\star + \tau_k d_k \in U_{\text{ad}}\ \forall k \right\}.$$

For illustration of (4.6), let $p = 2$, $U = L^2(D)$, $S(u^\star) = \mathbf{A}^{-1}(\mathbf{B}u^\star + \mathbf{b})$ and

$$J(y, \omega) = J(y) := \frac{1}{2}\|y - y_d\|^2_{L^2(D)} \text{ and } \wp = \frac{\nu}{2}\|u\|^2_{L^2(D)},$$

where $\mathbf{A}^{-1}$ is a linear isomorphism from $\mathcal{V}^*$ into $\mathcal{V}$, $\mathbf{B} \in \mathcal{L}(U, \mathcal{V}^*)$, and $\mathbf{b} \in \mathcal{V}^*$; then (4.6) unfolds into a somewhat more familiar form: If $u^\star$ is an optimal solution of (4.2), then there exists an adjoint state $p^\star \in \mathcal{V}^*$ and a subgradient $\vartheta^\star \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ such that

$$\begin{gathered}
\left(u^\star - \frac{1}{\nu}\mathbb{E}[\mathbf{B}^* p^\star \vartheta^\star], u - u^\star \right)_U \geq 0, \ \forall u \in U_{\text{ad}}, \\
\mathcal{R}[X] - \mathcal{R}[\mathcal{J}(y^\star)] - \mathbb{E}[\vartheta^\star(X - \mathcal{J}(y^\star))] \geq 0, \ \forall X \in L^1(\Omega, \mathcal{F}, \mathbb{P}), \\
\mathbf{A}y^\star - \mathbf{B}u^\star + \mathbf{b} = 0, \\
\mathbf{A}^* p^\star - y_d + y^\star = 0.
\end{gathered} \tag{4.7}$$

This provides us with the interesting fact that the optimal control is the projection onto U_{ad} of the expectation of adjoint term $\mathbf{B}^* p^\star$, where the expectation has been adjusted according to the risk preference expressed in $\mathcal{R}$ via the subgradient $\vartheta^\star$. The latter is often referred to as the "risk indicator" in the literature for obvious reasons. In the case of AVaR_β, the numerical experiments in [24] indicate that $\mathbb{P}(\text{supp}\,\vartheta^\star) = 1 - \beta$. Therefore, the majority of support is used to treat tail events. Note also that when designing first-order methods for such problems, this fact allows a significant reduction in the number of PDEs solved per iteration required to calculate the reduced gradient.

For a more challenging example, we recall the setting from [28] in (4.3) in more detail. Among the most difficult aspects of the assumptions used to prove existence of a solution and derive optimality conditions are the conditions placed on the solution mapping S. In [28], we postulate several verifiable assumptions. To this aim, we suppose that $S(u)$ is the solution of a general parametric operator equation: For each $u \in U$, find $y(\omega) = [S(u)](\omega) \in U$ such that

$$e(y, u; \omega) := \mathbf{A}(\omega)y + \mathbf{N}(y, \omega) - \mathbf{B}(\omega)u - \mathbf{b}(\omega) \ni 0 \quad \text{for a.a. } \omega \in \Omega. \tag{4.8}$$

We impose the following assumptions on the operators.

Assumption 4.8 (Pointwise Characterization of the Problem Data in (4.8))

1. Let $\mathbf{A} : \Omega \to \mathcal{L}(V, V^*)$ satisfy $\mathbf{A}(\omega)$ is monotone for a.a. $\omega \in \Omega$ and there exists $\gamma > 0$ and a random variable $C : \Omega \to [0, \infty)$ with $C > 0$ a.e. such that

$$\langle \mathbf{A}(\cdot)y, y\rangle_{U^*,U} \geq C\|y\|_V^{1+\gamma} \quad \text{a.e. } \forall\, y \in V. \tag{4.9}$$

2. Let $\mathbf{b} : \Omega \to V^*$.
3. Let $\mathbf{N} : V \times \Omega \rightrightarrows V^*$ satisfy $\mathbf{N}(\cdot, \omega)$ is maximal monotone with $\mathbf{N}(0, \omega) = \{0\}$ for a.a. $\omega \in \Omega$.
4. Let $\mathbf{B} : \Omega \to \mathcal{L}(U, V^*)$ be completely continuous for a.a. $\omega \in \Omega$.

Since these conditions are taken to be pointwise in ω, they can be viewed as the minimal data assumptions that are imposed when considering optimization of elliptic semilinear equations. The following assumption is essential for measurability issues. It is unclear if it can be weakened. Ultimately, the coefficients and mappings used to define $\mathbf{A}$, $\mathbf{N}$, etc. will dictate the integrability of $S(u)$.

Assumption 4.9 (Measurability and Integrability of the Operators in (4.3)) Let Assumption 4.8 hold and suppose there exists $s,\ t \in [1, \infty]$ with

$$1 + \frac{1}{\gamma} \leq s < \infty \quad \text{and} \quad t \geq \frac{s}{\gamma(s-1) - 1}$$

such that $\mathbf{A}(\cdot)y \in L^s(\Omega, \mathcal{F}, \mathbb{P}; V^*)$ for all $y \in V$, $\mathbf{N}(\cdot, \omega)$ is single-valued and continuous for a.a. $\omega \in \Omega$ and $\mathbf{N}(y, \cdot) \in L^s(\Omega, \mathcal{F}, \mathbb{P}; V^*)$ for all $y \in V$, $\mathbf{B} \in L^s(\Omega, \mathcal{F}, \mathbb{P}; \mathcal{L}(U, V^*))$, $\mathbf{b} \in L^s(\Omega, \mathcal{F}, \mathbb{P}; V^*)$ and $C^{-1} \in L^t(\Omega, \mathcal{F}, \mathbb{P})$.

Finally, we require assumptions on $\mathbf{N}$ to derive optimality conditions.

Assumption 4.10 (Differentiability of $\mathbf{N}(\cdot, \omega)$) In addition to Assumption 4.9, we assume that $\mathbf{N}(\cdot, \omega)$ is single-valued and continuously Fréchet differentiable from V into V^* for a.a. $\omega \in \Omega$ with partial derivative $\mathbf{N}'(y, \omega)$, which defines a bounded, nonnegative linear operator from V into V^* a.e. for all $y \in V$. Moreover, we assume that $\mathbf{A}$ and $y \mapsto \mathbf{N}(y, \cdot)$ are continuous maps from $\mathcal{V}$ into $L^s(\Omega, \mathcal{F}, \mathbb{P}; V^*)$ and $y \mapsto \mathbf{N}'(y, \cdot)$ is a continuous map from $\mathcal{V}$ into $L^{qs/(q-s)}(\Omega, \mathcal{F}, \mathbb{P}; \mathcal{L}(V, V^*))$.

We gather the main results in [28, Sec. 2.3] here for the reader's convenience.

Theorem 4.11 (Properties of the Solution Mapping $S(u)$) *Under the standing assumptions, the following statements hold.*

1. *If Assumption 4.8 holds, then $\mathbf{A}(\omega) + \mathbf{N}(\cdot, \omega)$ is surjective from V into V^* for a.a. $\omega \in \Omega$. In particular, there exists a unique solution $S(u)$ to (4.8) such that $[S(u)](\omega) \in V$ for a.a. $\omega \in \Omega$.*

2. *If in addition Assumption 4.9 holds and we let*

$$q := \frac{s\gamma}{1 + s/t}, \tag{4.10}$$

then $S(u) \in \mathcal{V} := L^q(\Omega, \mathcal{F}, \mathbb{P}; V)$ for all $u \in V$. Furthermore, if $u_k \rightharpoonup u$ in U, then $S(u_k) \to S(u)$ in V a.e. and $S(u_k) \to S(u)$ in $\mathcal{V}$, i.e., S is completely continuous.
3. *If in addition Assumption 4.10 holds, then $u \mapsto S(u)$ is continuously Fréchet differentiable from U into $\mathcal{V}$.*

We now return to a concrete example and cast (4.3) in the form (4.8).

Example Define the linear elliptic operator $\mathbf{A}(\omega)$ by

$$\langle \mathbf{A}(\omega)y, v\rangle_{V^*,V} = \int_D \{\kappa(\omega, x)\nabla y(x) \cdot \nabla v(x) + c(\omega, x)y(x)v(x)\}\,\mathrm{d}x,$$

for $y, v \in \mathcal{V}$. Analogously, we let $\mathbf{N}(\cdot, \omega)$ be the nonlinear operator given by

$$\langle \mathbf{N}(y, \omega), v\rangle_{V^*,V} = \int_D N(y(x), \omega, x)v(x)\,\mathrm{d}x,$$

where $N : \mathbb{R} \times \Omega \times D \to \mathbb{R}$. The right-hand side can be defined by

$$\langle \mathbf{B}(\omega)u, v\rangle_{V^*,V} = \int_D [B(\omega)u](x)v(x)\,\mathrm{d}x \quad \text{and} \quad \langle \mathbf{b}(\omega), v\rangle_{V^*,V} = \int_D b(\omega, x)v(x)\,\mathrm{d}x,$$

where $B : \Omega \to \mathcal{L}(U, L^2(D))$ and $b \in \mathcal{V}^*$.

Assuming that $\kappa(\omega, \cdot)$, $c(\omega, \cdot) \in L^\infty(D)$ for a.a. $\omega \in \Omega$ and for a.a. $\omega \in \Omega$, $x \in D$, satisfy: there exist $\kappa_0 > 0$ and $c_0 > 0$ such that

$$\kappa_0 \le \kappa(\omega, x) \text{ and } c_0 \le c(\omega, x),$$

then the conditions in Assumptions 4.8 and 4.9 on $\mathbf{A}$ are satisfied with $\gamma = 1$, $C = \min\{\kappa_0, c_0\}$, $s = 2$, $t = \infty$. For $\mathbf{N}$, we at least need $N(\cdot, \omega, x) : \mathbb{R} \to \mathbb{R}$ to be continuous and monotonically increasing with $N(0, \omega, x) = 0$ for a.a. $\omega \in \Omega$ and a.a. $x \in D$. This would yield the monotonicity requirement in Assumption 4.8, which would be the case for a nonlinearity of the type: $N(u, \omega, x) = c(\omega, x)(\sinh(u) - u)$. Otherwise, we can obtain continuity via the usual growth conditions of Krasnosel'skii as in, e.g., Theorems 1 and 4 in [15] or the comprehensive monograph [2]. Similarly, if we have $b(\omega, \cdot) \in L^r(D)$ with $r > d/2$ for a.a. $\omega \in \Omega$, then Assumption 4.8.2 holds and if B is, e.g., the canonical embedding operator from $L^2(D)$ into $H^1(D)^*$, then Assumption 4.8.3 also holds. For Assumption 4.9, we could require $b \in L^\infty(\Omega, \mathcal{F}, \mathbb{P}; L^2(D))$ and κ, $c \in L^\infty(\Omega, \mathcal{F}, \mathbb{P}; L^\infty(D))$. This assumption would not hold for $\mathbf{N}$ when generated by the hyperbolic sine unless V was replaced by a more regular space,

e.g., $H^2(D)$. However, if $d = 2$ and ∂D is sufficiently regular, then by the Sobolev embedding theorems we could still use $V = H^1(D)$ when N is generated by monotone polynomials of arbitrary degree.

Behind all of these technical details lie the hypotheses imposed by measurable selection theorems, e.g., Filippov's theorem, which generally require the random elements to map into separable spaces. The integrability conditions are then derived using the monotonicity of the operators. Therefore, one should be rather careful when generating new examples from deterministic PDE models as they may not always be well defined in the stochastic setting.

Finally, we conclude this section by noting that many example problems used in the literature consider linear elliptic PDE under uncertainty. This drastically simplifies the measurability, integrability, continuity, and differentiability issues for the solution mapping S. Building on the properties of the solution operator and requirements on the objective functionals J discussed above, one can derive similar measurability, integrability, and (weak) continuity results for the adjoint equations and ultimately an optimality system as in the linear case shown above.

4.4 A Risk-Averse PDE-Constrained Nash Equilibrium Problem

We may now formulate a model risk-averse PDE-constrained Nash equilibrium problem. Using the results of the previous section, we prove existence of a Nash equilibrium and derive optimality conditions. In what follows, we consider the following setting: For each $i = 1, \ldots, N$ ($N > 1$), we assume:

1. $U^i := L^2(D)$, $U^i_{\text{ad}} := \left\{ v \in U^i \mid a_i \leq v \leq b_i \text{ a.e. } D \right\}$, $a_i, b_i \in L^2(D) : a_i < b_i$.
2. $J_i(y, \omega) := \frac{1}{2}\|y - y^i_d\|^2_{L^2(D)}$, $y^i_d \in L^2(D)$; $\wp(u) := \frac{\nu_i}{2}\|u\|^2_{L^2(D)}$ $\nu_i > 0$.
3. $S : U^1 \times \cdots \times U^N \to \mathcal{V}$ is the solution mapping for the random PDE given by (4.8) under Assumptions 4.8 and 4.9 such that $\mathbf{A}$ is defined as in Example 4.3, i.e., uniformly elliptic with $\gamma = 1$, $C = \min\{\kappa_0, c_0\}$, $s = 2$, $t = \infty$; $\mathbf{N} \equiv 0$; $\mathbf{b} \in \mathcal{V}^*$; and $\mathbf{B} : U^1 \times \cdots \times U^N \to \mathcal{V}^*$ satisfies

$$\mathbf{B}u = \mathbf{B}_1 u_1 + \ldots \mathbf{B}_N u_N,$$

where $\mathbf{B}_i$, $i = 1, \ldots, N$, is defined as in Assumptions 4.8 and 4.9. In particular,

$$S(u) := \mathbf{A}^{-1}\left(\sum_i \mathbf{B}_i u_i + \mathbf{b}\right).$$

4. $\mathcal{R}_i : L^1(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ is a regular coherent measure of risk, e.g., AVaR_β.

Under these assumptions, we consider the associated risk-averse PDE-constrained Nash equilibrium problem (NEP) in which the i^{th} player's problem takes the form

$$\min_{u_i \in U^i_{\text{ad}}} \mathcal{R}_i(\mathcal{J}_i(S(u_i, u_{-i}))) + \wp(u_i) \text{ over } u_i \in U^i. \tag{4.11}$$

Using the Kakutani–Fan–Glicksberg fixed-point theorem (Theorem (2.1) above, [14]), we can demonstrate that this problem admits a Nash equilibrium.

Theorem 4.12 (Existence of a Risk-Averse Nash Equilibrium) *The Nash equilibrium problem whose individual players each solve a variant of* (4.11) *admits a solution in the form of a pure strategy Nash equilibrium.*

Proof We need to verify the conditions of Theorem 2.1. Since each U^i is infinite-dimensional, we view each U^i_{ad} as metrizable compact locally convex topological vector spaces as in [20, 21]. This is possible since U^i_{ad} is a norm bounded, closed, and convex set in a separable Hilbert space. Next, we define the best response mappings:

$$\mathcal{B}_i(u_{-i}) := \arg\min_{u_i \in U^i_{\text{ad}}} \mathcal{R}_i(\mathcal{J}_i(S(u_i, u_{-i}))) + \wp(u_i) \text{ over } u_i \in U^i.$$

We need to show that each $\mathcal{B}_i$ has nonempty, bounded, and convex images in U^i.

The risk measure $\mathcal{R}_i$ is proper, closed, convex, and monotonic. Since $\mathcal{R}_i$ is defined on all of $L^1(\Omega, \mathcal{F}, \mathbb{P})$, it is finite everywhere and therefore continuous and consequently subdifferentiable; in particular at $\mathcal{J}_i(S(u_i, u_{-i}))$ for any feasible strategy vector (u_i, u_{-i}). The tracking-type functional considered here can easily be shown to satisfy all the necessary assumptions outlined above; see [24] or [26]. Concerning S, we note that for any fixed $u_{-i} \in U^{-i}_{\text{ad}}$, we have $\mathbf{B}(0, u_{-i}) = \sum_{j \neq i} \mathbf{B}_j u_j$. The latter term can be taken on the right-hand side of the PDE as a perturbation of $\mathbf{b}$. Clearly, this "new" constant term is in $\mathcal{V}^*$. It light of this, we can readily verify the necessary assumptions for continuity and differentiability with respect to u_i required in Theorem 4.11.

It follows that $\mathcal{R} \circ \mathcal{J} \circ S : U^i \to \mathbb{R}$ is weakly lower semicontinuous (cf. Lemma 4.5). The existence of solutions results from the fact that $\wp$ is coercive and $\mathcal{R}_i \circ \mathcal{J}_i \circ S$ nonnegative (cf. Theorem 4.6). Furthermore, since $\mathcal{R}_i$ is a monotone risk measure, it preserves the pointwise convexity of the integrand $\mathcal{J} \circ S$. Therefore, the set of all optimal solutions is convex and, by hypothesis on U^i_{ad}, bounded. Therefore, we conclude that $\mathcal{B}_i$ has nonempty, convex, bounded images in U^i_{ad}.

Next, define $\mathcal{B} : U^1_{\text{ad}} \times \cdots \times U^N_{\text{ad}} \rightrightarrows U^1_{\text{ad}} \times \cdots \times U^N_{\text{ad}}$ by

$$\mathcal{B}(u) := \mathcal{B}_1(u_{-1}) \times \cdots \times \mathcal{B}_N(u_{-N}).$$

Suppose that $(u^k, v^k) \in \operatorname{gph} B$ such that $(u^k, v^k) \rightharpoonup (\bar{u}, \bar{v})$. This means in particular that for all k we have $v_i^k \in \mathcal{B}_i(u_{-i}^k)$, i.e.,

$$(\mathcal{R}_i \circ \mathcal{J}_i \circ S)(v_i^k, u_{-i}^k) + \wp(v_i^k) \leq (\mathcal{R}_i \circ \mathcal{J}_i \circ S)(w, u_{-i}^k) + \wp(w) \quad \forall w \in U_{\mathrm{ad}}^i.$$

In the current setting

$$S(u) = \mathbf{A}^{-1}\left(\sum_i \mathbf{B}_i u_i + \mathbf{b}\right) = \mathbf{A}^{-1}\mathbf{B}_i u_i + \mathbf{A}^{-1}\mathbf{b} + \sum_{j \neq i} \mathbf{A}^{-1}\mathbf{B}_j u_j.$$

As shown in Lemma 2.1 [28], each $\mathbf{B}_i$ is completely continuous from U^i into $L^2(\Omega, \mathcal{F}, \mathbb{P}; V^*) = \mathcal{V}^*$. Therefore, we have $\mathbf{B}_i v_i^k \to \mathbf{B}_i \bar{v}_i$ and $\mathbf{B}_j u_j^k \to \mathbf{B}_j \bar{u}_j$ strongly in $\mathcal{V}^*$ for each i and each $j \neq i$. It immediately follows from that $S(v_i^k, u_{-i}^k) \to S(\bar{v}_i, \bar{u}_{-i})$ and for any $w \in U_{\mathrm{ad}}^i$ $S(w, u_{-i}^k) \to S(w, \bar{u}_{-i})$.

Next, since $\mathcal{R}_i$ and $\mathcal{J}_i$ are continuous on their respective spaces, we have

$$(\mathcal{R}_i \circ \mathcal{J}_i \circ S)(v_i^k, u_{-i}^k) \to (\mathcal{R}_i \circ \mathcal{J}_i \circ S)(\bar{v}_i, \bar{u}_{-i})$$
$$(\mathcal{R}_i \circ \mathcal{J}_i \circ S)(w, u_{-i}^k) \to (\mathcal{R}_i \circ \mathcal{J}_i \circ S)(w, \bar{u}_{-i}).$$

Then due to the weak lower semicontinuity of $\wp$ on U^i, it follows that

$$(\mathcal{R}_i \circ \mathcal{J}_i \circ S)(\bar{v}_i, \bar{u}_{-i}) + \wp(\bar{v}_i) \leq (\mathcal{R}_i \circ \mathcal{J}_i \circ S)(w, \bar{u}_{-i}) + \wp(w) \quad \forall w \in U_{\mathrm{ad}}^i,$$

i.e., $\bar{v}_i \in \mathcal{B}_i(\bar{u}_{-i})$. Hence, the noncooperative game admits a Nash equilibrium. □

Remark 4.13 The previous proof can easily be extended to more complicated PDE models and objective functions. However, for nonlinear operators $\mathbf{N}$, we need to extend the results in Sect. 2 to the stochastic setting.

Given the explicit structure of the current setting, we can also derive optimality conditions for the NEP. Moreover, we can show that this specific problem reduces to a special kind of equilibrium problem in which the risk indicators are determined simultaneously by a single "risk trader."

Theorem 4.14 (Optimality Conditions) *Let $\bar{u}$ be a Nash equilibrium for (4.11). Then for each $i = 1, \ldots, N$ there exists a pair $(p_i^\star, \vartheta_i^\star) \in \mathcal{V} \times L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ such that the following conditions hold: $\vartheta^\star \in \partial \mathcal{R}_i[\mathcal{J}_i(y^\star)]$ and*

$$\begin{aligned} \left(u_i^\star - \frac{1}{\nu_i}\mathbb{E}[\mathbf{B}_i^* p_i^\star \vartheta_i^\star], w - u_i^\star\right)_{U^i} &\geq 0, \ \forall w \in U_{\mathrm{ad}}^i, \\ \mathbf{A}y^\star - \mathbf{B}u^\star + \mathbf{b} &= 0, \\ \mathbf{A}^* p_i^\star - y_d^i + y^\star &= 0. \end{aligned} \tag{4.12}$$

Proof This follows from Sect. 4.3 and the definition of a Nash equilibrium. □

System (4.12) leads to a useful reformulation. For each i, the adjoint states $p_i^\star$ split into the sum of a joint adjoint state $q^\star := \mathbf{A}^{-*}y^\star$ and a fixed i-dependent term $\tilde{y}_d^i := -\mathbf{A}^{-*}y_d^i$, where $\tilde{y}_d^i$ is now stochastic. Then, for each i, we have

$$\frac{1}{\nu_i}\mathbb{E}[\mathbf{B}_i^* p_i^\star \vartheta_i^\star] = \frac{1}{\nu_i}\mathbb{E}[\mathbf{B}_i^*(q^\star + \tilde{y}_d^i)\vartheta_i^\star] = \frac{1}{\nu_i}\mathbb{E}[\mathbf{B}_i^* q^\star \vartheta_i^\star] + \underbrace{\frac{1}{\nu_i}\mathbb{E}[\mathbf{B}_i^* \tilde{y}_d^i \vartheta_i^\star]}_{=:\widehat{c}_i}.$$

By defining $\mathbf{G}_i u := \frac{1}{\nu_i}\mathbf{B}_i^*\mathbf{A}^{-*}\mathbf{A}^{-1}\mathbf{B}u$ and $\mathbf{g}_i := \frac{1}{\nu_i}\mathbf{B}_i^*\mathbf{A}^{-*}\mathbf{A}^{-1}\mathbf{b}$, the variational inequality in (4.12) can be written as

$$(u_i^\star - (\mathbb{E}[\vartheta_i^\star \mathbf{G}_i u^\star] + c_i(\vartheta_i^\star)), v - u_i^\star)_{U^i} \geq 0 \quad \forall v \in U_{\text{ad}}^i,$$

where $c_i(\vartheta_i) := \mathbb{E}[\vartheta_i \mathbf{g}_i] - \widehat{c}_i$. Summing over i, we obtain

$$\sum_{i=1}^N (u_i^\star - (\mathbb{E}[\vartheta_i^\star \mathbf{G}_i u^\star] + c_i(\vartheta_i^\star)), v_i - u_i^\star)_{U^i} \geq 0 \quad \forall v \in U_{\text{ad}}. \tag{4.13}$$

Conversely, if the previous inequality holds, then by using the variations

$$(v_1^\star, \ldots, v_i, \ldots, v_N^\star) = v \in U_{\text{ad}} = U_{\text{ad}}^1 \times \cdots \times U_{\text{ad}}^N$$

for each $i = 1, \ldots, N$ (leaving only v_i to vary), we recover the individual inequalities. We will refer to (4.13) as the "aggregate player's problem." Letting $\text{Proj}_{U_{\text{ad}}^i}$ denote the metric projection onto U_{ad}^i, this can be formulated as a single nonsmooth equation in the product space $U = U^1 \times \cdots \times U^N$: Find $u^\star \in U$: $\forall i = 1, \ldots, N$

$$u_i^\star = \text{Proj}_{U_{\text{ad}}^i}\left[\mathbb{E}[\vartheta_i^\star \mathbf{G}_i u^\star] + c_I(\vartheta_i^\star)\right]. \tag{4.14}$$

Continuing, since $\mathcal{R}_i$ is assumed to be a coherent risk measure, we have

$$\vartheta_i^\star \in \underset{\vartheta \in \mathfrak{A}_i}{\text{argmax}}\ \mathbb{E}[\vartheta \mathcal{J}_i(y^\star)],$$

where $\mathfrak{A}_i := \text{dom}(\mathcal{R}_i^*)$ is the domain of the Fenchel conjugate $\mathcal{R}_i^*$ of $\mathcal{R}_i$. It is then easy to show that all of the subdifferential inequalities can be joined into a single maximization problem:

$$\max\left\{\sum_{i=1}^N \mathbb{E}[\vartheta_i \mathcal{J}_i(\mathbf{A}^{-1}(\mathbf{B}u^\star + \mathbf{b}))],\ \text{over } \vartheta \in \mathfrak{A}\right\}, \tag{4.15}$$

where $\mathfrak{A} := \mathfrak{A}_1 \times \cdots \times \mathfrak{A}_N$. Problem (4.15) always has a solution since the objective is a bounded linear functional and $\mathfrak{A}$ is a weakly-$*$ sequentially compact, closed, and convex set. Inspired by the terminology in [37], we will refer to (4.15) as the "risk trader's problem."

We have thus proven that the risk-averse PDE-constrained NEP can be understood as a type of MOPEC (multiple optimization problems with equilibrium constraints) comprising a single aggregate player, who solves a well-posed variational inequality in u given a fixed risk indicator vector ϑ, and a risk trader who spreads the risk of the decision vector u over the components of ϑ in light of the various objectives $\mathcal{J}_i$ and risk preferences $\mathfrak{A}_i$.

Even in this special case, it is difficult to immediately select an appropriate solution algorithm. Perhaps the main challenge lies in the fact that the risk trader's problem does not have a unique solution. One remedy for this to ensure a unique ϑ for a given u is to replace the objective in (4.15) by

$$\mathbb{E}[\vartheta_i \mathcal{J}_i(\mathbf{A}^{-1}(\mathbf{B}u^\star + \mathbf{b}))] - \frac{\varepsilon}{2}\mathbb{E}[\vartheta_i^2] \quad \varepsilon > 0. \tag{4.16}$$

This was suggested in [24] for treating the nonsmooth risk measure AVaR_β in the context of PDE-constrained optimization under uncertainty. It was later demonstrated that such a regularization is a special case of the deeper theory of epi-regularization of risk measures in [25]. We briefly discuss this notion below.

4.5 *Risk-Averse PDE-Constrained Decision Problems: Smooth Approximation*

As a means of circumventing the unacceptably slow performance of classical nonsmooth optimization algorithms such as subgradient methods or bundle methods, we proposed smoothing approaches in [24] and [25]. An alternative viewpoint can be found by exploiting the structure of a specific class of coherent risk measures and using an interior-point approach as in [13]. In addition, the analysis in the previous section indicates yet another reason to consider some form of variational smoothing in the context of stochastic PDE-constrained equilibrium problems.

We briefly give the details of epi-regularization as it has proven to be a versatile tool not only for smoothing risk measures but also for analyzing new optimization methods for risk-averse PDE-constrained optimization, cf. [27]. Let $\Psi : \mathcal{X} \to \overline{\mathbb{R}}$ be a proper, closed, and convex functional and $\mathcal{R}$ a regular measure of risk. Then for $\varepsilon > 0$, we define the epi-regularized measure of risk as

$$\mathcal{R}_\varepsilon^\Psi[X] = \inf_{Y \in \mathcal{X}} \left\{ \mathcal{R}[X - Y] + \varepsilon \Psi\left[\varepsilon^{-1} Y\right] \right\} = \inf_{Y \in \mathcal{X}} \left\{ \mathcal{R}[Y] + \varepsilon \Psi\left[\varepsilon^{-1}(X - Y)\right] \right\}.$$

As mentioned above, the regularization in (4.16) is equivalent to using the function $\Psi[X] = \frac{1}{2}\mathbb{E}[X^2]$. Another import example can be seen by setting $\mathcal{X} = L^2(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{R} = \text{AVaR}_\beta$, and $\Psi[X] := \mathbb{E}[X] + \frac{1}{2}\mathbb{E}[X^2]$. This results in

$$\mathcal{R}_\varepsilon^\Psi[X] = \inf_{t\in\mathbb{R}} \left\{ t + \mathbb{E}\left[v_{\beta,\epsilon}(X - t) \right] \right\},$$

which is continuously Fréchet differentiable and in which the scalar function $v_{\beta,\varepsilon}$ is given by

$$v_{\beta,\varepsilon}(x) = \begin{cases} -\frac{\varepsilon}{2}, & \text{if } x \le -\varepsilon \\ \frac{1}{2\varepsilon}x^2 + x, & \text{if } x \in \left(-\varepsilon, \frac{\varepsilon\beta}{1-\beta}\right) \\ \frac{1}{1-\beta}\left(x - \frac{\varepsilon\beta^2}{2(1-\beta)}\right), & \text{if } x \ge \frac{\varepsilon\beta}{1-\beta}. \end{cases}$$

Epi-regularization has a number of advantageous properties. For example, we can show that the sequence of functionals $\left\{\mathcal{R}_\varepsilon^\Psi\right\}_{\varepsilon>0}$ converges in the sense of Mosco to $\mathcal{R}$. Furthermore, under certain assumptions on $\mathcal{J}$ and $\wp$, we can show that weak accumulation points of approximate minimizers $z_\varepsilon^\star$ are optimal for (4.2) and weak accumulation points of approximate stationary points are stationary for (4.2). For more on this topic, we refer to the forthcoming publication [25].

4.6 *Risk-Averse PDE-Constrained Optimization: Solution Methods*

In this final section, we outline the main components of the recently proposed primal–dual risk minimization algorithm in [27]. This is an all purpose optimization algorithm for minimizing risk measures in the context of PDE-constrained optimization under uncertainty.

In general, the individual problems in our risk-averse setting have the form:

$$\min_{x\in X_{\text{ad}}} \{g(x) + \Phi(G(x))\}, \tag{4.17}$$

where g is a deterministic objective function, G is an uncertain objective function, and Φ is a functional that maps random variables into the real numbers. The functional Φ is typically convex, positively homogeneous, and monotonic with respect to the natural partial order on the space of random variables.

Let $\Phi : \mathcal{Y} \to \mathbb{R}$, where $\mathcal{Y} = L^2(\Omega, \mathcal{F}, \mathbb{P})$. As shown in [41, Th. 6.5], there exists a nonempty, convex, closed, and bounded set $\mathfrak{A} \subseteq \{\theta \in \mathcal{Y}^* \mid \theta \ge 0 \text{ a.s.}\}$ such that a convenient bi-dual representation of Φ is available:

$$\Phi(X) = \sup_{\theta\in\mathfrak{A}} \mathbb{E}[\theta X]. \tag{4.18}$$

Moreover, Φ is continuous and subdifferentiable, cf. [41, Prop. 6.6], and $\mathfrak{A} = \partial\Phi(0)$.

Using these facts, (4.17) exhibits a familiar structure in which, by introducing the Lagrangian-type function $\ell(x, \lambda) := g(x) + \mathbb{E}[\lambda G(x)]$, we can consider the minimax reformulation:

$$\min_{x \in X_{\text{ad}}} \sup_{\lambda \in \mathfrak{A}} \ell(x, \lambda). \tag{4.19}$$

We can then develop a method similar to the classical method of multipliers [16, 36].

To this end, we introduce the (dual) generalized augmented Lagrangian:

$$L(x, \lambda, r) := \max_{\theta \in \mathfrak{A}} \left\{ \ell(x, \theta) - \frac{1}{2r}\mathbb{E}[(\lambda - \theta)^2] \right\}. \tag{4.20}$$

Now, using several techniques from convex analysis, it can be shown that

$$L(x, \lambda, r) = g(x) + \min_{Y \in \mathcal{Y}} \left\{ \Phi(G(x) - Y) + \mathbb{E}[\lambda Y] + \frac{r}{2}\mathbb{E}[Y^2] \right\}. \tag{4.21}$$

In other words, L is the objective in (4.17) with Φ replaced by a multiplier-dependent epi-regularization, where the regularizer is

$$\Psi_{r,\lambda}(Y) = \mathbb{E}[\lambda Y] + \tfrac{r}{2}\mathbb{E}[Y^2].$$

Furthermore, letting

$$\Lambda(x, \lambda, r) := \text{Proj}_{\mathfrak{A}}(rG(x) + \lambda),$$

where $\text{Proj}_{\mathfrak{A}} : \mathcal{Y} \to \mathcal{Y}$ is the projection onto $\mathfrak{A}$, L attains the closed form

$$L(x, \lambda, r) = g(x) + \mathbb{E}[\lambda G(x)] + \frac{r}{2}\mathbb{E}[G(x)^2] - \frac{1}{2r}\mathbb{E}[\{(\text{Id} - \text{Proj}_{\mathfrak{A}})(rG(x) + \lambda)\}^2].$$

For many risk measures of interest, e.g., mean-plus-semideviation or convex combinations of mean and AVaR [27, Sec. 5.1], the optimization problem (4.17) can be rewritten so that $\Phi(Y) = \mathbb{E}[(Y)_+]$. Therefore, the projection operator $\text{Proj}_{\mathfrak{A}}$ can be easily evaluated. For more general coherent risk measures, $\mathfrak{A}$ can be split into box constraints and a simple normalizing constraint that is treatable with a Lagrange multiplier, cf. [27, Sec. 5.2].

The basic algorithm is given in Algorithm 1. A detailed implementable version allowing for inexact subproblem solves, and multiplier-update strategies can be found in [27] (Algorithm 2). A full convergence theory for the primal and dual updates in both convex and nonconvex settings in infinite-dimensional spaces is given in [27, Sec. 4]. Here, the convergence of the primal variables exploits a number

of powerful results arising in the theory of epi-regularization. For the dual variables, a regularity condition that postulates the existence of a saddle point is needed.

Algorithm 1 Primal–dual risk minimization

1. **Initialize:** Given $x_0 \in X_{\text{ad}}$, $r_0 > 0$, and $\lambda_0 \in \mathfrak{A}$.
2. **While("Not Converged")**
   (a) Compute $x_{k+1} \in X_{\text{ad}}$ as approximate minimizer of $L(\cdot, \lambda_k, r_k)$.
   (b) Set $\lambda_{k+1} = \Lambda(x_{k+1}, \lambda_k, r_k)$.
   (c) Update r_{k+1}.
3. **End While**

Returning to our game-theoretic setting in Sect. 4.4, we see a clear link to the risk trader's problem (4.15). As mentioned in Sect. 4.4, (4.15) does not admit a unique solution. This makes the numerical solution of the game, in its original form as well as the proposed reduced from, very challenging. The suggestion in (4.16) indicates that we could handle this aspect by applying an epi-regularization technique to the risk measures. Though the suggestion given there is viable, the favorable convergence behavior of Algorithm 1 given in [27, Sec. 4] indicates that the multiplier-dependent epi-regularization update in the primal–dual algorithm is probably better suited (clearly algorithmically motivated). We thus propose a method that successively solves the aggregate player's game using an update formula for ϑ similar to the Λ-operator in the primal–dual algorithm. This avenue of thought will be the focus of future work. Nevertheless, the epi-regularization technique does not rule out the possibility that the associated system of nonlinear and semismooth equations admits distinct solutions. A possible remedy to this issue can be found in the recent publication [10].

5 Outlook

Generalized Nash equilibrium problems with PDE constraints represent a challenging class of infinite-dimensional equilibrium problems. Beyond the deterministic convex setting involving linear elliptic or parabolic PDEs, major theoretical and algorithmic challenges arise. Nevertheless, we have shown that it is still possible to treat some GNEPs involving semilinear, nonsmooth, and even mutlivalued forward problems by appealing to the notions of generalized convexity and isotonic mappings. Due to a lack of convexity, we have chosen to derive stationarity conditions using the versatile limiting variational calculus in the sense of Mordukhovich. In doing so, we have been able to push the boundaries of existence and optimality theory in the deterministic setting beyond linear state systems. Therefore, we may now build upon these advances toward the development of function-space-based numerical methods similar to [20, 21]. The recent results in [23] on augmented

Lagrangian-type methods (also developed within the priority program) may also prove to be useful here.

As outlined above, the stochastic risk-averse setting is now poised to transfer the results from the newly developed theory of risk-averse PDE-constrained optimization [13, 24–28] to the setting of noncooperative strategic games. This will be the focus for the remainder of the project duration. In addition to the algorithmic strategy mentioned above, there are several open theoretical questions relating to variational convergence in the context of strategic games and asymptotic statistical properties of Nash equilibrium in the vein of [41, Chap. 5]. Some progress on related stability issues using probability metrics has been made in the recent Master's thesis [22]. In addition, the results from the deterministic nonlinear case can be folded into the stochastic setting by using the results in [27] for risk-averse control of semilinear equations. Finally, in order to treat even jointly convex state-constrained risk-averse PDE-constrained GNEPs, a sufficient theory of PDE-constrained optimization under uncertainty with state constraints is under development.

References

1. Adams, R. A., and Fournier, J. J.-F. *Sobolev Spaces*, second ed. Elsevier, Amsterdam, 2008.
2. Appell, J., and Zabrejko, P. P. *Nonlinear Superposition Operators*. Cambridge Tracts in Mathematics. Cambridge University Press, 2008.
3. Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. Coherent measures of risk. *Math. Finance 9*, 3 (1999), 203–228.
4. Blyth, T. *Lattices and Ordered Algebraic Structures*. Universitext. Springer London, 2006.
5. Bonnans, J. F., and Shapiro, A. *Perturbation Analysis of Optimization Problems*. Springer Verlag, Berlin, Heidelberg, New York, 2000.
6. Borwein, J., and Zhu, Q. *Techniques of Variational Analysis*. CMS Books in Mathematics. Springer New York, 2006.
7. Bredies, K., Clason, C., Kunisch, K., and Winckel, G. *Control and optimization with PDE constraints*, vol. 164. Springer Science & Business Media, 2013.
8. Eilenberg, S., and Montgomery, D. Fixed point theorems for multi-valued transformations. *American Journal of Mathematics 68*, 2 (1946), 214–222.
9. Facchinei, F., and Kanzow, C. Generalized Nash equilibrium problems. *4OR 5*, 3 (Sep 2007), 173–210.
10. Farrell, P. E., Croci, M., and Surowiec, T. M. Deflation for semismooth equations. *Optimization Methods and Software 0*, 0 (2019), 1–24. DOI: 10.1080/10556788.2019.1613655.
11. Ferris, M. C., and Pang, J. S. Engineering and economic applications of complementarity problems. *SIAM Review 39* (1997), 669–713.
12. Föllmer, H., and Schied, A. Convex measures of risk and trading constraints. *Finance Stoch. 6*, 4 (2002), 429–447.
13. Garreis, S., Surowiec, T. M., and Ulbrich, M. An interior-point approach for solving risk-averse PDE-constrained optimization problems with coherent risk measures. *SIAM Journal on Optimization* (2019). submitted.
14. Glicksberg, I. L. A further generalization of the Kakutani fixed theorem, with application to Nash equilibrium points. *Proc. Amer. Math. Soc. 3* (1952), 170–174.
15. Goldberg, H., Kampowsky, W., and Tröltzsch, F. On Nemytskij operators in Lp-spaces of abstract functions. *Mathematische Nachrichten 155*, 1 (1992), 127–140.

16. HESTENES, M. R. Multiplier and gradient methods. *Journal of Optimization Theory and Applications 4*, 5 (Nov 1969), 303–320.
17. HILLE, E., AND PHILLIPS, R. S. *Functional analysis and semi-groups*. American Mathematical Society Colloquium Publications, vol. 31. American Mathematical Society, Providence, R. I., 1957. rev. ed.
18. HINTERMÜLLER, M., AND STENGL, S.-M. Generalization of path-following methods for generalized Nash equilibrium problems (tentative title). Preprint, in preparation.
19. HINTERMÜLLER, M., AND STENGL, S.-M. On the convexity of optimal control problems involving nonlinear PDEs or VIs and applications to Nash games (tentative title). Preprint, in preparation.
20. HINTERMÜLLER, M., AND SUROWIEC, T. A PDE-constrained generalized Nash equilibrium problem with pointwise control and state constraints. *Pac. J. Optim. 9*, 2 (2013), 251–273.
21. HINTERMÜLLER, M., SUROWIEC, T., AND KÄMMLER, A. Generalized Nash equilibrium problems in Banach spaces: theory, Nikaido-Isoda-based path-following methods, and applications. *SIAM J. Optim. 25*, 3 (2015), 1826–1856.
22. HOFFHUES, M. Stabilität stochastischer Optimierungsprobleme in Hilberträumen. Master's thesis, Philipps-Universität Marburg, Fachbereich Mathematik und Informatik, Marburg, Germany, 2019.
23. KANZOW, C., KARL, V., STECK, D., AND WACHSMUTH, D. The multiplier-penalty method for generalized Nash equilibrium problems in Banach spaces. *SIAM Journal on Optimization 29*, 1 (2019), 767–793.
24. KOURI, D. P., AND SUROWIEC, T. M. Risk-averse PDE-constrained optimization using the conditional value-at-risk. *SIAM Journal on Optimization 26*, 1 (2016), 365–396.
25. KOURI, D. P., AND SUROWIEC, T. M. Epi-regularization of risk measures for PDE-constrained optimization. *Math. Oper. Res.* (2018). to appear.
26. KOURI, D. P., AND SUROWIEC, T. M. Existence and optimality conditions for risk-averse PDE-constrained optimization. *SIAM/ASA Journal on Uncertainty Quantification 6*, 2 (2018), 787–815.
27. KOURI, D. P., AND SUROWIEC, T. M. A primal-dual algorithm for risk minimization. *Math. Programm. Ser. A.* (2018). submitted.
28. KOURI, D. P., AND SUROWIEC, T. M. Risk-averse optimal control of semilinear elliptic PDEs. *ESAIM COCV* (2018). submitted.
29. KRASNOSEL'SKII, M. A. *Topological methods in the theory of nonlinear integral equations*. Translated by A. H. Armstrong; translation edited by J. Burlak. A Pergamon Press Book. The Macmillan Co., New York, 1964.
30. LEUGERING, G., ENGELL, S., GRIEWANK, A., HINZE, M., RANNACHER, R., SCHULZ, V., ULBRICH, M., AND ULBRICH, S. *Constrained optimization and optimal control for partial differential equations*, vol. 160. Springer Science & Business Media, 2012.
31. LIONS, J. L. *Some aspects of the optimal control of distributed parameter systems*, vol. 6. Siam, 1972.
32. MIGNOT, F. Contrôle dans les inéquations variationelles elliptiques. *Journal of Functional Analysis 22*, 2 (1976), 130–185.
33. MORDUKHOVICH, B. S. *Variational analysis and generalized differentiation I: Basic theory*, vol. 330. Springer Science & Business Media, 2006.
34. MOSER, J. A sharp form of an inequality by N. Trudinger. *Indiana University Mathematics Journal 20*, 11 (1971), 1077–1092.
35. NISAN, N., ROUGHGARDEN, T., TARDOS, E., AND VAZIRANI, V. *Algorithmic Game Theory*. Cambridge University Press, 2007.
36. POWELL, M. J. D. A method for nonlinear constraints in minimization problems. *Optimization* (1969), 283–298.
37. RALPH, D., AND SMEERS, Y. Risk trading and endogenous probabilities in investment equilibria. *SIAM J. Optim. 25*, 4 (2015), 2589–2611.
38. ROCKAFELLAR, R. T., AND URYASEV, S. Optimization of conditional value-at-risk. *Journal of Risk 2* (2000), 21–41.

39. ROCKAFELLAR, R. T., AND URYASEV, S. The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science 18*, 12 (2013), 33–53.
40. SHAPIRO, A. On concepts of directional differentiability. *J. Optim. Theory Appl. 66*, 3 (1990), 477–487.
41. SHAPIRO, A., DENTCHEVA, D., AND RUSZCZYNSKI, A. *Lectures on Stochastic Programming: Modeling and Theory, Second Edition*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, Philadelphia, 2014.
42. TRÖLTZSCH, F. *Optimal control of partial differential equations: theory, methods, and applications*, vol. 112. American Mathematical Soc., 2010.
43. VAINBERG, M. M. *Variational methods for the study of nonlinear operators*. Holden-Day, Inc., San Francisco, Calif.-London-Amsterdam, 1964. With a chapter on Newton's method by L. V. Kantorovich and G. P. Akilov. Translated and supplemented by Amiel Feinstein.
44. WERNER, D. *Funktionalanalysis*. Springer-Lehrbuch. Springer Berlin Heidelberg, 2011.
45. ZOWE, J., AND KURCYUSZ, S. Regularity and stability for the mathematical programming problem in Banach spaces. *Applied Mathematics and Optimization 5*, 1 (1979), 49–62.

Stability and Sensitivity Analysis for Quasi-Variational Inequalities

Amal Alphonse, Michael Hintermüller, and Carlos N. Rautenberg

Abstract We discuss various aspects of quasi-variational inequalities (QVIs) related to their sensitivity analysis and optimal control. Starting with the necessary functional framework and existence results for elliptic QVIs of obstacle type, we study stability of the solution map taking the source term onto the set of solutions: we show that certain realisations of the map have appropriate continuity properties. We then focus on showing that a notion of directional derivative exists for QVIs and we characterise this derivative as a monotone limit of directional derivatives associated to particular variational inequalities. The differentiability theory is illustrated with a novel application in thermoforming. Using the stability results, we discuss control problems with QVI constraints and prove existence of optimal controls.

Keywords Quasi-variational inequality · Obstacle problem · Directional differentiability · Set-valued analysis · Thermoforming · Sensitivity analysis

Mathematics Subject Classification (2020) Primary 47J20; Secondary 49K40

A. Alphonse
Weierstrass Institute, Berlin, Germany
e-mail: alphonse@wias-berlin.de; amal.alphonse@wias-berlin.de

M. Hintermüller (✉)
Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany
e-mail: hintermueller@wias-berlin.de; hint@math.hu-berlin.de

C. N. Rautenberg
Department of Mathematical Sciences and the Center for Mathematics and Artificial Intelligence (CMAI), George Mason University, Fairfax, VA, USA
e-mail: crautenb@gmu.edu

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_8

1 Introduction

Quasi-variational inequalities (QVIs) are generalisations of variational inequalities (VIs) where the constraint set associated to the inequality depends itself on the a priori unknown solution. In the elliptic setting, they have the general form

$$\text{given } f \in V^*, \text{find } y \in \mathbf{K}(y) : \langle A(y) - f, y - v \rangle \leq 0 \quad \forall v \in \mathbf{K}(y),$$

where $\mathbf{K}(\cdot)$ is the constraint set map (full details of all terms and spaces appearing in this inequality will be given in Sect. 2). As such, QVIs are generally much more complicated than VIs due to the fact that the constraint set is parametrised by the solution as well as due to the extra source of potential nonlinearity and non-smoothness. In general, solutions to such QVIs are also non-unique, giving rise to a set-valued solution map which causes additional technical difficulties.

QVIs were first studied by Bensoussan and Lions [11, 25] for stochastic impulse control problems, and they arise in a variety of other applications in the physical and social sciences and economics. These include generalised Nash equilibrium problems [13, 19, 31], magnetisation of superconductors [8, 22, 34, 37], the growth of sandpiles [9, 33, 35, 36], and networks of lakes and rivers [10, 33, 35], and thermoforming [2]. More generally, QVIs play an important role in the modelling of complex phenomena where compliancy and state-dependent bounds are important features.

The solution-dependent constraint set, which is the defining feature of QVIs, means that special attention must be paid for the development of existence results and solution algorithms. In this chapter, we shall outline (and in places detail) contributions that we have made [1–4] to the stability/sensitivity analysis and optimal control of QVIs of obstacle type. We will focus on elliptic QVIs here for simplicity, but we also shall mention where appropriate related parabolic results from our recent work [3].

We start in Sect. 2 with the fundamental functional framework and existence results for QVIs, and we also point out a common erroneous technique found in some literature. In Sect. 3, we discuss continuity and sensitivity properties of the solution map associated to QVIs. An application of QVIs and their sensitivity analysis in thermoforming is given in Sect. 3.4, and a numerical experiment will be presented. Section 4 discusses some optimal control problems and the existence thereof, making use of the stability results acquired in the previous section. We finish in Sect. 5 with a few words on the current and future work.

2 QVIs: Mathematical Setting and Existence

We focus on the following functional setting and problem class. We consider a Hilbert space V that is continuously and densely embedded into the Hilbert space $H := L^2(\Omega)$ where $\Omega \subset \mathbb{R}^N$ is a Lipschitz domain. We usually denote the inner

product on H by $(\cdot,\cdot)$. Define the convex cone $H_+ := L^2(\Omega)^+$, which is the set of almost everywhere non-negative elements of $L^2(\Omega)$. Every $x \in H$ admits a decomposition $x = x^+ - x^- \in H_+ - H_+$ with $(x^+, x^-) = 0$ where x^+ denotes the orthogonal projection of x onto H_+. The infimum and supremum of $x, y \in H$ are defined as $\sup(x, y) := x + (y - x)^+$ and $\inf(x, y) := x - (x - y)^+$, respectively. Finally, we assume that

$$y \in V \implies y^+ \in V \quad \text{and} \quad \exists \kappa > 0 : \|y^+\|_V \leq \kappa \|y\|_V \quad \forall y \in V$$

and that $\sup(0, \cdot)$ and $\inf(0, \cdot)$ are continuous with respect to the weak and strong topologies of V.

Note that the ordering in H induces an ordering in the dual space V^*: if $f, g \in V^*$, we say $f \leq g$ if $\langle f, \phi\rangle \leq \langle g, \phi\rangle$ for all $\phi \in V_+ := V \cap H_+$, and define $V_+^* := \{f \in V^* : f \geq 0\}$.

Suppose that $A\colon V \to V^*$ is an operator which is:

1. Homogeneous of order 1, i.e.,

$$A(\lambda u) = \lambda A(u) \quad \forall u \in V, \lambda > 0$$

2. Lipschitz continuous, i.e., there exists $C > 0$ such that

$$\|A(u) - A(v)\|_{V^*} \leq C\|u - v\|_V \quad \forall u, v \in V$$

3. Uniformly monotone, i.e., there exists $c > 0$ such that

$$\langle A(u) - A(v), u - v\rangle \geq c\|u - v\|_V^2 \quad \forall u, v \in V$$

4. Strictly T-monotone, i.e.,[1]

$$\langle A(u) - A(v), (u - v)^+\rangle > 0 \quad \forall u, v \in V : (u - v)^+ \neq 0.$$

Example 2.1 (Prototypical Sobolev Space Setting) The typical realisation of the above is based on the Sobolev space $V = H_0^1(\Omega)$, and $A = -\Delta$ is taken to be the Laplacian, which is defined for functions $u, v \in H_0^1(\Omega)$ through the action

$$\langle -\Delta u, v\rangle := \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x.$$

We can also consider $V = H^1(\Omega)$ with appropriate modifications to A.

[1] In particular, if A is linear, this is equivalent to $\langle Ay^-, y^+\rangle \leq 0$ for all $y \in V$, and we have the availability of maximum principles for A.

We define a constraint set map $\mathbf{K}\colon H \rightrightarrows V$ by

$$\mathbf{K}(v) := \{w \in V : w \leq \Phi(v)\},$$

where:

1. $\Phi : H \to H$ is a given increasing map.
2. There exist $f_{\min}, f_{\max} \in V^*$ such that

$$\text{the data } f \in V^* \text{ satisfies } f_{\min} \leq f \leq f_{\max} \quad \text{and} \quad \Phi(v) \geq A^{-1}(f_{\min}) \quad \forall v \in V.$$

We are now in position to state the precise problem we are interested in.

Problem ($\mathrm{P_{QVI}}$) Given $f \in V^*$ as above, find $y \in \mathbf{K}(y)$ such that

$$\langle A(y) - f, y - v\rangle \leq 0 \quad \forall v \in \mathbf{K}(y). \tag{$\mathrm{P_{QVI}}$}$$

2.1 Existence of Solutions: Order Approach

We describe now an approach based on order that was pioneered by Tartar to prove existence of solutions to a class of QVIs [38], [7, Chapter 15, §15.2].

We say that a map $R\colon H \to H$ has a *subsolution* y if $y \leq R(y)$; a *supersolution* is defined with the reverse inequality. The following general result for existence of fixed points for increasing maps is the fundamental tool to prove existence of solutions to problem ($\mathrm{P_{QVI}}$). The theorem essentially states that an increasing map that possesses a subsolution y_1 and a supersolution y_2 has a fixed point between y_1 and y_2, and furthermore, there are minimal and maximal fixed points in $[y_1, y_2]$.

Theorem 2.2 (Tartar–Birkhoff) *Let $R\colon H \to H$ be increasing, and suppose that there exist $\underline{y}, \overline{y} \in H$ such that*

$$\underline{y} \leq \overline{y}, \qquad \underline{y} \leq R(\underline{y}), \quad \textit{and} \quad R(\overline{y}) \leq \overline{y}.$$

Then the set of fixed points of R in the interval $[\underline{y}, \overline{y}]$ is non-empty, and there exist smallest and largest (defined through the ordering given above) fixed points.

By our assumptions on Φ and f, it follows that

$$\underline{y} := A^{-1}(f_{\min}) \qquad \text{and} \qquad \overline{y} := A^{-1}(f_{\max})$$

are sub- and supersolutions (due to the comparison principle for VIs), respectively, of the map

$$T(\cdot) := S(f, \mathbf{K}(\cdot)), \tag{1}$$

where, given a set $\mathbf{C} \subset V$, the notation $S(f, \mathbf{C})$ is defined as the unique solution to the variational inequality

$$\text{find } y \in \mathbf{C} : \langle A(y) - f, y - v \rangle \leq 0 \quad \forall v \in \mathbf{C}. \tag{2}$$

Clearly, T maps H into V and is increasing (thanks again to the comparison principle), and hence, the previous theorem can be applied and we deduce the existence of solutions to ($\mathrm{P_{QVI}}$).

Definition 2.3 If we define the set $\mathbf{A}_{\mathrm{ad}} := \{g \in V^* : f_{\min} \leq g \leq f_{\max}\}$, we then have the minimal and maximal solution operators

$$\mathsf{m} \colon \mathbf{A}_{\mathrm{ad}} \to V \qquad \text{and} \qquad \mathsf{M} \colon \mathbf{A}_{\mathrm{ad}} \to V$$

that take elements of $\mathbf{A}_{\mathrm{ad}}$ to minimal and maximal solutions to ($\mathrm{P_{QVI}}$) on the interval $[\underline{y}, \overline{y}]$.

We shall discuss the stability of these minimal and maximal solution maps later.

2.2 Existence of Solutions: Iteration Approach

Let (V, H, V^*) be now a Gelfand triple, so that $V \hookrightarrow H \hookrightarrow V^*$ with the first embedding continuous and dense and where H is identified with its dual H^*. Defining the sub- and supersolution $\underline{y}$ and $\overline{y}$ as above and the map T as (1), consider the iterations

$$m_{n+1} := T(m_n), \qquad M_{n+1} := T(M_n),$$
$$m_0 := \underline{y}, \qquad M_0 := \overline{y}.$$

We will write $a_n \uparrow a$ to mean that $a_n \to a$ and $a_n \geq a_{n-1}$ for all $n \in \mathbb{N}$; $a_n \downarrow a$ is defined with the obvious modification. The next result shows that the above sequences converge to the expected limits.

Proposition 2.4 (Theorem 4 in [4]) *Suppose that $\Phi \colon V \to V$ is completely continuous. Then $m_n \uparrow \mathsf{m}(f)$ and $M_n \downarrow \mathsf{M}(f)$ in H and $m_n \to \mathsf{m}(f)$ and $M_n \to \mathsf{M}(f)$ in V.*

Proof We only sketch the proof and refer the reader to the proof of [4, Theorem 4] for the full details. We find that $\{m_n\}$ and $\{M_n\}$ are monotonically increasing and decreasing, respectively, and additionally $m_n, M_n \in [\underline{y}, \overline{y}]$. Furthermore,

$$m_n \uparrow m^*, \quad M_n \downarrow M^* \text{ in } H \quad \text{and} \quad m_n \rightharpoonup m^*, \quad M_n \rightharpoonup M^* \text{ in } V$$

for some $M^*, m^* \in V$. Note that $M^* \leq \Phi(M_{n-1})$ so that

$$c\|M_n - M^*\|_V^2 \leq \langle AM_n - AM^*, M_n - M^* \rangle \leq \langle f - AM^*, M_n - M^* \rangle,$$

hence $M_n \to M$ in V. Provided that $\Phi : H \to H$ is continuous, it is not hard to prove that M^* is a solution to ($\mathrm{P_{QVI}}$), i.e., $M^* = S(f, \mathbf{K}(M^*))$. Since $\mathsf{M}(f)$ is the maximum solution to ($\mathrm{P_{QVI}}$) on $[\underline{y}, \overline{y}]$, $M^* \leq \mathsf{M}(f)$. Further, since $\mathsf{M}(f) \leq \overline{y}$, by repeated iteration of T on the previous inequality, we have that $\mathsf{M}(f) \leq M^*$, i.e., $\mathsf{M}(f) = M^*$.

In order to prove that $m^* = \mathsf{m}(f)$, the additional assumption that $\Phi : V \to V$ be completely continuous is required. With this, $v_n := \min(m^*, \Phi(m_{n-1}))$ satisfies $v_n \to m^*$ in V and $v_n \leq \Phi(m_{n-1})$. Hence,

$$c\|m_n - v_n\|_V^2 \leq \langle Am_n - Av_n, m_n - v_n \rangle \leq \langle f - Av_n, m_n - v_n \rangle,$$

where we have used that $m_n = S(f, \mathbf{K}(m_{n-1}))$; thus $m_n \to m^*$ in V. From $m_n \leq \Phi(m_{n-1})$, and since strong convergence in H preserves order, we have $m^* \leq \Phi(m^*)$. Choose $v \leq \Phi(m^*)$ arbitrary and define $v_n := \min(v, \Phi(m_{n-1}))$, so that $v_n \to m^*$ in V and $v_n \leq \Phi(m_{n-1})$. Then

$$\langle Am^* - f, m^* - v \rangle = \lim_{n \to \infty} \langle Am_n - f, m_n - v_n \rangle \leq 0,$$

that is, m^* is a solution to ($\mathrm{P_{QVI}}$) within $[\underline{y}, \overline{y}]$. Hence, by definition of $\mathsf{m}(f)$, we have $\mathsf{m}(f) \leq m^*$, and from $\underline{y} \leq \mathsf{m}(f)$ and the consecutive iteration of T on the previous inequality, we have $m^* \leq \mathsf{m}(f)$, i.e., $m^* = \mathsf{m}(f)$. □

2.3 *Miscellaneous: A Pitfall*

Before we proceed further, it is useful to warn the reader of a common mistake that appears in the literature which is based on trying to extend the theorem of Lions and Stampacchia in [26] to the QVI framework.

First observe the following. Provided that $\Phi : V \to V$ is Lipschitz, we can consider the change of variable $z = y : -\Phi(v)$, and it is straightforward to prove via the monotonicity of A that T satisfies

$$\|T(v_1) - T(v_2)\|_V \leq \frac{1}{c}\|A\Phi(v_1) - A\Phi(v_2)\|_{V'} \leq \frac{C}{c} L_\Phi \|v_1 - v_2\|_V.$$

Hence for

$$\frac{C}{c} L_\Phi < 1,$$

the map T has a unique fixed point and the iteration $y_{n+1} = T(y_n)$ converges to this fixed point for any initial $y_0 \in V$. The extent of the usage of this technique is limited to the exact case described here (also note that if $V = H_0^1(\Omega)$, then, the assumptions here also imply that $\Phi(v) = 0$ on $\partial\Omega$ in the sense of the trace).

Now, let $\mathrm{P}_{\mathbf{K}(y)} \colon V \to V \subset \mathbf{K}(y)$ be the projection map, i.e., for any $v \in V$, $\mathrm{P}_{\mathbf{K}(y)}(v)$ is the unique element in $\mathbf{K}(y)$ such that

$$\|\mathrm{P}_{\mathbf{K}(y)}(v) - v\|_V = \inf_{w \in \mathbf{K}(y)} \|w - v\|_V.$$

Let $i \colon V \to V^*$ denote the canonical isomorphism defined as $\langle iu, v\rangle_{V^*,V} := (u, v)_V$ (note that the inner product is in V), and its inverse $i^{-1} := j$ is the Riesz map for V. The solution to ($\mathrm{P}_{\mathrm{QVI}}$) is equivalently determined by $y \in V$ satisfying $y = B_\rho(y)$ where

$$B_\rho(y) := \mathrm{P}_{\mathbf{K}(y)}(y - \rho j(A(y) - f))$$

for any $\rho > 0$. In the case where $\Phi(y) \equiv \phi$, we have

$$\|B_\rho(v) - B_\rho(w)\|_V \leq \sqrt{1 - 2\rho c + \rho^2 C^2}\|v - w\|_V.$$

A significant proportion of the literature on QVIs is based on trying to extend this result to the quasi-variational setting. This approach relies on the hard assumption

$$\|\mathrm{P}_{\mathbf{K}(y)}(w) - \mathrm{P}_{\mathbf{K}(z)}(w)\|_V \leq \eta\|y - z\|_V \tag{3}$$

for some $0 < \eta < 1$ and all y, z, w in a bounded set in V (this should not be confused with the non-expansive nature of the map $z \mapsto \mathrm{P}_{\mathbf{K}(y)}(z)$). In general, (3) is **not** valid, and the only framework (in our setting) where it seems to work is in the obstacle-type case with $\Phi \colon V \to V$, in which case the projection map can be rewritten in simpler terms:

$$\mathrm{P}_{\mathbf{K}(y)}(w) = \Phi(y) + \mathrm{P}_{\{z \in V : z \leq 0\}}(w - \Phi(y)). \tag{4}$$

It is necessary for this representation that Φ preserves the V regularity, and for example, if $V = H_0^1(\Omega)$ and $\Phi \colon V \to L^2(\Omega) \setminus H_0^1(\Omega)$, this is no longer valid.

In the case (4) holds, the map B_ρ satisfies

$$\|B_\rho(v) - B_\rho(w)\|_V \leq (2L_\Phi + \sqrt{1 - 2\rho c + \rho^2 C^2})\|v - w\|_V,$$

and in order for this to be contractive, a first observation is that we require

$$\frac{C}{c} L_\Phi < \frac{1}{2}.$$

This is a much more restrictive and convoluted approach than the iterative approach described above where only $CL_\Phi/c < 1$ is required! Furthermore, the linear convergence rate (in case of a contraction) in this case is worse than the contraction approach, given by CL_Φ/c.

The reason why condition (3) fails in a general setting can be answered by a result of Attouch and Wets in [6]. For any closed, non-empty, and convex set $\mathbf{K}$ in V, we define the distance function of an element $y \in V$ to the set $\mathbf{K}$ as

$$\mathrm{d}(y, \mathbf{K}) := \inf_{z \in \mathbf{K}} \|z - y\|_V,$$

and for two closed, non-empty, and convex sets $\mathbf{K}_1, \mathbf{K}_2$, we define the excess function e as

$$\mathrm{e}(\mathbf{K}_1, \mathbf{K}_2) := \sup_{z \in \mathbf{K}_1} \mathrm{d}(z, \mathbf{K}_2).$$

For any $\rho \geq 0$, the ρ-Hausdorff distance between $\mathbf{K}_1$ and $\mathbf{K}_2$ is given by

$$\mathrm{haus}_\rho(\mathbf{K}_1, \mathbf{K}_2) := \sup(\mathrm{e}(\mathbf{K}_1^\rho, \mathbf{K}_2), (\mathrm{e}(\mathbf{K}_2^\rho, \mathbf{K}_1)),$$

where $\mathbf{K}_i^\rho := \mathbf{K}_i \cap \rho B$, $i = 1, 2$, and B is the open unit ball centred at zero.

Theorem 2.5 (Attouch–Wets [6, Proposition 5.3]) *Let V be a Hilbert space and $\mathbf{K}_1, \mathbf{K}_2$ any two closed, convex, non-empty subsets of V. For $y_0 \in V$, we have that*

$$\|\mathrm{P}_{\mathbf{K}_1}(y_0) - \mathrm{P}_{\mathbf{K}_2}(y_0)\|_V \leq \rho^{1/2} \mathrm{haus}_\rho(\mathbf{K}_1, \mathbf{K}_2)^{1/2}$$

for $\rho := \|y_0\| + \mathrm{d}(y_0, \mathbf{K}_1) + \mathrm{d}(y_0, \mathbf{K}_2)$.

The $1/2$ exponent in the right-hand side expression is optimal, and examples (even in finite dimensions) can be found where equality holds. In order to understand how this result fully translates into our class of maps $y \mapsto \mathbf{K}(y)$, consider the following example. Let $\Omega = (0, 1)$ and $V = \{v \in H^1(\Omega) : v(0) = 0\}$ with norm $\|v\|_V^2 := \int_\Omega |v'|^2 \, \mathrm{d}x$. Suppose that $\mathbf{K}_i := \{v \in V : |\nabla v| \leq \phi_i\}$ with $\phi_2 > \phi_1 > 0$ constants. Then, we see from the calculations in [1] that if $\Phi : V \to \mathbb{R}$ is Lipschitz, then we can only obtain

$$\|\mathrm{P}_{\mathbf{K}(y)}(y_0) - \mathrm{P}_{\mathbf{K}(w)}(y_0)\|_V \leq C \|y - w\|_V^{1/2}.$$

3 Sensitivities

Sensitivity analysis of QVIs refers to concepts such as the continuity, stability, and directional differentiability of solution maps associated to the QVI in consideration. Such questions of sensitivity are of prime importance for optimal control. Indeed,

in order to show existence of optimal controls for problems with QVI constraints, arguments related to the direct method of the calculus of variations require continuity of some kind for the control-to-state map (which typically is the map that takes the source term into solution of the QVI). Moreover, one must keep in mind that, as mentioned, solutions to QVIs are generally non-unique; hence, there are (at least) two routes of investigation here:

1. Study sensitivity of a particular selection mechanism that picks a particular solution
2. Study sensitivity of the whole set-valued solution map, e.g., via tools from set-valued analysis

We initially take the first approach and study the sensitivity of the minimal and maximal selection mechanisms (these were introduced in Sect. 2.1 and studied in Sect. 2.2).

3.1 Stability for Minimal and Maximal Solution Maps

We state our fundamental result concerning the behaviour of the maps $f \mapsto \mathsf{m}(f)$ and $f \mapsto \mathsf{M}(f)$ that we obtained in [4]. Given a constant $\nu > 0$, we define the space

$$L^\infty_\nu(\Omega) := \{z \in L^\infty(\Omega) : z \geq \nu \text{ a.e. in } \Omega\}.$$

Theorem 3.1 (Theorem 5 of [4]) *Let $\{f_n\}$ in $L^\infty_\nu(\Omega)$ be such that:*

1. $0 \leq f_n \leq F$ *for some* $F \in V^*$ *with* $F \geq 0$
2. $[\underline{y}, \overline{y}] = [0, A^{-1}F]$
3. $\lim f_n = f^*$ *in* $L^\infty(\Omega)$ *for some* f^*

Suppose also that $\Phi : V \cap H^+ \to H^+$ satisfies the following assumption: if $v_n \rightharpoonup v$ in V, then one of the following holds:

(a) $\Phi(v_n) \to \Phi(v)$ *in* $L^\infty(\Omega)$, *or* $\Phi(v_n) \to \Phi(v)$ *in* V
(b) $\Phi(v_n) \to \Phi(v)$ *in* H, *and if* $v \in V \cap H^+$, *then* $\Phi(v) \in V$ *and* $\mathscr{Q}\Phi(v) \geq 0$ *in* V, *for some strongly monotone* $\mathscr{Q} \in \mathscr{L}(V, V^*)$, *such that* $\langle \mathscr{Q}v^-, v^+\rangle \leq 0$ *for all* $v \in V$

Then if $\lambda\Phi(y) \geq \Phi(\lambda y)$ *for any* $\lambda > 1$ *and* $y \in V \cap H^+$, *we have*

$$\mathsf{m}(f_n) \to \mathsf{m}(f^*) \text{in} H \qquad \text{and} \qquad \mathsf{m}(f_n) \rightharpoonup \mathsf{m}(f^*) \text{in} V. \tag{5}$$

$$\mathsf{M}(f_n) \to \mathsf{M}(f^*) \text{in} H \qquad \text{and} \qquad \mathsf{M}(f_n) \rightharpoonup \mathsf{M}(f^*) \text{in} V. \tag{6}$$

This result states that under some compactness and homogeneity-type assumptions on Φ, the maps m and M are (weakly) continuous, that is, the minimal and maximal solution maps are stable with respect to perturbations in the source term. This is

essential for the existence of optimal control problems related to the maps m and M that we shall discuss later.

3.2 Directional Differentiability

Having studied continuity, in this section, we consider the *differential* stability of the solution map associated to ($\mathrm{P}_{\mathrm{QVI}}$). The framework for such results requires some additional structure on the spaces associated to the QVI that we now state.

Let X be a locally compact topological space, countable at infinity, with ξ a Radon measure on X. Suppose V is a Hilbert space and $H := L^2(X;\xi)$ and $|u| \in V$ whenever $u \in V$, and let $A\colon V \to V^*$ now be linear (in addition to the assumptions previously introduced) and denote by $a\colon V \times V \to \mathbb{R}$ the bilinear form generated by A. We further assume that

$$V \cap C_c(X) \subset C_c(X) \quad \text{and} \quad V \cap C_c(X) \subset V \quad \text{are dense embeddings,} \tag{7}$$

and thus (V, a) is a *regular* form [14, §1.1]. This setting means that we can define capacity, quasi-continuity, and related notions, see [27, §3] and [17, §3]. For some concrete examples of V and A, see [27, §3] and [2, §1.2].

In order to present the differential theory for QVIs, we first recall the corresponding theory for VIs which has been fully investigated in, e.g., [17, 27, 41]. Given an obstacle $\phi \in V_+$, define the set

$$\mathbf{K} := \{w \in V : w \leq \phi\},$$

and given a source term $f \in V^*$, define by $S\colon V^* \to V$ the map $S(f) := S(f, \mathbf{K})$ with the latter defined in (2). The well-known notions of the *tangent cone* and the *critical cone* (from convex analysis) associated to $\mathbf{K}$ are given, respectively, by

$$T_{\mathbf{K}}(y) := \{\varphi \in V : \varphi \leq 0 \text{ q.e. on } \{y = \phi\}\} \text{ and } \mathscr{K}_{\mathbf{K}}(y) := T_{\mathbf{K}}(y) \cap [f - Ay]^{\perp}. \tag{8}$$

The fundamental result of Mignot [28, Theorem 3.3] guarantees directional differentiability of S, and it reads as follows: given $f \in V^*$ and $d \in V^*$, there exists a function $S'(f)(d) \in V$ such that

$$S(f + td) = S(f) + tS'(f)(d) + o(t) \quad \forall t > 0$$

holds where $t^{-1}o(t) \to 0$ as $t \to 0^+$ in V and $\delta := S'(f)(d)$ satisfies the VI

$$\delta \in \mathscr{K}_{\mathbf{K}}(y) : \langle A\delta - d, \delta - v\rangle \leq 0 \quad \forall v \in \mathscr{K}_{\mathbf{K}}(y), \text{ where } y = S(f).$$

Furthermore, the directional derivative $\delta = \delta(d)$ is positively homogeneous in d.

One says that *strict complementarity* holds if the critical cone simplifies to the linear subspace

$$\mathscr{K}_{\mathbf{K}}(y) = \mathscr{S}_{\mathbf{K}}(y) := \{\varphi \in V : \varphi = 0 \text{ q.e. on } \{y = \phi\}\}. \tag{9}$$

In this case, δ satisfies not (just) a VI but a variational *equality* due to the relaxation of constraints on the test functions for the inequality. Formally, strict complementarity arises when the biactive set $\{Ay - f = 0\} \cap \{y - \phi = 0\}$ is empty; see [15, 16, 18] for some technical details regarding biactivity. Under strict complementarity, Mignot showed [27, Theorem 3.4] that the derivative δ above is in fact a Gâteaux derivative: $\delta = \delta(d)$ is linear in d and it satisfies

$$\delta \in \mathscr{S}_{\mathbf{K}}(y) : \langle A\delta - d, v - \delta\rangle = 0 \quad \forall v \in \mathscr{S}_{\mathbf{K}}(y).$$

Let us now return to the QVI setting and try to extend the above results to this case. Let $\Phi : H \to V \subset H$ be increasing with $\Phi(0) \geq 0$. Given $f \in V^*$, consider ($\mathrm{P}_{\mathrm{QVI}}$):

$$y \in \mathbf{K}(y) : \langle Ay - f, y - v\rangle \leq 0 \quad \forall v \in \mathbf{K}(y), \tag{10}$$

where $\mathbf{K}(y) := \{v \in V : v \leq \Phi(y)\}$. We study the map $\mathbf{Q} : V^*_+ \rightrightarrows V$, the set-valued solution map taking $f \mapsto y$. To show that this map is directionally differentiable (in some sense), the first idea that comes to mind is to rewrite (10) by relegating the obstacle onto the source term and then to apply the theory of Mignot. Indeed, the function $\hat{y} := (\mathrm{id} - \Phi)y$ solves

$$\hat{y} \in \mathbf{K}_0 : \langle A(\mathrm{id} - \Phi)^{-1}\hat{y} - f, \hat{y} - \phi\rangle \leq 0 \quad \forall \phi \in \mathbf{K}_0,$$

with $\mathbf{K}_0 := \{w \in V : w \leq 0\}$; however, in general, the operator $A(\mathrm{id} - \Phi)^{-1}$ is no longer linear, nor coercive nor T-monotone, so the VI theory is not applicable and another method is needed.

Our idea in [2] is the following: approximate the solution $q(t) \in \mathbf{Q}(f + td)$ of the perturbed QVI by a sequence $\{q_n(t)\}$ of solutions of VIs, obtain differential formulae for those VIs involving directional derivatives $\{\alpha_n\}$ and remainder terms $\{o_n(t)\}$, and then pass to the limit to obtain an expansion formula relating elements of $\mathbf{Q}(f + td)$ to elements of $\mathbf{Q}(f)$. Some finesse is needed in this procedure in order to handle the issues that arise, which we enumerate here:

1. The **derivation of the expansion formulae** for the above-mentioned VI iterates $q_n(t)$ must relate $q(t)$ to a solution $y \in \mathbf{Q}(f)$, and recursion plays a highly nonlinear role in the relationship between one iterate and the next.
2. **Obtaining uniform bounds on the directional derivatives**; the derivatives satisfy a VI, but this still requires the handling of a recurrence inequality (unless some regularity is available, see [2, §4.3]).

3. **Identifying the limit of the higher order terms as a higher order term**; this step involves the commutation of two limits: one as $t \to 0^+$ and one as $n \to \infty$, and such commutation generally requires uniform convergence.

The iteration scheme mentioned above requires some further restrictions on the data f and the direction d that the derivative is taken in. We take $f \in V_+^*$ and define $\bar{y} \in V$ as the (non-negative) weak solution of the unconstrained problem $A\bar{y} = f$. In a similar fashion, define $\bar{q}(t) \in V$ as the solution of the unconstrained problem with the perturbed right-hand side: $A\bar{q}(t) = f + td$.

Since the mapping Φ plays an inextricable role in defining the obstacle and we are considering sensitivity of QVIs, it is natural that further regularity is required of Φ. These further assumptions will be introduced below where we state the main theorem of [2], but first let us define

$$\mathscr{K}_{\mathbf{K}(y)}(y, \alpha) := \Phi'(y)(\alpha) + \mathscr{K}_{\mathbf{K}(y)}(y),$$

which can be thought of as a *translated* critical cone.

Theorem 3.2 (cf. Theorem 1 of [2]) *Let $f, d \in V_+^*$. Given $y \in \mathbf{Q}(f) \cap [0, \bar{y}]$, assume the following:*

(A1) The map $\Phi : V \to V$ is Hadamard directionally differentiable.[2]
(A2) Either

(A2a) $\Phi : V \to V$ is completely continuous, or
(A2b) $V = H^1(\Omega)$, $X = \overline{\Omega}$, where Ω is a bounded Lipschitz domain, $\Phi : L_+^\infty(\Omega) \to L_+^\infty(\Omega)$ and is concave with $\Phi(0) \geq c > 0$, and $f, d \in L_+^\infty(\Omega)$.[3]

(A3) The map $\Phi'(v) : V \to V$ is completely continuous (for fixed $v \in V$).
(L1) There exists $\epsilon > 0$ such that

$$\left\| \Phi'(z)(v) \right\|_V \leq C_\Phi \|v\|_V \quad \forall z \in B_\epsilon(y),\ \forall v \in V$$

with

$$C_\Phi < \frac{c}{c + C},$$

where C and c are the constants of boundedness and coercivity of A, respectively.

[2] In fact, (A1) can be weakened significantly by requiring Hadamard differentiability of Φ only around the point y, i.e., locally, as in assumption (L1).

[3] In this case, solutions of the QVI (10) are unique [23].

Then, there exists $q(t) \in \mathbf{Q}(f + td) \cap [y, \bar{q}(t)]$ *and* $\alpha = \alpha(d) \in V_+$ *such that*

$$q(t) = y + t\alpha + o(t) \quad \forall t > 0$$

holds where $t^{-1}o(t) \to 0$ *as* $t \to 0^+$ *in* V *and* α *satisfies the QVI*

$$\alpha \in \mathscr{K}_{\mathbf{K}(y)}(y, \alpha) : \langle A\alpha - d, v - \alpha \rangle \geq 0 \quad \forall v \in \mathscr{K}_{\mathbf{K}(y)}(y, \alpha).$$

The directional derivative $\alpha = \alpha(d)$ *is positively homogeneous in* d.

Note that the assumption (L1) depends on the specific function y, that is, it is a *local* condition. It imposes certain restrictions: when Φ is linear, it enforces a smallness condition on the (operator) norm of Φ that enforces uniqueness of solutions of the QVI. However, it does not necessarily rule out the multi-valued setting when Φ is nonlinear.

Remark 3.3 Assumption (L1) in Theorem 3.2 implies the following (used in the proof of [2, Lemma 5.7]): there exists $T_0 > 0$ such that for all n,

$$\left\| \Phi'(y + t\alpha_n + \lambda o_n(t))o_n(t) \right\|_V \leq C_\Phi \left\| o_n(t) \right\|_V \quad \forall t \leq T_0. \tag{11}$$

This is the necessary statement needed to prove that the limit of the higher order terms $o_n(t)$ is a higher order term itself. Let us see why it holds. For convenience, let $C_X := Cc^{-1}$, which is such that $C_\Phi C_X < 1$. Fix an $\epsilon > 0$ and take

$$t \leq \frac{c(1 - C_\Phi C_X)\epsilon}{2 \left\| d \right\|_{V^*}}. \tag{12}$$

The sequence $\{q_n(t)\}$ is defined as follows: set $q_0(t) \equiv y$ and for $n \geq 1$, $q_n(t)$ is the unique solution of the VI

$$q \in \mathbf{K}(q_{n-1}(t)) : \langle Aq - (f + td), q - v \rangle \leq 0 \quad \forall v \in \mathbf{K}(q_{n-1}(t)),$$

where $\mathbf{K}(w) := \{v \in V : v \leq \Phi(w)\}$ is as defined before. Let us show by induction that each $q_n(t)$ lies in a closed ball around y of radius $\epsilon/2$ (with respect to the norm in V) for such t. We have the estimate

$$\left\| q_1(t) - y \right\|_V \leq c^{-1} t \left\| d \right\|_{V^*} \leq \frac{\epsilon}{2},$$

i.e., $q_1(t) \in B_{\epsilon/2}(y)$. Suppose the claim holds for the $(n-1)$th term. Regarding $q_n(t)$, we estimate by testing the inequality for $q_n(t)$ with $y - \Phi(y) + \Phi(q_{n-1}(t))$ and the inequality for y with $q_n(t) - \Phi(q_{n-1}(t)) + \Phi(y)$:

$$\begin{aligned} \langle Aq_n(t) - (f + td), q_n(t) - y + \Phi(y) - \Phi(q_{n-1}(t)) \rangle &\leq 0, \\ \langle Ay - f, y - q_n(t) + \Phi(q_{n-1}(t)) - \Phi(y) \rangle &\leq 0, \end{aligned}$$

whence adding, we obtain

$$\langle A(q_n(t) - y) - td, q_n(t) - y + \Phi(y) - \Phi(q_{n-1}(t))\rangle \leq 0.$$

This leads to

$$\begin{aligned} c\,\|q_n(t) - y\|_V^2 &\leq C\,\|q_n(t) - y\|_V \left\|\Phi(y) - \Phi(q_{n-1}(t))\right\|_V + t\,\|d\|_{V^*}\,\|q_n(t) - y\|_V \\ &\quad + t\langle d, \Phi(y) - \Phi(q_{n-1}(t))\rangle \\ &\leq C\,\|q_n(t) - y\|_V \left\|\Phi(y) - \Phi(q_{n-1}(t))\right\|_V + t\,\|d\|_{V^*}\,\|q_n(t) - y\|_V \end{aligned}$$

since $d \geq 0$ and $y \leq q_{n-1}(t)$ and Φ is increasing, giving the bound

$$\|q_n(t) - y\|_V \leq \frac{C}{c}\,\|\Phi(y) - \Phi(q_{n-1}(t))\|_V + \frac{t}{c}\,\|d\|_{V^*}.$$

Now, recalling C_X and using the mean value theorem [32, §2, Proposition 2.29]

$$\begin{aligned} \|q_n(t) - y\|_V &\leq c^{-1}t\,\|d\|_{V^*} + C_X\left\|\Phi(q_{n-1}(t)) - \Phi(y)\right\|_V \\ &\leq c^{-1}t\,\|d\|_{V^*} + \sup_{\lambda\in(0,1)} C_X\left\|\Phi'(\lambda q_{n-1}(t) + (1-\lambda)y)(q_{n-1}(t) - y)\right\|_V \\ &\leq c^{-1}t\,\|d\|_{V^*} + C_\Phi C_X\left\|q_{n-1}(t) - y\right\|_V \\ &\qquad \text{(applying the assumption thanks to the induction hypothesis)} \\ &\leq c^{-1}t\,\|d\|_{V^*}\,(1 + C_\Phi C_X + (C_\Phi C_X)^2 + \ldots + (C_\Phi C_X)^{n-1}) \\ &\leq \frac{c^{-1}t\,\|d\|_{V^*}}{1 - C_\Phi C_X} \qquad \text{(by the formula for a geometric series)} \\ &\leq \frac{\epsilon}{2}, \end{aligned}$$

and hence, $q_n(t) \in B_{\epsilon/2}(y)$ for all n as long as t satisfies (12). Now the term inside the norm on the left-hand side of the desired inequality (11) is

$$\Phi'(u + t\alpha_n + \lambda o_n(t)) = \Phi'(\lambda q_n(t) + (1-\lambda)(u + t\alpha_n)). \tag{13}$$

Observe that for $\lambda \in (0, 1)$

$$\begin{aligned} \|\lambda q_n(t) + (1-\lambda)(u + t\alpha_n) - u\|_V &= \|\lambda(q_n(t) - u) + (1-\lambda)t\alpha_n\|_V \\ &\leq \frac{\epsilon}{2} + t\,\|\alpha_n\|_V \\ &\leq \frac{\epsilon}{2} + tC^*, \end{aligned}$$

where C^* is the uniform bound (see [2, §5.1]) on $\{\alpha_n\}$. Thus, if t satisfies (12) and satisfies

$$t \leq \frac{\epsilon}{2C^*},$$

then $\lambda q_n(t) + (1-\lambda)(u + t\alpha_n) \in B_\epsilon(u)$ for all n. This implies from (13), using the assumption, that (11) holds as long as

$$t \leq \min\left(\frac{c(1 - C_\Phi C_X)\epsilon}{2\,\|d\|_{V^*}}, \frac{\epsilon}{2C^*}\right). \tag{14}$$

Remark 3.4 The result in the general multi-valued setting given in Theorem 3.2 is a differentiability result for a specific selection mechanism that associates to a function $y \in \mathbf{Q}(f)$ a function $q(t) \in \mathbf{Q}(f + td)$ (the precise mechanism is given in [2, §3.2.1]). A useful variant of the theorem would be to obtain the result for the mapping that selects the minimal or maximal solution to the QVI, i.e., if $\mathsf{M}(f) \in \mathbf{Q}(f)$ is the maximal solution of the QVI with source term f, is M directionally differentiable? A difficulty lies in the approximation scheme we use; in the proof of Theorem 3.2, we chose $q_0 = y$; instead, we could choose $q_0 = y_0$, where $0 \leq y_0 \leq \bar{y}$, which leads to the equality

$$q_n(t) = y_n(t) + t\hat{\alpha}_n + \hat{o}_n(t)$$

where $y_n = S(f, \mathbf{K}(y_{n-1}))$. The main problem is in dealing with the limiting behaviour of the higher order terms $\hat{o}_n(t)$, which now depends on the base point y_n which depends on n. This fact constrains us in this direction. Further details can be found in [2, Remark 3.9].

Under a notion similar to strict complementarity, we obtain a regularity result on the directional derivative. In this setting, we say that *strict complementarity* holds if the set $\mathscr{K}_{\mathbf{K}(y)}(y, w)$ simplifies to

$$\mathscr{K}_{\mathbf{K}(y)}(y, w) = \mathscr{S}_{\mathbf{K}(y)}(y, w) := \{\varphi \in V : \varphi = \Phi'(y)(w) \text{ q.e. on } \{y = \Phi(y)\}\}.$$

Theorem 3.5 (Theorem 2 of [2]) *In the context of Theorem 3.2, if strict complementarity holds, then the derivative α satisfies*

$$\alpha \in \mathscr{S}_{\mathbf{K}(y)}(y, \alpha) : \langle A\alpha - d, \alpha - v\rangle = 0 \quad \forall v \in \mathscr{S}_{\mathbf{K}(y)}(y, \alpha).$$

In this case, if $h \mapsto \Phi'(v)(h)$ is linear, $\alpha = \alpha(d)$ satisfies $\alpha(c_1 d_1 + c_2 d_2) = c_1\alpha(d_1) + c_2\alpha(d_2)$ for constants $c_1, c_2 > 0$ and directions $d_1, d_2 \in V_+^*$.

It is worth restating Theorem 3.2 in the case when $\mathbf{Q}\colon V_+^* \rightrightarrows V$ is single-valued (i.e., the QVI problem has a unique solution).

Theorem 3.6 *Suppose* $\mathbf{Q}$ *is single-valued and let the hypotheses of Theorem 3.2 hold given* $f, d \in V_+^*$. *There exists a function* $\mathbf{Q}'(f)(d) \in V_+$ *such that*

$$\mathbf{Q}(f + td) = \mathbf{Q}(f) + t\mathbf{Q}'(f)(d) + o(t) \quad \forall t > 0$$

holds where $t^{-1}o(t) \to 0$ *as* $t \to 0^+$ *in* V *and* $\mathbf{Q}'(f)(d)$ *satisfies the QVI given in Theorem 3.2.*

We of course recover the results of Mignot in [27] in the case where Φ is a constant mapping (the VI setting).

3.3 Parabolic QVIs

The authors have recently studied the above issues for the case of parabolic QVIs in [3]. There, we consider time-dependent QVIs of the form

$$\text{find } z : z(t) \leq \Phi(z)(t) : \int_0^T \langle z'(t) + Az(t) - f(t), z(t) - v(t) \rangle \leq 0 \quad \forall v : v(t) \leq \Phi(z)(t)$$

and show that solutions exist in certain Bochner spaces under appropriate assumptions. These solutions have been shown to be given as a limit related to elliptic QVIs arising from the time discretisation of the problem [3, Theorem 2.9], or as limit of solutions of parabolic VIs [3, Theorems 3.8 and 3.10]. Directional differentiability is also proved [3, Theorem 5.15] in much the same way as in the elliptic case using the recently obtained results in [12] on the directional differentiability of parabolic VIs. More details can be found in [3] (we do not elaborate here for reasons of space).

3.4 Application to Thermoforming

We present an application of QVIs to thermoforming that was initially proposed by the authors in [2]. Thermoforming aims to manufacture products by heating a membrane or plastic sheet to its pliable temperature and then forcing the membrane onto a mould, commonly made of aluminium or an alloy of aluminium, which deforms the membrane and enables it to take on the shape of the mould. The process is applied to create both large structures (such as car panels) and microscopic products (such as microfluidic structures). Research into the modelling and accurate numerical simulation of thermoforming can be seen in [21] and [40].

The contact problem associated with the heated membrane and the mould can be described through a VI problem (assuming perfect sliding of the membrane with the mould as in [5]). However, a complex physical phenomenon occurs when the heated sheet is forced into contact with the mould: the mould is not at the

same temperature as the plastic sheet (it might be relatively cold with respect to membrane) and this triggers heat transfer with hard-to-predict consequences (e.g., it changes the polymer viscosity, see, for example, [24]). In practice, the thickness of the thermoformed piece can be controlled locally by the mould structure and its initial temperature distribution [24], and the non-uniform temperature distribution of the polymer sheet has a substantial effect on the results [29].

The size of the common mould material aluminium is highly sensitive to heat fluctuations; aluminium has a relatively high thermal expansion volumetric coefficient and this implies that there is a dynamic change in the mould (the obstacle) as the polymer sheet is forced in contact with it. This determines a compliant obstacle-type problem like in [30], and hence, the overall process is a QVI with underlying nonlinear PDEs determining the heat transfer and the volume change in the obstacle. In what follows, we consider this compliant obstacle behaviour whilst simultaneously making various simplifying assumptions in order to study a basic but nevertheless meaningful model.

3.4.1 The Model

We restrict the analysis to the 1D case for the sake of simplicity, but we provide 2D numerical tests. Let $\Phi_0 \colon [0,1] \to \mathbb{R}^+$ be the (parametrised) mould shape that we wish to reproduce through a sheet (or membrane). The membrane lies below the mould and is pushed upwards through some process f (such as vacuum and/or air pressure). We make the following simplifying physical assumptions:

1. The temperature for the membrane is a prescribed constant.
2. The mould grows affinely with respect to changes in its temperature.
3. The temperature of the mould is subject to diffusion, convection, and boundary conditions arising from the insulated boundary, and it depends on its vertical distance to the membrane.

The thermoforming process is a time evolution system, but the setting described by the previous assumptions is appropriate for one time step in the time semi-discretisation.

We denote the position of the mould and membrane by $\Phi(u)$ and u, respectively, and T will stand for the temperature of the mould. Let us define the spaces $W = H^1(0,1)$ and $H = L^2(0,1)$, and let either

$$A = -\Delta_N + I \text{ and } V = H^1(0,1) \quad \text{or} \quad A = -\Delta_D \text{ and } V = H_0^1(0,1)$$

in the case of Neumann or Dirichlet boundary conditions,[4] respectively, for the membrane u. The system we consider is the following:

$$u \in V : u \leq \Phi(u), \quad \langle Au - f, u - v\rangle \leq 0 \quad \forall v \in V : v \leq \Phi(u) \tag{15}$$

$$kT - \Delta T = g(\Phi(u) - u) \qquad \text{on } [0, 1] \tag{16}$$

$$\partial_\nu T = 0 \qquad \text{on } \{0, 1\} \tag{17}$$

$$\Phi(u) = \Phi_0 + LT, \qquad \text{on } [0, 1] \tag{18}$$

where $f \in H_+$ is given, $k > 0$ is a constant, $\Phi_0 \in V$, $L \colon W \to V$ is a bounded linear operator such that

$$\text{for every } \Omega_0 \subset \Omega, \text{ if } u \leq v \text{ a.e. on } \Omega_0 \text{ then } Lu \leq Lv \text{ a.e. on } \Omega_0,$$

and $g \colon \mathbb{R} \to \mathbb{R}$ is decreasing and C^2 with $g(0) = M > 0$ a constant, $0 \leq g \leq M$ and g' bounded. Thus when the membrane and mould are in contact or are close to each other, there is a maximum level of heat transfer onto the mould, whilst when they are sufficiently separated, there is no heat exchange. An example of g is a smoothing of the function

$$G(r) = \begin{cases} 1 & : \text{if } r \leq 0 \\ 1 - r & : \text{if } 0 < r < 1 \\ 0 & : \text{if } r \geq 1. \end{cases} \tag{19}$$

Note that the local increasing property of L stated above is equivalent to

$$vLv \geq 0 \text{ a.e. for all } v \in W. \tag{20}$$

The system above is derived as follows: consideration of the potential energy of the membrane will show that [5] u solves the VI (15) with the QVI nature arising from assuming that heat transfer occurs between the membrane and the mould. If we let $\hat{T} \colon \Gamma \to \mathbb{R}$ be the temperature of the mould defined on the curve

$$\Gamma := \{(r, \Phi(u)(r)) : r \in [0, 1]\} \subset \mathbb{R}^2$$

[4] Zero Dirichlet conditions arise from clamping the membrane at its ends.

(a 1D hypersurface in 2D), our modelling assumptions directly imply that

$$k\hat{T}(x) - \Delta_\Gamma \hat{T}(x) = g(x_2 - u(x_1)) \quad \text{for } x = (r, \Phi(u)(r)) \in \Gamma, \tag{21}$$

where the notation x_i means the ith component of x. We reparametrise by $T(r) = \hat{T}(r, \Phi(r))$ and simplify (21) to obtain (16).

3.4.2 Properties and Existence for the System

We now show that the system above has a solution, and we check that the QVI in the system fits into the framework described in Sects. 2 and 3, which in particular requires us to check a number of assumptions on Φ.

First, plugging (18) into (16), we obtain

$$\begin{aligned} kT - \Delta T &= g(LT + \Phi_0 - u) && \text{on } [0, 1], \\ \partial_\nu T &= 0 && \text{on } \{0, 1\}. \end{aligned} \tag{22}$$

Monotonicity properties of the right-hand side allow us to prove that for every $u \in H$, there exists a unique solution $T \in W$ to Eq. (22). From (18), this then implies that $\Phi : H \to V$.

Lemma 3.7 *It holds that $\Phi(0) \geq 0$ a.e.*

Proof Note that $\Phi(0) = \Phi_0 + LT|_{u=0} =: \Phi_0 + LT_0$, where

$$\begin{aligned} kT_0 - \Delta T_0 &= g(LT_0 + \Phi_0) && \text{on } [0, 1], \\ \partial_\nu T_0 &= 0 && \text{on } \{0, 1\}. \end{aligned}$$

Test this equation with T_0^- and use the sign on g to obtain $T_0 \geq 0$. The claim follows by the local increasing property of L. □

Lemma 3.8 *The map $\Phi : H \to H$ is increasing.*

Proof Since $\Phi(u) = LT(u) + \Phi_0$, it suffices to show that $u \mapsto T(u)$ is increasing. Take the solutions T_1 and T_2 of Eq. (22) corresponding to $u = u_1 \in H$ and $u = u_2 \in H$ with $u_1 \leq u_2$, take the difference of the equations and test with $(T_1 - T_2)^+$:

$$\begin{aligned} &\int k|(T_1 - T_2)^+|^2 + |\nabla(T_1 - T_2)^+|^2 \\ &\quad = \int (g(LT_1 + \Phi_0 - u_1) - g(LT_2 + \Phi_0 - u_2))(T_1 - T_2)^+ \\ &\quad = \int_{\{T_1 \geq T_2\}} (g(LT_1 + \Phi_0 - u_1) - g(LT_2 + \Phi_0 - u_2))(T_1 - T_2). \end{aligned}$$

On the area of integration, $LT_1 \geq LT_2$ and thus $LT_1 + \Phi_0 - u_1 \geq LT_2 + \Phi_0 - u_2$ pointwise a.e. Since g is decreasing, $g(LT_1+\Phi_0-u_1)-g(LT_2+\Phi_0-u_2) \leq 0$, and hence, the above integral is non-positive. Therefore, $(T_1 - T_2)^+ = 0$ in H giving $T_1 \leq T_2$ on Ω. Applying L to both sides and using the increasing property, we find the result. □

Theorem 3.9 (Theorem 7 of [2]) *There exists a solution $(u, T, \Phi(u))$ to the system (15), (16), (17), (18).*

Proof By Lemmas 3.7 and 3.8, we see that 0 is a subsolution and $A^{-1}f$ is a supersolution (since it is assumed that $f \geq 0$) for the map associated to the QVI (15), and thus by the Tartar–Birkhoff result, there exists a solution u to (15). This then uniquely determines $T(u)$ and thus $\Phi(u)$, and consequently, (16) has a solution T. □

Now, before we discuss compactness of Φ, let us give the following continuous dependence result for two solutions T_1, T_2 corresponding to data u_1, u_2:

$$\begin{aligned}&\min(k, 1)\, \|T_1 - T_2\|_W^2 \\ &\quad \leq \mathrm{Lip}(g) \left(\|L\|_{\mathscr{L}(W,H)}\, \|T_1 - T_2\|_W^2 + \|u_1 - u_2\|_H\, \|T_1 - T_2\|_H \right). \end{aligned} \tag{23}$$

We use the hypothesis

$$\mathrm{Lip}(g)\, \|L\|_{\mathscr{L}(W,H)} < \min(1, k) \tag{24}$$

at various points.

Lemma 3.10 *If (24) holds, $\Phi\colon V \to V$ is completely continuous.*

Proof Suppose that $u_n \rightharpoonup u$ in V and consider the solutions T_n and T corresponding to data u_n and u, respectively. The estimate (23) implies

$$\min(k, 1)\, \|T_n - T\|_W^2 \leq \mathrm{Lip}(g) \left(\|L\|_{\mathscr{L}(W,H)}\, \|T_n - T\|_W^2 + \|u - u_n\|_H\, \|T_n - T\|_H \right).$$

Thus under the condition in the lemma, we can move the first term on the RHS onto the LHS, divide by $\|T_n - T\|$, and then take the limit to see that $T_n \to T$ in W and by continuity of $L\colon W \to V$ that $\Phi(u_n) \to \Phi(u)$ in V. □

Theorem 3.11 (Theorem 8 of [2]) *If g'' is bounded from above, $\Phi\colon V \to V$ is Fréchet differentiable at a solution u given by Theorem 3.9. Furthermore, $\Phi'(u)(d) := -L\delta$, where δ satisfies the PDE*

$$(k - \Delta)\delta - g'(\Phi(u) - u)L\delta = g'(\Phi(u) - u)d.$$

The idea of the proof (which we skip here) is to apply the implicit function theorem to the map $\mathscr{F}\colon V \times W \to W^*$ defined by

$$\langle \mathscr{F}(u,T),\varphi\rangle_{W^*,W} = k\int T\varphi + \int \nabla T\nabla\varphi - \int g(LT+\Phi_0-u)\varphi.$$

Corollary 3.12 *Let (24) hold. Then assumptions (A1), (A2a), and (A3) are satisfied. If also*

$$\min(1,k)^{-1}\,\|L\|_{\mathscr{L}(W,V)}\,\|g'\|_\infty < \frac{1}{2},$$

then (L1) is satisfied.

Proof It remains for us to show the latter two assumptions. Let us see why the mapping $d \mapsto \Phi'(u)(d)$ is completely continuous. Let $d_n \rightharpoonup d$ in V. Using continuous dependence, we find

$$\min(1,k)\,\|\delta_n-\delta_m\|_W \le \|g'\|_\infty\,\|d_n-d_m\|_H\,,$$

and thus δ_n converges strongly in W.

Take $b \in V$ and $h\colon (0,T)\to V$ a higher order term. Then by boundedness of L,

$$\|\Phi'(u+tb+\lambda h(t))(h(t))\|_V \le C\min(1,k)^{-1}\,\|g'\|_\infty\,\|h(t)\|_H\,,$$

which vanishes in the limit after division by t. By the previous theorem, we see that

$$\|\Phi'(z)(d)\|_V \le \|L\|_{\mathscr{L}(W,V)}\,\|\delta\|_W \le \min(1,k)^{-1}\,\|L\|_{\mathscr{L}(W,V)}\,\|g'\|_\infty\,\|d\|_H\,,$$

which by assumption leads to (L1). □

3.4.3 Numerical Implementation Details

We simulate (15)–(18) on the 2D domain $[0,1]\times[0,1]$ with homogeneous Dirichlet conditions for the QVI. We approximate the QVI (15) by a penalised equation and numerically solve the system

$$\begin{aligned} Au + \alpha\max(0,u-y) - f &= 0\\ kT - \Delta T - g(y-u) &= 0\\ \partial_\nu T &= 0\\ y-\Phi_0-LT &= 0 \end{aligned} \tag{25}$$

for a large parameter α (as $\alpha \to \infty$, the solution of (25) converges to the solution of (15)–(18)). We use a finite difference scheme with N^2 uniformly distributed nodes and meshsize $h = 1/(N+1)$ with $N = 256$. The system (25) is discretised and solved via a semismooth Newton method applied to the mapping

$$\mathscr{F}\colon V \times W \times V \to V^* \times W^* \times V, \qquad (u, T, y) \mapsto \mathscr{F}(u, T, y)$$

defined by the left-hand side of (25). For this purpose, the derivative

$$\mathscr{F}'(u, T, y)(v, \tau, z) = \begin{pmatrix} Av + \alpha \max'(0, u - y)(v - z) \\ k\tau - \Delta\tau - g'(y - u)(z - v) \\ \partial_v \tau \\ z - L\tau \end{pmatrix}$$

is needed in order to obtain the next iterate through the Newton update scheme. Here, $\max'(0, u-y)$ denotes the Newton derivative of the maximum function, given by §8.3 in [20] as

$$\max'(0, u) = \begin{cases} 0 & \text{if } u < 0 \\ \delta_N & \text{if } u = 0 \\ 1 & \text{if } u > 0, \end{cases}$$

where $\delta_N \in [0, 1]$ is arbitrary; we pick $\delta_N = 0.1$. An approximation to the directional derivative of the QVI solution mapping was computed by first smoothing the nonlinearity in the first equation of (25) by a function $\max_g$ and differentiating it with respect to f in a direction d:

$$Au'(f)(d) + \alpha \max_g'(0, u - y)u'(f)(d) - d = 0.$$

This equation was solved and was checked to be within a tolerance of 10^{-4} of the difference quotient

$$\frac{u(f + 10^{-5}d) - u(f)}{10^{-5}}.$$

We take the source term $f \equiv 10^2$. The nonlinearity g appearing in the source term for the T equation is selected as a piecewise smooth smoothing of (19), see Fig. 1.

The operator L is chosen as the superposition mapping

$$(Lv)(x) = 5.25 \times 10^{-3} \rho(x) v(x) =: C_L \rho(x) v(x),$$

where $\rho\colon [0, 1] \times [0, 1] \to \mathbb{R}$ is a smooth bump function with $\|\rho\|_\infty = 1$ and $\|\nabla\rho\|_\infty \leq \sqrt{50}$. Let us check the condition of Corollary 3.12 that assures (L1). We

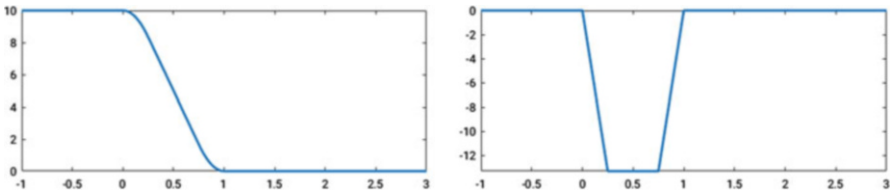

Fig. 1 Plot of the function g and its derivative

see that, given our choice of g and k,

$$\min(1,k)^{-1}\,\|L\|_{\mathscr{L}(W,V)}\,\|g'\|_\infty = \frac{40}{3}\,\|L\|_{\mathscr{L}(W,V)}\,,$$

where the operator norm on the right-hand side can be estimated by the calculation

$$\|Lv\|_V^2 = \int |C_L\varphi|^2|v|^2 + |\nabla(C_L\varphi v)|^2 \le C_L^2(\|\varphi\|_\infty^2 + \|\nabla\varphi\|_\infty^2)\,\|v\|_W^2\,,$$

so that the right-hand side does not exceed $40C_L\sqrt{51}/3$, and hence, assumption (L1) holds if $C_L < 3/80\sqrt{51} \approx 0.00525$. As for the "initial" mould Φ_0, we define $w\colon [0,1] \to \mathbb{R}$ by

$$w(r) = \begin{cases} 5(r/N - 1/10) & \text{if } N/10 \le r \le 3N/10 \\ 1 & \text{if } 3N/10 < r < 7N/10 \\ 1 - 5(r/N - 7/10) & \text{if } 7N/10 \le r \le 9N/10 \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

and set $\Phi_0(r,t) = w(r)w(t)$, see Fig. 2.

The directional derivative is taken in the direction χ_A, the characteristic function of the set $A = \{(x,y) : x > 1/2\}$. The remaining parameters appearing in the physical model are $k = 1$, $\alpha = 10^8$, $\kappa = 10$, and $s = 1$. The initial iterate $(u_h^0, T_h^0, y_h^0) = (0.9 \times \Phi_0, 0.2, 10)$ is used. The Newton iterates, (u_h^j, T_h^j, y_h^j), were assumed to converge if $\|\mathscr{F}(u_h^j, T_h^j, y_h^j)\|_{L^2} < 4 \times 10^{-9}$, for some j, and we denote the solution as (u_h, T_h, y_h).

3.4.4 Numerical Results

See Table 1 for the numerical results. The third column in the table refers to the error of the approximate solution (u_h, T_h, y_h) in that it measures the L^2 norm of $\mathscr{F}(u_h, T_h, y_h)$, and likewise for the fourth column. One can see that a relatively

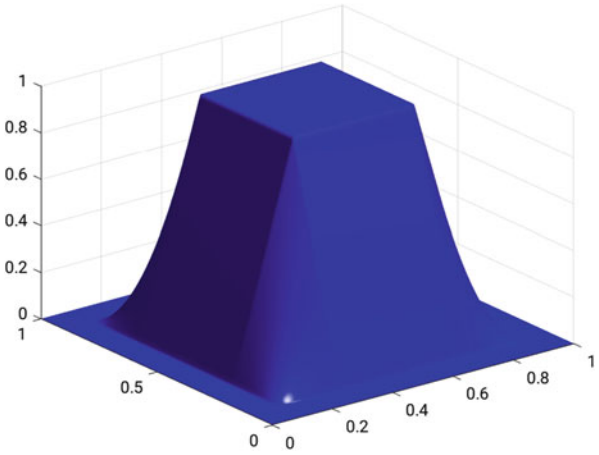

Fig. 2 The initial mould Φ_0

Table 1 Numerical results

No. of nodes	# Newton iterations to solve system (25)	L^2 error in solution of system	L^2 error in solution of derivative
256	14	3.96×10^{-9}	1.96×10^{-15}

low number of Newton iterations are performed to obtain an accurate solution. The results of the experiment are visualised in Fig. 3:

- The effect of the temperature interplay between the membrane and the mould can be immediately seen: Φ_0 (Fig. 2) grows and becomes more curved and smoothed out, which is natural given that the membrane is initially placed below the mould and is pushed upwards.
- The model produces a membrane u that appears to be a good fit for the thermoforming process; it can be observed to be similar to the final mould, which is confirmed by the images of the coincidence sets.
- The directional derivative is coloured yellow and red; red refers to the parts of the domain corresponding to the coincidence set $\{u = y\}$.

4 Control of QVIs

In applications, one is typically interested in confining the solution set $\mathbf{Q}(f)$ to a certain interval $[\underline{y}, \overline{y}]$ for some given $\underline{y}, \overline{y} \in H$. For example, consider again the thermoforming application of Sect. 3.4. Here, one may want to control the heating in order to produce a mould shape that is close to a desired mould but within some threshold of acceptability.

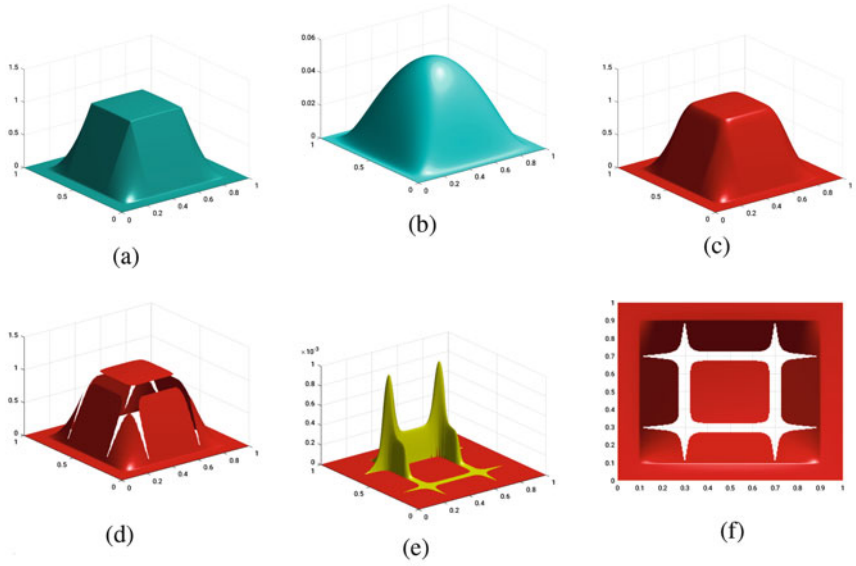

Fig. 3 Computation results. (**a**) Final mould y. (**b**) Difference between y and Φ_0. (**c**) Membrane u. (**d**) Membrane u on the coincidence set. (**e**) The directional derivative. (**f**) Top-down view of the coincidence set

Given a control force f, we consider the following optimal control problem:

$$\text{Minimise } J(\mathbf{O}, f) := J_1(T_{\sup}(\mathbf{O}), T_{\inf}(\mathbf{O})) + J_2(f) \text{ over } (\mathbf{O}, f) \in 2^H \times U,$$
$$\text{subject to } f \in U_{\text{ad}} \text{ and } y \in \mathbf{O} := \{z \in V : z \text{ solves } \mathrm{P_{QVI}}\}. \tag{$\mathbb{P}$}$$

Here, $U_{\text{ad}} \subset U \subset V^*$ is the set of admissible controls where U is a given Hilbert space, $J_1 \colon H \times H \to \mathbb{R}$ and $J_2 : U \to \mathbb{R}$, and for $\underline{y}, \overline{y} \in H$, we have the set-valued map

$$T_{\sup}(\mathbf{O}) := \begin{cases} \sup_{z \in \mathbf{O} \cap [\underline{y}, \overline{y}]} z, & \mathbf{O} \cap [\underline{y}, \overline{y}] \neq \emptyset, \\ \underline{y}, & \text{otherwise,} \end{cases}$$

and analogously

$$T_{\inf}(\mathbf{O}) := \begin{cases} \inf_{z \in \mathbf{O} \cap [\underline{y}, \overline{y}]} z, & \mathbf{O} \cap [\underline{y}, \overline{y}] \neq \emptyset, \\ \overline{y}, & \text{otherwise.} \end{cases}$$

Problems of type ($\mathbb{P}$) had not yet been considered in the literature before us, and they pose several formidable challenges. For instance, the existence of solutions is

highly delicate due to the dependence $y \mapsto \mathbf{K}(y)$ and the fact that $y = y(f)$. As a consequence, the direct method of the calculus of variations is only applicable if certain convergence properties of the constraint set can be guaranteed. Another delicacy is related to the potential set-valuedness of the solution of the QVI in the constraint system of ($\mathbb{P}$). This fact necessitates the identification of a suitable selection mechanism such as the maximal or minimal solution, if available at all. We note, however, that in the special case where $T_{\text{inf}}(\mathbf{Q}(f))$ and $T_{\text{sup}}(\mathbf{Q}(f))$ also belong to $\mathbf{Q}(f)$, they are the minimal and maximal solutions, respectively, to ($\mathrm{P}_{\mathrm{QVI}}$) in $V \cap [\underline{y}, \overline{y}]$. Then, the proof of existence of solutions to ($\mathbb{P}$) reduces to a stability result for this minimal and maximal solution to the QVI of interest. With the aid of Sect. 2, we can now formulate the result that proves the well-posedness of ($\tilde{\mathbb{P}}$). We assume that $U \subset L^\infty(\Omega)$ and in particular that

$$U_{\text{ad}} \subset \{f \in L^\infty_\nu(\Omega) : f \leq F\},$$

for some $F \in V^*$. As in previous sections $\underline{y} = 0$ and $\overline{y} = A^{-1}F$, so that $\mathsf{m}(f)$ and $\mathsf{M}(f)$ are defined as the minimal and maximal solutions, respectively, of the QVI in ($\mathrm{P}_{\mathrm{QVI}}$). Hence, the reduced version of ($\mathbb{P}$) is given by

$$\begin{aligned} &\text{minimise } J_1(\mathsf{m}(f), \mathsf{M}(f)) + J_2(f), \\ &\text{subject to } f \in U_{\text{ad}}. \end{aligned} \tag{$\tilde{\mathbb{P}}$}$$

The existence to ($\tilde{\mathbb{P}}$) (and hence, ($\mathbb{P}$)) is now shown in the following result, which, given Theorem 3.1, is just an application of the direct method of the calculus of variations.

Theorem 4.1 (Theorem 6 of [4]) *Suppose that $J_1 : V \times V \to \mathbb{R}$ is weakly lower semicontinuous and $J_2 : L^\infty(\Omega) \to \mathbb{R}$ is continuous, and both are bounded from below. In addition, suppose that for each $\alpha > 0$ the set*

$$\{f \in U_{\text{ad}} : J_2(f) \leq \alpha\}$$

is sequentially compact in $L^\infty(\Omega)$, and that Φ satisfies the assumptions of Theorem 3.1. Then, problem ($\tilde{\mathbb{P}}$) and (hence) problem ($\mathbb{P}$) admit solutions.

5 Outlook

A natural next step is to study the derivation of stationarity conditions for optimal control problems with QVI constraints, and indeed, this is an ongoing work by the authors. When trying to obtain strong stationarity conditions, the requirement of a signed source term and signed direction in the differentiability result of Sect. 3 implies that the associated optimal control problem necessarily has constraints on

the control, and it is known [39] that such conditions cannot be obtained in cases where the admissible control set does not satisfy some requirements that do not hold in this case. Therefore, the differentiability result has to be extended in some sense to cover more general source terms and directions, and as mentioned, this is a topic of work under preparation.

One may also then study optimal control problems related to parabolic QVIs making use of our work in [3].

References

1. Amal Alphonse, Michael Hintermüller, and Carlos N. Rautenberg. Recent trends and views on elliptic quasi-variational inequalities. Preprint SPP1962-071, 9 2018.
2. Amal Alphonse, Michael Hintermüller, and Carlos N. Rautenberg. Directional differentiability for elliptic quasi-variational inequalities of obstacle type. *Calc. Var. Partial Differential Equations*, 58(1):Art. 39, 47, 2019.
3. Amal Alphonse, Michael Hintermüller, and Carlos N. Rautenberg. Existence, iteration procedures and directional differentiability for parabolic QVIs. *arXiv e-prints*, page arXiv:1904.13230, Apr 2019.
4. Amal Alphonse, Michael Hintermüller, and Carlos N. Rautenberg. Stability of the solution set of quasi-variational inequalities and optimal control. Preprint SPP1962-107, 3 2019.
5. H Andrä, M. K. Warby, and J. R. Whiteman. Contact problems of hyperelastic membranes: Existence theory. *Mathematical Methods in the Applied Sciences*, 23:865–895, 2000.
6. Hédy Attouch and Roger J-B Wets. Quantitative stability of variational systems ii. A framework for nonlinear conditioning. *SIAM Journal on Optimization*, 3(2):359–381, 1993.
7. Jean-Pierre Aubin. *Mathematical methods of game and economic theory*, volume 7 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam-New York, 1979.
8. John W. Barrett and Leonid Prigozhin. A quasi-variational inequality problem in superconductivity. *Math. Models Methods Appl. Sci.*, 20(5):679–706, 2010.
9. John W. Barrett and Leonid Prigozhin. A quasi-variational inequality problem arising in the modeling of growing sandpiles. *ESAIM Math. Model. Numer. Anal.*, 47(4):1133–1165, 2013.
10. John W. Barrett and Leonid Prigozhin. Lakes and rivers in the landscape: a quasi-variational inequality approach. *Interfaces Free Bound.*, 16(2):269–296, 2014.
11. Alain Bensoussan, Maurice Goursat, and Jacques-Louis Lions. Contrôle impulsionnel et inéquations quasi-variationnelles stationnaires. *C. R. Acad. Sci. Paris Sér. A-B*, 276:A1279–A1284, 1973.
12. Constantin Christof. Sensitivity analysis and optimal control of obstacle-type evolution variational inequalities. *SIAM J. Control Optim.*, 57(1):192–218, 2019.
13. Francisco Facchinei and Christian Kanzow. Generalized Nash equilibrium problems. *4OR*, 5(3):173–210, Sep 2007.
14. Masatoshi Fukushima, Yoichi Oshima, and Masayoshi Takeda. *Dirichlet forms and symmetric Markov processes*, volume 19 of *De Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, extended edition, 2011.
15. A. Gaevskaya, M. Hintermüller, R. H. W. Hoppe, and C. Löbhard. Adaptive finite elements for optimally controlled elliptic variational inequalities of obstacle type. In *Optimization with PDE constraints*, volume 101 of *Lect. Notes Comput. Sci. Eng.*, pages 95–150. Springer, Cham, 2014.
16. Alexandra Gaevskaya. *Adaptive finite elements for optimally controlled elliptic variational inequalities of obstacle type*. PhD Thesis, Universität Augsburg, 2013.

17. A Haraux. How to differentiate the projection on a convex set in Hilbert space. Some applications to variational inequalities. *J. Math. Soc. Japan*, 29(4):615–631, 1977.
18. Felix Harder and Gerd Wachsmuth. Comparison of optimality systems for the optimal control of the obstacle problem. *GAMM-Mitt.*, 40(4):312–338, 2018.
19. Patrick T. Harker. Generalized Nash games and quasi-variational inequalities. *European Journal of Operational Research*, 54(1):81–94, 1991.
20. K. Ito and K. Kunisch. *Lagrange Multiplier Approach to Variational Problems and Applications*. SIAM, 2008.
21. W.-G. Jiang, M. K. Warby, J. R. Whiteman, S. Abbot, W. Shorter, P Warwick, T. Wright, A. Munro, and B. Munro. Finite element modelling of high air pressure forming processes for polymer sheets. *Computational Mechanics*, 31:163–172, 2001.
22. M. Kunze and J. F. Rodrigues. An elliptic quasi-variational inequality with gradient constraints and some of its applications. *Math. Methods Appl. Sci.*, 23(10):897–908, 2000.
23. T. H. Laetsch. A uniqueness theorem for elliptic q. v. i. *J. Functional Analysis*, 18:286–288, 1975.
24. J. K. Lee, T. L Virkler, and C. E. Scott. Effects of rheological properties and processing parameters on abs thermoforming. *Polymer Engineering and Science*, 41(2):240–261, 2001.
25. J.-L. Lions. Sur le côntrole optimal des systemes distribuées. *Enseigne*, 19:125–166, 1973.
26. J.-P. Lions and G. Stampacchia. Variational inequalities. *Commun. Pure Appl. Math.*, 20:493–519, 1967.
27. F. Mignot. Contrôle dans les inéquations variationelles elliptiques. *J. Functional Analysis*, 22(2):130–185, 1976.
28. F. Mignot and J. P. Puel. Inéquations variationelles et quasi variationelles hyperboliques du premier ordre. *J. Math Pures Appl.*, 55:353–378, 1976.
29. G. J. Nam, K. H. Ahn, and J. W. Lee. Three-dimensional simulation of thermoforming process and its comparison with experiments. *Polymer Engineering and Science*, 40(10):2232–2240, 2000.
30. J. Outrata, M. Kocvara, and J. Zowe. *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*. Kluwer Academic Publishers, 1998.
31. Jong-Shi Pang and Masao Fukushima. Quasi-variational inequalities, generalized Nash equilibria, and multi-leader-follower games. *Computational Management Science*, 2(1):21–56, Jan 2005.
32. Jean-Paul Penot. *Calculus without derivatives*, volume 266 of *Graduate Texts in Mathematics*. Springer, New York, 2013.
33. L. Prigozhin. Sandpiles and river networks: extended systems with non-local interactions. *Phys. Rev. E*, 49:1161–1167, 1994.
34. L. Prigozhin. On the Bean critical-state model in superconductivity. *European Journal of Applied Mathematics*, 7:237–247, 1996.
35. L. Prigozhin. Sandpiles, river networks, and type-ii superconductors. *Free Boundary Problems News*, 10:2–4, 1996.
36. Leonid Prigozhin. Variational model of sandpile growth. *European J. Appl. Math.*, 7(3):225–235, 1996.
37. José Francisco Rodrigues and Lisa Santos. A parabolic quasi-variational inequality arising in a superconductivity model. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)*, 29(1):153–169, 2000.
38. Luc Tartar. Inéquations quasi variationnelles abstraites. *CR Acad. Sci. Paris Sér. A*, 278:1193–1196, 1974.
39. Gerd Wachsmuth. Strong stationarity for optimal control of the obstacle problem with control constraints. *SIAM Journal on Optimization*, 24(4):1914–1932, 2014.
40. M. K. Warby, J. R. Whiteman, W.-G. Jiang, P Warwick, and Wright T. Finite element simulation of thermoforming processes for polymer sheets. *Mathematics and Computers in Simulation*, 61:209–218, 2003.
41. Eduardo H Zarantonello. Projections on Convex Sets in Hilbert Space and Spectral Theory. *Contributions to Nonlinear Functional Analysis*, pages 237–424, 1971.

Simulation and Control of a Nonsmooth Cahn–Hilliard Navier–Stokes System with Variable Fluid Densities

Carmen Gräßle, Michael Hintermüller, Michael Hinze, and Tobias Keil

Abstract We are concerned with the simulation and control of a two-phase flow model governed by a coupled Cahn–Hilliard Navier–Stokes system involving a nonsmooth energy potential. We establish the existence of optimal solutions and present two distinct approaches to derive suitable stationarity conditions for the bilevel problem, namely C- and strong stationarity. Moreover, we demonstrate the numerical realization of these concepts at the hands of two adaptive solution algorithms relying on a specifically developed goal-oriented error estimator. In addition, we present a model order reduction approach using proper orthogonal decomposition (POD-MOR) in order to replace high-fidelity models by low-order surrogates. In particular, we combine POD with space-adapted snapshots and address the challenges that are the consideration of snapshots with different spatial resolutions and the conservation of a solenoidal property.

Keywords Two-phase flow · POD model order reduction · Adaptivity · Nonsmooth systems · Mathematical programming with equilibrium constraints · Optimal control

Mathematics Subject Classification (2020) 35B65, 35J87, 35K55, 65K15, 49K20

C. Gräßle
University of Hamburg, Hamburg, Germany
e-mail: carmen.graessle@uni-hamburg.de; carmen.graessle@uni-hamburg.de

M. Hintermüller (✉)
Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany
e-mail: michael.hintermueller@wias-berlin.de

M. Hinze
University of Koblenz-Landau, Koblenz, Germany
e-mail: hinze@uni-koblenz.de

T. Keil
Weierstraß-Institut, Berlin, Germany
e-mail: tobias.keil@wias-berlin.de

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_9

1 Introduction

We consider the simulation and control for multiphase flows governed by a Cahn–Hilliard Navier–Stokes (CHNS) system with nonsmooth homogeneous free energy densities utilizing a diffuse interface approach. The free energy is a double-obstacle potential according to [14]. The resulting problem belongs to the class of mathematical programs with equilibrium constraints (MPECs) in function space.

Even in finite dimensions, this problem class is well known for its constraint degeneracy [52, 54]. Due to the presence of the variational inequality constraint, classical constraint qualifications (see, e.g., [67]) fail, which prevents the application of Karush–Kuhn–Tucker (KKT) theory in Banach space for the first-order characterization of an optimal solution by (Lagrange) multipliers. As a result, stationarity conditions for this problem class are no longer unique (in contrast to KKT conditions); compare [40, 41] in function space and, e.g., [60] in finite dimensions. They rather depend on the underlying problem structure and/or on the chosen analytical approach.

The simulation of two-phase flows with matched densities is rather well understood in the literature, see, e.g., [45]. In contrast, there exist different approaches to model the case of fluids with non-matched densities. These range from quasi-incompressible models with non-divergence-free velocity fields, see, e.g., [51], to possibly thermodynamically inconsistent models with solenoidal fluid velocities, cf. [21]. In this chapter, we study the incompressible and thermodynamically consistent model presented in [6]. We refer to [3, 11, 12, 27, 29] for additional analytical and numerical results for some of these models.

Stable numerical schemes for the thermodynamically consistent diffuse interface model according to [6] are developed in [29, 34]. A fully integrated adaptive finite-element approach for the numerical treatment of the Cahn–Hilliard system with a nonsmooth free energy is developed in [37]. This approach is extended in [35] to a fully practical adaptive solver for the coupled Cahn–Hilliard Navier–Stokes system.

While there are numerous publications concerning the optimal control of the phase separation process itself, i.e., the distinct Cahn–Hilliard system, see, e.g., [14, 18, 25, 37, 42, 65], there has been considerably less research on the control of the Cahn–Hilliard Navier–Stokes system. Some of the few publications in this field address the case of matched densities and a nonsmooth homogeneous free energy density (double-obstacle potential), see [43, 44]. We also mention the recent articles [26] that treat the control of a nonlocal Cahn–Hilliard Navier–Stokes system in two dimensions, [61] and [30], which include numerical convergence results for the optimal control of the model developed in [29].

From a numerical point of view, the simulation and especially the optimal control of the coupled Cahn–Hilliard Navier–Stokes system are challenging tasks with regard to the computation times and the storage effort. For this reason, we apply model order reduction using Proper Orthogonal Decomposition (POD-MOR) in order to replace the high-fidelity models by low-order surrogates. We follow a simulation-based approach according to [58], where the snapshots are

generated by finite-element simulations of the system. In particular, we utilize space-adapted snapshots, which leads to the challenge that, in a discrete formulation, the snapshots are vectors of different lengths due to the different spatial resolutions. A consideration of the problem setting from an infinite-dimensional view according to [28] allows the combination of POD with spatially adapted snapshots. Moreover, we utilize a Moreau–Yosida regularization of the Cahn–Hilliard system and observe that the accuracy of the reduced-order model depends on the smoothness of the approximated object. Finally, we consider POD-MOR for the Navier–Stokes part. The use of space-adapted finite elements has the consequence that a weak-divergence-free property only holds in the current adapted finite-element space. In order to guarantee stability of the resulting reduced-order model, in [32] two solution approaches are proposed.

Regarding physical applications, we point out that the CHNS system is used to model a variety of situations. These range from the aforementioned solidification process of liquid metal alloys, cf. [22], or the simulation of bubble dynamics, as in Taylor flows [1], or pinch-offs of liquid–liquid jets [48], to the formation of polymeric membranes [66] or to protein crystallization, see, e.g., [49] and the references therein. Furthermore, the model can be easily adapted to include the effects of surfactants such as colloid particles at fluid–fluid interfaces in gels and emulsions used in food, pharmaceutical, cosmetic, or petroleum industries [2, 55].

This chapter is organized as follows. After introducing the problem setting in Sect. 2, we formulate the associated optimal control problem with respect to a semi-discrete system in Sect. 3.1. We proceed by securing the existence of global solutions and characterizing these solutions via suitable stationarity conditions in Sects. 3.2, 3.3, and 3.4. A goal-oriented error estimator is derived in Sect. 3.5, and we present two distinct numerical solution algorithms based on our analytical results in Sect. 3.6 and 3.7, which incorporate an adaptive mesh refinement technique. In Sect. 4, we focus on model order reduction with Proper Orthogonal Decomposition. The POD method in Hilbert spaces is explained in Sect. 4.1 and comprises the case of space-adapted snapshots. In Sect. 4.2, we derive a POD reduced-order model for the Cahn–Hilliard system and provide a numerical example in Sect. 4.3. Moreover, in Sect. 4.4, we consider POD-MOR with space-adapted snapshots for the Navier–Stokes equations. We conclude this chapter with a brief outlook on associated future research topics in Sect. 5.

2 Problem Setting

Let us specify the problem setting. We denote by Ω an open bounded domain with Lipschitz boundary $\partial\Omega$, and $T > 0$ is a given end time. We are concerned with the coupled Cahn–Hilliard Navier–Stokes (CHNS) system according to [6]

given by

$$\begin{aligned}
&\partial_t(\rho(\varphi)v) + \operatorname{div}(v \otimes \rho(\varphi)v) - \operatorname{div}(2\eta(\varphi)\epsilon(v)) + \nabla p \\
&\qquad + \operatorname{div}\Big(v \otimes -\frac{\widehat{\rho_2} - \widehat{\rho_1}}{2} m(\varphi)\nabla\mu\Big) - \mu\nabla\varphi = 0 \qquad \text{in } (0,T)\times\Omega, \qquad (2.1a)\\
&\qquad \operatorname{div} v = 0 \qquad \text{in } (0,T)\times\Omega, \qquad (2.1b)\\
&\qquad \partial_t\varphi + v\nabla\varphi - \operatorname{div}(m(\varphi)\nabla\mu) = 0 \qquad \text{in } (0,T)\times\Omega, \qquad (2.1c)\\
&\qquad -\sigma\epsilon\Delta\varphi + \frac{\sigma}{\epsilon}(\partial\Psi_0(\varphi) - \kappa\varphi) - \mu \ni 0 \qquad \text{in } (0,T)\times\Omega, \qquad (2.1d)\\
&\qquad v = \partial_n\varphi = \partial_n\mu = 0 \qquad \text{on } (0,T)\times\partial\Omega, \qquad (2.1e)\\
&\qquad v(0,\cdot) = v_a \qquad \text{in } \Omega, \qquad (2.1f)\\
&\qquad \varphi(0,\cdot) = \varphi_a \qquad \text{in } \Omega. \qquad (2.1g)
\end{aligned}$$

We denote by v the velocity and by p the pressure of the fluid, which is governed by the Navier–Stokes equations (2.1a)–(2.1b). The density ρ depends on the order parameter φ given by the Cahn–Hilliard equations (2.1c)–(2.1d) via

$$\rho(\varphi) = \frac{\rho_1 + \rho_2}{2} + \frac{\rho_2 - \rho_1}{2}\varphi. \qquad (2.2)$$

The mobility m and the viscosity η are variable and depend on the phase field φ. By μ, we denote the chemical potential. The surface tension $\sigma > 0$, the interface parameter $\epsilon > 0$, and the parameter $\kappa > 0$ are given constants. Furthermore, initial conditions v_a and φ_a for the velocity and phase field are given, respectively. By Ψ_0, we denote the convex part of the free energy potential $\Psi(\varphi) := (\Psi_0(\varphi) - \frac{\kappa}{2}\varphi^2)$. Depending on the underlying applications, there exist different modeling choices for Ψ_0. In this chapter, we focus on the double-obstacle potential introduced in (3.2). Possible other choices include the double-well potential $\Psi(\varphi) = \frac{\kappa}{2}(1-\varphi^2)^2$ and the logarithmic potential $\Psi(\varphi) = (1+\varphi)\ln(1+\varphi) + (1-\varphi)\ln(1-\varphi) - \frac{\kappa}{2}\varphi^2$.

An important property of the above CHNS system is its thermodynamical consistency. It is possible to derive a (dissipative) energy estimate by testing (2.1a), (2.1b), (2.1c), and (2.1d) with v, p, μ, and $\partial_t\varphi$, which yields

$$\partial_t E(v,\varphi) + 2\int_\Omega \eta(\varphi)|\epsilon(v)|^2 dx + \int_\Omega m(\varphi)|\nabla\mu|^2 dx \le 0, \qquad (2.3)$$

where the total energy E is given by the sum of the kinetic and the potential energy, i.e.,

$$E(v,\varphi) = \int_\Omega \rho(\varphi)\frac{|v|^2}{2}dx + \frac{\sigma\epsilon}{2}\int_\Omega \frac{|\nabla\varphi|^2}{2}dx + \frac{\sigma}{\epsilon}\Psi(\varphi). \tag{2.4}$$

Besides mirroring the physical property that the total energy of a closed system is non-increasing, inequality (2.3) also serves as a very valuable analytical tool, e.g., to secure the boundedness of solutions to (2.1).

3 Optimal Control of the Semi-Discrete CHNS System

In the following, we study the optimal control of a semi-discrete variant of the Cahn–Hilliard Navier–Stokes system (2.1), where the free energy density is related to the double-obstacle potential, see (3.2) below. This yields an optimal control problem for a family of coupled systems in each time instant of a variational inequality of fourth order and the Navier–Stokes equations. The time discretization is chosen in such a way that the thermodynamical consistency of the system (cf. (2.3)) is maintained.

We ensure the existence of feasible and globally optimal points for the respective optimal control problem and provide a first characterization of those points via a stationarity system of limiting $\mathcal{E}$-almost C-stationary types. We proceed with a thorough analysis of the sensitivity and differentiability properties of the associated control-to-state operator that culminates in the presentation of a strong stationarity system.

Our analytical results are subsequently supplemented by the development and demonstration of two numerical solution algorithms, which compute discrete approximations of C-stationary or strong stationary points of the optimal control problem (3.2) below. In order to handle the tremendous computational effort caused by repeatedly solving the large-scale Navier–Stokes systems, we incorporate an adaptive mesh refinement strategy based on a goal-oriented error estimator.

3.1 The Semi-Discrete CHNS System and the Optimal Control Problem

Let us start by presenting the underlying time discretization of the CHNS system and by imposing some common assumptions on the related physical data. For this purpose, we choose an arbitrary time step size $\tau > 0$ and denote the total number of time instants by $K \in \mathbb{N}$. Moreover, we introduce a distributed force u on the right-hand side of the Navier–Stokes equations.

Definition 3.1 (Semi-Discrete CHNS System) For a given initial state $(\varphi_{-1}, v_0) = (\varphi_a, v_a) \in \left(H^2_{\partial_n}(\Omega) \cap \mathbb{K}\right) \times H^2_{0,\sigma}(\Omega; \mathbb{R}^n)$, we say that a triple

$$(\varphi, \mu, v) = ((\varphi_i)_{i=0}^{K-1}, (\mu_i)_{i=0}^{K-1}, (v_i)_{i=1}^{K-1})$$

in $H^2_{\partial_n}(\Omega)^K \times H^2_{\partial_n}(\Omega)^K \times H^1_{0,\sigma}(\Omega; \mathbb{R}^n)^{K-1}$ solves the semi-discrete CHNS system with respect to a given control $u = (u_i)_{i=1}^{K-1} \in L^2(\Omega; \mathbb{R}^n)^{K-1}$, if for all $\phi \in \overline{H}^1(\Omega)$ and $\psi \in H^1_{0,\sigma}(\Omega; \mathbb{R}^n)$ we have that

$$\left\langle \frac{\varphi_{i+1} - \varphi_i}{\tau}, \phi \right\rangle + \langle v_{i+1} \nabla \varphi_i, \phi \rangle + (m(\varphi_i) \nabla \mu_{i+1}, \nabla \phi) = 0, \tag{3.1a}$$

$$(\nabla \varphi_{i+1}, \nabla \phi) + \langle a_{i+1}, \phi \rangle - \langle \mu_{i+1}, \phi \rangle - \langle \kappa \varphi_i, \phi \rangle = 0, \tag{3.1b}$$

$$\begin{aligned} &\left\langle \frac{\rho(\varphi_i) v_{i+1} - \rho(\varphi_{i-1}) v_i}{\tau}, \psi \right\rangle_{H^{-1}_{0,\sigma}, H^1_{0,\sigma}} - (v_{i+1} \otimes \rho(\varphi_{i-1}) v_i, \nabla \psi) \\ &+ \left(v_{i+1} \otimes \frac{\rho_2 - \rho_1}{2} m(\varphi_{i-1}) \nabla \mu_i, \nabla \psi \right) + (2\eta(\varphi_i) \epsilon(v_{i+1}), \epsilon(\psi)) \\ &\qquad - \langle \mu_{i+1} \nabla \varphi_i, \psi \rangle_{H^{-1}_{0,\sigma}, H^1_{0,\sigma}} = \langle u_{i+1}, \psi \rangle_{H^{-1}_{0,\sigma}, H^1_{0,\sigma}}, \end{aligned} \tag{3.1c}$$

with $a_i \in \partial \Psi_0(\varphi_i)$. The first two equations are supposed to hold for every $0 \leq i + 1 \leq K - 1$, and the last equation holds for every $1 \leq i + 1 \leq K - 1$.

The corresponding solution operator is denoted by S_Ψ, i.e., $(\varphi, \mu, v) \in S_\Psi(u)$.

In the above definition, the boundary conditions (2.1e) and the solenoidality of the velocity field (2.1b) are integrated in the chosen function spaces

$$H^k_{0,\sigma}(\Omega; \mathbb{R}^n) := \left\{ f \in H^k(\Omega; \mathbb{R}^n) \cap H^1_0(\Omega; \mathbb{R}^n) : \operatorname{div} f = 0, \text{ a.e. on } \Omega \right\},$$

$$H^k_{\partial_n}(\Omega) := \left\{ f \in H^k(\Omega) : \partial_n f_{|\partial\Omega} = 0 \text{ on } \partial\Omega \right\}, \ k \geq 2,$$

for φ, μ and v. Furthermore, the definition already includes the inherent regularity properties of φ and μ, which anticipates the results of Theorem 3.4 below.

Moreover, the semi-discrete CHNS system involves three time instants $(i - 1, i, i + 1)$, and (φ_0, μ_0) is characterized in an initialization step by the (decoupled) Cahn–Hilliard system only. At the subsequent time instants, the strong coupling of the Cahn–Hilliard and Navier–Stokes system is maintained.

In this chapter, we consider non-degenerate mobility and viscosity coefficients $m, \eta \in C^2(\mathbb{R})$, i.e., $0 < c_1 \leq \min_{x \in \mathbb{R}} \{m(x), \eta(x)\}$. We further assume that m and η, as well as their derivatives up to second order, are bounded, which is typically satisfied if they originate from a practical application.

As noted above, the free energy density is related to the double-obstacle potential. In other words, the functional $\Psi_0 : H^1(\Omega) \to \mathbb{R}$ is given by $\Psi_0(\varphi) := \int_\Omega \iota_{[\psi_1;\psi_2]}(\varphi(x))dx$, where $\iota_{[\psi_1;\psi_2]}$ denotes the indicator function of $[\psi_1; \psi_2]$, i.e.,

$$\iota_{[\psi_1;\psi_2]} := \begin{cases} +\infty \text{ if } z < \psi_1, \\ 0 \quad \text{if } \psi_1 \le z \le \psi_2, \\ +\infty \text{ if } z > \psi_2, \end{cases} \qquad \psi_1 < 0 < \psi_2. \tag{3.2}$$

As a consequence, the inclusion (3.1b) ensures that the order parameter φ_i is contained in $[\psi_1; \psi_2]$ almost everywhere (a.e.) on Ω for every time instant $-1 \le i \le K-1$, assuming that the initial data is well posed in the sense that

$$\varphi_a \in \mathbb{K} := \left\{ v \in H^1(\Omega) : \psi_1 \le v \le \psi_2 \text{ a.e. in } \Omega \right\}. \tag{3.3}$$

In order to formulate the optimal control problem associated to (3.1), we introduce an objective functional $\mathcal{J} : \mathcal{X} \to \mathbb{R}$ defined on

$$\mathcal{X} := H^1(\Omega)^K \times H^1(\Omega)^K \times H^1_{0,\sigma}(\Omega;\mathbb{R}^n)^{K-1} \times L^2(\Omega;\mathbb{R}^n)^{K-1}$$

and assume that $\mathcal{J}$ is convex, weakly lower semi-continuous, Fréchet differentiable, and partially coercive.

Definition 3.2 We study the optimal control problem

$$\begin{aligned} &\min \mathcal{J}(\varphi, \mu, v, u) \text{ over } (\varphi, \mu, v, u) \in \mathcal{X} \\ &\text{s.t. } (\varphi, \mu, v) \in S_\Psi(u). \end{aligned} \tag{3.4}$$

For our numerical computations below, we consider the specific functional

$$\mathcal{J}(\varphi, \mu, v, u) := \frac{1}{2}\|\varphi_{K-1} - \varphi_d\|^2 + \frac{\xi}{2}\|u\|^2, \ \xi > 0, \tag{3.5}$$

where $\varphi_d \in L^2(\Omega)$ represents a desired state. The so-called tracking type functional, which is used in various applications, clearly satisfies the above assumptions.

3.2 Existence of Feasible and Globally Optimal Points

One of the main requirements for the existence of solutions to (3.4) is the boundedness of the state. In our setting, this property follows from the energetic stability of the chosen discretization in time. More precisely, we have the following

(dissipative) energy law for the total energy

$$E(v, \varphi, \varphi_{-1}) = \int_\Omega \rho(\varphi_{-1}) \frac{|v|^2}{2} dx + \int_\Omega \frac{|\nabla \varphi|^2}{2} dx + \Psi(\varphi), \tag{3.6}$$

associated with the semi-discrete CHNS system (3.1). For more details on the proof of Lemma 3.3 and the other results of this subsection, we refer the reader to [39].

Lemma 3.3 (Energy Estimate for a Single Time Step) *Let $\varphi_i, \varphi_{i-1} \in H^2_{\partial_n}(\Omega) \cap \mathbb{K}$, $\mu_i \in H^2_{\partial_n}(\Omega)$, $v_i \in H^1_{0,\sigma}(\Omega; \mathbb{R}^n)$ and $u_{i+1} \in (H^1_{0,\sigma}(\Omega; \mathbb{R}^n))^*$ be given.*

If $(\varphi_{i+1}, \mu_{i+1}, v_{i+1}) \in H^1(\Omega) \times H^1(\Omega) \times H^1_{0,\sigma}(\Omega; \mathbb{R}^n)$ satisfies the system (3.1), then the corresponding total energy is bounded by

$$\begin{aligned} E(v_{i+1}, \varphi_{i+1}, \varphi_i) + \int_\Omega \rho(\varphi_{i-1}) \frac{|v_{i+1} - v_i|^2}{2} dx + \int_\Omega \frac{|\nabla \varphi_{i+1} - \nabla \varphi_i|^2}{2} dx \\ + \tau \int_\Omega 2\eta(\varphi_i) |\epsilon(v_{i+1})|^2 dx + \tau \int_\Omega m(\varphi_i) |\nabla \mu_{i+1}|^2 dx + \int_\Omega \kappa \frac{(\varphi_{i+1} - \varphi_i)^2}{2} \\ \leq E(v_i, \varphi_i, \varphi_{i-1}) + \langle u_{i+1}, v_{i+1} \rangle_{H^{-1}_{0,\sigma}, H^1_{0,\sigma}}. \end{aligned} \tag{3.7}$$

It should be noted that the density is always positive, since φ_i is contained in $\mathbb{K}$ for every i. Consequently, all the terms of the left-hand side of the inequality are always nonnegative such that Lemma 3.3 indeed ensures that the energy of the next time step is non-increasing if the external force u_{i+1} is absent.

Lemma 3.3 allows us to verify the existence of solutions to the CHNS system (3.1) via the repeated application of Schaefer's fixed point theorem. The proof further involves arguments from PDE theory and monotone operator theory.

Theorem 3.4 (Existence of Feasible Points) *Let $u \in L^2(\Omega; \mathbb{R}^n)^{K-1}$ be given.*

Then the semi-discrete CHNS system admits a solution $(\varphi, \mu, v) \in H^2_{\partial_n}(\Omega)^K \times H^2_{\partial_n}(\Omega)^K \times H^2_{0,\sigma}(\Omega; \mathbb{R}^n)^{K-1}$.

The last theorem also ensures an additional regularity of the state, which is necessary to guarantee that the system (3.1) is well posed for each time step. The proof relies on the regularity theory for Navier–Stokes equations and variational inequalities.

By Theorem 3.4, the feasible set of problem (3.4) is non-empty. Then, the existence of globally optimal points can be verified via standard arguments from optimization theory.

Theorem 3.5 (Existence of Global Solutions) *The optimization problem (3.4) possesses a global solution.*

3.3 $\mathcal{E}$-Almost C-Stationary Points

After securing the existence of solutions to the optimal control problem (3.4), we target a more precise characterization of globally and/or locally optimal points via necessary optimality conditions. This lays the foundation to the development of efficient numerical solution methods in the subsequent subsections.

As a first step, we establish a limiting $\mathcal{E}$-almost C-stationarity system. For this purpose, we additionally assume that $\mathcal{J}'$ is a bounded mapping and that $\frac{\partial \mathcal{J}}{\partial u}$ satisfies the following weak lower-semicontinuity property

$$\left\langle \frac{\partial \mathcal{J}}{\partial u}(\hat{z}), \hat{u} \right\rangle \leq \liminf_{k\to\infty} \left\langle \frac{\partial \mathcal{J}}{\partial u}(\hat{z}^{(k)}), \hat{u}^{(k)} \right\rangle,$$

where the sequence $\hat{z}^{(k)}$ converges weakly in the space $H^2_{\partial_n}(\Omega)^K \times H^2_{\partial_n}(\Omega)^K \times H^1_{0,\sigma}(\Omega;\mathbb{R}^n)^{K-1} \times L^2(\Omega;\mathbb{R}^n)^{K-1}$ toward a limit point $\hat{z}$. Here and in the following, z represents the primal variables, i.e., $\hat{z} := (\hat{\varphi}, \hat{\mu}, \hat{v}, \hat{u})$.

The derivation is based on a penalization of the lower-level problem, where the double-obstacle potential is approximated by certain smooth double-well-type potentials Ψ_k, $k \in \mathbb{N}$. This gives rise to a family of smooth auxiliary nonlinear programs (P_{Ψ_k}) for which the following necessary optimality system can be derived via a well-known result from Zowe and Kurcyusz [67, Theorem 4.1].

Theorem 3.6 (First-Order Optimality Conditions for Smooth Potentials) *Let $\overline{z}$ be a minimizer of the auxiliary problem (P_{Ψ_k}).*

Then, there exist $(p, r, q, \lambda) \in H^1(\Omega)^K \times H^1(\Omega)^K \times H^1_{0,\sigma}(\Omega;\mathbb{R}^n)^{K-1} \times \overline{H}^1(\Omega)^{*\,K}$, with $\lambda_i := \Psi_k''(\varphi_{i+1})^* r_i$, such that

$$\begin{aligned}
&-\frac{1}{\tau}(p_i - p_{i-1}) + m'(\varphi_i)\nabla\mu_{i+1}\cdot\nabla p_i - \operatorname{div}(p_i v_{i+1}) \\
&-\Delta r_{i-1} + \lambda_{i-1} - \kappa r_{i+1} - \frac{1}{\tau}\rho'(\varphi_i)v_{i+1}\cdot(q_{i+1} - q_i) \\
&-\left(\rho'(\varphi_i)v_{i+1} - \frac{\rho_2-\rho_1}{2}m'(\varphi_i)\nabla\mu_{i+1}\right)(Dq_{i+1})^\top v_{i+2} \\
&\qquad +2\eta'(\varphi_i)\epsilon(v_{i+1}) : Dq_i + \operatorname{div}(\mu_{i+1}q_i) = \frac{\partial \mathcal{J}}{\partial \varphi_i}(\overline{z}),
\end{aligned} \tag{3.8}$$

$$\begin{aligned}
&-r_{i-1} - \operatorname{div}(m(\varphi_{i-1})\nabla p_{i-1}) \\
&-\operatorname{div}\left(\frac{\rho_2-\rho_1}{2}m(\varphi_{i-1})(Dq_i)^\top v_{i+1}\right) - q_{i-1}\cdot\nabla\varphi_{i-1} = \frac{\partial \mathcal{J}}{\partial \mu_i}(\overline{z}),
\end{aligned} \tag{3.9}$$

$$
\begin{aligned}
&-\frac{1}{\tau}\rho(\varphi_{j-1})(q_j - q_{j-1}) - \rho(\varphi_{j-1})(Dq_j)^\top v_{j+1} \\
&-(Dq_{j-1})\left(\rho(\varphi_{j-2})v_{j-1} - \frac{\rho_2 - \rho_1}{2} m(\varphi_{j-2})\nabla\mu_{j-1}\right) \\
&\qquad - \operatorname{div}(2\eta(\varphi_{j-1})\epsilon(q_{j-1})) + p_{j-1}\nabla\varphi_{j-1} \;=\; \frac{\partial \mathcal{J}}{\partial v_j}(\overline{z}),
\end{aligned}
\tag{3.10}
$$

$$
\frac{\partial \mathcal{J}}{\partial u_j}(\overline{z}) - q_{j-1} \;=\; 0 \tag{3.11}
$$

for all $i = 0, \dots, K-1$ *and* $j = 1, \dots, K-1$. *Here, we use the convention that* p_i, r_i, q_i *are equal to* 0 *for* $i \geq K-1$ *along with* q_{-1} *and* φ_i, μ_i, v_i *for* $i \geq K$.

A careful limit analysis with respect to a vanishing penalization parameter yields the following stationarity system for the optimal control problem (3.4), cf. [39].

Theorem 3.7 (Limiting $\mathcal{E}$-Almost C-Stationarity) *Let* $(\varphi^{(k)}, \mu^{(k)}, v^{(k)}, u^{(k)})$ *be a minimizer for* (P_{Ψ_k}), *and let further* $(p^{(k)}, r^{(k)}, q^{(k)}, \lambda^{(k)})$ *be given as in Theorem 3.6.*

Then, there exists a weakly convergent subsequence

$$
\begin{aligned}
&\left\{(\varphi^{(m)}, \mu^{(m)}, v^{(m)}, u^{(m)}, p^{(m)}, r^{(m)}, q^{(m)}, \lambda^{(m)})\right\}_{m\in\mathbb{N}} \\
&\subset H^2_{\partial_n}(\Omega)^K \times H^2_{\partial_n}(\Omega)^K \times H^1_{0,\sigma}(\Omega;\mathbb{R}^n)^{K-1} \times L^2(\Omega;\mathbb{R}^n)^{K-1} \\
&\quad \times H^1(\Omega)^K \times H^1(\Omega)^K \times H^1_{0,\sigma}(\Omega;\mathbb{R}^n)^{K-1} \times H^1(\Omega)^{*K},
\end{aligned}
\tag{3.12}
$$

and the limit point $(\varphi, \mu, v, u, p, r, q, \lambda)$ *satisfies the adjoint system* (3.8)–(3.11), *as well as*

$$
(a_i, r_{i-1})_{L^2} = 0, \qquad \liminf (\lambda_i^{(m)}, r_{i-1}^{(m)})_{L^2} \geq 0. \tag{3.13}
$$

Moreover, for every $\varepsilon > 0$, *there exist a measurable subset* M_i^ε *of* $M_i := \{x \in \Omega : \psi_1 < \varphi_i(x) < \psi_2\}$ *with* $|M_i \setminus M_i^\varepsilon| < \varepsilon$ *and*

$$
\langle \lambda_i, v\rangle = 0 \qquad \forall v \in \overline{H}^1(\Omega), \quad v|_{\Omega\setminus M_i^\varepsilon} = 0. \tag{3.14}
$$

The above stationarity conditions correspond to a function space version of C-stationarity, see, e.g., [40, 60]. The proof of the last condition (3.14) is based on the application of Egorov's theorem, cf. [9], which motivated the notion of $\mathcal{E}$-almost C-stationarity.

3.4 Strong Stationarity

Starting from the C-stationarity system of the previous section, it is possible to derive a more restrictive stationarity system for the problem (3.4) employing the directional differentiability of the control-to-state operator S_Ψ. In this subsection, we consider the control of the semi-discrete CHNS system for a single time step, i.e., $K = 2$, and $\varphi_{-1}, \varphi_0, \mu_0, v_0$ are given. This corresponds to an instantaneous control problem.

First, we verify that the solution operator S_Ψ of the semi-discrete CHNS system is Lipschitz continuous.

Theorem 3.8 (Lipschitz Continuity of S_Ψ) *The mapping $S_\Psi : H^{-1}_{0,\sigma}(\Omega) \to H^1(\Omega) \times H^1(\Omega) \times H^1_{0,\sigma}(\Omega; \mathbb{R}^N)$ is Lipschitz continuous.*

The proof follows a similar line of argumentation as Lemma 3.3. An immediate consequence of the above theorem is that the solutions to the constraint system are uniquely determined by the control u.

Although solution operators of variational inequalities are in general not Fréchet differentiable, we can now compute the directional derivative of S_Ψ via the following theorem.

Theorem 3.9 *The directional derivative of S_Ψ at $\hat{u} \in H^{-1}_{0,\sigma}(\Omega)$ with $S_\Psi(\hat{u}) = (\hat{\varphi}, \hat{\mu}, \hat{v})$ in direction $h \in H^{-1}_{0,\sigma}(\Omega)$ is the unique solution $(\chi, w, \zeta) \in H^1(\Omega) \times H^1(\Omega) \times H^1_{0,\sigma}(\Omega; \mathbb{R}^N)$ of the system*

$$\chi \in T_{\mathbb{K}}(\hat{\varphi}) \cap a^{+\perp} \cap a^{-\perp}, \tag{3.15a}$$

$$\langle -\Delta\chi - w, v - \chi \rangle \geq 0, \ \forall v \in T_{\mathbb{K}}(\hat{\varphi}) \cap a^{+\perp} \cap a^{-\perp}, \tag{3.15b}$$

$$\left\langle \frac{\chi}{\tau}, \phi \right\rangle + \langle \zeta \nabla\varphi_0, \phi \rangle + (m(\varphi_0)\nabla w, \nabla\phi) = 0, \tag{3.15c}$$

$$\begin{aligned} &\left\langle \frac{\rho(\varphi_0)\zeta}{\tau}, \psi \right\rangle_{H^{-1}_{0,\sigma}, H^1_{0,\sigma}} - (\zeta \otimes \rho(\varphi_{-1})v_0, \nabla\psi) \\ &+ \left(\zeta \otimes \frac{\rho_2 - \rho_1}{2} m(\varphi_{-1})\nabla\mu_0, \nabla\psi \right) + (2\eta(\varphi_0)\epsilon(\zeta), \epsilon(\psi)) \\ &- \langle w\nabla\varphi_0, \psi \rangle_{H^{-1}_{0,\sigma}, H^1_{0,\sigma}} - \langle h, \psi \rangle_{H^{-1}_{0,\sigma}, H^1_{0,\sigma}} = 0. \end{aligned} \tag{3.15d}$$

Here, $T_{\mathbb{K}}(\hat{\varphi})$ represents the tangent cone of $\mathbb{K}$ at $\hat{\varphi}$ and $a^{+/-\perp} := \{\phi \in H^1(\Omega) : \langle \phi, a^{+/-} \rangle = 0\}$ is the orthogonal space associated with $a^+(x) := \max\{a(x), 0\}$ and $a^-(x) := \min\{a(x), 0\}$.

Note that a^+ and a^- can be interpreted as the multipliers to the constraints $\varphi \leq 1$ and $\varphi \geq -1$, and the convex constraint set $T_{\mathbb{K}}(\hat{\varphi}) \cap a^{+\perp} \cap a^{-\perp}$ associated to the variational inequality (3.15b)–(3.15c) is also called the critical cone, cf. [53]. The proof of Theorem 3.9 combines arguments from Jarusek et al. in [47] and PDE theory.

With the help of the directional derivative of S_Ψ, we derive strong stationarity conditions for (3.4) by evaluating the B-stationarity condition of the reduced optimization problem

$$\min_{u \in L^2(\Omega;\mathbb{R}^N)} \overline{\mathcal{J}}(u) := \mathcal{J}(S_\Psi(u), u) \tag{3.16}$$

for suitable test directions.

Theorem 3.10 *If $\hat{u}$ is an optimal control of (3.4), then there exists an adjoint state $(p, r, q) \in H^1(\Omega) \times H^1(\Omega) \times H^1_{0,\sigma}(\Omega;\mathbb{R}^N)$ and $\lambda \in H^1(\Omega)^*$ such that for all $\phi \in H^1(\Omega)$ and $\psi \in H^1_{0,\sigma}(\Omega;\mathbb{R}^N)$ it holds that*

$$\left\langle D_\varphi \mathcal{J}[z_0] + \frac{r}{\tau}, \phi \right\rangle + (\nabla p, \nabla \phi) + \langle \lambda, \phi \rangle = 0, \tag{3.17}$$

$$(m(\varphi_0)\nabla r, \nabla \phi) - \langle p, \phi \rangle - \langle q \nabla \varphi_0, \phi \rangle = 0, \tag{3.18}$$

$$\left\langle \frac{\rho(\varphi_0)}{\tau} q, \psi \right\rangle_{H^{-1}_{0,\sigma}, H^1_{0,\sigma}} - \langle \nabla q v, \psi \rangle_{H^{-1}_{0,\sigma}, H^1_{0,\sigma}}$$
$$+ \langle 2\eta(\varphi_0)\epsilon(q), \epsilon(\psi) \rangle_{H^{-1}_{0,\sigma}, H^1_{0,\sigma}} - \langle r \nabla \varphi_0, \psi \rangle_{H^{-1}_{0,\sigma}, H^1_{0,\sigma}} = 0, \tag{3.19}$$

$$\langle -q, \psi \rangle_{H^{-1}_{0,\sigma}, H^1_{0,\sigma}} + \left\langle D_u \mathcal{J}[\hat{z}], \psi \right\rangle_{H^{-1}_{0,\sigma}, H^1_{0,\sigma}} = 0, \tag{3.20}$$

$$\lambda \in \left(T_{\mathbb{K}}(\hat{\varphi}) \cap a^{+\perp} \cap a^{-\perp} \right)^0, \tag{3.21}$$

$$q \in \left(\left[D \left(\left(T_{\mathbb{K}}(\hat{\varphi}) \cap a^{+\perp} \cap a^{-\perp} \right)^0 \times H^1_{0,\sigma}(\Omega;\mathbb{R}^N) \right) \right]_2 \right)^0, \tag{3.22}$$

where D is a specific linear operator and the subscript K^0 represents the polar cone of the cone K.

This concludes our analytical investigations. We point out that the strong stationarity conditions represent the most selective stationarity system available for the problem under consideration up to this point in time.

3.5 Adaptive Mesh Refinement

In the following subsections, we discuss efficient numerical solution methods for the problem (3.4), where the objective functional is given by (3.5), based on our analytical results. The main challenges hereby are imposed by the non-differentiability of the solution operator due to the Cahn–Hilliard system and the immense numerical expense caused by repeatedly solving the large-scale Navier–Stokes-type primal and dual systems.

We deal with the second challenge by developing a goal-oriented error estimator based on the dual-weighted residual approach, cf., e.g., [13]. This allows us to implement an adaptive mesh refinement strategy, which acknowledges the error contributions of the primal residuals, the dual residuals, and the mismatch in the complementarity terms, to reduce the computational effort.

The central idea of this approach is depicted by the subsequent theorem, which estimates the difference of the objective values at stationary points of the semi-discrete and the fully discretized problem with the help of the associated MPCC-Lagrangian $\mathcal{L}$, cf. [36].

Theorem 3.11 *Let $(y, u, \Phi, \pi, \lambda^+, \lambda^-)$ be a stationary point of the optimal control problem (3.4) and assume that $(y_h, u_h, \Phi_h, \pi_h, \lambda_h^+, \lambda_h^-) \in \mathcal{Y}_h$ satisfies the discretized stationarity system. Then it holds that*

$$\begin{aligned}
\mathcal{J}(\varphi_h, \mu_h, v_h, u_h) - \mathcal{J}(\varphi, \mu, v, u) = \frac{1}{2}\left(\sum_{i=0}^{K-1}\left\langle a_h^i, \pi^i\right\rangle - \sum_{i=0}^{K-1}\left\langle a^i, \pi_h^i\right\rangle\right) \\
-\frac{1}{2}\left(\sum_{i=0}^{K-1}\left\langle (\lambda^i)^+, \varphi_h^i - \psi_2\right\rangle - \sum_{i=0}^{K-1}\left\langle (\lambda_h^i)^+, \varphi^i - \psi_2\right\rangle\right) \\
+\frac{1}{2}\left(\sum_{i=0}^{K-1}\left\langle (\lambda^i)^-, \varphi_h^i - \psi_1\right\rangle - \sum_{i=0}^{K-1}\left\langle (\lambda_h^i)^-, \varphi^i - \psi_1\right\rangle\right) \\
+\frac{1}{2}\nabla_x \mathcal{L}(y_h, u_h, \Phi_h, \pi_h, \lambda_h^+, \lambda_h^-)((y_h, u_h, \Phi_h) - (y, u, \Phi)) \\
+O\left(\|(y_h, u_h, \Phi_h) - (y, u, \Phi)\|^3\right), \qquad (3.23)
\end{aligned}$$

where O denotes the Landau symbol Big-O.

This allows us to approximate the discretization error with respect to the objective function as follows:

$$\begin{aligned}\mathcal{J}(\varphi_h, \mu_h, v_h, u_h) - \mathcal{J}(\varphi, \mu, v, u) \\ \approx \sum_{i=0}^{K-1} (\eta_{CM1,i} + \eta_{CM2,i} + \eta_{CM3,i} + \eta_{CM4,i} + \eta_{CH1,i} \\ + \eta_{CH2,i} + \eta_{NS,i} + \eta_{AD\varphi,i} + \eta_{AD\mu,i} + \eta_{ADv,i}),\end{aligned} \tag{3.24}$$

where the complementarity error terms $\eta_{CM1,i}, .., \eta_{CM4,i}$, and the weighted primal residuals $\eta_{CH1,i}, \eta_{CH2,i}, \eta_{NS,i}$ and the weighted dual residuals $\eta_{AD\varphi,i}, \eta_{AD\mu,i}, \eta_{ADv,i}$ are defined as in [36, Section 4]. These individual error terms can be evaluated separately on each patch of the current mesh due to their integral structure. In order to obtain a fully a posteriori error estimator, the continuous quantities are approximated with the help of a local higher-order approximation based on the respective discrete variables.

3.6 Penalization Algorithm

A first approach to handle the non-differentiability of S_Ψ numerically is motivated by the penalization method of Sect. 3.3. Namely, we solve a sequence of auxiliary optimization problems, where we approximate Ψ_0 by

$$\Psi_{0,\alpha}(\varphi) := \frac{1}{2\alpha}\left(\max(0, \varphi - 1)^2 + \min(\varphi + 1)^2\right), \ \alpha > 0, \ \alpha \to 0.$$

The resulting nonlinear programs can be solved by a standard steepest descent method, and the calculated solution approximates a C-stationary point of (3.4) if the complementarity conditions of Theorem 3.7 are satisfied sufficiently well, i.e., up to a given tolerance tol_c. In combination with an outer adaptation loop based on the error estimator (3.24), this yields Algorithm 1.

Algorithm 1: The overall solution procedure

Data: Initial data: φ_a, v_a;
repeat
  repeat
    solve the regularized problem (P_{Ψ_α}) using a steepest descent method;
    decrease α;
  until complementarity conditions are satisfied up to a tolerance ϵ_{tol};
  calculate the error indicators and identify the sets $\mathcal{M}_r$,$\mathcal{M}_c$ of cells to refine/coarsen;
  adapt $(\mathcal{T}^i)_{i=1}^K$ based on $\mathcal{M}_r$ and $\mathcal{M}_c$;
until $\sum_{i=1}^K |\mathcal{T}^i| > \mathcal{A}_{\max}$;

Hereby, the outer adaptation loop relies on the Dörfler marking procedure. Hence, the error indicators from (3.24) are evaluated for all time steps i and for all cells $T \in \mathcal{T}^i$ of the current triangulation $(\mathcal{T}^i)_{i=1}^K$. Then we choose a set $\mathcal{M}_r$ of cells to be refined as the set with the smallest cardinality, which satisfies

$$\sum_{T \in \mathcal{M}_r} \eta_T \geq \theta^r \sum_{i=1}^{K} \sum_{T \in \mathcal{T}^i} \eta_T$$

for a given parameter $0 < \theta^r < 1$. Due to the movement of the interface, we also select cells for coarsening if the calculated error indicator is smaller than a certain fraction of the mean error, i.e.,

$$M_c := \left\{ T \in (\mathcal{T}^i)_{i=1}^K \mid \eta_T \leq \frac{\theta^c}{\mathcal{A}} \sum_{i=1}^{K} \sum_{T \in \mathcal{T}^i} \eta_T \right\},$$

where $0 < \theta^c < 1$ is fixed and $\mathcal{A} := \sum_{i=1}^K |\mathcal{T}^i|$. The mesh refinement process is terminated if a desired total number of cells $\mathcal{A}_{\max}$ is exceeded.

Moreover, the problem is discretized in space using Taylor–Hood finite elements, i.e., we utilize linear finite elements for φ, μ, and p and quadratic finite elements for v. For more details on the implementation of the algorithm and the numerical results, we refer to [36].

Let us briefly illustrate the performance of the proposed Algorithm 1 at the hands of a specific example. Our goal is to control the motion of a circular bubble to prevent it from rising and split it into two square-shaped bubbles. For this purpose, 2×4 locally supported ansatz functions of the control are distributed over the two-dimensional domain as depicted in Fig. 1. The figure further shows the initial state φ_a, the desired shape φ_d together with the zero level line of the phase field at final time if no control is applied. The corresponding objective functional is defined as in (3.5) with $\xi = 10^{-11}$.

The associated fluid parameters are given by $\rho_1 = 1000$, $\rho_2 = 100$, $\eta_1 = 10$, $\eta_2 = 1$, and $\sigma = 24.5 \cdot \frac{2}{\pi}$ and are taken from a benchmark problem for rising bubble dynamics in [46]. Furthermore, we incorporate a gravitational acceleration

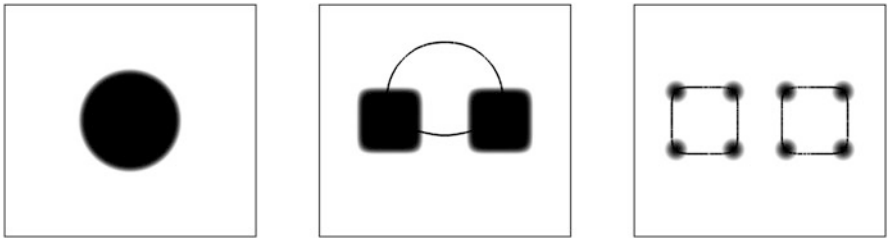

Fig. 1 The initial shape φ_0, the desired shape φ_d, the ansatz for the control u

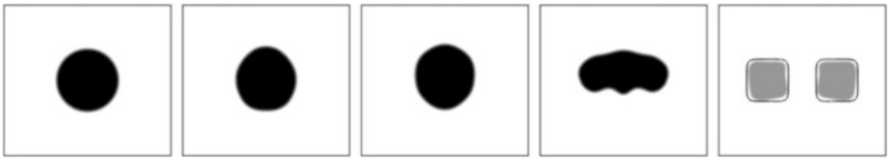

Fig. 2 The evolution of the phase field φ

$g = 0.981$ in the vertical direction and set $\epsilon = 0.02$, $m(\varphi) \equiv 4 \cdot 10^{-5}$. The time horizon is set to $T = 1.0$, and the time step size is $\tau = 125 \cdot 10^{-5}$.

For the marking procedure, we use the parameters $\theta^r = 0.7$ and $\theta^c = 0.01$. Furthermore, the stopping criteria use the tolerance $tol_c = 10^{-3}$ for the complementarity conditions and the maximum amount of cells $\mathcal{A}_{max} = 8 \cdot 10^6$ for the adaptation process, which relates to 10^4 cells in average per time instance.

The optimal solution on the first level and for the initial value for α is found after 26 steepest descent iterations, while the complete algorithm terminates after 419 steepest descent steps. Hereby, the algorithm solves the auxiliary optimization problems 10 times, i.e., line 3 of Algorithm 1 is executed 10 times. After the first two solves the Moreau–Yosida parameter was decreased, and after the next 8 solves the algorithm directly proceeded with the outer adaptation loop.

In Fig. 2, we depict the temporal evolution of the phase field φ corresponding to the optimal solution at the times $t = 0.00, 0.25, 0.50, 0.75, 1.00$. The figure additionally includes the zero level line of the desired shape φ_d for $t = 1.00$.

Regarding the mesh adaptation process, we observe that the cells are mainly refined in the interfacial region and, in particular, at the border of the diffuse interface. Such a behavior is typical for the numerical simulation of phase-field models. However, since our error estimator also contains terms from the Navier–Stokes and the adjoint equation, we further obtain significant mesh adaptations outside of the interface of the phases, which suggests that these errors should not be neglected, e.g., by a simple interface refinement technique. In Fig. 3, we depict the subdomain $\Omega_u = (0, 1) \times (0.5, 1.0) \subset \Omega$ at $t = 0.7$. On the left, we show $|v|$ in grayscale together with the isolines $\varphi \equiv \pm 1$ in black. On the right, we show the corresponding mesh. Note that the mesh is symmetric with respect to the central line.

3.7 Bundle-Free Implicit Programming Approach

Algorithm 1 can be further enhanced by exploiting the specific structure of the directional derivative of the control-to-state operator. Hereby, we apply the descent method directly to the problem (3.4) or (3.16) (instead of a regularized version) and compute a descent direction of $\overline{\mathcal{J}}$ at u^* with $(v^*, \varphi^*, \mu^*) = S(u^*)$ by solving the

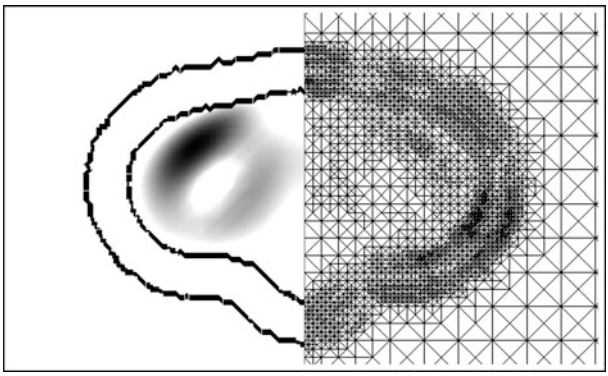

Fig. 3 The magnitude of v in grayscale and the isolines $\varphi \equiv \pm 1$ (left), and the associated triangulation (right)

optimization problem

$$\begin{aligned} \min_{h \in L^2(\Omega;\mathbb{R}^N)^{K-1}} & \overline{\mathcal{J}}'[u^*](h) + \|h\|^2 = (\varphi^* - \varphi_d, q) + \xi(u^*, h) + \|h\|^2, \\ & \text{s.t. } DS_\Psi[u^*](h) = (q, w, \zeta), \end{aligned} \tag{3.25}$$

where the stabilizing term $\|h\|^2_{L^2}$ ensures the existence of solutions. If a solution h of (3.25) equals zero, then u^* is a B-stationary point; otherwise, it is indeed a descent direction, since $\overline{\mathcal{J}}'[u^*](h) \leq -\|h\|^2 < 0$. In combination with a classical line search procedure, this leads to the following Algorithm 2.

Algorithm 2: The descent method for (3.4)

Data: Initial data: φ_a, v_a, u_0;

repeat
- Calculate a descent direction h_k by solving (3.25);
- Find a step size τ_k and a new iterate $u_{k+1} := u_k + \tau_k h_k$ by performing an Armijo line search along h_k;
- Set $k := k + 1$.

until $h_k \leq \epsilon_{tol}$;

The convergence of Algorithm 2 is ensured based on the arguments of [38].

Theorem 3.12 *The conceptual Algorithm 2 terminates after finitely many steps for any starting point u_0 if either $\tau_k \geq \underline{\tau} > 0$ for every $k \in \mathbb{N}$, or $\tau_k \to 0$ and*

$$\limsup_{k \to \infty} \frac{\overline{\mathcal{J}}(u_k + \overline{\tau}_k h_k) - \overline{\mathcal{J}}(u_k) - \overline{\tau}_k \overline{\mathcal{J}}'[u_k](h_k)}{\overline{\tau}_k} \leq 0, \tag{3.26}$$

where $\overline{\tau}_k > 0$ *represents the smallest step size for which the line search still fails at step* k.

Motivated by Theorem 3.12, we include an additional robustification step by performing one step of the penalization algorithm of Sect. 3.6, if the step size tends to zero. Thus, the resulting algorithm targets strong stationary points of (3.4) while guaranteeing at least C-stationarity of the computed solutions.

In order to solve the problem (3.25), we take advantage of the fact that it corresponds to a quadratic program, if strict complementarity holds, i.e., if the biactive set associated with the variational inequality (3.1b) is empty. Otherwise, we employ a regularization of the lower-level problem associated with (3.25).

As in the previous subsection, we utilize Taylor–Hood finite elements for the spacial discretization and supplement the algorithm with a similar adaptive mesh refinement strategy. Moreover, we solve the discretized CHNS system via a primal–dual active set method.

In the following example, we aim to transform a ring-shaped initial region into a curved tube, see Fig. 4. As seen on the right picture, the control acts via 16 locally supported ansatz functions.

The parameters for the physical model and the adaptation procedure are adopted from the previous example. In this example, the algorithm terminates at a C-stationary point after performing the Armijo line search (in line 3) 276 times. The maximum number of cells is exceeded after 6 mesh refinement steps.

Figure 5 presents the computed evolution of the phase field φ at the optimal solution along with the associated slack variable a emerging from the primal–dual active set method at the final time. In addition, we portray the magnitude of the velocity and the underlying mesh at final time.

Fig. 4 The initial shape φ_0, the desired shape φ_d, the ansatz for the control u

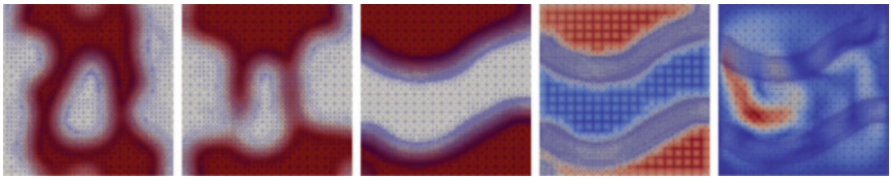

Fig. 5 The evolution of the phase field φ, the slack variable a, and the magnitude of v at the final time

4 Model Order Reduction with Proper Orthogonal Decomposition

From a numerical point of view, the simulation and in particular the optimal control of the coupled Cahn–Hilliard Navier–Stokes system (2.1) are computationally demanding tasks. Although the use of adaptive finite-element discretization concepts (see, e.g., [36]) makes numerical implementation feasible (in comparison to the use of a very fine, uniform discretization), the computational costs can be very large. For this reason, we apply model order reduction using Proper Orthogonal Decomposition (POD-MOR) in order to speed up computation times while ensuring a good approximation quality.

In order to construct a low-dimensional surrogate model, the usual POD framework first requires a so-called offline phase, in which high-fidelity solutions (snapshots) of the underlying dynamical system are generated by, e.g., finite-element simulations. From this snapshot set, the POD method finds a proper basis representation of the most relevant information encoded in the snapshots by computing a truncated singular value decomposition or by solving an associated eigenvalue problem. If the snapshots are discretized adaptively in space, the challenge arises that the snapshots are vectors of different lengths due to the different spatial resolutions at each time instance. This does not fit into the standard POD framework that assumes snapshots of the same length.

This section is concerned with POD reduced-order modeling using space-adapted snapshots. Section 4.1 describes the idea to consider the setting from an infinite-dimensional perspective that allows a broad spectrum of discretizations for the snapshots. Then, we derive a POD reduced-order model for the Cahn–Hilliard equations using space-adapted snapshots in Sect. 4.2 and present numerical results. Moreover, in Sect. 4.4, we consider POD-MOR with space-adapted snapshots for incompressible flow governed by the Navier–Stokes equations, where two strategies are proposed in order to ensure stability of the reduced-order model.

4.1 POD in Hilbert Spaces with Space-Adapted Snapshots

For a comprehensive study of the infinite-dimensional perspective on POD in a Hilbert space setting, we refer to [50], for example. Here, we recall main aspects and provide a practical implementation that is proposed in [28].

Let $\{y_h^0, \dots, y_h^{K-1}\} \subset X$ be a given set of snapshots, where X denotes a real, separable Hilbert space and y_h^i for $i = 0, \dots, K-1$ are high-fidelity-adapted finite-element solutions of the underlying dynamical system at different time instances. In particular, each of the snapshots belongs to a different discrete Galerkin space $y_h^i \in V_h^i$ with $V_h^0, \dots, V_h^{K-1} \subset X$. Then, a POD basis of rank ℓ is constructed by

solving the following equality constrained minimization problem:

$$\min_{\psi_1,\dots,\psi_\ell \in X} \sum_{i=0}^{K-1} \alpha_j \left\| y_h^i - \sum_{j=1}^{\ell} (y_h^i, \psi_j)_X \psi_j \right\|_X^2 \quad \text{s.t.} \quad (\psi_i, \psi_j)_X = \delta_{ij} \text{ for } 1 \le i, j \le \ell, \tag{4.1}$$

where α_i for $j = 0, \dots, K-1$ denote nonnegative weights and δ_{ij} is the Kronecker symbol. Since the snapshots are spatially adapted, the number of degrees of freedom and/or the location of the node points might differ such that it is not possible to build a corresponding snapshot matrix containing the finite-element Galerkin coefficients. For this reason, we assemble the snapshot Gramian defined by

$$\mathcal{K} \in \mathbb{R}^{K \times K}, \quad \mathcal{K}_{ij} := \sqrt{\alpha_i \alpha_j} (y_h^i, y_h^j)_X$$

for $i, j = 0, \dots, K-1$. In order to set up the matrix $\mathcal{K}$, we only require that the snapshots belong to the same Hilbert space X in order to evaluate the inner product $(\cdot, \cdot)_X$. Solving an eigenvalue problem for $\mathcal{K}$, i.e.,

$$\mathcal{K} \phi_i = \lambda_i \phi_i \quad \text{for } i = 1, \dots, \ell$$

delivers eigenvalues $\lambda_1 \ge \cdots \ge \lambda_\ell \ge 0$ and eigenvectors $\{\phi_1, \dots, \phi_\ell\} \subset \mathbb{R}^K$, which suffice to set up the POD reduced-order model, see [28, Section 4] for more details. The advantage of this perspective is that it allows a broad spectrum of discretization techniques and includes the case of r-adaptivity, for example. However, in this case, the evaluation of the inner products $(y_h^i, y_h^j)_X$ might get involved such that the necessity of, e.g., parallelization, becomes evident for practical implementations. In case of h-adapted snapshots using hierarchical, nested meshes, it is reasonable to express the snapshots with respect to a common finite-element space as proposed in [63].

4.2 POD Reduced-Order Modeling for the Cahn–Hilliard System

Let us consider the weak formulation of the Cahn–Hilliard equations (2.1c)–(2.1d) with boundary conditions (2.1e) and an initial condition for the phase field (2.1g), where we assume the velocity v to be given and fixed. The weak form reads as: Find a phase field $\varphi \in W(0, T; H^1(\Omega))$ with $\varphi|_{t=0} = \varphi_a$ and a chemical potential

$\mu \in L^2(0,T;H^1(\Omega))$ such that for all $\phi \in H^1(\Omega)$ it holds that

$$\frac{d}{dt}(\varphi(t),\phi)_{L^2(\Omega)} + (v\nabla\varphi(t),\phi)_{L^2(\Omega)} + m(\nabla\mu(t),\nabla\phi)_{L^2(\Omega)} = 0, \tag{4.2a}$$

$$\sigma\epsilon(\nabla\varphi(t),\nabla\phi)_{L^2(\Omega)} + \frac{\sigma}{\epsilon}(\Psi_0'(\varphi(t)) - \kappa\varphi(t),\phi)_{L^2(\Omega)} - (\mu(t),\phi)_{L^2(\Omega)} = 0. \tag{4.2b}$$

Note that in (4.2) we assume for simplicity a constant mobility $m > 0$ and sufficient regularity for Ψ_0. In order to derive an associated POD reduced-order model, we approximate the phase field φ and the chemical potential μ by a POD Galerkin ansatz given as $\varphi(t) \approx \varphi_\ell(t) = \sum_{j=1}^{\ell} c_j(t)\psi_j$ and $\mu(t) \approx \mu_\ell(t) = \sum_{j=1}^{\ell} w_j(t)\psi_j$. In [28, 31], we construct separate POD-reduced spaces for the phase field and the chemical potential, respectively. In contrast, here we compute the POD modes ψ_j for $j = 1,\dots,\ell$ according to (4.1) from space-adapted finite-element snapshots of the phase field and use the same POD modes in the Galerkin ansatz for both phase field and chemical potential. Using the POD space $V_\ell = \operatorname{span}\{\psi_1,\dots,\psi_\ell\} \subset H^1(\Omega)$ as trial and test space leads to the following POD reduced-order model for the Cahn–Hilliard equations: Find a phase field $\varphi_\ell \in V_\ell$ with $\varphi|_{t=0} = \mathcal{P}_\ell\varphi_a$ and a chemical potential $\mu_\ell \in V_\ell$ such that for all $\psi \in V_\ell$ it holds that

$$\frac{d}{dt}(\varphi_\ell(t),\psi)_{L^2(\Omega)} + (v\nabla\varphi_\ell(t),\psi)_{L^2(\Omega)} + m(\nabla\mu_\ell(t),\nabla\psi)_{L^2(\Omega)} = 0, \tag{4.3a}$$

$$\sigma\epsilon(\nabla\varphi_\ell(t),\nabla\psi)_{L^2(\Omega)} + \frac{\sigma}{\epsilon}(\Psi_0'(\varphi_\ell(t)) - \kappa\varphi_\ell(t),\psi)_{L^2(\Omega)} - (\mu_\ell(t),\psi)_{L^2(\Omega)} = 0. \tag{4.3b}$$

By $\mathcal{P}_\ell : V \to V_\ell$, we denote the orthogonal projection onto the POD space. Note that in (4.3), the evaluation of the nonlinear term $\Psi_0'(\varphi_\ell(t))$ is dependent on the full-order dimension. The treatment of nonlinearities is a well-known challenge within POD-MOR. In order to enable an efficient evaluation of the nonlinearity that is related to the low-order dimension ℓ of the reduced system, a linearization can be considered, compare [28] for more details. Alternatively, the so-called hyper-reduction methods like EIM [16], DEIM [20], or DMD [7] can be applied.

4.3 Numerical Example of POD-MOR for the Cahn–Hilliard System

In this section, we numerically investigate two major issues within POD-MOR for the Cahn–Hilliard equations:

(i) How does the regularity of the free energy Ψ_0 affect the accuracy of the POD reduced-order model?
(ii) How does the use of spatial adaptivity in the offline phase for snapshot generation influence the computation times and the accuracy of the POD reduced-order model?

The first aspect (i) is studied numerically in [4]: there, the initial phase field is given as a circle in a two-dimensional domain, which is transported in the horizontal direction over time. In this simulation, a uniform and static discretization in space is used to generate the snapshots and a POD basis is computed with respect to the $X = L^2(\Omega)$-inner product. The decay of the normalized eigenvalues is shown in Fig. 6. It compares the use of a smooth double-well potential $\Psi_0(\varphi) = \frac{1}{4}\varphi^4$ (pDWE) to the use of a Moreau–Yosida relaxation of the double-obstacle potential given as $\Psi_0(\varphi) = \frac{s}{r}(|\max(0, \varphi - 1)|^r + |\min(0, \varphi + 1)|^r)$ for different values of r (DOEr). We observe that the smoother the considered free energy is, the faster is the decay of the eigenvalues. This is similar to a well-known behavior in Fourier analysis, where the decay of the Fourier coefficients depends on the smoothness of the object. For POD reduced-order modeling, this means that if a potential with lower regularity is used, then more POD modes are needed for an adequate approximation than using a smooth potential.

In future research, we plan to apply POD model order reduction for the Cahn–Hilliard equations using a nonsmooth double-obstacle potential. This involves

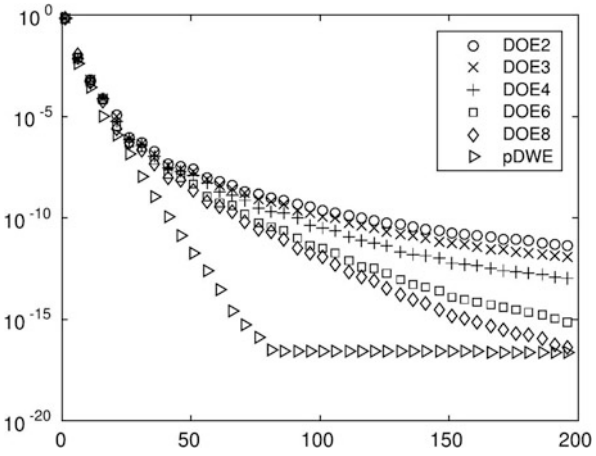

Fig. 6 Decay of the normalized eigenvalues for the phase field φ considering a Moreau–Yosida relaxation (DOEr) for different relaxation parameters r and a polynomial free energy (pDWE)

reduced-order modeling for variational inequalities, see, e.g., [15] for a reduced-order technique for Black–Scholes and Heston models.

For the second aspect (ii), let us consider the following setting: the spatial domain is $\Omega = (0, 2) \times (0, 1) \subset \mathbb{R}^2$, the mobility is $m = 1.0$, the interface parameter is $\epsilon = 0.02$, and the potential Ψ_0 is the smooth double-well energy. The initial condition has the shape of an ellipse. We consider a solenoidal velocity field $y = (y_1, y_2)$ given by

$$y_1(x) = c \sin(\pi x_0) \cdot \cos(\pi x_1), \;\; y_2(x) = -c \sin(\pi x_1) \cdot \cos(\pi x_0) \text{ for } x_0 \leq 1$$

and

$$y_1(x) = -c \sin(\pi x_0) \cdot \cos(\pi x_1), \;\; y_2(x) = c \sin(\pi x_1) \cdot \cos(\pi x_0) \text{ for } x_0 > 1,$$

where $x = (x_0, x_1)$. In this example, we choose $c = 70$, such that the velocity field leads to a break-up of the ellipse into two separate droplets. This topology change can be handled naturally due to the consideration of a diffuse interface approach.

For the temporal discretization, we use an unconditional gradient stable scheme based on a convex–concave splitting of the potential according to [23, 24]. As time step size, we use $\tau = 2.5 \cdot 10^{-5}$ and perform $K = 300$ time steps. For the spatial discretization, we use h-adapted piecewise linear and continuous finite elements. The solutions to the adaptive finite-element simulation at initial time, half time, and end time with the associated adapted meshes are shown in Fig. 7. The number of node points varies between 16,779 and 19,808, and the finite-element simulation time is 1674 s.

In order to construct a POD reduced-order model, we utilize the adapted finite-element solutions for the phase field as snapshots in (4.1), where we choose $X = L^2(\Omega)$ for the norm and inner products. The resulting solutions for a POD reduced-order model of dimension $\ell = 10$ and $\ell = 20$ are shown in Fig. 8 at the initial, half, and end times. In the approximations using $\ell = 10$ POD modes, we observe oscillations due to the transport term, which are smoothened out by enlarging the

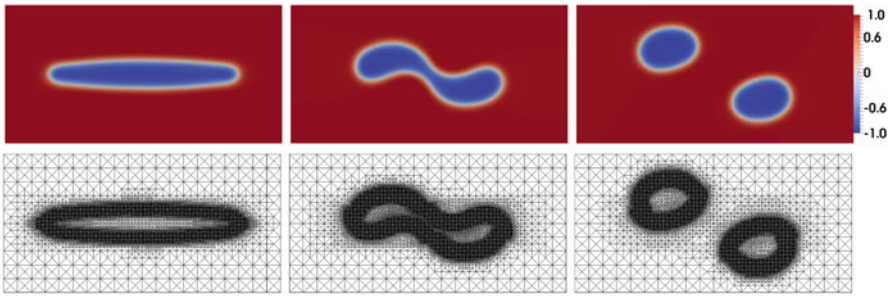

Fig. 7 Finite-element snapshots of the phase field at $t = 0$, $t = T/2$, and $t = T$ (top) with the associated adapted finite-element meshes (bottom)

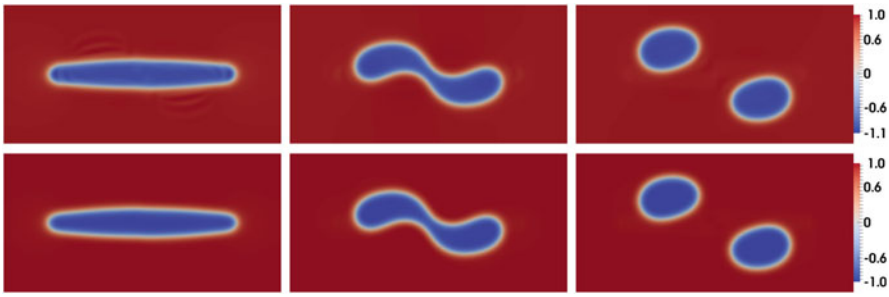

Fig. 8 POD reduced-order approximation of the phase field at $t = 0$, $t = T/2$, and $t = T$ using $\ell = 10$ POD modes (top) and $\ell = 20$ POD modes (bottom)

reduced dimension. We note that POD model order reduction for systems involving a dominant transport is challenging, and refer to [19, 56, 59, 64] for different solution concepts.

The relative $L^2(0, T; \Omega)$-error between the adaptive finite-element solution and the POD reduced-order solution using $\ell = 20$ POD modes is $2.793 \cdot 10^{-4}$. The solution time for the reduced-order simulation is 88 s, which leads to a speedup factor of 19 compared to the time needed for the adaptive finite-element simulation. Note that the reduced-order model still depends on the finite-element dimension, since an expansion of the reduced solution to the full-order model is needed for the evaluation of the nonlinearity. In order to enable an efficient evaluation of the nonlinearity that is related to the reduced-order dimension, the use of hyper-reduction methods like DEIM is needed. This leads to a further speedup, such that the solution of the reduced-order system takes only a fraction of seconds (compare, e.g., [28, Table 5]). However, especially in the case of lower regularity of the potential, we observe instabilities. In future research, we plan to derive a stable POD reduced-order model including hyper-reduction for systems with nonlinearities of low regularity. Moreover, we refer to [62] for an energy stable model order reduction for the Allen–Cahn equation.

For further details on POD with space-adapted snapshots and additional numerical test runs, we refer to [28, 31].

The speedup in the computation times when replacing the high-fidelity finite-element model by the POD reduced-order surrogate especially pays off in multi-query scenarios like optimal control. In this case, a repeated solution of the associated state and adjoint equations is necessary in order to find a minimum to a given cost functional. We refer to [33] for an optimal control of a Cahn–Hilliard system, where the control enters the equations as velocity in the transport term. A reduced-order model using space-adapted snapshot data is used. A different optimal control problem for the Cahn–Hilliard system is considered in [8], where the control enters as a right-hand side in (4.2a). Within a POD trust-region framework according to [5], the reduced-order model accuracy is evaluated by the Carter condition. This guarantees a relative gradient accuracy and indicates whether an enlargement of

the reduced dimension or a POD basis update with space-adapted snapshots at the current optimization iterate is necessary.

4.4 Stable POD Reduced-Order Modeling for Navier–Stokes with Space-Adapted Snapshots

Let us now consider the Navier–Stokes system (2.1a)–(2.1b) for a single-phase system in strong form, i.e.,

$$\partial_t v + (v \cdot \nabla)v - \frac{1}{Re}\Delta v + \nabla p = f \quad \text{in } (0,T)\times\Omega, \tag{4.4a}$$

$$\operatorname{div} v = 0 \quad \text{in } (0,T)\times\Omega, \tag{4.4b}$$

equipped with homogeneous Dirichlet boundary conditions $v = 0$ on $\partial\Omega$ and an initial condition for the velocity (2.1f). In order to derive a fully discrete formulation of (4.4), we first discretize in time using an implicit Euler scheme, which allows to use a different (adaptive) finite-element space at each time instance. Let $t_0 = 0 < t_1 < \ldots t_{K-1} = T$ denote a time grid with constant time step size τ, and let (V_h^i, Q_h^i) for $i = 0, \ldots, K-1$ denote inf–sup stable Taylor–Hood finite-element pairs. Then, the fully discrete Navier–Stokes systems reads as: for given $v_h^0 = v_a$, find $v_h^1 \in V_h^1, \ldots, v_h^{K-1} \in V_h^{K-1}$ and $p_h^1 \in Q_h^1, \ldots, p_h^{K-1} \in Q_h^{K-1}$ such that

$$\left(\frac{v_h^i - v_h^{i-1}}{\tau}, w\right) + ((v_h^i \cdot \nabla)v_h^i, w) + \frac{1}{Re}(\nabla v_h^i, \nabla w) + b(w, p_h^i) = \langle f(t_i), w\rangle \quad \forall w \in V_h^i, \tag{4.5a}$$

$$b(v_h^i, q) = 0 \quad \forall q \in Q_h^i, \tag{4.5b}$$

for $i = 1, \ldots, K-1$, where $(\cdot,\cdot)$ denotes the $L^2(\Omega)$-inner product and $\langle\cdot,\cdot\rangle$ is the duality pairing of $H_0^1(\Omega)$ with $H^{-1}(\Omega)$. Moreover, we introduce $b(w,q) := -(q, \nabla\cdot v)$ such that the strong-divergence-free condition (4.4b) is now postulated in a weak form in (4.5b). In order to derive the POD reduced-order model, we compute a POD basis from the space-adapted solutions from (4.5) according to Sect. 4.1. In particular, we introduce reduced spaces V_ℓ and Q_ℓ for the velocity and pressure and search for reduced approximations $\{v_\ell^1, \ldots, v_\ell^{K-1}\} \in V_\ell$ and $\{p_\ell^1, \ldots, p_\ell^{K-1}\}$

such that

$$\left(\frac{v_\ell^i - v_\ell^{i-1}}{\Delta t}, w\right) + ((v_\ell^i \cdot \nabla) v_\ell^i, w) + \frac{1}{Re}(\nabla v_\ell^i, \nabla w) + b(w, p_\ell^i) = \langle f(t_i), w \rangle \quad \forall w \in V_\ell, \tag{4.6a}$$

$$b(v_\ell^i, q) = 0 \quad \forall q \in Q_\ell. \tag{4.6b}$$

The difficulty consists in the fact that stability of (4.6) is not ensured for all choices of (V_ℓ, Q_ℓ). For this reason, in [32], we provide two solution concepts:

(i) A velocity ROM in the spirit of [58] using an optimal projection onto a weak-divergence-free space
(ii) A velocity–pressure ROM using a supremizer stabilization technique in the spirit of [17, 57]

In the first approach (i), we utilize the following optimal projection. For a given function $v \in X$, find a reference function $\tilde{v}$ in a reference velocity function space $\tilde{V}$ such that it fulfills

$$\min_{u \in \tilde{V}} \frac{1}{2}\|v - u\|_X^2 \quad \text{s.t.} \quad b(u, q) = 0 \quad \forall q \in \tilde{Q}.$$

This projection is computed either for each of the space-adapted velocity snapshots $\{v_h^1, \ldots, v_h^{K-1}\}$ or for each of the velocity POD basis functions $\{\psi_1^v, \ldots, \psi_\ell^v\}$ computed from velocity snapshots according to (4.1). Then, a common weak-divergence-free property is inherited in the reduced-order model, which leads to a cancelation of the pressure term and continuity equation from (4.6), such that the reduced system is stable by construction. Particular attention must be paid to the treatment of inhomogeneous boundary conditions, for which we refer to [32, Section 6] for details.

The second approach (ii) utilizes a supremizer enrichment technique. After computing separate POD bases $\{\psi_1^v, \ldots, \psi_\ell^v\}$ and $\{\psi_1^p, \ldots, \psi_\ell^p\}$ for the velocity and pressure, respectively, we enrich the reduced velocity space by stabilization functions. These are computed as follows: for a given $q \in L_0^2(\Omega)$ find $\mathbb{T}q \in \tilde{V}$ such that

$$(\mathbb{T}q, \phi)_{H_0^1(\Omega)} = b(\phi, q) \quad \forall \phi \in \tilde{V}.$$

Then, as supremizer functions, we choose $\{\mathbb{T}\psi_1^p, \ldots, \mathbb{T}\psi_\ell^p\}$. The inf–sup stability of the resulting velocity–pressure reduced-order model follows from the inf–sup stability of the finite-element model, see [32, Section 5.2] for the proof.

5 Outlook

In the second phase of the Priority Programme 1962, we consider shape optimization with instationary fluid flow in a diffuse interface setting. We will provide a well-posed formulation for shape optimization in instationary fluids with general cost functionals, which on the one hand allow for topological changes and impose no geometric constraints on the optimal shape, and on the other hand overcome some potential weaknesses of sharp interface models that are related to a loss of robustness. Moreover, a phase-field approach provides flexibility in data-driven model order reduction for efficient numerical shape optimization.

To achieve these goals, we combine the porous medium approach of [10] and a phase-field approach including a regularization by the Ginzburg–Landau energy. This results in a diffuse interface problem, which approximates a sharp interface problem for shape optimization in fluids that is penalized by a perimeter term. The related optimization problem then is a control in the coefficient optimal control problem where the phase field represents the control. For the fast numerical solution of those optimal control problems, we use POD-MOR techniques, which are based upon the findings and methods presented in Sects. 3 and 4.

Acknowledgments Many thanks to Christian Kahle for providing software libraries for the adaptive simulation and control of the Cahn–Hilliard Navier–Stokes system that we could use to build on.

References

1. S. Aland, S. Boden, A. Hahn, F. Klingbeil, M. Weismann, and S. Weller, *Quantitative comparison of Taylor flow simulations based on sharp-interface and diffuse-interface models*, Int. J. Numer. Meth. Fluids, **73** (2013), 344–361.
2. S. Aland, J. Lowengrub, and A. Voigt, *Particles at fluid-fluid interfaces: A new Navier-Stokes-Cahn-Hilliard surface-phase-field-crystal model*, Phys. Rev. E, **86** (2012), 046321.
3. S. Aland, and A. Voigt, *Benchmark computations of diffuse interface models for two-dimensional bubble dynamics*, Int. J. Numer. Meth. Fluids, **69** (2012), 747–761.
4. J. O. Alff, *Modellordnungsreduktion für das Cahn-Hilliard System*, Bachelorarbeit, Universität Hamburg (2015).
5. E. Arian, M. Fahl, and E. W. Sachs, *Trust-region Proper Orthogonal Decomposition for Flow Control*, ICASE Report No. 2000-25, ICASE, NASA Langley Research Center, Hampton (2000).
6. H. Abels, H. Garcke, and G. Grün, *Thermodynamically consistent, frame indifferent diffuse interface models for incompressible two-phase flows with different densities*, Math. Models Methods Appl. Sci. **22**(3) (2012).
7. A. Alla, and N. Kutz, *Nonlinear model order reduction via dynamic mode decomposition*, SIAM J. Sci. Comput., **39**(5) (2017), B778–B796.
8. J. O. Alff, *Trust Region POD for Optimal Control of Cahn-Hilliard Systems*, Master's Thesis, Universität Hamburg (2018).
9. V. Barbu, *Optimal control of variational inequalities*, Research Notes in Math., Pitman (Advanced Publishing Program), Boston, MA, **100** (1984).

10. T. Borrvall, and J. Petersson, *Topology optimization of fluids in Stokes flow*, Internat. J. Numer. Methods Fluids **41**(1) (2003), 77–107.
11. F. Boyer, *A theoretical and numerical model for the study of incompressible mixture flows*, Computers & fluids, **31** (2002), 41–68.
12. F. Boyer, L. Chupin, and P. Fabrie, *Numerical study of viscoelastic mixtures through a Cahn-Hilliard flow model*, Eur. J. Mech. B Fluids, **23** (2004), 759–780.
13. C. Brett, C. M. Elliott, M. Hintermüller, and C. Löbhard, *Mesh adaptivity in optimal control of elliptic variational inequalities with point-tracking of the state*, Interfaces Free Bound., **17**(1) (2015), 21–53.
14. J. Blowey, and C. Elliot, *The Cahn-Hilliard gradient theory for phase separation with non-smooth free energy. Part I: Mathematical analysis*, Eur. J. Appl. Math., **2** (1991), 233–280.
15. O. Burkovska, B. Haasdonk, J. Salomon, and B. Wohlmuth, *Reduced Basis Methods for Pricing Options with the Black-Scholes and Heston Models*, SIAM J. Financial Math., **6**(1) (2015), 685–712.
16. M. Barrault, Y. Maday, N. C. Nguyen, and A. T. Patera, *An "empirical interpolation" method: application to efficient reduced-basis discretization of partial differential equations*, C. R. Acad. Sci. Paris, **339**(9) (2004), 667–672.
17. F. Ballarin, A. Manzoni, A. Quarteroni, and G. Rozza, *Supremizer stabilization of POD-Galerkin approximation of parametrized steady incompressible Navier-Stokes equations*, Int. J. Numer. Methods Eng., **102**(5) (2015), 1136–1161.
18. P. Colli, M. H. Farshbaf-Shaker, G. Gilardi, and J. Sprekels, *Optimal boundary control of a viscous Cahn-Hilliard system with dynamic boundary condition and double obstacle potentials*, SIAM J. Control Optim., **53** (2015), 2696–2721.
19. N. Cagniart, Y. Maday, and B. Stamm, *Model Order Reduction with large Convection Effects*, Contributions to Partial Differential Equations, Springer, (2018), 131–150.
20. S. Chaturantabut, and D. C. Sorensen, *Nonlinear model order reduction via discrete empirical interpolation*, SIAM J. Sci. Comput., **32**(5) (2010), 2737–2764.
21. H. Ding, P. D. M. Spelt, and C. Shu, *Diffuse interface model for incompressible two-phase flows with large density ratios*, J. Comput. Phys., **226** (2007), 2078–2095.
22. S. Eckert, P. A. Nikrityuk, B. Willers, D. Räbiger, N. Shevchenko, H. Neumann-Heyme, V. Travnikov, S. Odenbach, A. Voigt, and K. Eckert, *Electromagnetic melt flow control during solidification of metallic alloys*, Eur. Phys. J-Spec. Top., **220** (2013), 123–137.
23. C. M. Elliott, and A. Stuart, *The global dynamics of discrete semilinear parabolic equations*, SIAM J. Numer. Anal., **30**(6) (1993), 1622–1663.
24. D. J. Eyre, *Unconditionally gradient stable time marching the Cahn-Hilliard equation*, MRS Proceedings, **529** (1998).
25. C. M. Elliott, and Z. Songmu, *On the Cahn-Hilliard equation*, Arch. Rational Mech. Anal., **96** (1986), 339–357.
26. S. Frigeri, E. Rocca, and J. Sprekels, *Optimal distributed control of a nonlocal Cahn-Hilliard/Navier-Stokes system in two dimensions*, SIAM J. Control Optim., **54** (2016), 221–250.
27. C. G. Gal, and M. Grasselli, *Asymptotic behavior of a Cahn-Hilliard-Navier-Stokes system in 2D*, Ann. Inst. H. Poincaré Anal. Non Linéaire, **27** (2010), 401–436.
28. C. Gräßle, and M. Hinze, *POD reduced-order modeling for evolution equations utilizing arbitrary finite element discretizations*, Adv. Comput. Math., **44**(6) (2018), 1941–1978.
29. H. Garcke, M. Hinze, and C. Kahle, *A stable and linear time discretization for a thermodynamically consistent model for two-phase incompressible flow*, Appl. Numer. Math., **99** (2016), 151–171.
30. H. Garcke, M. Hinze, and C. Kahle, *Optimal control of time-discrete two-phase flow driven by a diffuse-interface model*, ESAIM Control Optim. Calc. Var., **25** (2019), 2018006, 31.
31. C. Gräßle, and M. Hinze, *The combination of POD model reduction with adaptive finite element methods in the context of phase field models*, PAMM, **17**(1) (2017), 47–50.

32. C. Gräßle, M. Hinze, J. Lang, and S. Ullmann, *POD model order reduction with space-adapted snapshots for incompressible flows*, accepted for publication in Adv. Comput. Math. (2018), preprint available https://arxiv.org/abs/1810.03892.
33. C. Gräßle, M. Hinze, and N. Scharmacher, *POD for optimal control of the Cahn-Hilliard system using spatially adapted snapshots*, in Numerical Mathematics and Advanced Applications ENUMATH 2017 (2019), 703–711.
34. G. Grün, and F. Klingbeil, *Two-phase flow with mass density contrast: stable schemes for a thermodynamic consistent and frame-indifferent diffuse interface model*, J. Comput. Phys., **257** (2014), 708–725.
35. M. Hintermüller, M. Hinze, C. Kahle, *An adaptive finite element Moreau-Yosida-based solver for a coupled Cahn-Hilliard/Navier-Stokes system*, J. Comput. Phys., **235** (2013), 810–827.
36. M. Hintermüller, M. Hinze, C. Kahle, and T. Keil, *A goal-oriented dual-weighted adaptive finite element approach for the optimal control of a nonsmooth Cahn-Hilliard Navier-Stokes system*, Optim. Eng., **19**(3) (2018), 629–662.
37. M. Hintermüller, M. Hinze, and M. Tber, *An adaptive finite element Moreau-Yosida-based solver for a nonsmooth Cahn-Hilliard problem*, Optim. Method. Softw., **25** (2011), 777–811.
38. M. Hintermüller, and T. Surowiec, *A bundle-free implicit programming approach for a class of elliptic MPECs in function space*, Math. Program., **160**(1) (2016), 271–305.
39. M. Hintermüller, T. Keil, and D. Wegner, *Optimal control of a semidiscrete Cahn-Hilliard-Navier-Stokes system with nonmatched fluid densities*, SIAM J. Control Optim., **55**(3) (2017), 1954–1989.
40. M. Hintermüller, and I. Kopacka, *Mathematical programs with complementarity constraints in function space: C- and strong stationarity and a path-following algorithm*, SIAM J. Optim., **20** (2009), 868–902.
41. M. Hintermüller, B. S. Mordukhovich, and T. M. Surowiec, *Several approaches for the derivation of stationarity conditions for elliptic MPECs with upper-level control constraints*, Math. Program., **146** (2014), 555–582.
42. M. Hintermüller, and D. Wegner, *Distributed optimal control of the Cahn-Hilliard system including the case of a double-obstacle homogeneous free energy density*, SIAM J. Control Optim., **50** (2012), 388–418.
43. M. Hintermüller, and D. Wegner, *Distributed and boundary control problems for the semidiscrete Cahn-Hilliard/Navier-Stokes system with nonsmooth Ginzburg-Landau energies*, in Topological Optimization and Optimal Transport, M. Bergounioux, E. Oudet, M. Rumpf, G. Carlier, T. Champion, F. Santambrogio, eds., Radon Series on Computational and Applied Mathematics, De Gruyter **17** (2017), 40–63.
44. M. Hintermüller, and D. Wegner, *Optimal control of a semidiscrete Cahn-Hilliard-Navier-Stokes system*, SIAM J. Control Optim., **52** (2014), 747–772.
45. P. C. Hohenberg, and B. I. Halperin, *Theory of dynamic critical phenomena*, Rev. Mod. Phys., **49**(3) (1977), 435.
46. S. Hysing, S. Turek, D. Kuzmin, N. Parolini, E. Burman, S. Ganesan, and L. Tobiska, *Quantitative benchmark computations of two-dimensional bubble dynamics*, Int. J. Numer. Meth. Fluids, **60**(11) (2009), 1259–1288.
47. J. Jarušek, M. Krbec, M. Rao, and J. Sokołowski, *Conical differentiability for evolution variational inequalities*, J. Differ. Equations, **193** (2003), 131–146.
48. J. Kim, K. Kang, and J. Lowengrub, *Conservative multigrid methods for Cahn-Hilliard fluids*, J. Comput. Phys., **193** (2004), 511–543.
49. J. Kim, and J. Lowengrub, *Interfaces and multicomponent fluids*, Encyclopedia of Mathematical Physics, (2004), 135–144.
50. K. Kunisch, and S. Volkwein, *Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics*, SIAM J. Numer. Anal., **40**(2) (2002), 492–515.
51. J. Lowengrub, and L. Truskinovsky, *Quasi-incompressible Cahn-Hilliard fluids and topological transitions*, Proc. R. Soc. Lond. A., **454** (1998), 2617–2654.
52. Z.-Q. Luo, J.-S. Pang, and D. Ralph, *Mathematical programs with equilibrium constraints*, Cambridge University Press, Cambridge, 1996.

53. B. S. Mordukhovich, *Variational analysis and generalized differentiation II, Applications*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], **331**, Springer-Verlag, Berlin, 2006.
54. J. Outrata, M. Kočvara, and J. Zowe, *Nonsmooth approach to optimization problems with equilibrium constraints: theory, applications and numerical results*, Nonconvex Optimization and Its Applications, **28**, Springer US, 1998.
55. S. Praetorius, and A. Voigt, *A Navier–Stokes phase-field crystal model for colloidal suspensions*, The Journal of chemical physics, **142**(15) (2015), 154904.
56. J. Reiss, P. Schulze, J. Sesterhenn, and V. Mehrmann, *The shifted proper orthogonal decomposition: a mode decomposition for multiple transport phenomena*, SIAM J. Sci. Comput., **40**(3) (2018), A1322–A1344.
57. G. Rozza, and K. Veroy, *On the stability of the reduced basis method for Stokes equations in parametrized domains*, Comput. Method Appl. M., **196**(7) (2007), 1244–1260.
58. L. Sirovich, *Turbulence and the dynamics of coherent structures I-III*, Q. Appl. Math., **45**(3) (1987), 561–590.
59. M. Sieber, C. O. Paschereit, and K. Oberleithner, *Spectral proper orthogonal decomposition*, J. Fluid Mech., **792** (2016), 798–828.
60. H. Scheel, and S. Scholtes, *Mathematical programs with complementarity constraints: stationarity, optimality, and sensitivity*, Math. Oper. Res., **25** (2000), 1–22.
61. T. Tachim Medjo, *Optimal control of a Cahn-Hilliard-Navier-Stokes model with state constraints*, J. Convex Anal., **22**(4) (2015), 1135–1172.
62. M. Uzunca, and B. Karasözen, *Energy stable model order reduction for the Allen-Cahn equation*, in Model Reduction of Parametrized Systems, Springer, (2017), 403–419.
63. S. Ullmann, M. Rotkvic, and J. Lang, *POD-Galerkin reduced-order modeling with adaptive finite element snapshots*, J. Comput. Phys., **325** (2016), 244–258.
64. D. Wells, Z. Wang, X. Xie, and T. Iliescu, *An evolve-then-filter regularized reduced order model for convection-dominated flows*, Int. J. Numer. Meth. Fluids, **84** (10) (2017), 598–615.
65. J. M. Yong, and S. M. Zheng, *Feedback stabilization and optimal control for the Cahn-Hilliard equation*, Nonlinear Anal., **17** (1991), 431–444.
66. B. Zhou, *Simulations of polymeric membrane formation in 2D and 3D*, PhD thesis, Massachusetts Institute of Technology, 2006.
67. J. Zowe, and S. Kurcyusz, *Regularity and stability for the mathematical programming problem in Banach spaces*, Appl. Math. Optim., **5** (1979), 49–62.

Safeguarded Augmented Lagrangian Methods in Banach Spaces

Christian Kanzow, Veronika Karl, Daniel Steck, and Daniel Wachsmuth

Abstract This chapter presents a state-of-the-art survey for safeguarded augmented Lagrangian methods for constrained optimization problems in Banach spaces. The difference between the classical augmented Lagrangian method and its safeguarded version lies in the update of the multiplier estimates. The safeguarded method has significantly stronger global convergence properties than the classical algorithm. Local and rate-of-convergence results are also summarized. Some numerical results illustrate the practical behavior of the safeguarded augmented Lagrangian approach.

Keywords Constrained optimization · Augmented Lagrangian method · Multiplier-penalty method · Banach space · Global convergence · Local convergence · Robinson constraint qualification

Mathematics Subject Classification (2020) 49M20, 65K10, 90C48

1 Introduction

This chapter is dedicated to a thorough discussion of the augmented Lagrangian method (ALM) for constrained minimization problems of the form

$$(P) \qquad \underset{x \in C}{\text{minimize}}\ f(x) \quad \text{subject to} \quad G(x) \in K, \tag{1.1}$$

This research was supported by the German Research Foundation (DFG) within the priority program "Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization" (SPP 1962) under grant numbers KA 1296/24-1 and Wa 3626/3-1.

C. Kanzow (✉) · V. Karl · D. Steck · D. Wachsmuth
University of Würzburg, Institute of Mathematics, Würzburg, Germany
e-mail: kanzow@mathematik.uni-wuerzburg.de; veronika.karl@mathematik.uni-wuerzburg.de; daniel.wachsmuth@mathematik.uni-wuerzburg.de

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_10

where X and Y are real Banach spaces, $f : X \to \mathbb{R}$ and $G : X \to Y$ are continuously differentiable functions, and $C \subseteq X$ and $K \subseteq Y$ are nonempty closed convex sets. The feasible set of (P) will be denoted by

$$\Phi := \{x \in C : G(x) \in K\}.$$

To facilitate the application of the augmented Lagrangian technique, we assume that $i : Y \hookrightarrow H$ densely for some real Hilbert space H. This implies that we are working in the *Gel'fand triple* framework

$$Y \overset{i}{\hookrightarrow} H \cong H^* \overset{i^*}{\hookrightarrow} Y^*. \tag{1.2}$$

Furthermore, we assume that there is a closed convex set $\mathcal{K} \subseteq H$ such that $i^{-1}(\mathcal{K}) = K$. This allows us to interpret the constraint $G(x) \in K$ equivalently as $G(x) \in \mathcal{K}$. Note that we will usually suppress the embedding for the sake of brevity.

It should be stressed that the above framework is extremely general, and the resulting augmented Lagrangian method therefore covers a very broad spectrum of applications. Moreover, many prominent problem classes can be recovered as special cases of (P). Here, we apply the safeguarded augmented Lagrangian method in order to solve (P).

Historically, the augmented Lagrangian technique was first developed for nonlinear programs (in finite dimension). Indeed, the algorithm goes back to the seminal works by Hestenes [33] and Powell [65], and in its early days, it was commonly referred to as the *method of multipliers*. The technique was further developed by many authors in the later parts of the 20th century, including Rockafellar [68–70], Bertsekas [9], and Conn, Gould, and Toint [21–23], who created the well-known LANCELOT software package. The algorithm was rediscovered by Andreani, Birgin, Martínez, and co-authors in [1, 2, 11, 12], a series of publications which culminated in the book [13] and the corresponding ALGENCAN software package.

In today's nonlinear programming landscape, algorithms such as interior point methods [29, 31] or sequential quadratic programming [31, 45] are often preferred to methods of augmented Lagrangian type, mainly due to their fast local convergence characteristics. In contrast, the augmented Lagrangian method possesses very strong global convergence properties, and it has been found to work rather well on degenerate problem classes such as problems with complementarity constraints [46]. A state-of-the-art local convergence analysis of the ALM for nonlinear programming is given in [27]. More discussion on nonlinear programming in general, and on the corresponding algorithms, can be found in [9, 10, 24, 62] and in the encyclopedia [28].

One of the main motivations for the generalization of augmented Lagrangian methods to the level of generality represented by (P) is the advent of function space optimization problems. Some early references in this context include [6, 7, 39–42, 76] and the book [30]. Most of these publications are restricted to very specific problem settings such as convex optimization problems or finite-dimensional

constraints. In [8, 43], an augmented Lagrangian-type penalty scheme was proposed, in combination with a semismooth Newton method, for the solution of state-constrained optimal control problems. The resulting method came to be known as *Moreau–Yosida regularization*; it was further developed in [34, 35], and it is today considered a standard approach for state-constrained optimal control [37, 44, 75]. Some other techniques for such problems include Lavrentiev regularization [36, 59], interior point methods [56, 72], and the so-called virtual control approach [55], which is related to the augmented Lagrangian technique [54].

The purpose of this chapter is to collect the recent developments and to summarize the convergence theory of the safeguarded method applied to Banach space optimization problems in a uniform framework. To this end, we first recall some background material and state some preliminary results in Sect. 2. We then provide a self-contained motivation of the augmented Lagrangian method in Sect. 3. A state-of-the-art summary of the global and local convergence properties of the augmented ALM is then provided in Sects. 4 and 5, respectively. Numerical results for a variety of applications are given in Sect. 6. We then close the chapter with some final remarks in Sect. 7.

2 Background Material

This chapter summarizes several concepts and results from optimization theory, Banach spaces, and variational analysis which will be used later in our subsequent convergence theory. Most results are known, so we refer to the existing literature; occasionally, we provide a proof if either this proof is very short or we were not able to find an explicit reference.

2.1 Cones

This section is dedicated to the study of some basic objects that are useful when characterizing the geometric structure of sets in Banach spaces. Many aspects of the geometry of sets can be characterized through the so-called cones (see below), and these play a major role in variational analysis, convex analysis, and optimization theory. The material discussed here incorporates elements from multiple books, e.g., [5, 14, 16].

Let $S \subseteq X$ be a nonempty set. We say that S is a *cone* if $\alpha S \subseteq S$ for all $\alpha > 0$. Given an arbitrary set $S \subseteq X$, we denote by

$$S^\circ := \{\phi \in X^* : \langle \phi, s \rangle \leq 0 \text{ for every } s \in S\}$$

the *polar cone* of S. Note that $S^\circ \subseteq X^*$. If X is a real Hilbert space, we treat S° as a subset of X.

Definition 2.1 (Tangent and Normal Cones) Let $C \subseteq X$ be an arbitrary set and $x \in C$. Then, we define

(a) the *tangent cone* $\mathcal{T}_C(x)$ as

$$\mathcal{T}_C(x) := \big\{d \in X : \exists \{x^k\} \subseteq C,\ t_k \downarrow 0 \text{ such that } x^k \to x \text{ and } (x^k - x)/t_k \to d\big\}.$$

(b) the *normal cone* $\mathcal{N}_C(x)$ as

$$\mathcal{N}_C(x) := \{\phi \in X^* : \langle \phi, y - x\rangle \le 0\ \forall y \in C\}.$$

For $x \notin C$, both cones are defined as the empty set.

If X is a real Hilbert space, we treat $\mathcal{N}_C(x)$ as a subset of X instead of X^*. The normal cone is always a closed set and satisfies the polarity relation

$$\mathcal{N}_C(x) = \mathcal{T}_C(x)^\circ,$$

which is sometimes also taken as the definition of the normal cone and makes sense also for possibly nonconvex sets C. It should be noted, however, that there are a variety of different normal cones for general sets (see, for instance, [60]). Therefore, to avoid any ambiguity, we will reserve the symbol $\mathcal{N}_C$ for the case where C is convex.

The normal cone can be used to formulate a simple Fermat-type optimality condition.

Theorem 2.2 (Necessary Optimality Condition, [16]) *Let $f : X \to \mathbb{R}$ be a continuously differentiable mapping and $C \subseteq X$ a nonempty closed convex set. If $\bar{x}$ is a local minimizer of f on C, then $0 \in f'(\bar{x}) + \mathcal{N}_C(\bar{x})$.*

The following is a famous decomposition theorem involving a closed convex cone in a Hilbert space and its polar.

Lemma 2.3 (Moreau Decomposition, [61]) *Let H be a real Hilbert space and $K \subseteq H$ a nonempty closed convex cone. Then, every $y \in H$ admits a unique decomposition $y = y_1 + y_2$ with $K \ni y_1 \perp y_2 \in K^\circ$. Indeed, $y_1 = P_K(y)$ and $y_2 = P_{K^\circ}(y)$.*

We now turn to another object that describes some aspects of the geometric structure of convex sets.

Definition 2.4 (Recession Cone) Let $C \subseteq X$ be a nonempty convex set. Then, the *recession cone* of C is the set $C_\infty := \{x \in X : x + C \subseteq C\}$.

The recession cone is always nonempty (since $0 \in C_\infty$) and a convex cone. Moreover, if C is closed, then so is C_∞. If the set C is a convex *cone*, then it is easy to see that $C_\infty = C$. On the other hand, if C is not a cone, then the recession cone can often be used as a substitute for C in situations where a conical structure

is necessary. This is the case, for instance, in the context of (partial) order relations, which closely correspond to convex cones, see Sect. 2.2.

The following result provides some information on the polar cone $C_\infty^\circ := (C_\infty)^\circ$.

Lemma 2.5 *Let H be a real Hilbert space and $C \subseteq H$ a nonempty convex set. Then, $\{y \in H : \sup_{w\in C}(w, y) < +\infty\} \subseteq C_\infty^\circ$. In particular, $\mathcal{N}_C(y) \subseteq C_\infty^\circ$ for all $y \in C$.*

Proof Let $y \in H$ be a point with $(w, y) \leq c$ for some $c \in \mathbb{R}$ and all $w \in C$. Let $x \in C_\infty$, and choose an arbitrary $x_0 \in C$. Then, $x_0 + tx \in C$ for all $t > 0$, and hence $(x_0 + tx, y) \leq c$. This cannot hold for all $t > 0$ if $(x, y) > 0$. Hence, $(x, y) \leq 0$, and $y \in C_\infty^\circ$. □

The set $\{y \in H : \sup_{w\in C}(w, y) < +\infty\}$ in the statement of Lemma 2.5 is often called the *barrier cone* of C. Note that the inclusion stated in the lemma can be strict. In particular, there are situations where the barrier cone is not closed, and this makes it a priori impossible for it to equal C_∞°, which is always a closed cone by virtue of polarity. An example for this phenomenon can be found in [5, Exercise 6.23].

2.2 Convex Functions and Concave Operators

Convex functions play a central role in optimization theory. Occasionally, we write $\partial f(x)$ for the subdifferential of a convex function f in x, but most of the time the underlying mapping f will be differentiable. One of the most fundamental examples of a convex function is the distance function $d_C : X \to \mathbb{R}$ to a convex set $C \subseteq X$. Note that the following result holds for an arbitrary Banach space X, not necessarily a Hilbert space.

Lemma 2.6 (Distance Function, [5, 64]) *Let $C \subseteq X$ be a nonempty convex set. Then, the function $d_C : X \to \mathbb{R}$, $d_C(x) := \inf_{y\in C} \|x - y\|_X$, is convex and nonexpansive.*

It is easy to see that the square of a nonnegative convex function is again convex. Thus, in the setting of Lemma 2.6, the squared distance function d_C^2 is also a convex function. If the space X is a real Hilbert space, then the squared distance function enjoys a much stronger form of regularity.

Lemma 2.7 ([5, Cor. 12.31]) *Let X be a real Hilbert space and $C \subseteq X$ a nonempty closed convex set. Then, the squared distance function d_C^2 is convex and continuously differentiable on X with $(d_C^2)'(x) = 2(x - P_C(x))$ for all $x \in X$.*

Recall that there exist several different continuity notions in infinite-dimensional spaces, based on the topology used within these spaces or whether a (weak) sequential continuity or (weak) lower semicontinuity is considered. The following well-known result states that several continuity properties coincide within the class of convex functions.

Proposition 2.8 ([5, Thm. 9.1]) *Let $C \subseteq X$ be a closed convex set and $f : C \to \mathbb{R}$ a convex function. Then, the following are equivalent:*

(i) f is lower semicontinuous,
(ii) f is weakly lower semicontinuous, and
(iii) f is weakly sequentially lower semicontinuous.

The theory of convex functions is useful for a wide variety of application problems. There are, however, certain practical scenarios where convexity properties of nonlinear *operators* $G : X \to Y$ are necessary, with X and Y real Banach spaces. More specifically, assume that we are dealing with an inclusion of the form

$$G(x) \in K, \quad K \subseteq Y \text{ a closed convex set.} \tag{2.1}$$

Ideally, we would like to work with a generalized notion of convexity which takes into account the mapping G and the geometry of the set K. To this end, assume for the moment that the set K in (2.1) is a closed convex *cone*. Then, K induces the order relation

$$a \leq_K b \ :\Longleftrightarrow\ b - a \in K, \tag{2.2}$$

and K itself can be regarded as the nonnegative cone with respect to $\leq_K$. Thus, (2.1) can be rewritten as $G(x) \geq_K 0$, which suggests that the appropriate convexity notion in this case is a generalized type of concavity with respect to the order relation $\leq_K$. This property takes on the form

$$G\big((1-t)x + ty\big) \geq_K (1-t)G(x) + tG(y) \quad \text{for all } x, y \in X,\ t \in [0,1].$$

The above property is usually called *K-concavity*, and it is in fact a special case of the general concept which we define below. In the case where K is not a cone, the recession cone K_∞ turns out to be a useful substitute to define the order relation (2.2).

Definition 2.9 (Concave Operator) Let $G : X \to Y$ be an arbitrary mapping and $K \subseteq Y$ a closed convex set with recession cone K_∞. We say that G is *K_∞-concave* if

$$G\big((1-t)x + ty\big) \geq_K (1-t)G(x) + tG(y) \quad \text{for all } x, y \in X,\ t \in [0,1],$$

where $\leq_K$ is the order relation defined by $a \leq_K b :\Longleftrightarrow b - a \in K_\infty$.

Let us now discuss the analytical consequences of generalized convexity (or concavity) in the sense of Definition 2.9. The resulting properties can be deduced by discussing situations in which the K_∞-concavity of G yields the (ordinary) convexity of a suitable composite mapping involving G.

We say that a mapping $m : Y \to \mathbb{R}$ is *K_∞-decreasing* if it is monotonically decreasing with respect to the order $\leq_K$, i.e., if $m(y_1) \leq m(y_2)$ whenever $y_1 \geq_K y_2$.

Theorem 2.10 *Let X and Y be real Banach spaces, $K \subseteq Y$ a nonempty closed convex set, and $G : X \to Y$ a K_∞-concave operator. Then,*

(a) If $m : Y \to \mathbb{R}$ is convex and K_∞-decreasing, then $m \circ G$ is convex.
(b) The function $d_K \circ G : X \to \mathbb{R}$ is convex.
(c) If $\lambda \in K_\infty^\circ$, then $x \mapsto \langle \lambda, G(x) \rangle$ is convex.
(d) The set $M := \{x \in X : G(x) \in K\}$ is convex.

Proof The proof can be found in [49, Lemma 2.1]. □

2.3 Pseudomonotone Operators

We first recall the following notion of *pseudomonotonicity* in the sense of Brezis [17].

Definition 2.11 (Pseudomonotonicity) We say that an operator $F : X \to X^*$ is *pseudomonotone* if whenever

$$\{x^k\} \subseteq X, \quad x^k \rightharpoonup x, \quad \text{and} \quad \limsup_{k\to\infty} \langle F(x^k), x^k - x \rangle \leq 0,$$

then

$$\langle F(x), x - y \rangle \leq \liminf_{k\to\infty} \langle F(x^k), x^k - y \rangle \quad \text{for all } y \in X.$$

Despite its somewhat peculiar appearance, the notion of pseudomonotonicity will play a fundamental role in the subsequent theory. Some sufficient conditions for pseudomonotone operators are summarized in the following lemma. This result illustrates that the class of pseudomonotone operators is quite large.

Lemma 2.12 (Sufficient Conditions for Pseudomonotonicity) *Let X be a real Banach space and $T, U : X \to X^*$ given operators. Then,*

(a) If T is monotone and continuous, then T is pseudomonotone.
(b) If, for every $y \in X$, the mapping $x \mapsto \langle T(x), x - y \rangle$ is weakly sequentially lsc, then T is pseudomonotone.
(c) If T is completely continuous, then T is pseudomonotone.
(d) If T is continuous and $\dim(X) < +\infty$, then T is pseudomonotone.
(e) If T and U are pseudomonotone, then $T + U$ is pseudomonotone.

Proof (b) is obvious. The remaining assertions can be found in [77, Prop. 27.6]. □

It follows from the above observations that the concept of pseudomonotone operators provides a unified approach to different classes of operators, including monotone and completely continuous ones. Property (b) in the above lemma is occasionally referred to as *Ky–Fan hemicontinuity*.

2.4 KKT-Type Conditions

We define the *Lagrange function* or *Lagrangian* of (P) as the mapping

$$\mathcal{L} : X \times Y^* \to \mathbb{R}, \quad \mathcal{L}(x, \lambda) := f(x) + \langle \lambda, G(x) \rangle \tag{2.3}$$

and denote by $\mathcal{L}'$ the derivative of the Lagrangian with respect to x alone. Note that we do not include the abstract constraint C into the Lagrangian. The Lagrangian can be used to formulate the KKT system of (P) in the following way.

Definition 2.13 (KKT Point) A point $(\bar{x}, \bar{\lambda}) \in X \times Y^*$ is a *KKT point* of (P) if

$$-\mathcal{L}'(\bar{x}, \bar{\lambda}) \in \mathcal{N}_C(\bar{x}) \quad \text{and} \quad \bar{\lambda} \in \mathcal{N}_K(G(\bar{x})).$$

We say that $\bar{x} \in X$ is a *stationary point* of (P) if $(\bar{x}, \bar{\lambda})$ is a KKT point for some multiplier $\bar{\lambda} \in Y^*$ and denote by $\Lambda(\bar{x})$ the set of such multipliers.

For the KKT conditions to be necessary optimality conditions of (P), certain *constraint qualifications* are required; they ensure that the feasible set is well behaved and that, roughly speaking, the reconstruction of its geometry from first-order information is possible. One of the most fundamental constraint qualifications in infinite dimensions is the following one.

Definition 2.14 (Robinson Constraint Qualification) Let $x \in X$ be a feasible point for (P). We say that the *Robinson constraint qualification (RCQ)* holds in x if

$$0 \in \operatorname{int}\big[G(x) + G'(x)(C - x) - K\big].$$

The above condition was introduced by Robinson in [67] in the context of certain stability properties of nonlinear inclusions. In the context of finite-dimensional nonlinear programs, RCQ turns out to be equivalent to the well-known Mangasarian–Fromovitz constraint qualification. Under RCQ, the following first-order optimality condition holds.

Theorem 2.15 (KKT Conditions Under RCQ, [14, Thm. 3.9]) *Let $\bar{x}$ be a local minimizer of (P), and assume that RCQ holds in $\bar{x}$. Then, the set of Lagrange multipliers $\Lambda(\bar{x})$ is nonempty, closed, convex, and bounded in Y^*.*

In order to verify feasibility of (weak) limit points in our global convergence analysis, we will also need the following straightforward generalization of RCQ to possibly infeasible points. To keep a clear distinction, we call the resulting condition the *extended* Robinson constraint qualification, though its definition is essentially the same as for RCQ itself.

Definition 2.16 (Extended Robinson Constraint Qualification) Let $x \in X$ be an arbitrary, not necessarily feasible point. We say that the *extended Robinson*

constraint qualification (extended RCQ, ERCQ) holds in x if

$$0 \in \operatorname{int}\big[G(x) + G'(x)(C - x) - K\big].$$

An important property of ERCQ is that it guarantees that, whenever x is a stationary point of a certain measure of infeasibility, then x is actually a feasible point. We formulate this result in a slightly more general framework. The proof can be found in [15, Lemma 5.2].

Proposition 2.17 *Let $i : Y \hookrightarrow H$ densely for some real Hilbert space H, and let $\mathcal{K} \subseteq H$ be a closed convex set with $i^{-1}(\mathcal{K}) = K$. Let $\bar{x} \in X$ be a stationary point of the problem $\min_{x \in C} d^2_{\mathcal{K}}(G(x))$, and assume that ERCQ holds in $\bar{x}$ with respect to the constraint system of (P). Then, $G(\bar{x}) \in K$.*

Assume now that we have a point $\hat{x}$ which is "almost" a solution of (P). A popular definition in this context is that of *ε-minimizers*: given $\varepsilon > 0$, we say that $\hat{x} \in \Phi$ is an ε-minimizer of (P) if $f(\hat{x}) \leq f(x) + \varepsilon$ for all $x \in \Phi$. For such approximate minimizers, it is indeed possible to obtain an inexact analog of the KKT conditions. This result is usually called Ekeland's variational principle.

Proposition 2.18 (Ekeland's Variational Principle, [14, Thm. 3.23]) *Let $\bar{x} \in \Phi$ be an ε-minimizer of (P), let $\delta := \varepsilon^{1/2}$, and assume that RCQ holds at every $x \in B_\delta(\bar{x}) \cap \Phi$. Then, there exist another ε-minimizer $\hat{x}$ of (P) and $\lambda \in Y^*$ such that $\|\hat{x} - \bar{x}\|_X \leq \delta$,*

$$\operatorname{dist}\big(-\mathcal{L}'(\hat{x}, \lambda), \mathcal{N}_C(\hat{x})\big) \leq \delta, \quad \text{and} \quad \lambda \in \mathcal{N}_K(G(\hat{x})).$$

Many practical algorithms for constrained optimization iteratively construct a primal–dual sequence $\{(x^k, \lambda^k)\}$, which satisfies the KKT conditions in an asymptotic sense. This motivates to analyze such "sequential" analogues of the KKT conditions in more detail. The subsequent notion is also used by similar approaches in finite dimensions, see [3, 4, 13].

Definition 2.19 (Asymptotic KKT Sequence) We say that a sequence $\{(x^k, \lambda^k)\} \subseteq C \times Y^*$ is an *asymptotic KKT sequence* for (P) if there exist null sequences $\{\varepsilon^k\} \subseteq X^*$ and $\{r_k\} \subseteq \mathbb{R}$ such that, for all k,

$$\varepsilon^k - \mathcal{L}'(x^k, \lambda^k) \in \mathcal{N}_C(x^k) \quad \text{and} \quad \big\langle \lambda^k, y - G(x^k) \big\rangle \leq r_k \ \forall y \in K. \tag{2.4}$$

Our main aim in this section is to give sufficient conditions which guarantee that, if $\{(x^k, \lambda^k)\}$ is an asymptotic KKT sequence and $\bar{x}$ is a (possibly weak) limit point of $\{x^k\}$, then $\bar{x}$ is a stationary point of (P). In this context, it is worth mentioning that Definition 2.19 imposes no conditions on the attainment of feasibility. This aspect is left unspecified for the sake of flexibility; indeed, we will mainly be concerned with scenarios where $\bar{x}$ is some kind of limit point of $\{x^k\}$, and we already know from a preliminary analysis that $\bar{x}$ is a feasible point.

Note that, while the conditions posed in Definition 2.19 seem reasonably weak, it is possible to generalize the asymptotic KKT concept even further. In particular, in our formulation, the second inequality in (2.4) is assumed to hold *uniformly* on K. If K is unbounded, then it may be more natural to require some kind of uniformness of the inequality on bounded subsets of K. In any case, however, the augmented Lagrangian method that we will discuss later satisfies the uniform bound from (2.4), and a more general analysis is therefore not necessary for our purposes.

3 Motivation and Statement of the Algorithm

This section first recalls the original method of multipliers for equality constraints. It then presents a self-contained and simple approach for its generalization to abstract inequality constraints (in a Banach space setting). Finally, we give a formal statement of the overall method for a general problem of the form (P) and prove some preliminary properties of this method.

3.1 The Original Method of Multipliers

In its initial form, the method of multipliers is an algorithm for the solution of equality-constrained minimization problems in finite dimensions. Here, we present this original method in a slightly more general framework. Consider an equality-constrained optimization problem of the form

$$\underset{x \in C}{\text{minimize}}\, f(x) \quad \text{subject to} \quad h(x) = 0, \tag{3.1}$$

where $f : X \to \mathbb{R}$, $C \subseteq X$ is a closed convex set, and $h : X \to H$. We assume that X is a real Banach space and that H is a real Hilbert space. In the special case of the original method of multipliers, we have $X := \mathbb{R}^n$, $H := \mathbb{R}^m$ with $m, n \in \mathbb{N}$, and $C := X$.

The basic idea is to tackle (3.1) by combining elements of Lagrangian theory with a penalty-type scheme. Recall that the Lagrangian of the problem takes on the form $\mathcal{L}(x, \lambda) = f(x) + \big(\lambda, h(x)\big)$. By adding a positive multiple of $\|h(x)\|_H^2$, we penalize the violation of the equality constraint, thus ending up with the *augmented Lagrangian*

$$\mathcal{L}_\rho(x, \lambda) := f(x) + \big(\lambda, h(x)\big) + \frac{\rho}{2}\|h(x)\|_H^2. \tag{3.2}$$

From an algorithmic perspective, we now proceed as follows. Given a penalty parameter ρ_k and a current estimate λ^k of the Lagrange multiplier, we compute x^{k+1} as a minimizer (or approximate minimizer) of (3.2) on C so that, ideally,

x^{k+1} is close to feasibility (if ρ_k is large) and close to being a minimizer of the Lagrangian $\mathcal{L}(\cdot, \lambda^k)$. Let us assume, for the moment, that the functions f and h are continuously differentiable and that x^{k+1} is an exact minimizer of $\mathcal{L}_{\rho_k}(\cdot, \lambda^k)$ on C. Then, the standard first-order optimality conditions yield the inclusion

$$\mathcal{N}_C(x^{k+1}) \ni -\mathcal{L}'_{\rho_k}(x^{k+1}, \lambda^k) = -f'(x^{k+1}) - h'(x^{k+1})^*(\lambda^k + \rho_k h(x^{k+1})).$$

This immediately suggests $\lambda^{k+1} := \lambda^k + \rho_k h(x^{k+1})$ as the new estimate of the Lagrange multiplier, which is often called the *Hestenes–Powell multiplier update*.

After the above procedure is completed, the penalty parameter is updated based on a heuristic test. The most common option is to keep ρ_k if the constraint violation has decreased sufficiently, and to increase it otherwise. We thus end up with the following overall algorithm.

Algorithm 3.1 Original method of multipliers

Let $(x^0, \lambda^0) \in X \times H$, $\rho_0 > 0$, let $\gamma > 1$, $\tau \in (0, 1)$, and set $k := 0$.

Step 1. If (x^k, λ^k) satisfies a suitable termination criterion, STOP.
Step 2. Compute an approximate solution x^{k+1} of the problem

$$\underset{x \in C}{\text{minimize}}\ \mathcal{L}_{\rho_k}(x, \lambda^k). \tag{3.3}$$

Step 3. Update the vector of multipliers to $\lambda^{k+1} := \lambda^k + \rho_k h(x^{k+1})$.
Step 4. If $\|h(x^{k+1})\|_H \le \tau \|h(x^k)\|_H$ holds, set $\rho_{k+1} := \rho_k$; otherwise, set $\rho_{k+1} := \gamma \rho_k$.
Step 5. Set $k \leftarrow k + 1$, and go to Step 1.

3.2 Inequality Constraints and Slack Variables

Having established the classical multiplier method for equality-constrained problems, we now outline how the algorithm can be extended to the inequality-constrained case. To this end, we consider an optimization problem of the form (P), that is,

$$\underset{x \in C}{\text{minimize}}\ f(x) \quad \text{subject to} \quad G(x) \in K,$$

where, as before, $f : X \to \mathbb{R}$ and $G : X \to Y$ are given mappings, and $C \subseteq X$ and $K \subseteq Y$ are nonempty closed convex sets. Moreover, H is a real Hilbert space with $i : Y \hookrightarrow H$ densely, and $\mathcal{K} \subseteq H$ is a closed convex set with $i^{-1}(\mathcal{K}) = K$. In this setting, we can restate (P) as the problem

$$(P_H) \qquad \underset{x \in C}{\text{minimize}}\ f(x) \quad \text{subject to} \quad G(x) \in \mathcal{K}. \tag{3.4}$$

We can transform this problem into an equality-constrained problem by adding an artificial variable $s \in \mathcal{K}$, also called a *slack variable*. This results in the equality-constrained problem

$$\underset{(x,s)\in C\times\mathcal{K}}{\text{minimize}}\ f(x) \quad \text{subject to} \quad G(x) - s = 0.$$

In the context of the equality-constrained framework (3.1) from the previous section, this essentially amounts to defining the mapping $h : X \times H \to H$, $h(x, s) := G(x) - s$. The new problem is now an equality-constrained optimization problem on the space $X \times H$, and its augmented Lagrangian in the sense of (3.2) is given by

$$\mathcal{L}^s_\rho(x, s, \lambda) = f(x) + \big(\lambda, h(x, s)\big) + \frac{\rho}{2}\|h(x, s)\|^2_H. \tag{3.5}$$

In order to transform the augmented Lagrangian into a form where s is eliminated, observe that we can rewrite $\mathcal{L}^s_\rho$ as

$$\mathcal{L}^s_\rho(x, s, \lambda) = f(x) + \frac{\rho}{2}\left\|G(x) + \frac{\lambda}{\rho} - s\right\|^2_H - \frac{\|\lambda\|^2_H}{2\rho}. \tag{3.6}$$

Taking into account the constraint $s \in \mathcal{K}$, we can now minimize this formula with respect to s for each fixed $x \in X$. Since s occurs only in the middle term, the result involves, by definition, the squared distance function $d^2_{\mathcal{K}}$.

Definition 3.2 (Augmented Lagrange Function) For $\rho > 0$, the *augmented Lagrange function* or *augmented Lagrangian* of (P) is the function

$$\mathcal{L}_\rho : X \times H \to \mathbb{R}, \quad \mathcal{L}_\rho(x, \lambda) := f(x) + \frac{\rho}{2} d^2_{\mathcal{K}}\left(G(x) + \frac{\lambda}{\rho}\right) - \frac{\|\lambda\|^2_H}{2\rho}. \tag{3.7}$$

Before discussing some other observations and consequences of the slack variable approach, we first give some general properties of the augmented Lagrangian.

Proposition 3.3 *Let $\mathcal{L}_\rho : X \times H \to \mathbb{R}$ be the augmented Lagrangian (3.7). Then,*

(a) $\mathcal{L}_\rho$ is concave and continuously differentiable with respect to λ.
(b) If f is convex and G is $\mathcal{K}_\infty$-concave, then $\mathcal{L}_\rho$ is convex with respect to x.
(c) If f and G are continuously differentiable, then $\mathcal{L}_\rho$ is so with respect to x.
(d) If $x \in X$ is a feasible point, then $\mathcal{L}_\rho(x, \lambda) \le f(x)$ for all $x \in X$ and $\lambda \in H$.

Proof

(a) The concavity follows from the fact that $\mathcal{L}_\rho(x, \cdot)$ is an infimum of affine functions by (3.5), and the continuous differentiability follows from that of $d^2_{\mathcal{K}}$.
(b) This is a consequence of Theorem 2.10.
(c) This follows again from the continuous differentiability of $d^2_{\mathcal{K}}$.

(d) If $G(x) \in \mathcal{K}$, then $d_{\mathcal{K}}(G(x) + \lambda/\rho) \leq \|\lambda\|_H/\rho$ by the nonexpansiveness of the distance function. Hence, $\mathcal{L}_\rho(x, \lambda) \leq f(x) + (\rho/2)\|\lambda\|_H^2/\rho^2 - \|\lambda\|_H^2/(2\rho) = f(x)$.

□

Let us close this section by mentioning some byproducts of the slack variable approach. For fixed λ and ρ, the minimizing value of s in (3.6) is given by $\bar{s}(x) := P_{\mathcal{K}}(G(x) + \lambda/\rho)$. It follows that

$$h(x, \bar{s}(x)) = G(x) - P_{\mathcal{K}}\left(G(x) + \frac{\lambda}{\rho}\right). \tag{3.8}$$

Recall that, in the original method of multipliers (Algorithm 3.1), the norm of the equality constraint was used to determine whether the penalty parameter ρ_k should be increased after a given iteration. The above calculations suggest that (3.8) should be used to control ρ_k in the general case.

Another byproduct of the slack variable technique is a natural candidate for the Lagrange multiplier update. Assume that $\lambda^k \in H$ is a given estimate of the Lagrange multiplier of (P_H), that $\rho_k > 0$, and x^{k+1} is the next primal iterate (typically, some kind of minimizer of $\mathcal{L}_{\rho_k}(\cdot, \lambda^k)$). Taking into account the update rule in Algorithm 3.1, the next dual iterate is given by

$$\lambda^{k+1} = \lambda^k + \rho_k h(x^{k+1}, \bar{s}(x^{k+1})) = \rho_k \left[G(x^{k+1}) + \frac{\lambda^k}{\rho_k} - P_{\mathcal{K}}\left(G(x^{k+1}) + \frac{\lambda^k}{\rho_k}\right)\right].$$

This formula will play a fundamental role in the subsequent algorithms. Note that the above updating scheme can also be motivated (in the differentiable case) by looking at the stationarity condition of $\mathcal{L}_{\rho_k}(\cdot, \lambda^k)$, evaluated in x^{k+1}.

3.3 The Algorithm

This section presents the main algorithmic framework for the remainder of this chapter. It is based on the method of multipliers from Sect. 3.1 and the slack variable transformation from Sect. 3.2, but it differs from the original multiplier method in one key aspect: the use of a safeguarded multiplier sequence. This will be the main tool to obtain much sharper (global) convergence assertions than those that are possible for the traditional algorithm.

Recall that we are dealing with a problem of the form (P), that we are working in the Gel'fand triple framework (1.2), and that $\mathcal{K} \subseteq H$ is a nonempty closed convex set with $i^{-1}(\mathcal{K}) = K$. The algorithm now proceeds by augmenting the constraint $G(x) \in \mathcal{K}$ in the space H. This means that, in a sense, we are not really attempting to solve (P) but the transformed problem (P_H). Nevertheless, we will see that many convergence properties of the augmented Lagrangian method can be stated

accurately in terms of (P) (using, for instance, constraint qualifications for that problem).

For the precise specification of the method below, we will need a means of controlling the penalty parameter ρ. Motivated by (3.8), it is natural to use the function

$$V(x, \lambda, \rho) = \left\| G(x) - P_{\mathcal{K}}\left(G(x) + \frac{\lambda}{\rho}\right)\right\|_H, \tag{3.9}$$

which can be seen as a composite measure of feasibility and complementarity at the current iterates. Using this function, the augmented Lagrangian method can be given as follows.

Algorithm 3.4 ALM for constrained optimization

Let $(x^0, \lambda^0) \in X \times H$, $\rho_0 > 0$, let $B \subseteq H$ be a nonempty bounded set, $\gamma > 1$, $\tau \in (0, 1)$, and set $k := 0$.

Step 1. If (x^k, λ^k) satisfies a suitable termination criterion, STOP.

Step 2. Choose $w^k \in B$, and compute an approximate solution x^{k+1} of the problem

$$\underset{x \in C}{\text{minimize}}\ \mathcal{L}_{\rho_k}(x, w^k). \tag{3.10}$$

Step 3. Update the vector of multipliers to

$$\lambda^{k+1} := \rho_k \left[G(x^{k+1}) + \frac{w^k}{\rho_k} - P_{\mathcal{K}}\left(G(x^{k+1}) + \frac{w^k}{\rho_k}\right)\right]. \tag{3.11}$$

Step 4. Let $V_{k+1} := V(x^{k+1}, w^k, \rho_k)$, and set

$$\rho_{k+1} := \begin{cases} \rho_k, & \text{if } k = 0 \text{ or } V_{k+1} \leq \tau V_k, \\ \gamma \rho_k, & \text{otherwise.} \end{cases} \tag{3.12}$$

Step 5. Set $k \leftarrow k + 1$, and go to Step 1.

Some remarks are in order. First among them is the fact that we have not specified what constitutes an "approximate solution" in Step 2. There are multiple options in this regard. For instance, we could require that x^{k+1} is an (approximate) global minimizer of $\mathcal{L}_{\rho_k}(\cdot, w^k)$. This is probably the simplest assumption from a theoretical point of view, but it is effectively restricted to problems where some form of convexity is present. On the other hand, we could also require that x^{k+1} is some kind of approximate stationary point of (3.10). This is more realistic in the nonconvex case, but it is also more intricate to deal with in theoretical terms. We will analyze both these approaches individually in the subsequent sections.

In practical terms, the augmented subproblems are typically solved by applying an appropriate generalized Newton method. The necessity for such methods stems

from the fact that the augmented Lagrangian is once but in general not twice continuously differentiable with respect to x.

The second remark pertains to the sequence $\{w^k\}$, which will occasionally be referred to as the *safeguarded (Lagrange) multiplier sequence*. The presence of w^k can be seen as the distinctive feature of the algorithm, and it separates the method from traditional augmented Lagrangian schemes. Indeed, in Algorithm 3.4, we use w^k in certain places where conventional algorithms simply use λ^k. The main motivation is that w^k is always a bounded sequence (it is specifically required to be so), and this is the main ingredient to obtain sharper global convergence results. As a consequence, the above algorithm has strictly stronger convergence properties than its traditional counterpart. An actual example demonstrating this fact is somewhat involved and given in [47]; see also the discussion at the end of Sect. 4. Note that, despite the boundedness of $\{w^k\}$, the sequence $\{\lambda^k\}$ in Algorithm 3.4 can still be unbounded. The actual choice of w^k allows for a certain degree of freedom. For instance, we could always choose $w^k := 0$, thus obtaining an algorithm that is essentially a quadratic penalty method. In practice, it is usually advantageous to keep w^k as close as possible to λ^k, for instance, by choosing the set B as a simple but large bounded set, and taking

$$w^k := P_B(\lambda^k)$$

for all k. This choice has the advantage that, if the sequence $\{\lambda^k\}$ is indeed bounded and the set B is large enough, then we can expect to have $w^k = \lambda^k$ for all k. On the other hand, if $\{\lambda^k\}$ is unbounded, then the safeguarding scheme will prevent w^k from escaping to infinity.

Finally, let us remark that the penalty updating scheme in (3.12) makes a distinction between the cases $k = 0$ and $k \geq 1$. This is because the value V_0 is formally undefined since we do not have w^{-1} and ρ_{-1}. In practice, it is often beneficial to treat this initial step differently, for instance, by simply setting $w^{-1} := w^0$, $\rho_{-1} := \rho_0$ and performing the penalty update in the same way as for $k \geq 1$. In any case, the treatment of this initial step has no impact on the convergence theory. The nature of the multiplier update allows us to state two assertions that hold completely independently of x^{k+1}, cf. [50].

Lemma 3.5 *We have $\lambda^k \in \mathcal{K}^\circ_\infty$ for all k. Moreover, there is a null sequence $\{r_k\} \subseteq \mathbb{R}_+$ such that $\big(\lambda^k, y - G(x^k)\big) \leq r_k$ for all $y \in \mathcal{K}$ and $k \in \mathbb{N}$.*

Remark 3.6 (Cone Constraints) If the set $\mathcal{K}$ is a closed convex *cone*, then the multiplier update (3.11) in Algorithm simplifies to $\lambda^{k+1} = P_{\mathcal{K}^\circ}(w^k + \rho_k G(x^{k+1}))$. This follows immediately from the Moreau decomposition, cf. Lemma 2.3.

Remark 3.7 (Dual Interpretation) The dual update of the classical augmented Lagrangian is known to be equivalent to the proximal-point iteration applied to the dual optimization problem. A similar interpretation is possible for the safeguarded augmented Lagrangian where the dual update can be seen as a shifted Tikhonov regularization method; see [48] for more details.

Remark 3.8 (Nonlinear Programs) Consider the nonlinear program

$$\min f(x) \quad \text{s.t.} \quad h(x) = 0,\ g(x) \leq 0$$

with continuously differentiable functions $f, g_i, h_j : \mathbb{R}^n \to \mathbb{R}$ for all components $i = 1, \ldots, m$ and $j = 1, \ldots, p$. This nonlinear program can be viewed as a special case of our general framework (P) by taking, e.g.,

$$X = \mathbb{R}^n, C = \mathbb{R}^n, G := \begin{pmatrix} h \\ g \end{pmatrix}, K := \{0\}^p \times (-\infty, 0]^m.$$

In this case, writing $\lambda =: (\mu, \eta)$ for the multipliers of the equality and inequality constraints, the squared distance function is given by

$$d_{\mathcal{K}}^2\left(G(x) + \frac{\eta}{\rho}\right) = \sum_{j=1}^{p} \left(h_j(x) + \frac{\mu_j}{\rho}\right)^2 + \sum_{i=1}^{m} \max^2\left\{0, g_i(x) + \frac{\eta_i}{\rho}\right\}.$$

Plugging this into the definition of the augmented Lagrangian, an elementary calculation shows that this function simplifies to

$$\mathcal{L}_\rho(x, \mu, \eta) = f(x) + \frac{\rho}{2}\|h(x)\|_2^2 + \mu^T h(x) + \frac{1}{2\rho} \sum_{i=1}^{m} \left[\max^2\left\{0, \eta_i + \rho g_i(x)\right\} - \eta_i^2\right],$$

which is the usual augmented Lagrangian for nonlinear programs with equality and inequality constraints.

Remark 3.9 (Simplified Augmented Lagrangian) In each iteration, Algorithm 3.4 minimizes the augmented Lagrangian with respect to x, for fixed w^k. Since this minimization procedure does not depend on the last term of our augmented Lagrangian, we would obtain the same sequence using the simplified Lagrangian

$$f(x) + \frac{\rho}{2} d_{\mathcal{K}}^2\left(G(x) + \frac{\lambda}{\rho}\right).$$

In fact, this is precisely the augmented Lagrangian used in [52]. On the other hand, this simplification changes the dual point of view completely and also gives a different function for finite-dimensional nonlinear programs, cf. Remarks 3.7 and 3.8.

Remark 3.10 (Moreau–Yosida Regularization) Algorithm 3.4 allows to take $\{w^k\}$ as the null sequence. This choice corresponds to the classical quadratic penalty approach and is better known under the name *Moreau–Yosida regularization* in the current context, cf. [34, 35]. The multiplier update in the Moreau–Yosida regularization usually allows a shift. In any case, the subsequent convergence theory also covers this (shifted) Moreau–Yosida regularization.

4 Global Convergence Theory

In this section, we present the global convergence characteristics of Algorithm 3.4. To this end, we first establish a result regarding the existence of solutions of the penalized subproblems in Sect. 4.1. The next two sections consider the convergence to global minimizers and stationary points, respectively, depending on the degree by which we solve the penalized subproblems. The results are taken from the recent paper [15] and can be viewed as improvements from those presented in [52], where suitable feasibility and stationarity results were shown for strong limit points.

4.1 Existence of Penalized Solutions

In most situations, the augmented Lagrangian $\mathcal{L}_\rho(\cdot, w)$ is bounded from below on C. This is satisfied, in particular, if f itself is already bounded from below on C, or if, roughly speaking, the penalty parameter is sufficiently large to make $\mathcal{L}_\rho$ coercive on the infeasible set. In any case, if $\mathcal{L}_\rho(\cdot, w)$ is bounded from below on C, then the augmented subproblems necessarily admit approximate minimizers. In the following, $\hat{x} \in C$ is called an ε-minimizer of a function $L : X \to \mathbb{R}$ on C if $L(\hat{x}) \leq L(x) + \varepsilon$ for all $x \in C$.

Proposition 4.1 *Let $w \in H$, $\rho > 0$, and assume that the augmented Lagrangian $\mathcal{L}_\rho(\cdot, w)$ is bounded from below on C. Then, the following assertions hold:*

(a) For any $\varepsilon > 0$, there is an ε-minimizer $x_\varepsilon \in C$ of $\mathcal{L}_\rho(\cdot, w)$ on C.
(b) If the functions f and G are continuously differentiable, then we can choose x_ε so that it additionally satisfies $\operatorname{dist}(-\mathcal{L}'_\rho(x_\varepsilon, w), \mathcal{N}_C(x_\varepsilon)) \leq \varepsilon^{1/2}$.

Proof The first assertion follows from the lower boundedness assumption. The second property is a consequence of Ekeland's variational principle. □

We now discuss the existence of exact minimizers. The main proof technique is the direct method of the calculus of variations. For this, we need an appropriate kind of lower semicontinuity of the augmented Lagrangian. The following lemma provides two sufficient conditions for this property.

Lemma 4.2 *Assume that f is weakly sequentially lsc and G is either*

(i) continuous and $\mathcal{K}_\infty$-concave or
(ii) weakly sequentially continuous.

Then, for each $\rho > 0$ and $w \in H$, the augmented Lagrangian $\mathcal{L}_\rho(\cdot, w)$ is weakly sequentially lsc on X.

Proof Let $w \in H$ and $\rho > 0$. It suffices to verify the weak sequential lower semicontinuity of the function $h(x) := d^2_{\mathcal{K}}(G(x) + w/\rho)$. Observe that $d_{\mathcal{K}}$ is weakly

sequentially lsc by Proposition 2.8. Hence, under (ii), we immediately obtain the same for h.

Consider now (i). In that case, the function h is convex (by Theorem 2.10) and continuous, thus again weakly sequentially lsc by Proposition 2.8. □

The weak sequential lower semicontinuity of the augmented Lagrangian yields the existence of penalized solutions if we assume either the weak compactness of the set C or an appropriate growth condition. We say that a function $J : X \to \mathbb{R}$ is *coercive* if $J(x^k) \to +\infty$ whenever $\{x^k\} \subseteq X$ and $\|x^k\|_X \to +\infty$.

Corollary 4.3 *Let $w \in H$, $\rho > 0$, and let one of the conditions in Lemma 4.2 be satisfied. If either*

(i) C is weakly compact or
(ii) X is reflexive and $\mathcal{L}_\rho(\cdot, w)$ is coercive,

then the problem $\min_{x \in C} \mathcal{L}_\rho(x, w)$ *admits a global minimizer.*

Clearly, a sufficient condition for the coercivity of the augmented Lagrangian is that of the objective function f. Even if this property does not hold, then it is common for $\mathcal{L}_\rho(\cdot, w)$ to be coercive if, roughly speaking, the objective function is coercive on the feasible set Φ and not too badly behaved outside of it. In that case, the penalty term in (3.7) yields the coercivity of $\mathcal{L}_\rho(\cdot, w)$ on the complement of Φ.

4.2 Convergence to Global Minimizers

In this section, we analyze the convergence properties of Algorithm 3.4 under the assumption that we can solve the subproblems in an (essentially) global sense. This is of course a rather restrictive requirement and can, in general, only be expected under certain convexity assumptions. However, the resulting theory is still appealing due to its simplicity. Indeed, the results below merely require some rather mild form of continuity (no differentiability) and can easily be extended to the case where the function f is extended valued, i.e., it is allowed to take on the value $+\infty$.

Assumption 4.4 (Global Minimization) We assume that f and $d_{\mathcal{K}} \circ G$ are weakly sequentially lsc on C and that $x^k \in C$ for all k. Moreover, for every $x \in C$, there is a null sequence $\{\varepsilon_k\} \subseteq \mathbb{R}$ such that $\mathcal{L}_{\rho_k}(x^{k+1}, w^k) \leq \mathcal{L}_{\rho_k}(x, w^k) + \varepsilon_{k+1}$ for all k.

Recall that, for convex functions, weak sequential lower semicontinuity is implied by ordinary continuity. Thus, if f is a continuous convex function, then f is weakly sequentially lsc.

A similar comment applies to the weak sequential lower semicontinuity of the function $d_{\mathcal{K}} \circ G$. Indeed, there are two rather general situations in which this condition is satisfied: if G is weakly sequentially continuous, then $d_{\mathcal{K}} \circ G$ is weakly sequentially lsc since $d_{\mathcal{K}}$ is so by Proposition 2.8. On the other hand, if G is continuous and $\mathcal{K}_\infty$-concave in the sense of Definition 2.9, then $d_{\mathcal{K}} \circ G$ is a

continuous convex function (by Theorem 2.10) and thus again weakly sequentially lsc. Let us also remark that, if G is continuous and affine, then both the above cases apply.

Finally, another salient feature of Assumption 4.4 is the dependence of the sequence $\{\varepsilon_k\}$ on the comparison point $x \in C$. The motivation behind this is that, if (P) is a smooth convex problem and the point x^{k+1} is "nearly stationary" in the sense that $\operatorname{dist}(-\mathcal{L}'_{\rho_k}(x^{k+1}, w^k), \mathcal{N}_C(x^{k+1})) \leq \delta$ for some (small) $\delta > 0$, then, by convexity, we obtain an estimate of the form

$$\begin{aligned}\mathcal{L}_{\rho_k}(x, w^k) &\geq \mathcal{L}_{\rho_k}(x^{k+1}, w^k) + \mathcal{L}'_{\rho_k}(x^{k+1}, w^k)(x - x^{k+1}) \\ &\geq \mathcal{L}_{\rho_k}(x^{k+1}, w^k) - \delta \|x^{k+1} - x\|_X.\end{aligned}$$

This suggests that we should allow the sequence $\{\varepsilon_k\}$ in Assumption 4.4 to depend on the point x. In any case, the stated assumption is satisfied automatically if x^{k+1} is a global ε_{k+1}-minimizer of $\mathcal{L}_{\rho_k}(\cdot, w^k)$ for some null sequence $\{\varepsilon_k\}$.

We now turn to the convergence analysis of Algorithm 3.4 under Assumption 4.4. The theory is divided into separate analyses of feasibility and optimality. Since the augmented Lagrangian method is, at its heart, a penalty-type algorithm, the attainment of feasibility is particularly important for the success of the algorithm. A closer look at the definition of the augmented Lagrangian suggests that, if ρ is large, then the minimization of $\mathcal{L}_\rho$ essentially reduces to that of the infeasibility measure $d^2_{\mathcal{K}}(G(x))$. Hence, we can expect (weak) limit points of the sequence $\{x^k\}$ to be minimizers of this auxiliary function, which means that, roughly speaking, these points are "as feasible as possible." A precise statement of this assertion can be found in the following lemma.

Lemma 4.5 *Let $\{x^k\}$ be generated by Algorithm 3.4, let Assumption 4.4 hold, and let $\bar{x}$ be a weak limit point of $\{x^k\}$. Then, $\bar{x}$ is a global minimizer of the function $d_{\mathcal{K}} \circ G$ on C. In particular, if the feasible set of (P) is nonempty, then $\bar{x}$ is feasible.*

Let us now turn to the optimality part.

Theorem 4.6 *Let $\{x^k\}$ be generated by Algorithm 3.4, let Assumption 4.4 hold, and assume that the feasible set of (P) is nonempty. Then, $\limsup_{k\to\infty} f(x^{k+1}) \leq f(x)$ for every $x \in \Phi$. Moreover, every weak limit point of $\{x^k\}$ is a global solution of (P).*

If the problem is convex with strongly convex objective, then it is possible to considerably strengthen the results of the previous theorem. Recall that, in this case, the weak sequential lower semicontinuity of f from Assumption 4.4 is implied by (ordinary) continuity. Recall also that a sufficient condition for the convexity of the feasible set Φ is the $\mathcal{K}_\infty$-concavity of G. Moreover, if G is $\mathcal{K}_\infty$-concave, then the distance function $d_{\mathcal{K}} \circ G$ is convex, and thus the weak sequential lower semicontinuity from Assumption 4.4 is implied by (ordinary) continuity of G.

Corollary 4.7 *Let $\{x^k\}$ be generated by Algorithm 3.4, and let Assumption 4.4 hold. Assume that X is reflexive, f is strongly convex on C, and the feasible set of (P) is nonempty and convex. Then, $\{x^k\}$ converges strongly to the unique solution of (P).*

4.3 Stationarity of Limit Points

The theory on global minimization in the preceding section is certainly appealing from a theoretical point of view. However, the practical relevance of the corresponding results is essentially limited to problems where some form of convexity is present. It therefore seems natural to conduct a dedicated analysis for the augmented Lagrangian method, which, instead of global minimization, takes into account suitable stationary concepts.

The present section is dedicated to precisely the approach described above. To that end, we assume that the functions defining the optimization problem are continuously differentiable and that we are able to compute local minimizers or stationary points of the subproblems (3.10), which occur in the algorithm. Recall that the first-order optimality conditions of these problems are given by

$$-\mathcal{L}'_{\rho_k}(x, w^k) \in \mathcal{N}_C(x).$$

Similarly to the previous section, we will allow for certain inexactness terms. A natural way of doing this is by considering the inexact first-order optimality condition

$$\varepsilon^{k+1} - \mathcal{L}'_{\rho_k}(x, w^k) \in \mathcal{N}_C(x),$$

where $\varepsilon^{k+1} \in X^*$ is an error term. For $k \to \infty$, the degree of inexactness should vanish in the sense that $\varepsilon^k \to 0$. Hence, we arrive at the following assumption.

Assumption 4.8 (Convergence to KKT Points) We assume that

(i) f and G are continuously differentiable on X,
(ii) the derivative f' is bounded and pseudomonotone,
(iii) G and G' are completely continuous on C, and
(iv) $x^{k+1} \in C$ and $\varepsilon^{k+1} - \mathcal{L}'_{\rho_k}(x^{k+1}, w^k) \in \mathcal{N}_C(x^{k+1})$ for all k, where $\varepsilon^k \to 0$.

Recall that $\mathcal{L}_{\rho_k}$ is continuously differentiable by Proposition 3.3. The derivative $\mathcal{L}'_{\rho_k}$ (with respect to x) is given by

$$\mathcal{L}'_{\rho_k}(x, w^k) = f'(x) + \rho_k G'(x)^* \left[G(x) + \frac{w^k}{\rho_k} - P_{\mathcal{K}}\left(G(x) + \frac{w^k}{\rho_k} \right) \right]. \tag{4.1}$$

In particular, it holds that $\mathcal{L}'_{\rho_k}(x^{k+1}, w^k) = \mathcal{L}'(x^{k+1}, \lambda^{k+1})$.

As in the previous section, we treat the questions of feasibility and optimality in a separate manner. For the feasibility part, we relate the augmented Lagrangian to the infeasibility measure $d^2_{\mathcal{K}} \circ G$.

Lemma 4.9 *Let $\{x^k\}$ be generated by Algorithm 3.4 under Assumption 4.8, and let $\bar{x}$ be a weak limit point of $\{x^k\}$. Then, $\bar{x}$ is a stationary point of the problem $\min_{x \in C} d^2_{\mathcal{K}}(G(x))$.*

The above lemma indicates that weak limit points of the sequence $\{x^k\}$ have a strong tendency to be feasible points. Apart from the heuristic appeal of the result, there are several nontrivial cases where Lemma 4.9 automatically implies the feasibility of the limit point $\bar{x}$. Here, two cases in particular deserve a special mention: first, let us assume that the mapping G is $\mathcal{K}_\infty$-concave in the sense of Definition 2.9 (for instance, G could be affine). In this case, the function $d^2_{\mathcal{K}} \circ G$ is convex by Theorem 2.10, and it follows that $\bar{x}$ is a global minimizer of this function. Hence, if the feasible set Φ is nonempty, then $\bar{x} \in \Phi$. The second interesting case arises if the point $\bar{x}$ satisfies the extended Robinson constraint qualification from Definition 2.16. In this case, the feasibility of $\bar{x}$ follows from Proposition 2.17.

We now analyze the optimality properties of limit points. The main result in this direction is the following.

Theorem 4.10 *Let $\{(x^k, \lambda^k)\}$ be generated by Algorithm 3.4 under Assumption 4.8, let $x^{k+1} \rightharpoonup_I \bar{x}$ for some index set $I \subseteq \mathbb{N}$, and let $\bar{x}$ satisfy ERCQ with respect to the constraint system of (P). Then, $\bar{x}$ is a stationary point of (P), the sequence $\{\lambda^{k+1}\}_{k \in I}$ is bounded in Y^*, and each of its weak-* limit points belongs to $\Lambda(\bar{x})$.*

Observe that the sequence $\{\lambda^k\}$ is only bounded in Y^* and not necessarily in H. If the extended RCQ holds with respect to the transformed constraint $G(x) \in \mathcal{K}$ (instead of the original condition $G(x) \in K$), then the result remains true with Y^* replaced by H. However, this assumption is too restrictive for many applications, in particular, those where (P) is regular (in the constraint qualification sense) with respect to the original space Y, but not with respect to the larger space H.

Remark 4.11 If we know from the specific problem structure or from some other convergence result (e.g., Corollary 4.7) that the sequence $\{x^k\}$ or one of its subsequences is strongly convergent, then we can dispense with the pseudomonotonicity and complete continuity assumptions. In this case, the assertions of Lemma 4.9 and Theorem 4.10 remain true under Assumption 4.8 (i) and (iv) only.

We now return to the general case and provide two additional results that can be useful to obtain convergence in certain special cases. First, let us consider the case of convex constraints. The resulting theorem requires neither the complete continuity of G or G' nor any constraint qualification.

Proposition 4.12 *Let $\{x^k\}$ be generated by Algorithm 3.4, let Assumption 4.8 (i), (ii), and (iv) hold, let G be $\mathcal{K}_\infty$-concave on C, and assume that Φ is nonempty. Then, every weak limit point $\bar{x}$ of $\{x^k\}$ satisfies $\bar{x} \in \Phi$ and $f'(\bar{x})d \geq 0$ for all $d \in \mathcal{T}_\Phi(\bar{x})$.*

Proof Let $x^{k+1} \rightharpoonup_I \bar{x}$ for some subset $I \subseteq \mathbb{N}$. The feasibility of $\bar{x}$ follows from Lemma 4.9 and the discussion below. For the optimality, let $y \in \Phi$ be any feasible point. Then, $\langle \mathcal{L}'_{\rho_k}(x^{k+1}, w^k), y - x^{k+1} \rangle \geq \langle \varepsilon^{k+1}, y - x^{k+1} \rangle$ by Assumption 4.8 and using $\mathcal{L}'_{\rho_k}(x^{k+1}, w^{k+1}) = \mathcal{L}'(x^{k+1}, \lambda^{k+1})$, we obtain

$$\begin{aligned}\left\langle \varepsilon^{k+1}, y - x^{k+1} \right\rangle &\leq \left\langle f'(x^{k+1}) + G'(x^{k+1})^* \lambda^{k+1}, y - x^{k+1} \right\rangle \\ &= \left\langle f'(x^{k+1}), y - x^{k+1} \right\rangle + \left(\lambda^{k+1}, G'(x^{k+1})(y - x^{k+1}) \right) \\ &\leq \left\langle f'(x^{k+1}), y - x^{k+1} \right\rangle + \left(\lambda^{k+1}, G(y) - G(x^{k+1}) \right),\end{aligned}$$

where we used the fact that $x \mapsto (\lambda^{k+1}, G(x))$ is convex by Theorem 2.10 and Lemma 3.5. Using again Lemma 3.5, we now obtain $\langle f'(x^{k+1}), y - x^{k+1} \rangle \geq \langle \varepsilon^{k+1}, y - x^{k+1} \rangle + r_{k+1}$ with a null sequence $\{r_k\} \subseteq \mathbb{R}$. Since $\bar{x} \in \Phi$, we obtain in particular that $\liminf_{k \to \infty} \langle f'(x^k), \bar{x} - x^k \rangle \geq 0$. The pseudomonotonicity of f' therefore implies that

$$\langle f'(\bar{x}), y - \bar{x} \rangle \geq \limsup_{k \to \infty} \langle f'(x^k), y - x^k \rangle \geq 0 \quad \forall y \in \Phi,$$

and the proof is complete. □

Another special case arises if $C = X$ and the operator $G'(\bar{x})$ is surjective, where $\bar{x}$ is again a weak limit point of the sequence $\{x^k\}$. If we already know (e.g., by Proposition 4.12) that $\bar{x}$ is a stationary point of (P), then it is possible to prove the weak-* convergence of a subsequence of $\{\lambda^k\}$ under weaker assumptions than those in Theorem 4.10. Indeed, it is possible to obtain a convergence result for asymptotic KKT sequences under only the convergence $G'(x^k) \to G'(x)$, with no convergence of the values $G(x^k)$. We will see later that this is crucial for obtaining convergence for Bratu's obstacle problem, see Sect. 6.2, where $G \colon H_0^1(\Omega) \to H_0^1(\Omega)$, $G(x) := x - \psi$, with $\psi \in H_0^1(\Omega)$. In particular, G is obviously not completely continuous, but for $x^k \rightharpoonup \bar{x}$, it holds $G'(x^k) \to G'(\bar{x})$. We need the following auxiliary results. The first theorem is a slightly more general version of the Banach open mapping theorem.

Theorem 4.13 (Uniform Open Mapping Theorem) *Let X and Y be real Banach spaces and $A \in L(X, Y)$ a surjective linear operator. Then, there exists $r > 0$ such that $B_r^Y \subseteq A(B_1^X)$ and, whenever $T \in L(X, Y)$ and $\delta := \|T - A\|_{L(X,Y)} < r$, then $B_{r-\delta}^Y \subseteq T(B_1^X)$.*

Proof The first assertion is the Banach open mapping theorem. For the proof of the second assertion, we refer the reader to [25, Thm. 1.2] or [26, Thm. 5D.2]. □

The second theorem states a convergence result for asymptotic KKT sequences under only the convergence $G'(x^k) \to G'(x)$, with no convergence of the values $G(x^k)$. We state this result in a slightly more general framework.

Proposition 4.14 *Let $\{x^k\} \subseteq X$, $\{T_k\} \subseteq L(X,Y)$, and $\{\lambda^k\} \subseteq Y^*$ be sequences such that $F(x^k)+T_k^*\lambda^k \overset{*}{\rightharpoonup} 0$. Assume that $x^k \rightharpoonup \bar{x}$ for some $\bar{x} \in X$, $F(x^k) \overset{*}{\rightharpoonup} F(\bar{x})$, $T_k \to T$ for some $T \in L(X,Y)$, and that T is surjective. Then, $\{\lambda^k\}$ converges weak-* in Y^* to the unique solution of $F(\bar{x}) + T^*\lambda = 0$.*

Proof We first show that $\{\lambda^k\}$ is weak-* convergent. Let $\hat{y} \in Y$ be an arbitrary point. It suffices to show that $\langle \lambda^k, \hat{y}\rangle$ is convergent. Let $r > 0$ be as in the uniform version of the Banach open mapping theorem (Theorem 4.13), so that $B_r^Y \subseteq T(B_1^X)$. Assume, without loss of generality, that $\hat{y} \in B_r^Y$, and let $\hat{w} \in B_1^X$ be a point such that $T\hat{w} = \hat{y}$. Set $\delta_k := \|T_k - T\|_{L(X,Y)}$, and let k be sufficiently large so that $\delta_k < r$. Then, $\|\hat{y} - T_k\hat{w}\|_Y \le \delta_k$, and, by Theorem 4.13, there are points $d^k \in X$ such that $T_k d^k = \hat{y} - T_k\hat{w}$ and

$$\|d^k\|_X \le \frac{\|\hat{y} - T_k\hat{w}\|_Y}{r - \delta_k} \le \frac{\delta_k}{r - \delta_k}.$$

Define $w^k := \hat{w} + d^k$. Then, $w^k \to \hat{w}$ and $T_k w^k = \hat{y}$ by definition. Hence,

$$0 \leftarrow \left\langle F(x^k) + T_k^*\lambda^k, w^k \right\rangle = \langle F(\bar{x}), \hat{w}\rangle + o(1) + \left\langle \lambda^k, \hat{y}\right\rangle.$$

Thus, we obtain $\langle \lambda^k, \hat{y}\rangle \to -\langle F(\bar{x}), \hat{w}\rangle$. Since $\hat{y} \in Y$ was arbitrary, this implies that $\{\lambda^k\}$ is weak-* convergent in Y^*.

Let $\bar{\lambda}$ denote the weak-* limit of $\{\lambda^k\}$. Using $F(x^k) + T_k^*\lambda^k \overset{*}{\rightharpoonup} 0$, it follows that $F(\bar{x}) + T^*\bar{\lambda} = 0$, and $\bar{\lambda}$ is unique since T^* is injective. □

Proposition 4.15 *Let $\{x^k\}$ be generated by Algorithm 3.4, and let $x^{k+1} \rightharpoonup_I \bar{x}$ for some $I \subseteq \mathbb{N}$ and $\bar{x} \in X$. Assume that $\bar{x}$ is a stationary point of (P), that $C = X$, f' is weak-* sequentially continuous, G' is completely continuous, and that $G'(\bar{x})$ is surjective. Then, $\{\lambda^{k+1}\}_{k\in I}$ converges weak-* to the unique element in $\Lambda(\bar{x})$.*

Proof Recall that $\mathcal{L}'_{\rho_k}(x^{k+1}, w^k) = \mathcal{L}'(x^{k+1}, \lambda^{k+1})$. Combining Assumption 4.4 and Lemma 3.5, we obtain the asymptotic conditions (for $k \ge 1$).

$$\varepsilon^k - \mathcal{L}'(x^k, \lambda^k) \in \mathcal{N}_C(x^k) \quad \text{and} \quad \langle \lambda^k, y - G(x^k)\rangle \le r_k \quad \forall y \in K.$$

Hence, the result follows from Proposition 4.14. □

In the context of optimality properties, it is worthwhile to briefly discuss the case of bounded penalty parameters. This is particularly interesting because any assertion made under this assumption is a *necessary* condition for the boundedness of $\{\rho_k\}$. It turns out that no constraint qualifications are needed in the bounded case, and the algorithm produces a Lagrange multiplier in H.

Corollary 4.16 *Let $\{(x^k, \lambda^k)\}$ be generated by Algorithm 3.4, let Assumption 4.8 hold, and let $\bar{x}$ be a weak limit point of $\{x^k\}$. If $\{\rho_k\}$ remains bounded, then $\{\lambda^k\}$*

has a bounded subsequence in H, and $\bar{x}$ satisfies the KKT conditions of (P) with a multiplier in H.

The above result implies that $\{\rho_k\}$ can only remain bounded if (P) admits a multiplier in H.

We close this section by noting that the nice global convergence properties of the safeguarded augmented Lagrangian method do not hold for the classical augmented Lagrangian approach, which is the main reason for the modification of the updating rule of the multipliers. In fact, a counterexample in [47] shows that the classical method may generate limit points that have no meaning from the point of view of satisfying a suitable stationarity measure, whereas the safeguarded method has the desired behavior. The counterexample provided in [47] is one dimensional and convex in the sense that its objective function is convex (even linear) and the feasible set is also convex, though represented by a nonconvex function. The authors are not aware of a "fully" convex counterexample where the objective function and the inequality constraints are all convex, and the equality constraints are linear. This leads to the following open problem.

Open Problem 4.17 Are the global convergence properties of the classical augmented Lagrangian method identical (or very similar) to the safeguarded Lagrangian method for fully convex problems?

5 Local Convergence

Here, we discuss the local convergence properties of Algorithm 3.4. We first discuss in Sect. 5.1 the existence of local minima and the (strong!) convergence of such minima. These properties are based on a second-order sufficiency condition, whereas constraint qualifications are not required. This is interesting since it allows applications of our results to problems with a complicated structure of the feasible set. Additional conditions are necessary, however, in order to verify the rate-of-convergence results, see Sect. 5.2. The results from this section are taken from the recent papers [15, 49].

5.1 Existence of Local Minima und Strong Convergence

Before we formulate the second-order sufficiency condition, we note that, as with constraint qualifications and KKT conditions, second-order conditions for (P) can be formulated either with respect to Y or with respect to H. In this section, to avoid unnecessary notational overhead, we will simply formulate the second-order condition and its consequences with respect to Y. The results below all remain true when Y is replaced by H (note that the choice $Y := H$ is even admissible in our framework).

Let $(\bar{x}, \bar{\lambda}) \in X \times Y^*$ be a KKT point of (P). Throughout this section, we assume that f and G are twice continuously differentiable in a neighborhood of $\bar{x}$. Then, consider, for $\eta > 0$, the *extended critical cone*

$$\mathcal{C}_\eta(\bar{x}) := \left\{ d \in \mathcal{T}_C(\bar{x}) : \begin{array}{l} f'(\bar{x})d \le \eta \|d\|_X, \\ \operatorname{dist}(G'(\bar{x})d, \mathcal{T}_K(G(\bar{x}))) \le \eta \|d\|_X \end{array} \right\}. \tag{5.1}$$

Note that $\mathcal{C}_\eta$ depends on $\bar{x}$ only. The following is the general form of a second-order sufficient condition, which we will use throughout this section.

Definition 5.1 (Second-Order Sufficient Condition) We say that the *second-order sufficient condition (SOSC)* holds in a KKT point $(\bar{x}, \bar{\lambda}) \in X \times Y^*$ of (P) if there are $\eta, c > 0$ such that

$$\mathcal{L}''(\bar{x}, \bar{\lambda})(d, d) \ge c\|d\|_X^2 \quad \text{for all } d \in \mathcal{C}_\eta(\bar{x}).$$

As mentioned before, the extended critical cone and SOSC can also be formulated with respect to $\mathcal{K}$ and H for KKT pairs $(\bar{x}, \bar{y}) \in X \times H$.

The above should be considered the "basic" second-order condition, which can be stated without any assumptions on the specific structure of (P). For many problem classes, it is possible to state more refined second-order conditions that are either equivalent to Definition 5.1 or turn out to have similar implications. Some information in this direction can be found, for instance, in [14, Section 3.3].

It turns out that SOSC implies the existence of local minimizers of the penalized subproblems in Algorithm 3.4 as well as strong convergence of the corresponding iterates. Our approach is motivated by a recent analysis in [27] for finite-dimensional nonlinear programming. Here, we extend the corresponding results to our general setting from (P) and show the existence of minimizers using only the proximity of x^k to $\bar{x}$, whereas no assumption regarding the proximity of the multipliers λ^k is required.

As a first step in the local convergence analysis, we consider a local minimizer of (P) and ask whether the augmented Lagrangian admits local minimizers near this point. As we shall see, the answer to this question is closely linked to the fulfillment of second-order sufficient conditions (SOSCs) of the form given in Definition 5.1. When using the second-order condition, special care needs to be taken because the embedding $Y \hookrightarrow H$ allows us to interpret the constraint in (P) either in Y or in H. We have already seen that this makes a strong difference for constraint qualifications, and the situation for SOSC is quite similar. The second-order condition in H, for instance, requires the existence of Lagrange multipliers in H, which in itself is already a restriction. Nevertheless, this is in a sense the more "natural" second-order condition for the augmented Lagrangian method since the augmentation is performed in H. Thus, for the most part of this section (with the exception of Proposition 5.4), we will make the following assumption.

Assumption 5.2 (Local Convergence) There is a KKT point $(\bar{x}, \bar{\lambda}) \in X \times H$ of (P), which satisfies the SOSC from Definition 5.1 with respect to the space H.

This assumption yields the following local existence and (strong) convergence result.

Theorem 5.3 *Let Assumption 5.2 hold, and let $B \subseteq H$ be a bounded set. Then, there are $\bar{\rho}, \bar{\varepsilon}, r > 0$ such that, for all $w \in B$, $\rho \geq \bar{\rho}$, and $\varepsilon \in (0, \bar{\varepsilon})$, there is a point $x = x_{\rho,\varepsilon}(w) \in C$ with $\|x - \bar{x}\|_X < r$ and the following properties:*

(i) x is an ε-minimizer of $\mathcal{L}_\rho(\cdot, w)$ on $B_r(\bar{x}) \cap C$,
(ii) x satisfies $\operatorname{dist}\big(-\mathcal{L}'_\rho(x, w), \mathcal{N}_C(x)\big) \leq \varepsilon^{1/2}$, and
(iii) $x = x_{\rho,\varepsilon}(w) \to \bar{x}$ uniformly on B as $\rho \to \infty$ and $\varepsilon \to 0$.

If X is reflexive and the augmented Lagrangian $\mathcal{L}_\rho(\cdot, w)$ is weakly sequentially lsc, then the assertions of the above theorem remain valid if we replace the ε-minimizers by exact minimizers. In this case, we obtain points $x = x_\rho(w)$, which satisfy (i) and (ii) with $\varepsilon := 0$ and which converge to $\bar{x}$ uniformly on B as $\rho \to \infty$. Sufficient conditions for the weak sequential lower semicontinuity of $\mathcal{L}_\rho(\cdot, w)$ were given in Lemma 4.2.

If the mapping G is completely continuous, then it is possible to prove a similar result under the second-order sufficient condition with respect to the space Y. This result is a generalization of a theorem from [53].

Proposition 5.4 *Let $(\bar{x}, \bar{\lambda}) \in X \times Y^*$ be a KKT point of (P), which satisfies SOSC with respect to the space Y, and $B \subseteq H$ a bounded set. Assume that*

(i) the space X is reflexive,
(ii) f is weakly sequentially lsc on X, and
(iii) G is completely continuous from X into Y.

Then, there are $\bar{\rho}, r > 0$ such that, for every $w \in B$ and $\rho \geq \bar{\rho}$, the problem $\min_{x \in C} \mathcal{L}_\rho(x, w)$ admits a local minimizer $x = x_\rho(w)$ in $B_r(\bar{x}) \cap C$, and $x_\rho \to \bar{x}$ uniformly on B as $\rho \to \infty$.

5.2 Rate of Convergence

We are now in a position to discuss the convergence of Algorithm 3.4 from a quantitative point of view. Throughout this section, we assume that the space X is a real Hilbert space, that there is a local minimizer $\bar{x} \in X$ of (P) with a unique Lagrange multiplier $\bar{\lambda} \in H$, and that the following local error bound condition

$$c_1 \Theta(x, \lambda) \leq \|x - \bar{x}\|_X + \|\lambda - \bar{\lambda}\|_H \leq c_2 \Theta(x, \lambda) \tag{5.2}$$

holds for all $(x, \lambda) \in X \times H$ with x near $\bar{x}$ and $\Theta(x, \lambda)$ sufficiently small, where Θ is the residual

$$\Theta(x, \lambda) := \|x - P_C(x - \mathcal{L}'(x, \lambda))\|_X + \|G(x) - P_{\mathcal{K}}(G(x) + \lambda)\|_H.$$

The regularity assumptions mentioned above may seem rather stringent in view of the Gel'fand triple framework $Y \hookrightarrow H \hookrightarrow Y^*$. Indeed, a sufficient condition for the local error bound is a combination of the second-order sufficient condition and the strict Robinson condition (SRC), both with respect to the space H. This effectively rules out certain applications where the embedding $Y \hookrightarrow H$ is too weak, but the underlying issue is that we simply cannot expect the results in this section to hold if the constraint system of (P) is only regular with respect to the space Y. This is also evidenced by the fact that the rate-of-convergence analysis will enable us to prove the boundedness of the penalty sequence $\{\rho_k\}$, and this actually *implies* the existence of a Lagrange multiplier in H under certain assumptions; see Corollary 4.16 and the discussion after Corollary 5.8 below.

Despite these restrictions, the theory we develop here is still applicable to a fair amount of nontrivial problems such as control-constrained optimal control, elliptic parameter estimation problems, and of course optimization in finite dimensions.

Assumption 5.5 (Rate of Convergence) We assume that

(i) X is a real Hilbert space with f and G continuously differentiable on X,
(ii) $(\bar{x}, \bar{\lambda}) \in X \times H$ is a KKT point of (P), which satisfies the error bound (5.2),
(iii) the primal–dual sequence $\{(x^k, \lambda^k)\}$ converges strongly to $(\bar{x}, \bar{\lambda})$ in $X \times H$,
(iv) the safeguarded multiplier sequence satisfies $w^k := \lambda^k$ for k sufficiently large, and
(v) $x^{k+1} \in C$ and $\varepsilon^{k+1} - \mathcal{L}'_{\rho_k}(x^{k+1}, w^k) \in \mathcal{N}_C(x^{k+1})$ for all k, where $\varepsilon^k \to 0$.

Two assumptions that may require some elaboration are (iii) and (iv). Note that we already know, by Theorem 5.3, that the augmented Lagrangian admits approximate local minimizers and stationary points in a neighborhood of $\bar{x}$. We shall now see that, if the algorithm chooses these local minimizers (or any other points sufficiently close to $\bar{x}$), then we automatically obtain the convergence $(x^k, \lambda^k) \to (\bar{x}, \bar{\lambda})$ in $X \times H$. In this case, the sequence $\{\lambda^k\}$ is necessarily bounded in H, so it is reasonable to assume that the safeguarded multipliers are eventually chosen as $w^k = \lambda^k$. The following result can therefore be considered as (retrospective) justification for Assumption 5.5.

Proposition 5.6 *Let Assumption 5.5 (i), (ii), and (v) hold, and let RCQ hold in $\bar{x}$ with respect to the space H. Then, there exists $r > 0$ such that, if $x^k \in B_r(\bar{x})$ for sufficiently large k, then $\Theta(x^k, \lambda^k) \to 0$ and $(x^k, \lambda^k) \to (\bar{x}, \bar{\lambda})$ strongly in $X \times H$.*

We will now state convergence rates for the primal–dual sequence $\{(x^k, \lambda^k)\}$.

Theorem 5.7 *Let Assumption 5.5 hold, and assume that $\varepsilon^{k+1} = o(\theta_k)$. Then,*

(a) for every $q \in (0, 1)$*, there exists* $\bar{\rho}_q > 0$ *such that, if* $\rho_k \geq \bar{\rho}_q$ *for sufficiently large* k*, then* $(x^k, \lambda^k) \to (\bar{x}, \bar{\lambda})$ *Q-linearly in* $X \times H$ *with rate* q*;*
(b) if $\rho_k \to \infty$*, then* $(x^k, \lambda^k) \to (\bar{x}, \bar{\lambda})$ *Q-superlinearly in* $X \times H$*.*

The assumption $\varepsilon^{k+1} = o(\theta_k)$ in the above theorem says that, roughly speaking, the degree of inexactness should be small enough to not affect the rate of convergence. Note that we are comparing ε^{k+1} to the optimality measure θ_k of the previous iterates (x^k, λ^k). Hence, it is easy to ensure this condition in practice, for instance, by always computing the next iterate x^{k+1} with a precision $\|\varepsilon^{k+1}\|_X \leq z_k \theta_k$ for some fixed null sequence z_k.

Corollary 5.8 *Let Assumption 5.5 hold, and assume that the subproblems occurring in Algorithm 3.4 are solved exactly, i.e., that* $\varepsilon^k = 0$ *for all* k*. Then,* $\{\rho_k\}$ *remains bounded.*

The boundedness of $\{\rho_k\}$ obviously rules out the Q-superlinear convergence of Theorem 5.7 (b). However, the former is usually considered more significant in practice since it prevents the subproblems from becoming excessively ill-conditioned.

Remark 5.9 If inexact solutions are allowed for the augmented Lagrangian subproblems, then the boundedness of $\{\rho_k\}$ requires a slightly modified updating rule for the penalty parameter since the one used in Algorithm 3.4 does not take into account the current measure of optimality. Indeed, if we replace the function V from (3.9) by

$$\tilde{V}(x, \lambda, \rho) := V(x, \lambda, \rho) + \|x - P_C(x - \mathcal{L}'(x, \lambda))\|_X,$$

then it is possible to show that $\{\rho_k\}$ remains bounded under the assumptions of Theorem 5.7. A proof for the case $C = X$ can be found in [49], and the extension to the general case is straightforward (see also [11, 13]).

Remark 5.10 In the case of finite-dimensional nonlinear programming, it is possible to obtain similar rate of convergence results to those above under the second-order sufficient condition only. In this case, one obtains that $(x^k, \lambda^k) \to (\bar{x}, \lambda)$ Q-linearly for some $\lambda \in \Lambda(\bar{x})$, which is not necessarily equal to $\bar{\lambda}$. This result can be found in [27]. The reason why this is possible is that, for nonlinear programming, the set $\mathcal{K}$ is polyhedral and, therefore, the second-order condition implies a local primal–dual error bound without any constraint qualification.

Remark 5.11 A specification of the previous results in the Banach space setting to nonlinear semi-definite programs, second-order cone programs, and related problems is given in [51]. Though these results were essentially obtained from the general theory, the resulting convergence conditions may still be viewed as generalizations of previous results known for semi-definite programs, etc., cf. [73]. Though these problems are finite dimensional, they have a non-polyhedral feasible set, and hence SOSC-type conditions alone were not enough in order to establish the rate-of-convergence results.

6 Numerical Results

Since the safeguarded augmented Lagrangian method discussed in this chapter is identical to the one from the recent paper [15] and since that paper already presents numerical results on a variety of different optimization problems, there is, formally, no need to provide additional material here. For illustrative reasons, however, we report some numerical results also in this chapter using some other test examples. The implementation of our numerical examples has been done with FEniCS [57] using the DOLFIN [58] Python interface.

6.1 State-Constrained Optimal Control Problems

PDE-constrained optimal control problems describe a rather popular class of optimization problems. For our numerical test, we adapted a linear elliptic example with known solution from [71] to the semilinear setting, see also [53].

Let $\Omega := (-1, 2)^2$. We aim at minimizing the objective function $f\colon L^2(\Omega) \to \mathbb{R}$

$$f(u) := \frac{1}{2}||Su - y_d||^2_{L^2(\Omega)} + \frac{\alpha}{2}||u||^2_{L^2(\Omega)}, \tag{6.1}$$

subject to the pointwise inequality constraints

$$Su \leq \psi \quad \text{in } \overline{\Omega}.$$

Here, $\alpha > 0$ is a positive parameter, and $y_d \in L^2(\Omega)$ and $\psi \in C(\overline{\Omega})$ are given functions. The solution operator $S\colon L^2(\Omega) \to H^1_0(\Omega) \cap C(\overline{\Omega})$ maps the control u to the state $y := Su$, which is the uniquely determined weak solution of the underlying semilinear partial differential equation

$$\begin{aligned} -\Delta y + y^5 &= u + f \quad && \text{in } \Omega, \\ y &= 0 && \text{on } \partial\Omega, \end{aligned}$$

where $f \in L^2(\Omega)$. In this setting, the operator S is completely continuous [19, Theorem 2.1] and Fréchet differentiable [19, Theorem 2.4]. We set

$$X := L^2(\Omega), \quad C := L^2(\Omega), \quad Y := C(\overline{\Omega}), \quad G(u) := Su - \psi, \quad K := C(\overline{\Omega})_-$$
$$H := L^2(\Omega), \quad \mathcal{K} := L^2(\Omega)_-,$$

where $C(\overline{\Omega})_-$ denotes the closed convex cone of non-positive continuous functions and $L^2(\Omega)_-$ the non-positive functions in $L^2(\Omega)$. Applying standard arguments, we obtain that problem (6.1) admits at least one solution; see, for instance, [38]. Let $\bar{u}$

denote a local solution, and let us assume that there exists $\hat{u} \in L^2(\Omega)$ such that the linearized Slater condition

$$G(\bar{u})+G'(\bar{u})(\hat{u}-u) \in \operatorname{int}\left(C(\overline{\Omega})_-\right) \quad \Leftrightarrow \quad S\bar{u}+S'(\bar{u})(\hat{u}-\bar{u}) \leq \psi-\sigma \text{ in } \bar{\Omega},\ \sigma > 0$$

is satisfied. Since the interior of $C(\overline{\Omega})_-$ is nonempty, the linearized Slater condition is, for feasible points, equivalent to the Robinson constraint qualification [14, Lemma 2.99], and we obtain existence of a multiplier $\lambda \in C(\overline{\Omega})_-^\circ$. Hence, λ is an element of the space of regular Borel measures $C(\overline{\Omega})^* = \mathcal{M}(\overline{\Omega})$ [19, Theorem 3.1]. Introducing the state $y = Su$ and the adjoint state $\bar{p} \in W_0^{1,s}(\Omega)$, $s \in (1, 2)$, it is well known that the first-order necessary optimality conditions for the original problem (6.1) are given by

$$\begin{cases} -\Delta \bar{y} + \bar{y}^5 = \bar{u} + f & \text{in } \Omega, \\ \bar{y} = 0 & \text{on } \partial\Omega, \end{cases} \qquad \begin{cases} -\Delta \bar{p} + 5\bar{y}^4 \bar{p} = \bar{y} - y_d + \bar{\lambda} & \text{in } \Omega, \\ \bar{p} = 0 & \text{on } \partial\Omega, \end{cases}$$

$$\bar{p} + \alpha \bar{u} = 0, \tag{6.2}$$

$$\bar{\lambda} \in C(\overline{\Omega})_-^\circ, \quad \langle \bar{\lambda}, \psi - \bar{y} \rangle_{\mathcal{M}(\overline{\Omega}), C(\overline{\Omega})} = 0.$$

The low regularity of $\bar{\lambda}$ complicates the direct numerical solution of the optimal control problem. However, by augmenting the objective function, we eliminate the state constraints from the set of explicit constraints. Due to our choice of $\mathcal{K}$, we obtain $d_{\mathcal{K}}^2(\cdot) = ||(\cdot)_+||_{L^2(\Omega)}^2$, where $(\cdot)_+ := \max(0, \cdot)$. Following Algorithm 3.4, we have to solve a sequence of unconstrained subproblems of the type

$$\min_{u^k} f(u^k) + \frac{\rho_k}{2} \left\| \left(Su^k - \psi + \frac{w^k}{\rho_k} \right)_+ \right\|_{L^2(\Omega)}^2. \tag{6.3}$$

Since these problems are control constrained only, it is straightforward to show existence of solutions and derive the corresponding optimality conditions [74, Theorem 4.20]. However, due to the nonlinearity of the solution operator S, the functional f is not convex. Accordingly, we can only expect to compute stationary points of the augmented subproblems, which are not necessarily local or global minimizers. In order to apply our convergence results from Sect. 4.3, we need to verify Assumption 4.8.

- The mapping $G' \colon X \to L(X, Y)$ is completely continuous. In the present setting, since $X = L^2(\Omega)$ is reflexive and $G'(u) \in L(X, Y)$ is completely continuous for all u, this is equivalent to the following property: whenever $u^k \rightharpoonup u$ and $h^k \rightharpoonup h$ in X, then $G(u^k)h^k \to G(u)h$ strongly in Y. A proof of this statement (for the Neumann case) can be found in [53, Lem. 4.7].

- The mapping $f'\colon X \to X^*$ is bounded and pseudomonotone. Note that $f'(u) := S'(u)(S(u) - y_d) + \alpha u$ for all $u \in X$. The operators S and S' are completely continuous and hence bounded (since X is reflexive). This implies the boundedness of f'. The pseudomonotonicity follows from the fact that the first term in f' is completely continuous and the second term is monotone and continuous, see Lemma 2.12.

In this scenario, it follows from Theorem 4.10 that every weak limit point u^* of the sequence $\{u^k\}$ is a stationary point of the problem. Moreover, the corresponding subsequence of multipliers $\{\lambda^k\}$ converges weak-* in $\mathcal{M}(\overline{\Omega})$ to a Lagrange multiplier in u^*.

For the sake of completeness, let us state the optimality system of (6.3) that has to be solved in every iteration of Algorithm 3.4. Let $\bar{u}^k$ denote a local solution of the subproblem (6.3) and $\bar{y}^k \in H_0^1(\Omega) \cap C(\overline{\Omega})$ the corresponding state $\bar{y}^k := S\bar{u}^k$. Then, there exists an adjoint state $\bar{p}^k \in H_0^1(\Omega)$ such that the following system is satisfied:

$$\begin{cases} -\Delta \bar{y}^k + \bar{y}^{k^5} = \bar{u}^k + f & \text{in } \Omega, \\ \bar{y}^k = 0 & \text{on } \partial\Omega, \end{cases} \qquad \begin{cases} -\Delta \bar{p}^k + 5\bar{y}^{k^4}\bar{p}^k = \bar{y}^k - y_d + \bar{\lambda}^k & \text{in } \Omega, \\ \bar{p}^k = 0 & \text{on } \partial\Omega, \end{cases}$$

$$\bar{p}^k + \alpha \bar{u}^k = 0, \tag{6.4}$$

$$\bar{\lambda}^k = (w^k + \rho_k(\bar{y}^k - \psi))_+.$$

In this system, the approximation of the multiplier $\bar{\lambda}^k$ enjoys a much stronger regularity. In fact, it is an $L^2(\Omega)$-function, which allows us to apply efficient solution algorithms. We use the notation $r := r(x_1, x_2) := \sqrt{x_1^2 + x_2^2}$ with $x_1, x_2 \in \Omega$ to set

$$\bar{y}(r) := -\frac{1}{2\pi\alpha}\chi_{r\le 1}\left(\frac{r^2}{4}(\log r - 2) + \frac{r^3}{4} + \frac{1}{4}\right), \qquad \psi(r) := -\frac{1}{2\pi\alpha}\left(\frac{1}{4} - \frac{r}{2}\right)$$

$$\bar{u}(r) := \frac{1}{2\pi\alpha}\chi_{r\le 1}(\log r + r^2 - r^3), \qquad y_d(r) := \tilde{y}_d(r) - 5\bar{y}^4\bar{p},$$

$$\bar{p}(r) := -\alpha\bar{u}(r), \qquad f(r) := \tilde{f}(r) - \bar{y}^5,$$

$$\bar{\lambda}(r) := \delta_0(r),$$

where $\tilde{y_d}(r)$ and $\tilde{f}(r)$ are given auxiliary functions

$$\tilde{y}_d(r) := \bar{y}(r) - \frac{1}{2\pi}\chi_{r\le 1}(4 - 9r), \qquad \tilde{f}(r) := -\frac{1}{8\pi}\chi_{r\le 1}(4 - 9r + 4r^2 - 4r^3).$$

Then, it can be shown that $(\bar{y}, \bar{u}, \bar{p}, \bar{\lambda})$ is a KKT point of (6.1). We used the parameters

$$\alpha := 1,\ \lambda^0 := 0,\ \rho_0 := 1,\ w_{\max} := 10^5,\ \gamma := 10,\ \tau := 0.2$$

and initialized our starting points equal to zero. To obtain a sequence of safeguarded multipliers $\{w^k\}$, we chose $w^k := \min(\lambda^k, w_{\max})$. We solved the arising subproblem with a semismooth Newton method up to the precision 10^{-6}. We stop the algorithm as soon as $||\min\{\lambda^k, \psi - y^k\}||_\infty \leq 10^{-6}$ was satisfied. The computed results can be seen in Figs. 1 and 2 for 256 grid points per dimension.

The $L^2(\Omega)$-error of the computed solution (y_h, u_h) to the constructed solution $(\bar{y}, \bar{u})$ in dependence of the degrees of freedom is shown on the right-hand side of Fig. 2. Table 1 shows the iteration numbers of outer and inner iterations as well as the final value of the penalty parameter $\rho_{\max}$ with respect to the number of grid points per dimension.

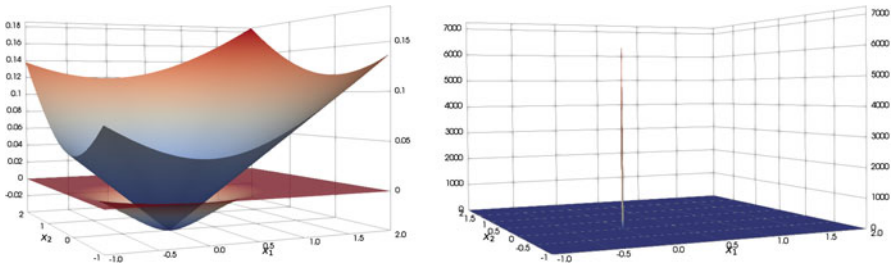

Fig. 1 (Example 1) Left: computed discrete optimal state y_h (transparent) with state constraint ψ. Right: Lagrange multiplier μ_h

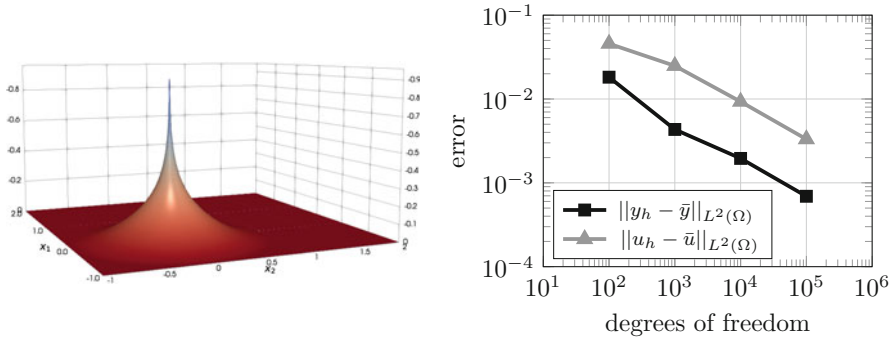

Fig. 2 (Example 1) Left: computed control u_h. Right: errors $||u_h - \bar{u}||_{L^2(\Omega)}$ and $||y_h - \bar{y}||_{L^2(\Omega)}$ vs. DOFs

Table 1 (Example 1) Iteration numbers

n	16	32	64	128	256
Outer iteration	10	9	10	11	12
Inner iteration	21	23	27	33	38
$\rho_{\max}$	10^5	10^5	10^7	10^7	10^8

6.2 Bratu's Obstacle Problem

Bratu's obstacle problem is a non-quadratic nonconvex problem, which is an efficient tool to model nonlinear diffusion phenomena. Let $\Omega \subseteq \mathbb{R}^2$ be a bounded domain. Bratu's obstacle problem is given by the minimization problem

$$\min_{u \in H_0^1(\Omega)} J(u) := ||\nabla u||^2_{L^2(\Omega)} - \alpha \int_\Omega \exp(-u(x))\, \mathrm{d}x \quad \text{s.t.} \quad u \geq \psi, \tag{6.5}$$

where $\alpha > 0$ is a positive parameter and $\psi \in H_0^1(\Omega)$ denotes the given fixed obstacle. To satisfy our general framework, we set

$$X := Y := H_0^1(\Omega), \quad C := H_0^1(\Omega), \quad G(u) := u - \psi, \quad K := H_0^1(\Omega)_+,$$
$$H := L^2(\Omega), \quad \mathcal{K} := L^2(\Omega)_+.$$

Due to [52, Lemma 7.1], we know that J is well defined, continuously Fréchet differentiable, and weakly sequentially lower semicontinuous from $H_0^1(\Omega)$ into $\mathbb{R}$. Due to the constraint $u \geq \psi$, the functional J is coercive on the feasible set. By standard arguments, we obtain existence of a solution $\bar{u} \in X$. Moreover, the surjectivity of the derivative $G'(\bar{u}) = \mathrm{Id}_X$ from X to Y implies the Robinson constraint qualification and, hence, the existence of a unique Lagrange multiplier $\bar{\lambda} \in H_0^1(\Omega)^* = H^{-1}(\Omega)$. The corresponding KKT system is given by

$$J'(\bar{u}) + \bar{\lambda} = 0$$
$$\langle \bar{\lambda}, \bar{u} - \psi \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} = 0, \quad \bar{\lambda} \in \left(H_0^1(\Omega)_+ \right)^\circ .$$

By definition of the polar cone, we obtain that $\langle \bar{\lambda}, u \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \leq 0$ for all $u \in H_0^1(\Omega)$ with $u \geq 0$. Since the objective function J is not convex, one can only expect to compute stationary points of the augmented subproblems

$$\min_{u^k} J(u^k) + \frac{\rho_k}{2} || \left(u^k - \psi + \frac{w^k}{\rho_k} \right)_- ||^2_{L^2(\Omega)},$$

which are not necessarily local or global solutions.

Lemma 6.1 *If $\Omega \subseteq \mathbb{R}^2$, then the derivative $J' \colon H_0^1(\Omega) \to H^{-1}(\Omega)$ is bounded and pseudomonotone.*

Proof We split the objective function $J(u) := J_1(u) - J_2(u)$, where

$$J_1(u) := ||\nabla u||^2_{L^2(\Omega)}, \quad J_2(u) := \alpha \int_\Omega \exp(-u(x))\, \mathrm{d}x.$$

The proof of [52, Lemma 7.1] shows that the integral term J_2 in the definition of J is weakly sequentially continuous, uniformly differentiable on bounded subsets of $H_0^1(\Omega)$, and J_2' is bounded on bounded subsets of $H_0^1(\Omega)$. It follows that J' is also a bounded operator. Since J_2 is completely continuous and uniformly differentiable on bounded subsets of X, it follows that J_2' is completely continuous [63] and in particular pseudomonotone. The monotonicity of $-\Delta$ yields that J_1' is monotone (and continuous). Thus, J' is pseudomonotone (Lemma 2.12). □

Due to Lemma 6.1, it follows from Proposition 4.12 that every weak limit point u^* of the sequence $\{u^k\}$ is a stationary point of the problem. Moreover, the corresponding subsequence of multipliers λ^k converges weak-* in $H^{-1}(\Omega)$ to the unique Lagrange multiplier in u^* (Proposition 4.15).

In particular, for $\alpha := 0$, problem (6.5) is reduced to the very well-known obstacle problem. Opposed to Bratu's problem, this problem is linear quadratic with a (strongly) convex objective function. The strong convexity of J not only implies uniqueness of the solution of the obstacle problem and its corresponding subproblem, but it also implies that the primal sequence $\{u^k\}$ converges strongly to $\bar{u}$ in X (Corollary 4.7) and the dual sequence $\{\lambda^k\}$ converges weak-* in $H^{-1}(\Omega)$ by Theorem 4.10 (see Remark 4.11) or Proposition 4.15.

In order to test our example, we chose the domain $\Omega := (0, 1)^2$. We implemented the Bratu problem for the obstacle

$$\psi(x_1, x_2) := \sum_{i=1}^{3} q_i \exp\left(-500\left((x_1 - z_i)^2 + (x_2 - z_i)^2\right)\right) - 1,$$

where $q := (60, 80, 60)$, $z := (0.25, 0.5, 0.75)$. We chose the parameters

$$\alpha := 2,\ \lambda^0 := 0,\ \rho_0 := 1,\ w_{\min} := -10^5,\ \gamma := 10,\ \tau := 0.1$$

and initialized our starting points equal to zero. We obtain a sequence of safeguarded multipliers $\{w^k\}$ by choosing $w^k := \max(\lambda^k, w_{\min})$. We solve the unconstrained subproblems with a semismooth Newton method with the precision 10^{-6} and stop the algorithm as soon as $||\max\{\lambda^k, \psi - u^k\}||_\infty \leq 10^{-6}$ is satisfied. The computed results can be seen for 128 grid points per dimension in Fig. 3. Furthermore, some iteration numbers are given in Table 2.

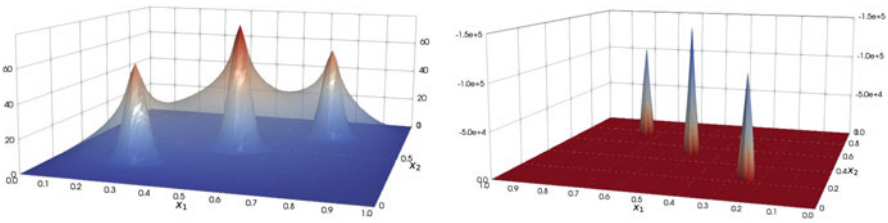

Fig. 3 (Example 2) Left: computed discrete optimal solution u_h (transparent) with constraint ψ. Right: Lagrange multiplier μ_h

Table 2 (Example 2) Iteration numbers

n	16	32	64	128	256
Outer iteration	9	9	12	12	13
Inner iteration	14	17	25	32	34
$\rho_{\max}$	10^4	10^5	10^{10}	10^{10}	10^{10}

6.3 $C(\overline{\Omega})$-Minimization

We consider an optimal control problem with an objective functional containing an $C(\overline{\Omega})$ norm term, namely

$$\underset{y\in H^1(\Omega)\cap C(\overline{\Omega}),\ u\in L^2(\Omega)}{\text{minimize}} \ \frac{1}{2}||y-y_d||^2_{C(\overline{\Omega})}+\frac{\alpha}{2}||u||^2_{L^2(\Omega)}, \tag{6.6}$$

where the state y has to satisfy the semilinear partial differential equation

$$\begin{aligned} -\Delta y+\exp(y) &= u+f \quad && \text{in } \Omega \\ \partial y &= 0 && \text{on } \partial\Omega, \end{aligned}$$

where ∂y denotes the normal derivative of y on $\partial\Omega$ and f a function in $L^2(\Omega)$. Functionals including an $C(\overline{\Omega})$ norm term are not differentiable and therefore difficult to handle. We introduce the control-to-state mapping $S\colon L^2(\Omega) \to H^1(\Omega)\cap C(\overline{\Omega})$, which maps the control u on the associated, uniquely determined, state $S: u\mapsto y$ [18, Theorem 3.1]. The original problem is now substituted by an equivalent problem with a differentiable function given by

$$\min_{z\in\mathbb{R},u\in L^2(\Omega)} f(u,z):=\frac{1}{2}z^2+\frac{\alpha}{2}||u||^2_{L^2(\Omega)} \quad \text{subject to} \quad |Su-y_d|\le z. \tag{6.7}$$

Clearly, problem (6.7) is related to state-constrained optimal control problems. However, now z is a free variable. Consequently, we aim at finding the smallest $z\in\mathbb{R}$ and $u\in L^2(\Omega)$ such that the pointwise inequality constraint is satisfied and the objective function f is minimized. Problems of this type have already been investigated in [32, 66]. Moreover, in [20], pointwise constraints on the state variable on a specified subdomain of Ω under piecewise constant controls were investigated. To satisfy our general framework, we set $x:=(u,z)\in L^2(\Omega)\times\mathbb{R}$, and

$$X:=L^2(\Omega)\times\mathbb{R}, \quad C:=L^2(\Omega)\times\mathbb{R}, \quad Y:=C(\overline{\Omega})\times C(\overline{\Omega}),$$

$$K:=C(\overline{\Omega})_-\times C(\overline{\Omega})_+, \quad H:=L^2(\Omega)\times\mathbb{R}, \quad \mathcal{K}:=L^2(\Omega)_-\times L^2(\Omega)_+$$

as well as

$$G(x) := \begin{pmatrix} Su - y_d - z \\ Su - y_d + z \end{pmatrix}.$$

This leads us again to a minimization problem of the type $\min_x \ f(x)$ such that $G(x) \in K$. Like in Example 1, the solution operator S [18], and hence G, is completely continuous and continuously Fréchet differentiable [74, Theorem 4.17]. Thus, we obtain by standard arguments the existence of an optimal solution $(\bar{y}, \bar{u}) \in H^1(\Omega) \times L^2(\Omega)$ of (6.6). Hence, defining $\bar{z} := ||\bar{y} - y_d||_{C(\overline{\Omega})}$, we can conclude that $(\bar{u}, \bar{z})$ is a solution of (6.7). Let $\bar{x} := (\bar{u}, \bar{z}) \in (L^2(\Omega) \times \mathbb{R})$ denote a local solution. Then, it is easy to see that the Robinson constraint qualification is satisfied. Indeed, the first line of the inclusion

$$0 \in \operatorname{int}\left[G(\bar{u}, \bar{z}) + G'(\bar{u}, \bar{z}) \begin{pmatrix} L^2(\Omega) - \bar{u} \\ \mathbb{R} - \bar{z} \end{pmatrix} - (K_- \times K_+) \right]$$

can be written as

$$0 \in \operatorname{int}\Big[(S\bar{u} - y_d - \bar{z}) + S'(\bar{u})(L^2(\Omega) - \bar{u}) - (\mathbb{R} - \bar{z}) - K_- \Big],$$

which is fulfilled as $\mathbb{R} + K_- = C(\bar{\Omega})$. Then, there exist Lagrange multipliers $\bar{\lambda}_1, \bar{\lambda}_2 \in C(\overline{\Omega})^* = \mathcal{M}(\overline{\Omega})$. Moreover, it is easy to see that $\bar{\lambda}_1 + \bar{\lambda}_2 \in \partial\left(\frac{1}{2}||\cdot||^2_{C(\overline{\Omega})}\right)(S\bar{u} - y_d)$, where ∂ denotes the convex subdifferential. It remains to verify Assumption 4.8. Following the same argumentation as in Example 1, we can deduce that $S'(u) \in L(L^2(\Omega), C(\overline{\Omega}))$ is completely continuous and, thus, $G' : X \to L(X, Y)$ is completely continuous. Furthermore, the mapping $f' \colon X \to X^*$, $f'(x) = (\alpha u, z)^T$ is bounded and by Lemma 2.12 pseudomonotone.

According to Algorithm 3.4, we have to solve the following unconstrained subproblem in every iteration of the algorithm:

$$\begin{aligned} \underset{u^k, z^k}{\text{minimize}} \ f(u^k, z^k) &+ \frac{\rho_k^1}{2} || \left(Su^k - y_d - z^k + \frac{w_1^k}{\rho_k^1} \right)_+ ||^2_{L^2(\Omega)} \\ &+ \frac{\rho_k^2}{2} || \left(Su^k - y_d + z^k + \frac{w_2^k}{\rho_k^2} \right)_- ||^2_{L^2(\Omega)}. \end{aligned}$$

Reintroducing the state $y = Su$ and the adjoint state $p \in H^1(\Omega)$, we obtain the corresponding optimality system by standard arguments

$$\begin{cases} -\Delta \bar{y}^k + \exp(\bar{y}^k) = \bar{u}^k + f \ \text{ in } \Omega, \\ \partial \bar{y}^k = 0 \qquad \text{on } \partial\Omega, \end{cases} \qquad \begin{cases} -\Delta \bar{p}^k + \exp(\bar{y}^k)\bar{p}^k = \lambda_1^k + \lambda_2^k \ \text{ in } \Omega, \\ \partial \bar{p}^k = 0 \qquad \text{on } \partial\Omega, \end{cases}$$

$$\alpha \bar{u}^k + \bar{p}^k = 0, \tag{6.8}$$

$$z^k - \int_\Omega \lambda_1^k + \int_\Omega \lambda_2^k = 0,$$

where

$$\lambda_1^k := (w_1 + \rho_k^1(\bar{y}^k - y_d - z^k))_+, \qquad \lambda_2^k := (w_2 + \rho_k^2(\bar{y}^k - y_d + z^k))_-.$$

To test our example, we took $\Omega := (0, 1)^2$, set our starting points equal to zero, and chose the parameters

$$\alpha := 10^{-4}, \quad \lambda^0 := 0, \quad w_{\max} := 10^{-5}, \quad \gamma := 10, \quad \tau := 0.1.$$

Furthermore, we chose $y_d := 0$ and $f := 8\sin(\pi x_1)\sin(\pi x_2) - 4$, where $(x_1, x_2) \in \Omega$. We solved the optimality system (6.8) with a semismooth Newton method with the precision 10^{-6} and stop the algorithm as soon as

$$||\min\{\lambda_1^k, -Su^k + y_d + z^k\}||_\infty + ||\max\{\lambda_2^k, -Su^k + y_d - z^k\}||_\infty \le 10^{-6}$$

is satisfied. Figures 4 and 5 depict the computed results for $n = 128$ grid points per dimension. The corresponding optimal value of z has been computed as $\bar{z} = 6.7 \cdot 10^{-3}$. Some iteration numbers are shown in Table 3.

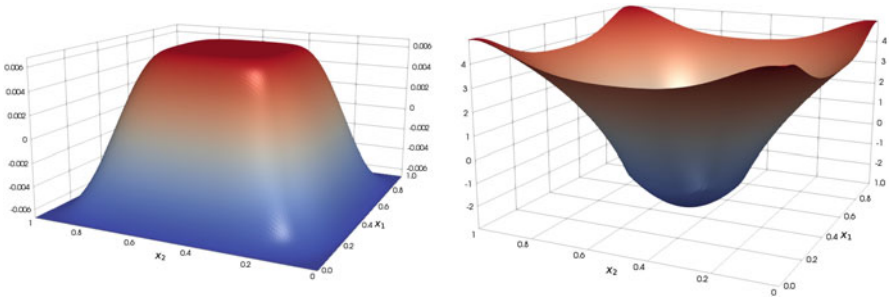

Fig. 4 (Example 3) Computed discrete optimal state y_h (left) with optimal control u_h (right)

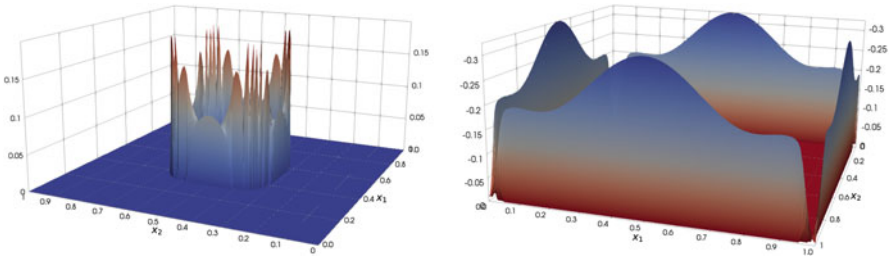

Fig. 5 (Example 3) Computed discrete Lagrange multipliers $\mu_{h,1}$ and $\mu_{h,2}$

Table 3 (Example 3) Iteration numbers

n	16	32	64	128
Outer iteration	8	7	8	8
Inner iteration	17	19	26	26
$\rho_{\max}$	10^4	10^5	10^6	10^6

7 Final Remarks

The previous survey shows that the safeguarded augmented Lagrangian approach has a very strong global and local convergence theory which allows its application to a wide variety of different applications. The numerical results in this and some related papers by the authors indicate that the approach also works quite successfully from a numerical point of view. Nevertheless, there are plenty of possible modifications that might be interesting to investigate. For example, in finite dimensions, the augmented Lagrangian approach converges under much weaker assumptions than the Robinson CQ, but these weaker assumptions currently do not exist in Banach spaces simply because there is not counterpart of the corresponding constraint qualifications in infinite dimensions. Another interesting generalization might be a relaxation of the second-order sufficiency condition, which is currently assumed to hold at a KKT point, but the existence of such a KKT point might be too strong an assumption for some difficult classes of optimization problems like mathematical programs of with complementarity constraints.

References

1. R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt. On augmented Lagrangian methods with general lower-level constraints. *SIAM J. Optim.*, 18(4):1286–1309, 2007.
2. R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt. Augmented Lagrangian methods under the constant positive linear dependence constraint qualification. *Math. Program.*, 111(1-2, Ser. B):5–32, 2008.
3. R. Andreani, G. Haeser, and J. M. Martínez. On sequential optimality conditions for smooth constrained optimization. *Optimization*, 60(5):627–641, 2011.

4. R. Andreani, J. M. Martínez, and B. F. Svaiter. A new sequential optimality condition for constrained optimization and algorithmic consequences. *SIAM J. Optim.*, 20(6):3533–3554, 2010.
5. H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, Cham, second edition, 2017. With a foreword by Hédy Attouch.
6. M. Bergounioux. Augmented Lagrangian method for distributed optimal control problems with state constraints. *J. Optim. Theory Appl.*, 78(3):493–521, 1993.
7. M. Bergounioux and K. Kunisch. Augmented Lagrangian techniques for elliptic state constrained optimal control problems. *SIAM J. Control Optim.*, 35(5):1524–1543, 1997.
8. M. Bergounioux and K. Kunisch. Primal-dual strategy for state-constrained optimal control problems. *Comput. Optim. Appl.*, 22(2):193–224, 2002.
9. D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1982.
10. D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific Optimization and Computation Series. Athena Scientific, Belmont, MA, third edition, 2016.
11. E. G. Birgin, D. Fernández, and J. M. Martínez. The boundedness of penalty parameters in an augmented Lagrangian method with constrained subproblems. *Optim. Methods Softw.*, 27(6):1001–1024, 2012.
12. E. G. Birgin, C. A. Floudas, and J. M. Martínez. Global minimization using an augmented Lagrangian method with variable lower-level constraints. *Math. Program.*, 125(1, Ser. A):139–162, 2010.
13. E. G. Birgin and J. M. Martínez. *Practical Augmented Lagrangian Methods for Constrained Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2014.
14. J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research. Springer-Verlag, New York, 2000.
15. E. Börgens, C. Kanzow, and D. Steck. Local and global analysis of multiplier methods for constrained optimization in Banach spaces. *SIAM J. Cont. Optim.* 57(6):3694–3722, 2019.
16. J. M. Borwein and Q. J. Zhu. *Techniques of Variational Analysis*, volume 20 of *CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC*. Springer-Verlag, New York, 2005.
17. H. Brezis. Équations et inéquations non linéaires dans les espaces vectoriels en dualité. *Ann. Inst. Fourier (Grenoble)*, 18(fasc. 1):115–175, 1968.
18. E. Casas. Boundary control of semilinear elliptic equations with pointwise state constraints. *SIAM J. Control Optim.*, 31(4):993–1006, 1993.
19. E. Casas, J. C. de los Reyes, and F. Tröltzsch. Sufficient second-order optimality conditions for semilinear control problems with pointwise state constraints. *SIAM J. Optim.*, 19(2):616–643, 2008.
20. C. Clason, K. Ito, and K. Kunisch. Minimal invasion: an optimal L^∞ state constraint problem. *ESAIM Math. Model. Numer. Anal.*, 45(3):505–522, 2011.
21. A. R. Conn, N. Gould, A. Sartenaer, and P. L. Toint. Convergence properties of an augmented Lagrangian algorithm for optimization with a combination of general equality and linear constraints. *SIAM J. Optim.*, 6(3):674–703, 1996.
22. A. R. Conn, N. I. M. Gould, and P. L. Toint. A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM J. Numer. Anal.*, 28(2):545–572, 1991.
23. A. R. Conn, N. I. M. Gould, and P. L. Toint. *LANCELOT. A Fortran Package for Large-Scale Nonlinear Optimization (Release A)*, volume 17 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1992.
24. A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust-Region Methods*. MPS/SIAM Ser. Optim. SIAM, Philadelphia, 2000.
25. A. L. Dontchev. The Graves theorem revisited. *J. Convex Anal.*, 3(1):45–53, 1996.

26. A. L. Dontchev and R. T. Rockafellar. *Implicit Functions and Solution Mappings. A View from Variational Analysis*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2014.
27. D. Fernández and M. V. Solodov. Local convergence of exact and inexact augmented Lagrangian methods under the second-order sufficient optimality condition. *SIAM J. Optim.*, 22(2):384–407, 2012.
28. C. A. Floudas and P. M. Pardalos, editors. *Encyclopedia of Optimization*. Springer, New York, second edition, 2009.
29. A. Forsgren, P. E. Gill, and M. H. Wright. Interior methods for nonlinear optimization. *SIAM Rev.*, 44(4):525–597 (2003), 2002.
30. M. Fortin and R. Glowinski. *Augmented Lagrangian Methods. Applications to the Numerical Solution of Boundary-Value Problems*, volume 15 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1983. Translated from the French by B. Hunt and D. C. Spicer.
31. N. Gould, D. Orban, and P. Toint. Numerical methods for large-scale nonlinear optimization. *Acta Numer.*, 14:299–361, 2005.
32. T. Grund and A. Rösch. Optimal control of a linear elliptic equation with a supremum norm functional. *Optim. Methods Softw.*, 15(3–4):299–329, 2001.
33. M. R. Hestenes. Multiplier and gradient methods. *J. Optimization Theory Appl.*, 4:303–320, 1969.
34. M. Hintermüller and K. Kunisch. Feasible and noninterior path-following in constrained minimization with low multiplier regularity. *SIAM J. Control Optim.*, 45(4):1198–1221, 2006.
35. M. Hintermüller, A. Schiela, and W. Wollner. The length of the primal-dual path in Moreau-Yosida-based path-following methods for state constrained optimal control. *SIAM J. Optim.*, 24(1):108–126, 2014.
36. M. Hinze and C. Meyer. Variational discretization of Lavrentiev-regularized state constrained elliptic optimal control problems. *Comput. Optim. Appl.*, 46(3):487–510, 2010.
37. M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, New York, 2009.
38. M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, New York, 2009.
39. K. Ito and K. Kunisch. The augmented Lagrangian method for equality and inequality constraints in Hilbert spaces. *Math. Programming*, 46(3, (Ser. A)):341–360, 1990.
40. K. Ito and K. Kunisch. The augmented Lagrangian method for parameter estimation in elliptic systems. *SIAM J. Control Optim.*, 28(1):113–136, 1990.
41. K. Ito and K. Kunisch. An augmented Lagrangian technique for variational inequalities. *Appl. Math. Optim.*, 21(3):223–241, 1990.
42. K. Ito and K. Kunisch. Augmented Lagrangian methods for nonsmooth, convex optimization in Hilbert spaces. *Nonlinear Anal.*, 41(5-6, Ser. A: Theory Methods):591–616, 2000.
43. K. Ito and K. Kunisch. Semi-smooth Newton methods for state-constrained optimal control problems. *Systems Control Lett.*, 50(3):221–228, 2003.
44. K. Ito and K. Kunisch. *Lagrange Multiplier Approach to Variational Problems and Applications*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.
45. A. F. Izmailov and M. V. Solodov. Stabilized SQP revisited. *Math. Program.*, 133(1-2, Ser. A):93–120, 2012.
46. A. F. Izmailov, M. V. Solodov, and E. I. Uskov. Global convergence of augmented Lagrangian methods applied to optimization problems with degenerate constraints, including problems with complementarity constraints. *SIAM J. Optim.*, 22(4):1579–1606, 2012.
47. C. Kanzow and D. Steck. An example comparing the standard and safeguarded augmented Lagrangian methods. *Oper. Res. Lett.*, 45(6):598–603, 2017.
48. C. Kanzow and D. Steck. A generalized proximal-point method for convex optimization problems in Hilbert spaces. *Optimization*, 66(10):1667–1676, 2017.
49. C. Kanzow and D. Steck. On error bounds and multiplier methods for variational problems in Banach spaces. *SIAM J. Control Optim.*, 56(3):1716–1738, 2018.

50. C. Kanzow and D. Steck. Quasi-variational inequalities in Banach spaces: theory and augmented Lagrangian methods. *SIAM J. Control Optim.* 29(4):3174–3200, 2019.
51. C. Kanzow and D. Steck. Improved local convergence results for augmented Lagrangian methods in C^2-cone reducible constrained optimization. *Math. Program.* 177(1–2):425–438, 2019.
52. C. Kanzow, D. Steck, and D. Wachsmuth. An augmented Lagrangian method for optimization problems in Banach spaces. *SIAM J. Control Optim.*, 56(1):272–291, 2018.
53. V. Karl, I. Neitzel, and D. Wachsmuth. A Lagrange multiplier method for semilinear elliptic state constrained optimal control problems. *Computational Optim. Appl.*, 77:831–869, 2020.
54. K. Krumbiegel, I. Neitzel, and A. Rösch. Sufficient optimality conditions for the Moreau–Yosida-type regularization concept applied to semilinear elliptic optimal control problems with pointwise state constraints. *Ann. Acad. Rom. Sci. Ser. Math. Appl.*, 2(2):222–246, 2010.
55. K. Krumbiegel, I. Neitzel, and A. Rösch. Regularization for semilinear elliptic optimal control problems with pointwise state and control constraints. *Comput. Optim. Appl.*, 52(1):181–207, 2012.
56. F. Kruse and M. Ulbrich. A self-concordant interior point approach for optimal control with state constraints. *SIAM J. Optim.*, 25(2):770–806, 2015.
57. A. Logg, K.-A. Mardal, G. N. Wells, et al. *Automated Solution of Differential Equations by the Finite Element Method*. Springer, 2012.
58. A. Logg and G. N. Wells. Dolfin: Automated finite element computing. *ACM Transactions on Mathematical Software*, 37(2), 2010.
59. C. Meyer, A. Rösch, and F. Tröltzsch. Optimal control of PDEs with regularized pointwise state constraints. *Comput. Optim. Appl.*, 33(2–3):209–228, 2006.
60. B. S. Mordukhovich. *Variational Analysis and Generalized Differentiation. I: Basic Theory*, volume 330 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2006.
61. J.-J. Moreau. Décomposition orthogonale d'un espace hilbertien selon deux cônes mutuellement polaires. *C. R. Acad. Sci. Paris*, 255:238–240, 1962.
62. J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, second edition, 2006.
63. K. J. Palmer. On the complete continuity of differentiable mappings. *J. Austral. Math. Soc.*, 9:441–444, 1969.
64. R. R. Phelps. *Convex Functions, Monotone Operators and Differentiability*, volume 1364 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, second edition, 1993.
65. M. J. D. Powell. A method for nonlinear constraints in minimization problems. In *Optimization (Sympos., Univ. Keele, Keele, 1968)*, pages 283–298. Academic Press, London, 1969.
66. U. Prüfert and A. Schiela. The minimization of a maximum-norm functional subject to an elliptic PDE and state constraints. *ZAMM Z. Angew. Math. Mech.*, 89(7):536–551, 2009.
67. S. M. Robinson. Stability theory for systems of inequalities. II. Differentiable nonlinear systems. *SIAM J. Numer. Anal.*, 13(4):497–513, 1976.
68. R. T. Rockafellar. A dual approach to solving nonlinear programming problems by unconstrained optimization. *Math. Programming*, 5:354–373, 1973.
69. R. T. Rockafellar. Augmented Lagrange multiplier functions and duality in nonconvex programming. *SIAM J. Control*, 12:268–285, 1974.
70. R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Math. Oper. Res.*, 1(2):97–116, 1976.
71. A. Rösch and D. Wachsmuth. A-posteriori error estimates for optimal control problems with state and control constraints. *Numer. Math.*, 120(4):733–762, 2012.
72. A. Schiela. An interior point method in function space for the efficient solution of state constrained optimal control problems. *Math. Program.*, 138(1-2, Ser. A):83–114, 2013.
73. D. Sun, J. Sun, and L. Zhang. The rate of convergence of the augmented Lagrangian method for nonlinear semidefinite programming. *Math. Program.*, 114(2, Ser. A):349–391, 2008.
74. F. Tröltzsch. *Optimal Control of Partial Differential Equations*. American Mathematical Society, Providence, RI, 2010.

75. M. Ulbrich. *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, volume 11 of *MOS-SIAM Ser. Optim.* SIAM, Philadelphia, 2011.
76. A. P. Wierzbicki and S. Kurcyusz. Projection on a cone, penalty functionals and duality theory for problems with inequality constraints in Hilbert space. *SIAM J. Control Optimization*, 15(1):25–56, 1977.
77. E. Zeidler. *Nonlinear Functional Analysis and its Applications. II/B: Nonlinear Monotone Operators*. Springer-Verlag, New York, 1990. Translated from the German by the author and Leo F. Boron.

Decomposition and Approximation for PDE-Constrained Mixed-Integer Optimal Control

Mirko Hahn, Christian Kirches, Paul Manns, Sebastian Sager, and Clemens Zeile

Abstract Using partial outer convexification, we can reformulate MINLPs constrained by ODEs or PDEs such that all integer control variables are binaries. We can obtain the canonical continuous relaxation of such problems by replacing the binary control variables with [0, 1]-valued ones. The relaxation is generally easier to solve. The two-step approach of computing a relaxed solution and approximating it using binary controls afterward is called Combinatorial Integral Approximation (CIA) decomposition. We survey recent developments concerning this methodology.

There are several well-behaved algorithmic approaches that approximate the relaxed controls with binary ones. For these algorithms, driving the mesh size of the rounding mesh to zero induces convergence of the binary control with the relaxed one in the weak-$*$ topology of L^∞. Such approximation results for one-dimensional domains transfer to multidimensional ones under a mild condition on the rounding mesh refinement. If the solution operator of the state equation exhibits sufficient regularity, i.e., compactness properties, the state vector corresponding to the rounded binary control converges in norm to the state vector of the relaxed problem. Variations of these algorithms allow additional pointwise constraints that involve the discrete controls without sacrificing these convergence properties.

As a test case, we present a multidimensional model problem that compares two recently investigated algorithmic approaches, which are transferred to the multidimensional setting using iterates of the Sierpinski curve.

Keywords Mixed-integer optimal control · Approximation theory

This work was completed with the support of DFG Priority Programme 1962.

M. Hahn · S. Sager · C. Zeile
Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany
e-mail: mirhahn@ovgu.de; sager@ovgu.de; clemens.zeile@ovgu.de

C. Kirches (✉) · P. Manns
Technische Universität Carolo-Wilhelmina zu Braunschweig, Braunschweig, Germany
e-mail: c.kirches@tu-bs.de; paul.manns@tu-bs.de

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_11

Mathematics Subject Classification (2020) Primary 90C11; Secondary 49M20, 65K10

1 Introduction

We consider *partial outer convexification* reformulations, see [21, 22], of optimal control problems with mixed control inputs, i.e., control problems of the form

$$\begin{aligned}
&\min_{y,\omega} J(y) \\
&\text{s.t.} \quad Ay = \sum_{i=1}^{M} \omega_i f_i(y), \\
&\qquad 0 \leq \omega_i(s) c_i(y(s)) \qquad \text{for a.a. } s \in \Omega_T, i \in \{1, \ldots, M\}, \\
&\qquad \omega(s) \in \{0,1\}^M \qquad \text{for a.a. } s \in \Omega_T, \\
&\qquad \sum_{i=1}^{M} \omega_i(s) = 1 \qquad \text{for a.a. } s \in \Omega_T.
\end{aligned} \tag{BC}$$

Here, the quantity M denotes the number of different control realizations (or right-hand sides in the differential equation context), y denotes the state variable, and ω denotes the binary control input of the problem. $Ay = \sum_{i=1}^{M} \omega_i f_i(y)$ is the state equation of the optimized process. We assume that it is defined on a bounded domain or space-time cylinder Ω_T, A is a suitable differential operator, and the f_i are suitable nonlinearities. The functions c_i are pointwise a.e. defined constraint functions. The function $\omega : \Omega_T \to \{0,1\}^M$ activates the different right-hand sides $f_1, \ldots, f_M$ of the state equation, i.e., $\omega_i(s) = 1$ for exactly one $i \in \{1, \ldots, M\}$ and $\omega_j(s) = 0$ for $j \neq i$ a.e. The continuous relaxation of (BC) reads

$$\begin{aligned}
&\min_{y,\alpha} J(y) \\
&\text{s.t.} \quad Ay = \sum_{i=1}^{M} \alpha_i f_i(y), \\
&\qquad 0 \leq \alpha_i(s) c_i(y(s)) \qquad \text{for a.a. } s \in \Omega_T, i \in \{1, \ldots, M\}, \\
&\qquad \alpha(s) \in [0,1]^M \qquad \text{for a.a. } s \in \Omega_T, \\
&\qquad \sum_{i=1}^{M} \alpha_i(s) = 1 \qquad \text{for a.a. } s \in \Omega_T.
\end{aligned} \tag{RC}$$

We note that additional continuous control inputs into J, the f_i, and c_i would be possible here if we added additional assumptions. However, we omit them to keep the article concise. We note that the constraint $0 \leq \alpha_i c_i(y)$ implies that versions of (RC) with discretized differential equations exhibit the so-called *vanishing constraints*. For further information on optimality conditions and algorithmic approaches for the class of optimization problems exhibiting such constraints, *Mathematical Programs with Vanishing Constraints (MPVCs)*, we refer to the articles [1, 8–10].

Let Y be a Banach space that serves as the state space for the state equation. We will make use of the abbreviations

$$\mathcal{F}_{(\mathrm{BC})} \coloneqq \left\{(y,\omega) \in Y \times L^\infty(\Omega_T, \mathbb{R}^M) : (y,\omega) \text{ feasible for (BC)}\right\},$$

$$\mathcal{F}_{(\mathrm{RC})} \coloneqq \left\{(y,\alpha) \in Y \times L^\infty(\Omega_T, \mathbb{R}^M) : (y,\alpha) \text{ feasible for (RC)}\right\},$$

for the feasible sets of (BC) and (RC). The following definition applies the naming convention of relaxed and binary control to α and ω, see [15].

Definition 1.1 (Binary and Relaxed Control) Let $d \in \mathbb{N}$. Let $\Omega_T \subset \mathbb{R}^d$ be a bounded domain. We call a measurable function $\omega : \Omega_T \to \{0,1\}^M$ with $\sum_{i=1}^M \omega_i = 1$ a.e. in Ω_T a *binary control* and a measurable function $\alpha : \Omega_T \to [0,1]^M$ with $\sum_{i=1}^M \alpha_i = 1$ a.e. in Ω_T a *relaxed control*.

We split the process of solving (BC) into the following two steps:

1. solve the relaxation (RC) to obtain an optimal relaxed control α^* and
2. derive a binary control ω from α^* as an approximate solution for (BC).

We call the second step *rounding* and stress that this is different from pointwise rounding to the nearest integer. This splitting methodology is described in detail in [23] and sometimes called *Combinatorial Integral Approximation (CIA) decomposition*. Several algorithmic approaches exist to compute the binary control in the second step. For instance, Sum-Up Rounding (SUR) [20] and Next-Forced Rounding (NFR) [11] provide guaranteed bounds on the so-called *integrality gap*, $\sup_t \left\| \int_0^t \alpha - \omega \right\|$ in the one-dimensional case $\Omega_T = (0,T)$, which behave linearly with respect to the mesh size of the rounding mesh. Here, the term mesh size refers to the maximum cell volume of the mesh cells, which is different from its use in the literature on PDE numerics. As the mesh size may be fixed prior to the solution process, it is also sometimes suggested to compute the binary control by directly minimizing the integrality gap for a given rounding mesh, see [23]. We later refer to the resulting optimization problem as the CIA problem.

As noted in [7, 14], similar convexification and approximation properties have been studied in the optimal control community in contexts other than mixed-integer

optimization. We reference the important Filippov–Ważewski theorem, see [6, 24]. This theorem states that the solutions of the differential inclusion

$$\frac{\mathrm{d}}{\mathrm{d}t} y(t) \in F(y(t)), t \in [0, T],$$
$$y(0) = y_0$$

are dense in the solutions of the differential inclusion

$$\frac{\mathrm{d}}{\mathrm{d}t} y(t) \in \overline{\mathrm{conv}\{F(y(t))\}}, t \in [0, T],$$
$$y(0) = y_0$$

for a Lipschitz continuous set-valued function F, which maps into compact subsets of a Euclidean space and a uniformly bounded solution set of the second differential inclusion.

Our rounding algorithms can be interpreted as constructive means to compute the approximation in a mixed-integer optimal control setting. We note that similar considerations are used for model order reduction using Koopman operators; see the recent publication [18].

1.1 Outline of the Remaining Sections

Section 2 summarizes sufficient conditions on the rounding meshes and algorithms as well as the approximation arguments to obtain norm convergence of the state vector associated with the rounded controls that are obtained in the second step of the CIA decomposition. Section 3 presents two algorithms that can be used in the second step of the CIA decomposition, i.e., both satisfy the prerequisites for the aforementioned convergence argument. The first is very resource efficient but not optimal with respect to the integrality gap. The second yields an optimal integrality gap and can be modified to incorporate additional combinatorial constraints on the control. Section 4 presents an algorithmic framework to perform the rounding step. Section 5 compares the two basic rounding algorithms computationally in terms of state vector and objective approximation error for an optimal control problem that is governed by an elliptic state equation on a two-dimensional domain. Finally, we summarize our findings in Sect. 6.

1.2 Notation

For an integer k, we use the abbreviating notation $[k] := \{1, \ldots, k\}$. For a Banach space X, we denote its topological dual by X^*. As we have done up to this point,

we use the abbreviated forms "a.e." and "for a.a." for "almost everywhere" and "for almost all," respectively.

2 Approximation Arguments for the CIA Decomposition

Independent of the actual rounding algorithms, this section summarizes the argument that a decaying integrality gap implies convergence of the control and state vectors. This later factors into the optimality and feasibility of the approximations. We begin by introducing required properties of rounding meshes and the output of the rounding algorithm. We continue by describing the convergence properties that result from these properties and show how they factor into optimality and feasibility. Finally, we point out the differences, i.e., our loss in approximation quality, if mixed constraints of the form $0 \leq \omega_i c_i(y)$ are present.

2.1 Properties of Rounding Meshes and Algorithms

The rounding algorithms presented later operate on controls discretized on meshes. We refer to these as rounding meshes.

Definition 2.1 (Rounding Mesh and Mesh Size) Let $d \in \mathbb{N}$. Let $\Omega_T \subset \mathbb{R}^d$ be a bounded domain. A set of mesh cells $\{\mathcal{T}_1, \dots, \mathcal{T}_N\} \subset \mathcal{B}(\Omega_T)$ is called a *rounding mesh* if the cells make up a finite partition of Ω_T. The quantity N denotes the number of mesh cells, and the quantity $h \coloneqq \max_{k\in[N]} \lambda(\mathcal{T}_k)$ denotes the *mesh size* of the rounding mesh.

We highlight again that, in contrast to PDE numerics literature, we have defined *mesh size* as the maximum cell volume and not as the maximum cell diameter of the mesh cells. Although these quantities are connected on the considered meshes, they are of course not equivalent.

The convergence results in this section require the following assumptions on the binary control vector ω produced during the rounding step. This will be justified for SUR in Sect. 3.1.

Assumption 2.2 There exists a constant $C > 0$ such that for all relaxed controls α and rounding meshes $\{\mathcal{T}_1, \dots, \mathcal{T}_N\}$ with mesh size h, the rounding ω satisfies

$$\max_{k\in[N]} \left\| \int_{\bigcup_{\ell=1}^{k} \mathcal{T}_\ell} \alpha(s) - \omega(s) \mathrm{d}s \right\|_\infty \leq Ch. \tag{2.1}$$

2.2 Weak Control Approximation

Assumption 2.2 implies convergence of ω to α in the weak-$*$ topology of $L^\infty(\Omega_T, \mathbb{R}^M)$ by means of a density argument. We refer to [14] for the proof. In case Ω_T is one dimensional, i.e., $\Omega_T = (0, T)$, this arises straightforwardly if the mesh cells are intervals.

Theorem 2.3 *Let $(\{\mathcal{T}_1^n, \dots, \mathcal{T}_{N_n}^n\})_n$ be a sequence of rounding meshes with the cells $\mathcal{T}_k^n$ being consecutive (closed, open, and half-closed) intervals for all $n \in \mathbb{N}$ and $k \in [N_n]$. Let $(h_n)_n$ denote the corresponding sequence of mesh sizes and $(\omega^n)_n$ the corresponding sequence of binary controls by a rounding algorithm satisfying Assumption 2.2. Then,*

$$\sup_{t\in[0,T]} \left\| \int_0^t \alpha(s) - \omega^n(s) \mathrm{d}s \right\|_\infty \leq Ch^n.$$

If $h^n \to 0$, we have

$$\omega^n \rightharpoonup^* \alpha \text{ in } L^\infty((0, T), \mathbb{R}^M).$$

The density argument to prove Theorem 2.3 makes use of the one-dimensional domain of integration, namely the integration by parts formula before Assumption 2.2, is applied. This procedure does not generalize to the multidimensional setting as there is no multidimensional analog to the forward progression along the single coordinate axis in one dimension. To overcome this, we impose a regularity condition on the refinement strategy of the sequence of rounding meshes to obtain weak-$*$ convergence of the sequence $(\omega^n)_n$.

Theorem 2.3 demonstrates that refining the meshes uniformly and satisfying a condition on the progression of the SUR algorithm through cells of consecutive meshes give the desired convergence. This condition is satisfied by space-filling curves, e.g., the Hilbert curve. For a short proof, we refer to [16].

Fortunately, it is possible to obtain the weak-$*$ approximation property independently of chosen progressions through the mesh cells, i.e., independent of the indexing of the mesh cells in the estimate (2.1). However, we still require a regularity condition to avoid a degeneration of the eccentricity of the mesh cells during the successive refinement of the rounding meshes. The regularity condition is given in Definition 2.4 below and is introduced in [15].

Definition 2.4 Let $d \in \mathbb{N}$ and $\Omega_T \subset \mathbb{R}^d$ be a bounded domain. Let $\left(\left\{\mathcal{T}_1^n, \dots, \mathcal{T}_{N^n}^n\right\}\right)_n$ be a sequence of rounding meshes with corresponding sequence of mesh sizes $(h^n)_n$. Then, we call the sequence $\left(\left\{\mathcal{T}_1^n, \dots, \mathcal{T}_{N^n}^n\right\}\right)_n$ an *admissible sequence of refined rounding meshes* if

1. $h^n \to 0$,

2. for all $n \in \mathbb{N}$ and all $k \in [N^{n+1}]$, there exists $\ell \in [N^n]$ such that $\mathcal{T}_k^{n+1} \subset \mathcal{T}_\ell^n$, and
3. the cells $\mathcal{T}_k^n$ *shrink regularly*, i.e., there exists $C > 0$ such that for each $\mathcal{T}_k^n$, there exists a ball B_k^n such that $\mathcal{T}_k^n \subset B_k^n$ and $\lambda(\mathcal{T}_k^n) \geq C\lambda(B_k^n)$.

As in [15], we note that the last condition, which limits the eccentricity of the cells along the refinements, is similar to requirements on finite-element triangulations, namely refining with an isotropic strategy on quasi-uniform triangulations, see [3]. We state the weak-$*$ convergence, which is proven in [15].

Theorem 2.5 *Let $d \in \mathbb{N}$ and $\Omega_T \subset \mathbb{R}^d$ be a bounded domain. Let $\left(\{\mathcal{T}_1^n, \ldots, \mathcal{T}_{N^n}^n\}\right)_n$ be an admissible sequence of refined rounding meshes and $(\omega^n)_n$ be the corresponding sequence of binary controls computed by means of a rounding algorithm satisfying Assumption 2.2. Then,*

$$\omega^n \rightharpoonup^* \alpha \text{ in } L^\infty(\Omega_T, \mathbb{R}^M).$$

2.3 State Vector Approximation

Let $y(\alpha)$ denote the solution of the state equation for the relaxed control α and $y(\omega^n)$ the solution of the state equation for the binary control ω^n. To obtain $y(\omega^n) \to y(\alpha)$ in the state space Y, we need compactness of the solution mapping to transform the weak-$*$ convergence into convergence in norm. We state two results. The first is for a class of semi-linear evolution equations with Lipschitz continuous nonlinear part and unbounded linear part, which generates a strongly continuous semigroup. It is proven in [14] and extends the results in [7].

Theorem 2.6 *Let X be a Banach space. Let $\alpha : [0, T] \to \mathbb{R}^M$ be a relaxed control. Let $y \in Y := C([0, T], X)$ solve*

$$\partial_t y + Ay = \sum_{i=1}^M \alpha_i f_i(y), \quad y(0) = y_0$$

with A being the generator of a strongly continuous semigroup on X and f_i being Lipschitz continuous with respect to y for $i \in [M]$. Let $(\omega^n)_n$ be a sequence of binary controls computed by means of a rounding algorithm satisfying Assumption 2.2 on a sequence of rounding meshes as demanded in Theorem 2.3 with $h^n \to 0$, and let $(y^n)_n \subset Y$ be the sequence of state vectors that solve

$$\partial_t y + Ay = \sum_{i=1}^M \omega_i^n f_i(y), \quad y(0) = y_0$$

for $n \in \mathbb{N}$. Then,

$$y^n \to y \text{ in } Y.$$

The second result is developed in [15] for PDEs governed by elliptic operators of second order, for which it follows immediately from the Lax–Milgram theorem.

Theorem 2.7 *Let X and Y be Banach spaces satisfying the dense and compact embedding $X \hookrightarrow^c Y$. Let $\alpha : \Omega_T \to \mathbb{R}^M$ be a relaxed control. Let $y \in Y$ be the solution of*

$$Ay = \sum_{i=1}^{M} \alpha_i f_i(y)$$

with the restriction A having a bounded inverse $A^{-1} : X^* \to X$. Let $(\omega^n)_n$ be a sequence of binary controls computed by means of a rounding algorithm satisfying Assumption 2.2 on an admissible sequence of refined rounding meshes, and let $(y^n)_n \subset X$ be the sequence of state vectors that solve

$$Ay = \sum_{i=1}^{M} \omega_i^n f_i(y)$$

for $n \in \mathbb{N}$. Let $\omega_i^n f_i(y) \rightharpoonup \alpha_i f(y^n)$ in Y^*. Then,

$$y^n \to y \text{ in } X.$$

One should have the Dirichlet–Laplacian with the Hilbert space setting $X = H_0^1(\Omega)$, $X^* = H^{-1}(\Omega)$ and $Y = Y^* = L^2(\Omega)$ in mind for this case. If the f_i do not depend on the state vector, the condition $\omega_i^n f_i(y^n) \rightharpoonup \alpha_i f(y^n)$ is trivially true in this case.

2.4 Optimality and Feasibility in the Absence of Mixed Constraints

Again, we denote the state space by the symbol Y. Regardless of the presence of the mixed constraint or not, we can deduce the following from continuity of the objective J with respect to the state vector.

Lemma 2.8 *Let* (y, α) *solve* (RC)*, and let* $(y^n, \omega^n)_n \subset Y \times L^\infty(\Omega_T)$ *satisfy* $y^n \to y$*. Then,*

$$\lim J(y^n) = \min_{(y,\alpha)\in\mathcal{F}_{(\mathrm{RC})}} J(y).$$

Now, assume that the mixed constraints c_i are not present, i.e., $c_i \equiv 0$ holds for all $i \in [M]$. Then, we even obtain the following theorem.

Theorem 2.9 *Let the prerequisites of Lemma 2.8 hold. Then,*

$$\min_{(y,\alpha)\in\mathcal{F}_{(\mathrm{RC})}} J(y) = \inf_{(y,\omega)\in\mathcal{F}_{(\mathrm{BC})}} J(y).$$

These statements are proven in [15] and guarantee algorithmic well-definedness and finite termination if we refine the rounding mesh successively in the sense of Definition 2.4 until an acceptable approximation error between the objective value of the current iterate and the optimal objective value of (RC) is reached.

2.5 Optimality and Feasibility in the Presence of Mixed Constraints

As mentioned before, in the presence of mixed constraints, we need to take some extra care, and unfortunately, the decomposition approach may not be able to produce a feasible point of (BC), in contrast to Theorem 2.9, but only one exhibiting an arbitrarily small constraint violation.

Applying a rounding algorithm in the presence of the constraints $0 \le \alpha_i c_i(y)$ without any modifications might lead to arbitrary low values of the term $\omega_i c_i(y)$. To see this, let $i \in [M]$ be fixed and remember that the functions c_i are assumed to be continuous. The problem arises from the bilinear structure of the constraint $0 \le \alpha_i c_i(y)$. If $\alpha_i = 0$ on a set of nonzero measure, the value of $c_i(y)$ may be arbitrarily low for $(y, \alpha) \in \mathcal{F}_{(\mathrm{RC})}$. If the algorithm does not prevent the rounding of ω_i^n to 1 on this particular set of nonzero measure, this may lead to an arbitrarily high violation of the constraint $0 \le \omega_i^n c_i(y^n)$ on this particular set.

To overcome this problem, the following assumption restricts the indices that are admissible for rounding in a particular cell $\mathcal{T}_k^n$ to the ones satisfying $\int_{\mathcal{T}_k^n} \alpha_i > 0$.

Assumption 2.10 For all relaxed controls α and rounding meshes $\{\mathcal{T}_1, \dots, \mathcal{T}_N\}$, the rounding ω satisfies

$$\int_{\mathcal{T}_k} \alpha_i = 0 \Rightarrow \int_{\mathcal{T}_k} \omega_i = 0$$

for all $k \in [N]$ and all $i \in [M]$.

Now, the continuity of the c_i and Assumption 2.10 yield the following result.

Theorem 2.11 *Let the prerequisites of Lemma 2.8 hold. Let the binary controls $(\omega^n)_n$ be computed by means of a rounding algorithm that satisfies Assumption 2.10. Then,*

$$\lim J(y^n) = \min_{(y,\alpha)\in\mathcal{F}_{(\mathrm{RC})}} J(y)$$

as well as

$$0 \leq \liminf \omega_i^n c_i(y^n) \text{ for all } i \in [M].$$

Note that this asymptotic feasibility of the constraints in particular holds for the special case of continuous path constraints. Further classes of constraints on states and controls are discussed in [21].

3 Approximation Quality of Roundings

The rounding step of the CIA decomposition can be performed using different algorithmic approaches. Section 3.1 focuses on variants of the SUR algorithm, while the explicit minimization of the integrality gap using mixed-integer linear programs (MILPs) is the subject of Sect. 3.2. Note that other approaches like Next-Forced Rounding, see [11], exist for the second step of the CIA decomposition.

3.1 Sum-up Rounding Algorithms

We introduce two variants of the SUR algorithm, see [13, 20], below and discuss their basic properties and the difference between them.

Definition 3.1 (SUR Algorithms) Let α be a relaxed control, and let $\{\mathcal{T}_1, \ldots, \mathcal{T}_N\}$ be a rounding mesh. We define the function ω iteratively for $k = 1, \ldots, N$ as

$$\omega(s) := \sum_{k=1}^{N} \chi_{\mathcal{T}_k}(s) W_k,$$

$$W_k(i) := \begin{cases} 1 \text{ if } i = \arg\max\limits_{j\in F_k} \int_{\mathcal{T}_k} \alpha_j - \int_{\bigcup_{\ell=1}^{k-1}\mathcal{T}_\ell} \alpha_j - \omega_j, \\ 0 \text{ else} \end{cases} \quad \text{for } i \in [M].$$

If a tie arises with respect to the maximizing index k, the smallest of the maximizing indices is chosen. We define the two variants, which differ in the sets of *admissible* indices for rounding in the cells of the rounding mesh:

$$F_k \coloneqq \{1, \dots, M\} \text{ for all } k \in [N], \tag{SUR}$$

$$F_k \coloneqq \left\{ i \in [M] : \int_{\mathcal{T}_k} \alpha_i > 0 \right\} \text{ for all } k \in [N]. \tag{SUR-VC}$$

The algorithm (SUR) is the original SUR algorithm introduced in [20], and the algorithm (SUR-VC) is a variant introduced in [13] that works properly in the presence of mixed constraints. We restate the approximation property that establishes Assumption 2.2 below. It is proven in [13, 22] for (SUR) and in [13, 17] for (SUR-VC).

Proposition 3.2 *The algorithms* (SUR) *and* (SUR-VC) *produce binary controls ω for all relaxed controls α and rounding meshes. There exists a constant $C > 0$ such that for a relaxed control α and ω computed by means of* (SUR) *or* (SUR-VC) *on a rounding mesh with mesh size h, we have the estimate*

$$\max_{k \in [N]} \left\| \int_{\bigcup_{\ell=1}^{k} \mathcal{T}_\ell} \alpha(s) - \omega(s) \mathrm{d}s \right\|_\infty \leq Ch.$$

In particular, Assumption 2.2 holds true.

Due to the integration domain being an increasing union of rounding mesh cells, this estimate depends on the ordering of the mesh cells. However, if the sequence of mesh cells is constructed such that Definition 2.4 is satisfied, the reasoning in Sect. 2.2 guarantees convergence.

Example We illustrate the necessity for making the rounding algorithm aware of the mixed constraints, see Sect. 2.5, for the algorithm (SUR). Let $M = 3$ and $\Omega_T = (0, 2)$, and let α be the relaxed control given by

$$\alpha_1 \coloneqq .5\chi_{[0,2]}, \quad \alpha_2 \coloneqq .5\chi_{[0,1]}, \quad \alpha_3 \coloneqq .5\chi_{[1,2]}.$$

Assume that, in mesh iteration n, Ω_T is discretized into $N_n = 2 \cdot 3^n$ equidistant intervals, i.e., $h_n = 3^{-n}$. By applying (SUR), we obtain $\omega_1^n(s) = 1$ on the intervals with odd indices and $\omega_2^n(s) = 1$ on the intervals with even indices. This implies

$$\begin{aligned}
\textstyle\int_0^1 \alpha_1 - \omega_1^n &= \textstyle\int_{\bigcup_{k=1}^{3^n} \mathcal{T}_k^n} \alpha_1 - \omega_1^n = -0.5 \cdot 3^{-n}, \\
\textstyle\int_0^1 \alpha_2 - \omega_2^n &= \textstyle\int_{\bigcup_{k=1}^{3^n} \mathcal{T}_k^n} \alpha_2 - \omega_2^n = 0.5 \cdot 3^{-n}, \\
\textstyle\int_0^1 \alpha_3 - \omega_3^n &= \textstyle\int_{\bigcup_{k=1}^{3^n} \mathcal{T}_k^n} \alpha_3 - \omega_3^n = 0.
\end{aligned}$$

Thus, for the $3^k + 1$-th interval, we have

$$\begin{aligned}
\int_{\mathcal{T}^n_{3^n+1}} \alpha_1 + \int_{\bigcup_{k=1}^{3^n} \mathcal{T}^n_k} \alpha_1 - \omega_1^n &= 0., \\
\int_{\mathcal{T}^n_{3^n+1}} \alpha_2 + \int_{\bigcup_{k=1}^{3^n} \mathcal{T}^n_k} \alpha_2 - \omega_2^n &= 0.5 \cdot 3^{-n} \\
\int_{\mathcal{T}^n_{3^n+1}} \alpha_3 + \int_{\bigcup_{k=1}^{3^n} \mathcal{T}^n_k} \alpha_3 - \omega_3^n &= 0.5 \cdot 3^{-n},
\end{aligned}$$

and (SUR) gives $\omega_2^n = 1$ on the interval $[1, 1 + h_n]$. Thus, $\|\omega_2^n|_{[1,2]}\|_{L^\infty} = 1$ for all $n \in \mathbb{N}$. Now, assume $c_2(y^n) \to c_2(y)$ and $c_2(y) \equiv -1$ on $[1, 2]$. Then,

$$\operatorname{ess\,inf} \omega_2^n c_2(y^n) \to -1 \text{ on } [1, 2].$$

The restriction of the set of admissible indices for rounding, F_k for $k \in [N]$, in the definition of (SUR-VC) ensures that Assumption 2.10 is satisfied as well, and the problem illustrated above cannot occur, see [13].

Proposition 3.3 *Algorithm* (SUR-VC) *satisfies Assumption 2.10.*

We note that a similar modification is not possible for the algorithm Next-Forced Rounding (NFR) from [11] mentioned above as this may lead to an empty set of indices admissible for rounding.

3.2 Combinatorial Integral Approximation Problems

In this subsection, we discuss the minimization problem

$$\min_{\omega} \max_{k \in [N]} \left\| \int_{\bigcup_{\ell=1}^{k} \mathcal{T}_\ell} \alpha(s) - \omega(s) \mathrm{d}s \right\|_\infty,$$

which defines binary controls ω that minimize the integrality gap. By introducing an additional variable $\theta \geq 0$ and adding inequality constraints for all control realizations and mesh cells, we are able to define an equivalent mixed-integer linear program (MILP) that aims at solving the above problem. We refer to the latter as *Combinatorial Integral Approximation Problem*, see [23], and provide its definition below.

Definition 3.4 (CIA-MILP) Let the prerequisites of Definition 3.1 hold. Based on the relaxed controls and the rounding mesh, we introduce the average values

$$A_k(i) := \frac{1}{\lambda(\mathcal{T}_k)} \int_{\mathcal{T}_k} \alpha_i(s) \mathrm{d}s, \qquad \text{for } i \in [M],\ k \in [N].$$

We define further the CIA-MILP to be

$$\min_{\theta, W} \theta \text{ s.t.} \qquad \text{(CIA-MILP)}$$

$$\begin{aligned} \theta &\geq \pm \sum_{l \in [k]} (A_l(i) - W_l(i))\, \lambda(\mathcal{T}_l), && \text{for } i \in [M],\ k \in [N], \\ W_k(i) &\in \{0, 1\} && \text{for } i \in [M],\ k \in [N], \\ 1 &= \sum_{i \in [M]} W_k(i) && \text{for } k \in [N]. \end{aligned}$$

The solution of (CIA-MILP) is used to construct a piecewise constant binary control function as already sketched in Definition 3.1:

$$\omega(s) \coloneqq \sum_{k=1}^{N} \chi_{\mathcal{T}_k}(s) W_k, \quad s \in \Omega_T.$$

We note that the family of SUR algorithms has linear complexity in the total number of mesh cells N. In contrast, using an MILP in the rounding step increases the computational burden exponentially with N but may construct solutions with smaller integrality gap. In fact, one can interpret SUR as a heuristic way to solve (CIA-MILP) or at least construct a feasible point. Since (SUR) provides a feasible point for (CIA-MILP), the following proposition, which asserts Assumption 2.2, follows directly from Proposition 3.2.

Proposition 3.5 *The solution of* (CIA-MILP) *yields a binary control ω for all relaxed controls α and rounding meshes. There exists a constant $C > 0$ such that for α being a relaxed control and ω being computed by solving* (CIA-MILP) *on a mesh with mesh size h, we have the estimate*

$$\max_{k \in \{1, \ldots, N\}} \left\| \int_{\bigcup_{\ell=1}^{k} \mathcal{T}_\ell} \alpha(s) - \omega(s) \mathrm{d}s \right\|_\infty \leq Ch.$$

In particular, Assumption 2.2 holds true.

(CIA-MILP) represents the CIA problem based on the ∞-norm, whereas there is a whole family of MILPs to carry out the binary approximation problem. A generalization of CIA problems with respect to different norms, the order of the accumulated control difference and different scaling of the latter, is proposed in [25]. For instance, we may scale the approximation inequality for the CIA problem with the evaluated right-hand side f_i after solving (RC).

Another aspect of using an MILP in the rounding step is the opportunity to include general combinatorial constraints on the binary controls. Real-world problems on a time domain, i.e., $\Omega_T \subset \mathbb{R}$, see e.g., [4, 19], often require a limited number of switches occurring between the system modes or the presence of so-called *minimum dwell time constraints* that describe the necessity of activating a

control ω_i for at least a given minimal duration if at all. Similar constraints can be introduced for deactivation periods. To impose a maximum number of switches $\sigma \in \mathbb{N}$ on the time horizon, we would add

$$\sigma \geq \frac{1}{2} \sum_{i \in [M]} \sum_{l \in [N-1]} |W_{l+1}(i) - W_l(i)| \tag{3.1}$$

to (CIA-MILP). The dwell time constraints for a given dwell time $C_D \in \mathbb{N}$, an assumed equidistant mesh, as well as $l \in [N-2]$, $k = l+1, \ldots, \min\{l+1+C_D, N\}$ would read

$$\begin{aligned} W_{k+1}(i) &\geq W_{l+1}(i) - W_l(i), \quad \text{for } i \in [M], \\ 1 - W_{k+1}(i) &\geq W_l(i) - W_{l+1}(i), \quad \text{for } i \in [M], \end{aligned}$$

and can also be addressed by (CIA-MILP). In contrast to the one-dimensional case, it is not immediately clear how to interpret such constraints on multidimensional domains. Here, the *total max-up constraint* is an example of a meaningful combinatorial condition, which limits the total number of activations on all mesh cells for certain controls by a constant $C_L(i) \in \mathbb{N}$:

$$C_L(i) \geq \sum_{l \in [N]} W_l(i), \quad \text{for } i \in [M].$$

Combinatorial conditions have in common that Assumption 2.2 cannot generally be satisfied in their presence, and hence the convergence argument in Sect. 2.2 may fail. The following example illustrates this issue.

Example Let us again consider the case $\Omega_T = [0, 2]$ with two discrete control realizations, i.e., $M = 2$, and with the presence of the constraint (3.1) that limits the number of switches with the choice $\sigma = 1$. We further assume that the relaxed control is given by

$$\alpha_1 := .5\chi_{[0,2]}, \quad \alpha_2 := .5\chi_{[0,2]}.$$

Then, we recognize that the optimal solution of (CIA-MILP) approximates α by setting the values $W_l(1) = 1$ on a minimal set covering $\cup_l \mathcal{T}_l$ of $[0, 1]$ and $W_l(1) = 0$ else. Therefore, (CIA-MILP) exhibits an objective, i.e., an integrality gap, of at least $\frac{1}{2}$ *independent* of the discretization of Ω_T. In particular, Assumption 2.2 is not satisfied.

This example can be adapted analogously to cases where $\sigma > 1$ is given or $M > 2$ holds.

4 Solving the CIA Problem

The open-source software package `pycombina`[1] contains an implementation for various rounding algorithms, e.g., for the presented SUR from Sect. 3. Sophisticated MILP solvers such as `Gurobi` struggle to solve (CIA-MILP) efficiently, see [11]. This may be due to the fact that its canonical linear programming relaxation, i.e., (CIA-MILP) with $W_k(i) \in [0, 1]$, yields only trivial lower bounds in case of absent additional combinatorial constraints. (CIA-MILP) can be solved more efficiently by means of a tailored Branch and Bound scheme, see [23]; an efficient version is also implemented in `pycombina`. Algorithm 1 describes the main steps. The algorithm exploits that an evaluation of the objective function up to the current mesh cell yields a valid lower bound due to the maximization operator over all intermediate steps in the objective function. This lower bound is extremely cheap to compute and is tighter than canonical relaxations [12]. We select nodes from a queue Q until it is empty or a termination criterion is reached, such as a maximum number of iterations or a time limit (line 2). The selected node `n` is pruned if its lower bound θ is greater than the global upper bound UB (lines 4–5) or we update the currently best node $\mathtt{n}^*$ to be `n`, if its depth equals the number of mesh cells N (lines 6–7). We branch forward with respect to the mesh index $k \in [N]$, whereby for each child node creation, all control entries $W_k(i)$ become fixed with exactly one index set to be active (line 9). Nodes contain information on their depth, which is the mesh cell index, their so far largest accumulated control deviation θ and the accumulated deviation for each control realization θ_i. Depending on the imposed combinatorial constraints, we save also information about previous $W_k(i)$ values in the nodes and add their child nodes only if they satisfy these constraints (line 10). For further details and numerical examples benchmarking Algorithm 1 with MILP solvers, we refer to [4, 11].

5 Illustration of the Multidimensional Control Approximation

As noted in Sect. 2.2, weak convergence of the control function can be ensured for elliptic PDEs with both algorithms, if we use an admissible sequence of refined rounding meshes. As shown in [16], this can be achieved by iterating over the mesh cells along approximants of a space-filling curve such as the Hilbert curve. In this section, we demonstrate the bare SUR algorithm and the MILP approach described above by applying them to a simple distributed inverse problem for the Poisson equation. We use a finite-element method with continuous first-order Lagrange

[1] Available at https://github.com/adbuerger/pycombina.

Algorithm 1: Branch and Bound for solving (CIA-MILP)

Input : Relaxed control values $A_k(i)$, mesh size volumes $\lambda(\mathcal{T}_k)$, $k \in [N]$, termination criterion, parameters for combinatorial constraints.

Output: (Optimal) solution (θ^*, W^*) of (CIA-MILP).

Initialize node queue Q with empty node and set upper bound UB.
while $Q \neq \emptyset$ *and termination criterion not reached* **do**
  Choose $\mathrm{n} \in Q$ according to node selection strategy.
  if $n.\theta > UB$ **then**
    Prune node n.
  else if $n.depth = N$ **then**
    Set new best node $\mathrm{n}^* \leftarrow \mathrm{n}$ and $UB = \mathrm{n}.\theta$
  else
    Create M child nodes c_i with

$$\begin{aligned} \mathrm{c}_i.depth &\leftarrow d := \mathrm{n}.depth + 1, \\ W_{\mathrm{c}_i,d}(j) &\leftarrow \begin{cases} 1 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases}, \\ \mathrm{c}_i.\theta_j &\leftarrow \mathrm{n}.\theta_j + (A_d(j) - W_d(j)) \cdot \lambda(\mathcal{T}_d) \\ \mathrm{c}_i.\theta &\leftarrow \max\left(\{\mathrm{n}.\theta\} \cup \left\{|\mathrm{c}_i.\theta_j| \;\middle|\; j \in [M]\right\}\right). \end{aligned}$$

    Add c_i to Q if and only if it satisfies all combinatorial constraints.
  end
end
return: $(\theta^*, W^*) = (\mathrm{n}^*.\theta, \mathrm{n}^*.W)$;

elements on a structured triangular mesh which we will iterate over according to the Sierpinski curve.

5.1 Test Problem

Our test problem is based on the Poisson equation, which is an inhomogeneous, uniformly elliptic second-order linear PDE system used to find stationary solutions to diffusion and heating problems. Due to its theoretical simplicity, the Poisson equation is often used as a test bed for mixed-integer PDE-constrained optimization. We solve the Poisson equation in two dimensions on the unit square $\Omega = [0, 1]^2$ using Robin boundary conditions, which guarantees the uniqueness and Fréchet differentiability of the PDE solution with respect to our controls, which select one

of the five discrete source term values for each point in the domain. Our objective is an L^2 tracking objective. Thus, the problem can be stated as

$$\begin{aligned}
&\min_{y,\omega} \|y - \bar{y}\|^2_{L^2(\Omega)} \\
&\text{s.t.} \qquad -\Delta y = \sum_{i=1}^{5} v_i \omega_i \qquad \text{a.e. in } \Omega, \\
&\qquad \frac{\partial y}{\partial \nu} - y = 0 \qquad \text{a.e. in } \partial\Omega, \\
&\qquad \sum_{i=1}^{5} \omega_i(x) = 1 \qquad \text{a.e. in } \Omega, \\
&\qquad \omega_i(x) \in \{0, 1\} \qquad \text{a.e. in } \Omega \ \forall i \in [5],
\end{aligned} \tag{P}$$

where $\nu \colon \partial\Omega \to \mathbb{R}^2$ is the outer unit normal of Ω and $\bar{y} \in L^2(\Omega)$ is the unique weak solution of the boundary value problem for the right-hand side given by

$$\bar{f}(x) \coloneqq \sum_{i=1}^{5} v_i \frac{\bar{\alpha}_i(x)}{\sum_{j=1}^{5} \bar{\alpha}_j(x)} \qquad \forall x \in \Omega$$

with a set of known control functions

$$\bar{\alpha}_i(x) \coloneqq \exp\left(-100\big(\min\left\{\|x - m_{*,1}\|, \|x - m_{*,2}\|\right\} - r_i\big)^2\right).$$

The additional parameters are

$$\begin{aligned}
v &\coloneqq \left(-2, -\frac{1}{2}, \frac{1}{4}, 1, 2\right)^T, \\
r &\coloneqq (0.25, 0.2, 0.15, 0.1, 0.05)^T, \\
m &\coloneqq \begin{pmatrix} 0.25 & 0.75 \\ 0.25 & 0.75 \end{pmatrix}.
\end{aligned}$$

After normalization, the functions $\bar{\alpha}$ sum up to one everywhere. Therefore, they are optimal controls for the relaxed problem with objective function value 0.

5.2 Mesh Structure and Sierpinski Curve

We use the finite-element package `FEniCS` [2] to generate meshes and solve the boundary value problem. Meshes are generated using a `RectangleMesh` with `cross- ed` diagonals, meaning that at refinement level $l \in \mathbb{N}_0$, the unit square is subdivided into 4^l equally sized squares, each of which is again subdivided into four congruent triangles along its diagonals. This is equivalent to subdividing each triangle into four congruent sub-triangles on each refinement level as illustrated in Fig. 1.

In order to generate an order approximating the Sierpinski curve, we generate the vertices of a Sierpinski curve at the l-th iteration, starting at the point $\left(\frac{1}{2^{l+1}}, \frac{\sqrt{2}-1}{2^{l+1}}\right)$ which is located in the leftmost triangle that has an edge contained entirely within the x_1 axis. The first step is made at an angle of $\frac{\pi}{4}$ and all steps have length $\frac{\sqrt{2}-1}{2^l}$. This produces one vertex within each triangle. We then iterate over the triangles in the mesh according to the order of the vertices.

For a more detailed description of the Sierpinski curve, we refer to [5, Section 2.10.3]. The procedure is illustrated for refinement levels 0, 1, and 2 in Fig. 2.

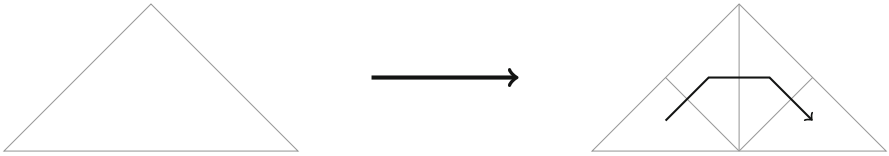

Fig. 1 Refinement of a single triangle

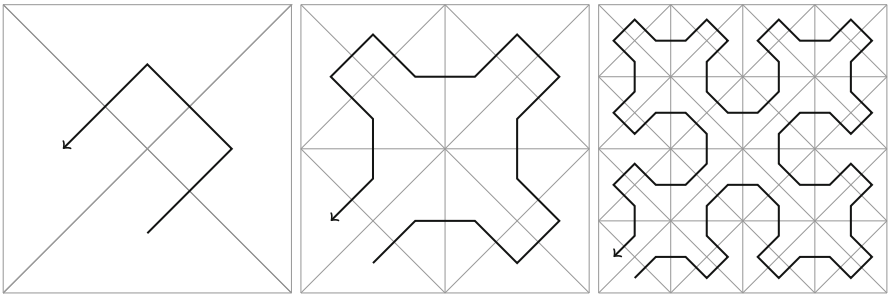

Fig. 2 First three refinement levels in an admissible sequence of rounding meshes using the Sierpinski curve

5.3 Numerical Results Obtained with the CIA Decomposition

For practical problems, we suggest to calculate optimal relaxed and derived binary controls iteratively on refined meshes. However, both the convergence of the relaxed solutions and of the rounding strategies have an impact and overlap, complicating the analysis of the overall convergence behavior. In our setting and due to the way the test problem is stated, the optimal relaxed control function is known in advance. This allows us to highlight the convergence of the rounded solutions to the optimal relaxed solution in function space. We use continuous first-order Lagrange elements to approximate weak PDE solutions and piecewise constant functions to approximate control functions. We coarsen the optimal relaxed control for lower refinement levels by taking a weighted average over each cell of the coarse mesh and approximate it using both sum-up rounding and `pycombina`'s specialized branch-and-bound algorithm. The latter is limited to 10^8 explored nodes and up to one CPU hour of computation time. We compare both approximation methods using the absolute error in the objective function value as well as the objective they achieve in the CIA problem (CIA-MILP). The latter approaching zero indicates weak-$*$ convergence of the control function.

We note that `pycombina` terminates early on account of exceeding the explored node limit for levels 3, 4, 5, and 6. However, it does so in less than 20 CPU minutes in all cases. By contrast, if we try to solve the CIA problem (CIA-MILP) using Gurobi, the CPU time limit of one hour is already exceeded at level 3.

Table 1 summarizes the outcome of our experiment. Despite early and possibly suboptimal termination, we see that the branch-and-bound algorithm always achieves a CIA objective that is at least as good as or better than that achieved by sum-up rounding, though this does not always translate into a smaller error in the actual objective function value. For levels 1, 3, and 5, we plot the right-hand side function and PDE solution for sum-up rounding and branch-and-bound alongside their relaxed counterparts in Figs. 3 and 4, respectively.

6 Conclusion

In this chapter, we surveyed recent improvements of the CIA decomposition for solving PDE-constrained mixed-integer optimal control problems. This approach consists of solving first the problem with relaxed controls before approximating these values with binary ones as part of a rounding problem. We summarized our findings with respect to convergence results in the weak-$*$ topology of L^∞ and discussed two rounding algorithms together with their efficient numerical implementation. Finally, these two algorithmic approaches were compared on a test problem based on the Poisson equation, where we used the space-filling Sierpinski curve to iterate over a structured triangular mesh.

Table 1 Results of numerical experiments

Level	Cells	h	Abs. Err. SUR	Abs. Err. BnB	CIA Obj. SUR	CIA Obj. BnB
0	4	2.500000×10^{-1}	1.825637×10^{-3}	1.825637×10^{-3}	1.487897×10^{-1}	1.487897×10^{-1}
1	16	6.250000×10^{-2}	8.382177×10^{-4}	1.734637×10^{-4}	4.562038×10^{-2}	4.562038×10^{-2}
2	64	1.562500×10^{-2}	1.110449×10^{-5}	6.927478×10^{-6}	9.355154×10^{-3}	9.355154×10^{-3}
3	256	3.906250×10^{-3}	7.461304×10^{-6}	7.232266×10^{-6}	3.395206×10^{-3}	2.910952×10^{-3}
4	1024	9.765625×10^{-4}	2.725262×10^{-7}	3.082747×10^{-7}	8.505270×10^{-4}	7.388801×10^{-4}
5	4096	2.441406×10^{-4}	2.005071×10^{-8}	1.848401×10^{-8}	2.377537×10^{-4}	2.053501×10^{-4}
6	16,384	6.103516×10^{-5}	2.574303×10^{-9}	4.133702×10^{-9}	7.280519×10^{-5}	7.280519×10^{-5}

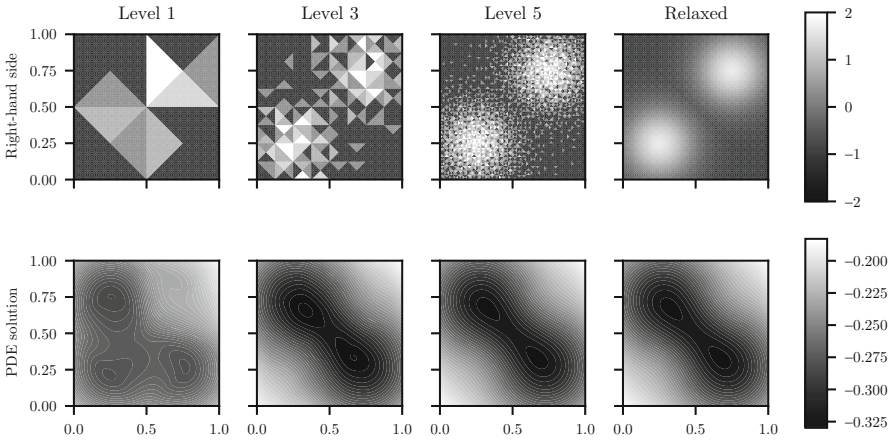

Fig. 3 Solutions for SUR at levels 1, 3, and 5

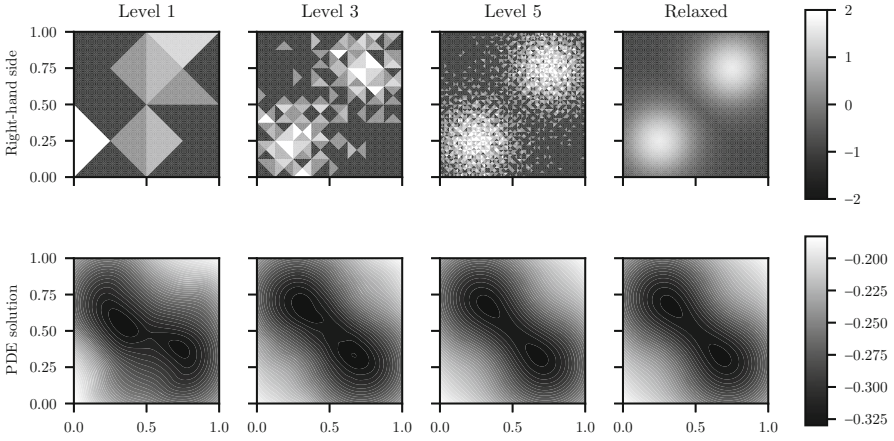

Fig. 4 Solutions for branch-and-bound at levels 1, 3, and 5

Acknowledgments C. Kirches, S. Sager, and P. Manns acknowledge funding by Deutsche Forschungsgemeinschaft through Priority Programme 1962, grant KI1839/1-1. C. Kirches acknowledges financial support by the German Federal Ministry of Education and Research, program "Mathematics for Innovations in Industry and Service," grants 05M17MBA-MOPhaPro and 05M18MBA-MOReNet, and program "IKT 2020: Software Engineering," grant 01/S17089C-ODINE. S. Sager, M. Hahn, and C. Zeile have received funding from the European Research Council (ERC), grant agreement no. 647573, from German Research Foundation—314838170, GRK 2297 MathCoRe and from German Federal Ministry of Education and Research, program "Mathematics for Innovations," grant P2Chem.

References

1. W. Achtziger and C. Kanzow. Mathematical programs with vanishing constraints: optimality conditions and constraint qualifications. *Mathematical Programming*, 114(1):69–99, 2008.
2. M. S. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, and G. N. Wells. The FEniCS Project Version 1.5. *Archive of Numerical Software*, 3(100), 2015.
3. T. Apel and G. Lube. Anisotropic mesh refinement in stabilized Galerkin methods. *Numerische Mathematik*, 74(3):261–282, 1996.
4. A. Buerger, C. Zeile, A. Altmann-Dieses, S. Sager, and M. Diehl. Design, implementation and simulation of an MPC algorithm for switched nonlinear systems under combinatorial constraints. *Journal of Process Control*, 81:15–30, September 2019.
5. H. Martyn Cundy and A. P. Rollett. *Mathematical models / by H. Martyn Cundy and A. P. Rollett*. Clarendon Press Oxford, 2d ed. edition, 1961.
6. A. Filippov. On some problems of optimal control theory. *Vestnik Moskowskovo Universiteta, Math*, 2:25–32, 1958. English version: On Certain Questions in the Theory of Optimal Control, J. SIAM Ser. A Control, Vol. 1 (1962), no. (1).
7. F. M. Hante and S. Sager. Relaxation methods for mixed-integer optimal control of partial differential equations. *Computational Optimization and Applications*, 55(1):197–225, 2013.
8. T. Hoheisel and C. Kanzow. First-and second-order optimality conditions for mathematical programs with vanishing constraints. *Applications of Mathematics*, 52(6):495–514, 2007.
9. T. Hoheisel and C. Kanzow. On the Abadie and Guignard constraint qualifications for mathematical programmes with vanishing constraints. *Optimization*, 58(4):431–448, 2009.
10. A. F. Izmailov and M. V. Solodov. Mathematical programs with vanishing constraints: optimality conditions, sensitivity, and a relaxation method. *Journal of Optimization Theory and Applications*, 142(3):501–532, 2009.
11. M. Jung. *Relaxations and approximations for mixed-integer optimal control*. PhD thesis, Heidelberg University, 2013.
12. M. Jung, G. Reinelt, and S. Sager. The Lagrangian Relaxation for the Combinatorial Integral Approximation Problem. *Optimization Methods and Software*, 30(1):54–80, 2015.
13. C. Kirches, F. Lenders, and P. Manns. Approximation properties and tight bounds for constrained mixed-integer optimal control. *SIAM Journal on Control and Optimization*, 58(3):1371–1402, 2020.
14. P. Manns and C. Kirches. Improved regularity assumptions for partial outer convexification of mixed-integer PDE-constrained optimization problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 26:32, 2020.
15. P. Manns and C. Kirches. Multidimensional sum-up rounding for elliptic control systems. *SIAM Journal on Numerical Analysis*, 58(6):3427–3447, 2020.
16. P. Manns, and C. Kirches. Multi-dimensional Sum-Up Rounding using Hilbert curve iterates. *PAMM*, 19(1):e201900065, 2019.
17. P. Manns, C. Kirches, and F. Lenders. Approximation properties of sum-up rounding in the presence of vanishing constraints. *Mathematics of Computation*, 90(329):1263–1296, 2021.
18. S. Peitz and S. Klus. Koopman operator-based model reduction for switched-system control of PDEs. *Automatica*, 106:184–191, 2019.
19. N. Robuschi, C. Zeile, S. Sager, F. Braghin, and F. Cheli. Multiphase mixed-integer nonlinear optimal control of hybrid electric vehicles. *Automatica*, 123:109325, 2021.
20. S. Sager. *Numerical methods for mixed-integer optimal control problems*. Der andere Verlag, Tönning, Lübeck, Marburg, 2005.
21. S. Sager. Reformulations and Algorithms for the Optimization of Switching Decisions in Nonlinear Optimal Control. *Journal of Process Control*, 19(8):1238–1247, 2009.
22. S. Sager, H.G. Bock, and M. Diehl. The Integer Approximation Error in Mixed-Integer Optimal Control. *Mathematical Programming, Series A*, 133(1–2):1–23, 2012.

23. S. Sager, M. Jung, and C. Kirches. Combinatorial Integral Approximation. *Mathematical Methods of Operations Research*, 73(3):363–380, 2011.
24. T. Ważewski. On an optimal control problem. In *Differential Equations and Their Applications*, pages 229–242. Publishing House of the Czechoslovak Academy of Sciences, 1963.
25. C. Zeile, T. Weber, and S. Sager. Combinatorial integral approximation decompositions for mixed-integer optimal control. Preprint 6472, Optimization Online, February 2018. http://www.optimization-online.org/DB_HTML/2018/02/6472.html.

Strong Stationarity for Optimal Control of Variational Inequalities of the Second Kind

Constantin Christof, Christian Meyer, Ben Schweizer, and Stefan Turek

Abstract This chapter is concerned with necessary optimality conditions for optimal control problems governed by variational inequalities of the second kind. The so-called strong stationarity conditions are derived in an abstract framework. Strong stationarity conditions are regarded as the most rigorous ones, since they imply all other types of stationarity concepts and are equivalent to purely primal optimality conditions. The abstract framework is afterward applied to four application-driven examples.

Keywords Optimal control of variational inequalities · Sensitivity analysis · Strong stationarity

Mathematics Subject Classification (2020) 49K27, 35J86, 49J40, 90C31, 49J27

1 Introduction

This chapter is concerned with optimal control problems governed by variational inequalities (VIs) of the second kind. Optimal control problems of this type arise in various applications, for instance, in the optimization of elastoplastic deformation processes, type-II semiconductors, or rheological fluids, see [7, 13, 29].

Optimal control problems governed by VIs provide the particular challenge that the control-to-state mapping, i.e., the solution mapping of the VI under consideration, is frequently not Gâteaux-differentiable. Therefore, the standard

C. Christof
Fakultät für Mathematik, TU München, Garching bei München, Germany
e-mail: christof@ma.tum.de

C. Meyer (✉) · B. Schweizer · S. Turek
Fakultät für Mathematik, TU Dortmund, Dortmund, Germany
e-mail: cmeyer@math.tu-dortmund.de; ben.schweizer@math.tu-dortmund.de; stefan.turek@math.tu-dortmund.de

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_12

adjoint approach for the derivation of Karush–Kuhn–Tucker (KKT) conditions that is widely used in optimal control is not applicable when it comes to optimization problems constrained by VIs. For this reason, the derivation of qualified optimality systems involving dual variables is all but elementary in the context of optimal control of VIs. There are multiple strategies to overcome this issue, among them are various smoothing and exact penalization techniques. We only refer to [8] and the various references therein for a broad overview. All these approaches yield necessary optimality conditions of different strength; see [15] for a survey of the multiple stationarity concepts. The most rigorous notion of stationarity is called *strong stationarity*. The characteristic feature of a system of strong stationarity is that it implies all the other stationarity conditions and is moreover equivalent to a purely primal optimality condition, called Bouligand(B)-stationarity.

To the best of our knowledge, there are in principle two ways to establish strong stationarity conditions for optimal control problems governed by VIs. Both approaches presume that the control-to-state map is at least directionally differentiable. The first approach was initiated by Mignot [19] and is to some extent based on the idea to take the "linearized" states as variations and not feasible controls as usually done in the derivation of first-order conditions. The alternative approach for the derivation of strong stationarity conditions originates from finite dimensional programs with complementarity constraints and works as follows: first, one "linearizes" the optimal control problem the control-to-state map by means of the directional derivative of S. Then, one defines auxiliary problems by fixing variables in the "linearized" problem. The latter are standard optimal control problems, which, under suitable assumptions, allow the derivation of KKT conditions. The latter then imply the desired strong stationarity conditions. For details on this approach, we refer to [14, 26, 27].

Here, we follow the first approach of [19] and generalize it for the optimal control of VIs of the second kind. Mignot's approach has mostly been applied to optimal control of VIs of the first kind, see, e.g., [1, 16, 20]. However, it turns out that the method of proof is essentially based on a particular structure of the directional derivative of the control-to-state mapping, which is also frequently observed in case of VIs of the second kind. Therefore, by slightly generalizing Mignot's approach, we construct a general framework for the derivation of strong stationarity conditions. We then apply this general result to four application-driven problems. First, we show that the obstacle problem fits into our general framework, which allows us to deduce the classical results by Mignot. As a second example, we consider the optimal control of static elastoplasticity in primal formulation, which is a VI of the second kind. As in case of the obstacle problem, the control-to-state map of the VI of static elastoplasticity provides a directional derivative with the desired structure without any further assumptions. This differs from our last two examples, which cover VIs of the second kind in the Sobolev space $H^1(\Omega)$ involving L^1-norms. Here, we need additional assumptions that ensure the existence of directional derivatives with certain properties in order to apply our general framework. In case of the so-called (generalized) lasso problem, these assumptions can be directly verified once a solution of the VI is given. In contrast to this, in our last example, which stems

from an application in rheological fluid mechanics, these assumptions are of rather intrinsic nature and may be hard to verify in practice.

It should be mentioned that this chapter is based on the PhD thesis of Constantin Christof, which was written within project P16 of the DFG priority program 1962.

Notation
The dual of a linear normed space X is denoted by X^*. If $x \in X$ and $g \in X^*$, we write for the dual pairing $g(x) = \langle g, x \rangle_X$. In case that the context is clear, we sometimes neglect the index and simply write $\langle \cdot, \cdot \rangle$ for a dual pairing. If X is a Hilbert space, we denote the corresponding scalar product by $(\cdot, \cdot)_X$. The space of linear and bounded operators from X to another linear normed space Y is denoted by $\mathcal{L}(X, Y)$. If X is continuously embedded in Y in the sense of [22, Definition 4.19], then we write $X \hookrightarrow Y$. If this embedding is dense, we write $X \hookrightarrow^d Y$.

2 Strong Stationarity in an Abstract Framework

Throughout this section, we consider the following abstract optimal control problem:

$$\left.\begin{aligned} \min \quad & J(y, u) \\ \text{s.t.} \quad & (y, u) \in Y \times U, \\ & y = S(u), \quad u \in U_{\text{ad}}. \end{aligned}\right\} \tag{P}$$

On the data in (P), we impose the following:

Assumption 2.1 (Standing Assumptions) *For the data in (P), we assume the following:*

- *U is a Hilbert space,*
- *Y is a linear normed space,*
- *the objective $J : Y \times U \to \mathbb{R}$ is Fréchet-differentiable,*
- *$S : U \to Y$ is a continuous mapping, and*
- *the set of admissible controls $U_{\text{ad}} \subset U$ is nonempty, closed, and convex.*

For the rest of this section, we tacitly assume that the above assumption is fulfilled without mentioning it every time. Moreover, in all what follows, let $(\bar{y}, \bar{u}) \in Y \times U_{\text{ad}}$ with $\bar{y} = S(\bar{u})$ be an arbitrary, but fixed *local minimizer* of (P).

As indicated in the introduction, we are interested in the derivation of necessary optimality conditions. The particular challenge in case of (P) is that we do not assume S to be Gâteaux-differentiable. Therefore, standard techniques cannot be applied to establish an optimality condition involving dual variables. In contrast to this, purely primal optimality conditions can be derived by classical arguments.

Proposition 2.2 (Bouligand Stationarity) *Suppose that S is directionally differentiable at $\bar{u}$ in all directions $h \in \mathrm{cone}(U_{\mathrm{ad}} - \bar{u})$. Then, there holds*

$$\langle \partial_y J(\bar{y}, \bar{u}), S'(\bar{u}; h)\rangle_Y + \langle \partial_u J(\bar{y}, \bar{u}), h\rangle_U \geq 0 \quad \forall h \in \mathrm{cone}(U_{\mathrm{ad}} - \bar{u}), \tag{2.1}$$

where $\mathrm{cone}(U_{\mathrm{ad}} - \bar{u}) := \{\alpha(u - \bar{u}) : u \in U_{\mathrm{ad}}, \alpha > 0\}$ *denotes the conic hull of* $U_{\mathrm{ad}} - \bar{u}$.

Proof Since S is directionally differentiable and J is Fréchet-differentiable and thus Hadamard directionally differentiable, we can apply the chain rule, which immediately gives the assertion (see [2, Prop. 2.47]). □

Throughout this chapter, the condition in (2.1) is termed *B-stationarity* and, accordingly, a point $\bar{u} \in U_{\mathrm{ad}}$ fulfilling this condition is called *B-stationary*.

As indicated above, our aim is to deduce an optimality system containing dual variables from (2.1). Since $h \mapsto S'(\bar{u}; h)$ is in general not linear, the standard adjoint calculus cannot be applied from the shelf. In order to cope with this challenge, we need the following additional assumption on the structure of the directional derivative of S at $\bar{u}$:

Assumption 2.3 (Directional Differentiability of the Control-to-State Map) *The map S is directionally differentiable in $\bar{u}$ in every direction $h \in \mathrm{cone}(U_{\mathrm{ad}} - \bar{u})$, and its directional derivative $\delta = S'(\bar{u}; h)$ in direction $h \in \mathrm{cone}(U_{\mathrm{ad}} - \bar{u})$ is characterized as the solution of the following VI of the first kind:*

$$\delta \in \mathcal{K}(\bar{y}), \quad \langle A(\bar{y})\delta, v - \delta\rangle_{V_{\bar{y}}} \geq (h, v - \delta)_U \quad \forall v \in \mathcal{K}(\bar{y}), \tag{2.2}$$

where

- *$V_{\bar{y}}$ is a Hilbert space such that $V_{\bar{y}} \hookrightarrow Y$ and $V_{\bar{y}} \hookrightarrow^d U$,*
- *$A(\bar{y}) \in \mathcal{L}(V_{\bar{y}}, V_{\bar{y}}^*)$ is a strongly monotone operator, and*
- *$\mathcal{K}(\bar{y}) \subset V_{\bar{y}}$ is a nonempty, closed, and convex cone.*

As indicated by the subscript, $V_{\bar{y}}$ as well as $A(\bar{y})$ and $\mathcal{K}(\bar{y})$ may well depend on $\bar{y} = S(\bar{u})$.

In the following, we will identify U with its dual (by the Riesz theorem), which gives rise to the Gelfand triple:

$$V_{\bar{y}} \hookrightarrow U \cong U^* \hookrightarrow V_{\bar{y}}^*.$$

In this spirit, we will frequently interpret elements in $V_{\bar{y}}$ as elements in U without mentioning the respective embedding operator. Similarly, we neglect the embedding operator, when $U \cong U^*$ is treated as a subset of $V_{\bar{y}}^*$. In addition, an element g of Y^* is considered as an element of $V_{\bar{y}}^*$ via $E^* g$, where $E \in \mathcal{L}(V_{\bar{y}}, Y)$ is the embedding operator from Assumption 2.3. For ease of notation, we will also omit E^* in the following.

Unfortunately, strong stationarity conditions are in general not necessary for local optimality, as the counterexamples in [18, 25] show. Therefore, additional conditions are required, and, in our case, we rely on the following:

Assumption 2.4 (Critical Constraint Qualification) *The conic hull* $\mathrm{cone}(U_{\mathrm{ad}}-\bar{u})$ *is dense in* U*, i.e.,* $\overline{\mathrm{cone}(U_{\mathrm{ad}}-\bar{u})}^{U} = U$.

Remark 2.5 Assumption 2.4 is rather restrictive. For instance, it does in general not allow us to consider classical box constraints for the control, unless additional assumptions are fulfilled, which cannot be checked a priori (as usual for constraint qualifications), see, e.g., [25]. Note that, in case of box constraints, Assumption 2.4 is fulfilled, if $\bar{u}$ does not touch the bounds a.e. in the domain.

Since $A(\bar{y})$ is strongly monotone, the VI in (2.2) does possess a solution not only for every right-hand side in $\mathrm{cone}(U_{\mathrm{ad}}-\bar{u})$ but also for inhomogeneities in $V_{\bar{y}}^*$. We denote the associated solution operator by $G_{\bar{y}} : V_{\bar{y}}^* \to V_{\bar{y}}$ so that $G_{\bar{y}}|_{\mathrm{cone}(U_{\mathrm{ad}}-\bar{u})}(\cdot) = S'(\bar{u};\cdot)$. Due to the strong monotonicity of $A(\bar{y})$, this solution operator is globally Lipschitz continuous, i.e.,

$$\|G_{\bar{y}}(g) - G_{\bar{y}}(h)\|_{V_{\bar{y}}} \le L\,\|g-h\|_{V_{\bar{y}}^*} \quad \forall\, g,h \in V_{\bar{y}}^* \tag{2.3}$$

with a Lipschitz constant $L > 0$, whose potential dependency on $\bar{y}$ is suppressed for ease of notation.

As indicated above, we cannot define an adjoint state by means of the adjoint operator associated with the derivative of the control-to-state map, since the latter is nonlinear w.r.t. the direction. Instead, we introduce an adjoint state by extending the partial derivative $\partial_u J(\bar{y},\bar{u})$ to the dual of $V_{\bar{y}}$, which implies that the gradient equation coupling adjoint state and optimal control is automatically fulfilled.

Lemma 2.6 *There exists* $p \in V_{\bar{y}}$ *such that*

$$(p,h)_U + \langle \partial_u J(\bar{y},\bar{u}), h\rangle_U = 0 \quad \forall\, h \in U. \tag{2.4}$$

Moreover, for all $h \in V_{\bar{y}}^*$*, it holds*

$$\langle \partial_y J(\bar{y},\bar{u}), G_{\bar{y}}(h)\rangle_{V_{\bar{y}}} - \langle h,\, p\rangle_{V_{\bar{y}}} \ge 0. \tag{2.5}$$

Proof From (2.1), we know that

$$\langle \partial_y J(\bar{y},\bar{u}), G_{\bar{y}}(h)\rangle_{V_{\bar{y}}} + \langle \partial_u J(\bar{y},\bar{u}), h\rangle_U \ge 0 \quad \forall\, h \in \mathrm{cone}(U_{\mathrm{ad}}-\bar{u}), \tag{2.6}$$

and, consequently, the global Lipschitz continuity of $G_{\bar{y}}$ implies in view of $G_{\bar{y}}(0) = S'(\bar{u};0) = 0$ that

$$\langle -\partial_u J(\bar{y},\bar{u}), h\rangle_U \le \langle \partial_y J(\bar{y},\bar{u}), G_{\bar{y}}(h)\rangle_{V_{\bar{y}}} \le c\,\|\partial_y J(\bar{y},\bar{u})\|_{Y^*}\,\|h\|_{V_{\bar{y}}^*}$$

for all $h \in \operatorname{cone}(U_{\mathrm{ad}} - \bar{u})$. Since this set is $V_{\bar{y}}^*$-dense in U by Assumption 2.4 and $U \hookrightarrow V_{\bar{y}}^*$, this implies the existence of a constant $c > 0$ so that

$$|\langle -\partial_u J(\bar{y}, \bar{u}), h\rangle_U| \leq c\, \|h\|_{V_{\bar{y}}^*} \quad \forall\, h \in U.$$

Hence, by the Hahn–Banach theorem, $-\partial_u J(\bar{y}, \bar{u})$ can be extended to an element of $V_{\bar{y}}^{**}$, which we identify with an element $p \in V_{\bar{y}}$ by the reflexivity of $V_{\bar{y}}$. Since p is the extension of $-\partial_u J(\bar{y}, \bar{u})$, we immediately deduce (2.4). Inserting this in (2.6) and using that $\operatorname{cone}(U_{\mathrm{ad}} - \bar{u})$ is $V_{\bar{y}}^*$-dense in U and that U is dense in $V_{\bar{y}}^*$ by Lemma A.1 then give the second claim. □

Based on the previous lemma, we are now in the position to state our main result.

Theorem 2.7 (Strong Stationarity) *Let $\bar{u} \in U$ be locally optimal for (P) with associated optimal state $\bar{y} = S(\bar{u})$. Suppose that Assumptions 2.3 and 2.4 are satisfied at $\bar{u}$. Then, there exist $p \in V_{\bar{y}}$ and $\mu \in V_{\bar{y}}^*$ so that the following optimality condition is fulfilled:*

$$p + \partial_u J(\bar{y}, \bar{u}) = 0 \quad \text{in } U^* \tag{2.7a}$$

$$A(\bar{y})^* p + \mu = \partial_y J(\bar{y}, \bar{u}) \quad \text{in } V_{\bar{y}}^*, \tag{2.7b}$$

$$p \in \mathcal{K}(\bar{y}), \quad \langle \mu, v\rangle_{V_{\bar{y}}} \geq 0 \quad \forall\, v \in \mathcal{K}(\bar{y}). \tag{2.7c}$$

Proof From Lemma 2.6, we already know that there exists a $p \in V_{\bar{y}}$ such that (2.7a) holds. Let us now show that $p \in \mathcal{K}(\bar{y})$. For this purpose, define $\eta := G_{\bar{y}}\big(A(\bar{y})p\big)$ and $\zeta := G_{\bar{y}}\big(A(\bar{y})(p-\eta)\big)$, i.e., η and ζ solve

$$\eta \in \mathcal{K}(\bar{y}), \quad \langle A(\bar{y})\eta, v - \eta\rangle_{V_{\bar{y}}} \geq \langle A(\bar{y})p, v - \eta\rangle_{V_{\bar{y}}} \qquad \forall\, v \in \mathcal{K}(\bar{y}), \tag{2.8}$$

$$\zeta \in \mathcal{K}(\bar{y}), \quad \langle A(\bar{y})\zeta, v - \zeta\rangle_{V_{\bar{y}}} \geq \langle A(\bar{y})(p-\eta), v - \zeta\rangle_{V_{\bar{y}}} \quad \forall\, v \in \mathcal{K}(\bar{y}). \tag{2.9}$$

Since $\mathcal{K}(\bar{y})$ is a closed cone by assumption, we can insert $0 \in \mathcal{K}(\bar{y})$ and $2\eta \in \mathcal{K}(\bar{y})$ as test elements in (2.8), which results in

$$\langle A(\bar{y})(p-\eta), \eta\rangle_{V_{\bar{y}}} = 0 \quad \text{and} \quad \langle A(\bar{y})(p-\eta), v\rangle_{V_{\bar{y}}} \leq 0 \quad \forall\, v \in \mathcal{K}(\bar{y}). \tag{2.10}$$

The latter inequality implies for (2.9) tested with $v = 0$ that $\langle A(\bar{y})\zeta, \zeta\rangle_{V_{\bar{y}}} \leq 0$, which, thanks to the strong monotonicity of $A(\bar{y})$, in turn gives $\zeta = 0$. Next, we insert $h = A(\bar{y})(p-\eta) \in V_{\bar{y}}^*$ as a test elements in (2.5), which, due to $G_{\bar{y}}(A(\bar{y})(p-\eta)) = \zeta = 0$, results in

$$\langle A(\bar{y})(p-\eta), p\rangle_{V_{\bar{y}}} \leq 0.$$

Together with the first equation in (2.10), this yields $\langle A(\bar{y})(p-\eta), p-\eta\rangle_{V_{\bar{y}}} \leq 0$ so that the strong monotonicity of $A(\bar{y})$ gives $p = \eta \in \mathcal{K}(\bar{y})$ as claimed.

Next, we simply define $\mu \in V_{\bar{y}}^*$ by setting $\mu := \partial_y J(\bar{y}, \bar{u}) - A(\bar{y})^* p \in V_{\bar{y}}^*$ so that (2.7b) is fulfilled, as well. It remains to verify the last condition in (2.7c). To this end, let $v \in \mathcal{K}(\bar{y})$ be arbitrary. Then, by construction of $G_{\bar{y}}$, the feasibility of v yields $v = G_{\bar{y}}(A(\bar{y})v)$. Therefore, if we insert $h = A(\bar{y})v$ in (2.5), then

$$\langle \mu, v \rangle_{V_{\bar{y}}} = \langle \partial_y J(\bar{y}, \bar{u}), v \rangle_{V_{\bar{y}}} - \langle A(\bar{y})v, p \rangle_{V_{\bar{y}}} \geq 0$$

follows. Since $v \in \mathcal{K}(\bar{y})$ was arbitrary, this completes the proof. □

Proposition 2.8 *Let Assumption 2.3 hold at a (not necessarily locally optimal) point $\bar{u} \in U_{\mathrm{ad}}$, and assume that $p \in V_{\bar{y}}$ and $\mu \in V_{\bar{y}}^*$ exist such that (2.7) is fulfilled. Then, $\bar{u}$ satisfies the B-stationarity condition in (2.1).*

Proof Let $h \in \mathrm{cone}(U_{\mathrm{ad}} - \bar{u})$ be arbitrary, and write again $\delta = S'(\bar{u}; h)$. Similarly to the beginning of the proof of Theorem 2.7, we test the VI in (2.2) with $v = 0$ and $v = 2\delta$ to obtain

$$\langle A(\bar{y})\delta, v \rangle_{V_{\bar{y}}} \geq (h, v)_U \quad \forall\, v \in \mathcal{K}(\bar{y}).$$

Due to $p \in \mathcal{K}(\bar{y})$, this inequality also holds for $v = p$, which, in combination with the adjoint equation in (2.7b), gives

$$\begin{aligned} \langle \partial_y J(\bar{y}, \bar{u}), S'(\bar{u}; h) \rangle_Y &= \langle \partial_y J(\bar{y}, \bar{u}), \delta \rangle_{V_{\bar{y}}} \\ &= \langle A(\bar{y})^* p + \mu, \delta \rangle_{V_{\bar{y}}} \\ &= \langle A(\bar{y})\delta, p \rangle_{V_{\bar{y}}} + \langle \mu, \delta \rangle_{V_{\bar{y}}} \geq (h, p)_U . \end{aligned}$$

In view of (2.7a), this gives the assertion. □

Remark 2.9 Theorem 2.7 and Proposition 2.8 (or the proofs of these results, to be more precise) demonstrate that, under Assumption 2.3 and the constraint qualification in Assumption 2.4, the optimality system in (2.7) and the B-stationarity condition in (2.1) are equivalent. We have thus found an optimality system involving dual variables, which is equivalent to the purely primal optimality condition. This motivates the notion *strong stationarity* for the optimality condition in (2.7).

Remark 2.10 The result of Theorem 2.7 can be substantially generalized by allowing for a more general structure of the directional derivative of S, see [3, Section 6].

3 Application to Concrete Settings

3.1 The Obstacle Problem

To keep the discussion concise, we restrict ourselves to the classical obstacle problem governed by the Laplacian, i.e.,

$$y \in K, \quad \int_\Omega \nabla y \cdot \nabla(v - y)\,\mathrm{d}x \geq \langle u, v - y\rangle \quad \forall\, v \in K. \tag{3.1}$$

Herein, $\Omega \subset \mathbb{R}^d$, $d \geq 1$, is a bounded domain and

$$K := \{v \in H_0^1(\Omega) : \psi_1 \leq v \leq \psi_2 \text{ a.e. in } \Omega\},$$

with two given functions $\psi_1, \psi_2 \in H^1(\Omega)$ such that $K \neq \emptyset$. Clearly, for every $u \in H^{-1}(\Omega)$, (3.1) admits a unique solution $y \in H_0^1(\Omega)$, and the associated solution operator $S : H^{-1}(\Omega) \to H_0^1(\Omega)$ is globally Lipschitz continuous. Moreover, there holds the following:

Proposition 3.1 ([19, Théorème 3.3]) *The solution operator S of (3.1) is directionally differentiable from $H^{-1}(\Omega)$ to $H_0^1(\Omega)$. Its directional derivative at $u \in H^{-1}(\Omega)$ in direction $h \in H^{-1}(\Omega)$ is given by the unique solution of the following VI of the first kind:*

$$\delta \in \mathcal{K}(y), \quad \int_\Omega \nabla\delta \cdot \nabla(v - \delta)\mathrm{d}x \geq \langle h, v - \delta\rangle \quad \forall\, v \in \mathcal{K}(y), \tag{3.2}$$

where $y = S(u)$ and $\mathcal{K}(y) \subset H_0^1(\Omega)$ is the closed and convex cone defined by

$$\mathcal{K}(y) := \{v \in H_0^1(\Omega) : v \leq 0 \text{ q.e., where } y = \psi_2,\ v \geq 0 \text{ q.e., where } y = \psi_1,\ \langle \Delta y + u, v\rangle = 0\}. \tag{3.3}$$

Note that the pointwise properties in the definition of $\mathcal{K}(y)$ are required quasi-everywhere, i.e., the quasi-continuous representative satisfies the respective property up to sets of zero $H^1(\mathbb{R}^d)$-capacity. The proof of the above proposition is based on the *polyhedricity* of the set K, which means that, for all $v \in K$ and all $g \in H^{-1}(\Omega)$, there holds

$$\overline{\mathrm{cone}(K - v)}^{H_0^1} \cap \mathrm{Ker}(g) = \overline{\mathrm{cone}(K - v) \cap \mathrm{Ker}(g)}^{H_0^1},$$

see [19, 28] and [3, Section 3.3].

Our optimal control problem governed by the obstacle problem now reads as follows:

$$\left.\begin{aligned} \min \quad & J(y,u) \\ \text{s.t.} \quad & (y,u) \in H_0^1(\Omega) \times L^2(\Omega), \\ & (y,u) \text{ satisfy (3.1)}, \quad u \in U_{\text{ad}}, \end{aligned}\right\} \tag{P_{obst}}$$

where $J : H_0^1(\Omega) \times L^2(\Omega) \to \mathbb{R}$ is a given Fréchet-differentiable objective and $U_{\text{ad}} \subset L^2(\Omega)$ is a nonempty, closed, and convex set. Again, we consider a fixed but arbitrary local minimizer of (P_{obst}), denoted by $(\bar{y}, \bar{u}) \in H_0^1(\Omega) \times L^2(\Omega)$. Then, by setting

$$U := L^2(\Omega), \quad Y = V_{\bar{y}} := H_0^1(\Omega), \quad A(\bar{y}) := -\Delta,$$

and $\mathcal{K}(\bar{y})$ as defined in (3.3) (with $y = \bar{y}$), problem (P_{obst}) fits into our general setting. Hence, the general theory from Sect. 2 can be applied to this example:

Theorem 3.2

1. *Suppose that $\bar{u} \in L^2(\Omega)$ is locally optimal for (P_{obst}) with associated state $\bar{y} = S(\bar{u})$. Moreover, let the critical constraint qualification in Assumption 2.4 be fulfilled, i.e.,*

$$\overline{\text{cone}(U_{\text{ad}} - \bar{u})}^{L^2} = L^2(\Omega). \tag{3.4}$$

Then, there exist $p \in H_0^1(\Omega)$ and $\mu \in H^{-1}(\Omega)$ such that

$$p + \partial_u J(\bar{y}, \bar{u}) = 0 \quad \text{a.e. in } \Omega, \tag{3.5a}$$

$$-\Delta p + \mu = \partial_y J(\bar{y}, \bar{u}) \quad \text{in } H^{-1}(\Omega), \tag{3.5b}$$

$$p \in \mathcal{K}(\bar{y}), \quad \langle \mu, v \rangle \geq 0 \quad \forall\, v \in \mathcal{K}(\bar{y}). \tag{3.5c}$$

2. *Assume that $\bar{u} \in U_{\text{ad}}$ with associated state $\bar{y} = S(\bar{u})$ is such that $p \in H_0^1(\Omega)$ and $\mu \in H^{-1}(\Omega)$ exist so that the system in (3.5) is fulfilled. Then, $\bar{u}$ is B-stationary for (P_{obst}).*

Remark 3.3 The first assertion of Theorem 3.2 concerning the necessary optimality condition was already proven in [20]. There, the critical constraint qualification (3.4) is ensured by simply setting $U_{\text{ad}} = U = L^2(\Omega)$.

Another rather implicitly given condition is the assumption that $V_{\bar{y}} = H_0^1(\Omega)$ must embed into $U = L^2(\Omega)$. The injectivity of the embedding operator thus prevents the derivation of strong stationarity conditions in case of boundary controls, where $U = L^2(\partial\Omega)$, as the counterexample in [18] demonstrates. Similarly, controls that only act on parts of the domain Ω can also not be treated by our analysis. This shows that the assumptions concerning the set of admissible controls are indeed rather restrictive.

3.2 Static Elastoplasticity

Now, we turn to a VI of the second kind and consider an optimal control problem governed by the system of static elastoplasticity with linear kinematic hardening. Strictly speaking, the problem of static elastoplasticity is physically not meaningful, but it may be regarded as the stationary problem that has to be solved in one time step of an implicit time discretization of the quasi-static elastoplastic evolution. The model under consideration is the primal formulation of static elastoplasticity with linear kinematic hardening and the von Mises yield condition and reads as follows:

$$\left.\begin{aligned} &(u,\, p) \in \mathscr{V} \times \mathscr{P}, \\ &\int_\Omega (\boldsymbol{\varepsilon}(u) - p) : \mathbb{C}\big(\boldsymbol{\varepsilon}(v-u) - (q-p)\big) + p : \mathbb{H}(q-p)\,\mathrm{d}x \\ &\qquad + \sigma_0 \int_\Omega |q|_F \,\mathrm{d}x - \sigma_0 \int_\Omega |p|_F \,\mathrm{d}x \\ &\qquad\qquad \geq \langle \ell, v-u\rangle + \int_\Omega \mathscr{L} : (q-p)\,\mathrm{d}x \quad \forall\, (v,q) \in \mathscr{V} \times \mathscr{P}. \end{aligned}\right\} \tag{3.6}$$

Herein, $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, is a given bounded Lipschitz domain, whose boundary is split into two disjoint parts $\partial\Omega = \Gamma_D \cup \Gamma_N$. The part Γ_D is assumed to have positive measure. Moreover, $u : \Omega \to \mathbb{R}^d$ denotes the displacement field, while $p : \Omega \to \mathbb{R}^{d\times d}_{\mathrm{dev}}$ is the plastic strain tensor. Herein, $\mathbb{R}^{d\times d}_{\mathrm{dev}}$ is the space of symmetric matrices with zero trace, equipped with the Frobenius norm, which is denoted by $|\cdot|_F$. The associated scalar product is denoted by $\mathbb{R}^{d\times d} \times \mathbb{R}^{d\times d} \ni (a, b) \mapsto a : b \in \mathbb{R}$. The spaces in (3.6) are defined as follows:

$$\mathscr{P} := L^2(\Omega; \mathbb{R}^{d\times d}_{\mathrm{dev}}), \quad \mathscr{V} := \{v \in W^{1,2}(\Omega; \mathbb{R}^d) : v = 0 \text{ a.e. on } \Gamma_D\}.$$

Moreover, $\mathbb{C}, \mathbb{H} \in L^\infty(\Omega; \mathcal{L}(\mathbb{R}^{d\times d}, \mathbb{R}^{d\times d}))$ are two given symmetric and uniformly coercive mappings (the elasticity and hardening tensor). In addition, $\boldsymbol{\varepsilon} = \frac{1}{2}(\nabla + \nabla^\top)$ denotes the linearized strain. Furthermore, $\sigma_0 > 0$ is the uni-axial yield stress, a constant material parameter. There are other equivalent formulations of the system in (3.6), for instance, in terms of a VI of the first kind via convex duality, which leads to the so-called dual formulation. A more detailed description of the model can be found in [11].

The variables $\ell \in L^2(\Omega; \mathbb{R}^d)$ and $\mathscr{L} \in \mathscr{P}$ serve as controls in our setting. While ℓ has a well-defined physical meaning as the loads applied to the body occupying Ω, the physical interpretation of $\mathscr{L}$ is rather critical. It may be seen as a pre-strain, but it is mainly motivated by the mathematical analysis, since it allows us to fulfill Assumption 2.3 in this case, as we will see below.

The optimal control problem considered in this subsection reads as follows:

$$\left.\begin{aligned} \min \quad & J(u, p, \ell, \mathscr{L}) \\ \text{s.t.} \quad & (u, p, \ell, \mathscr{L}) \in \mathscr{V} \times \mathscr{P} \times L^2(\Omega; \mathbb{R}^d) \times \mathscr{P}, \\ & (u, p, \ell, \mathscr{L}) \text{ satisfy (3.6)} \end{aligned}\right\} \tag{$\mathrm{P}_{\mathrm{plast}}$}$$

with a Fréchet-differentiable objective $J : \mathscr{V} \times \mathscr{P} \times L^2(\Omega; \mathbb{R}^d) \times \mathscr{P} \to \mathbb{R}$. We could also consider additional control constraints (and would then again need an additional assumption of the form in Assumption 2.4), but, in order to keep the discussion concise, we restrict to the case without control constraints.

Thanks to Korn's inequality and the coercivity of $\mathbb{C}$ and $\mathbb{H}$, the VI in (3.6) admits a unique solution in $\mathscr{V} \times \mathscr{P}$ for every right-hand side in $\mathscr{V}^* \times \mathscr{P}$, see, e.g., [11]. The next result shows that the associated solution mapping is directionally differentiable and was established in [3, Section 4.3].

Proposition 3.4 ([3, Corollary 4.3.5]) *The solution operator of (3.6), denoted by $S : (\ell, \mathscr{L}) \mapsto (u, p)$, is directionally differentiable from $\mathscr{V}^* \times \mathscr{P}$ to $\mathscr{V} \times \mathscr{P}$. Its directional derivative at $(\ell, \mathscr{L}) \in \mathscr{V}^* \times \mathscr{P}$ in direction $(h, G) \in \mathscr{V}^* \times \mathscr{P}$ is given by the unique solution (u', p') of*

$$\begin{aligned} &(u', p') \in \mathscr{V} \times K(u, p), \\ &\langle \mathcal{A}(p)(u', p'), (v, q) - (u', p') \rangle \\ &\qquad \geq \langle h, v - u' \rangle + \int_\Omega G : (q - p') \, \mathrm{d}x \quad \forall (v, q) \in \mathscr{V} \times K(u, p), \end{aligned}$$

where $(u, p) = S(\ell, \mathscr{L})$,

$$\mathscr{P}_p := \Big\{ q \in \mathscr{P} : \int_{\{p \neq 0\}} |p|_F^{-3} \big(|p|_F^2 \, |q|_F^2 - (p : q)^2 \big) \mathrm{d}x < \infty \Big\},$$

$$K(u, p) := \{ q \in \mathscr{P}_p : \big(\mathscr{L} + \mathbb{C}\boldsymbol{\varepsilon}(u) \big) : q = \sigma_0 |q|_F \text{ a.e., where } p = 0 \},$$

and

$$\begin{aligned} &\mathcal{A}(p) : \mathscr{V} \times \mathscr{P}_p \to \mathscr{V}^* \times \mathscr{P}_p^* \\ &\langle \mathcal{A}(p)(w, r), (v, q) \rangle := \int_\Omega (\boldsymbol{\varepsilon}(w) - r) : \mathbb{C}\big(\boldsymbol{\varepsilon}(v) - q \big) + r : \mathbb{H} q \, \mathrm{d}x \\ &\qquad\qquad + \sigma_0 \int_{\{p \neq 0\}} |p|_F^{-3} \big(|p|_F^2 \, r : q - (p : r)(p : q) \big) \mathrm{d}x. \end{aligned} \tag{3.7}$$

The directional differentiability of the solution operator associated with the (equivalent) dual formulation of (3.6) was investigated in [1, 14]. Via convex duality, one shows that these results are in accordance with the above proposition.

Let us now again consider a fixed but arbitrary local minimizer of (P_{plast}), denoted by $(\bar{\ell}, \bar{\mathscr{L}}) \in L^2(\Omega; \mathbb{R}^d) \times \mathscr{P}$ with associated state $\bar{y} := (\bar{u}, \bar{p}) = S(\bar{\ell}, \bar{\mathscr{L}})$. Then, with the setting

$$U = L^2(\Omega; \mathbb{R}^d) \times \mathscr{P}, \quad Y = \mathscr{V} \times \mathscr{P}, \quad V_{\bar{y}} := \mathscr{V} \times \mathscr{P}_{\bar{p}},$$
$$\mathcal{K}(\bar{y}) := \mathscr{V} \times K(\bar{u}, \bar{p}), \quad A(\bar{y}) := \mathcal{A}(\bar{p}),$$

the problem of static elastoplasticity fits into our general framework, as we will see in the following. Equipped with the scalar product

$$\big((w, r), (v, q)\big)_{V_{\bar{y}}} := (w, v)_{\mathscr{V}} + (r, q)_{\mathscr{P}} + \int_{\{\bar{p} \neq 0\}} |\bar{p}|_F^{-3} \big(|\bar{p}|_F^2 \, r : q - (\bar{p} : r)(\bar{p} : q)\big) \mathrm{d}x,$$

the space $V_{\bar{y}}$ becomes a Hilbert space as required. Since $\mathscr{P}_{\bar{p}}$ is a dense subset of $\mathscr{P}$, $V_{\bar{y}}$ is dense in U. We point out that the presence of the additional (and rather artificial) control variable $\mathscr{L}$ is crucial for the embedding of $V_{\bar{y}}$ in U (with the injective pointwise identity as embedding operator). Furthermore, by using Korn's inequality, one shows that $A(\bar{y}) = \mathcal{A}(\bar{p})$ as defined in (3.7) is strongly monotone. Therefore, all conditions in Assumption 2.3 are fulfilled. Since, in addition, Assumption 2.4 is trivially satisfied as $U_{\mathrm{ad}} = U$, one deduces the following:

Theorem 3.5 ([3, Corollary 6.1.13])

1. *Let $(\bar{\ell}, \bar{\mathscr{L}}) \in L^2(\Omega; \mathbb{R}^d) \times \mathscr{P}$ with associated state $\bar{y} = (\bar{u}, \bar{p}) \in \mathscr{V} \times \mathscr{P}$ be locally optimal for (P_{plast}). Then, there exist an adjoint state $(w, r) \in \mathscr{V} \times \mathscr{P}_{\bar{p}}$ and a multiplier $\mu \in \mathscr{P}$ such that*

$$w + \partial_\ell J(\bar{u}, \bar{p}, \bar{\ell}, \bar{\mathscr{L}}) = 0 \quad \text{a.e. in } \Omega, \tag{3.8a}$$

$$r + \partial_{\mathscr{L}} J(\bar{u}, \bar{p}, \bar{\ell}, \bar{\mathscr{L}}) = 0 \quad \text{a.e. in } \Omega, \tag{3.8b}$$

$$\left.\begin{aligned} &\langle \mathcal{A}(\bar{p})(w, r), (v, q) \rangle \\ &\quad = \langle \partial_u J(\bar{u}, \bar{p}, \bar{\ell}, \bar{\mathscr{L}}), v \rangle_{\mathscr{V}} + \int_\Omega \partial_p J(\bar{u}, \bar{p}, \bar{\ell}, \bar{\mathscr{L}}) : q \, \mathrm{d}x \\ &\qquad\qquad - \int_\Omega \mu : q \, \mathrm{d}x \quad \forall (v, q) \in \mathscr{V} \times \mathscr{P}_{\bar{p}} \end{aligned}\right\} \tag{3.8c}$$

$$\big(\bar{\mathscr{L}} + \mathbb{C}(\boldsymbol{\varepsilon}(\bar{u}) - \bar{p}) - \mathbb{H}\bar{p}\big) : r = \sigma_0 |r|_F \text{ a.e., where } \bar{p} = 0, \tag{3.8d}$$

$$\mu = 0 \text{ a.e., where } \bar{p} \neq 0, \tag{3.8e}$$

$$\left.\begin{array}{l}\mu : q \geq 0 \quad \forall q \in \mathbb{R}^{d\times d}_{\mathrm{dev}} \text{ with } \left(\bar{\mathscr{L}} + \mathbb{C}(\boldsymbol{\varepsilon}(\bar{u}) - \bar{p}) - \mathbb{H}\bar{p}\right) : q = \sigma_0 |q|_F \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{a.e., where } \bar{p} = 0\end{array}\right\}. \tag{3.8f}$$

2. *If a couple* $(\bar{\ell}, \bar{\mathscr{L}}) \in L^2(\Omega;\mathbb{R}^d) \times \mathscr{P}$ *together with its associated state* $(\bar{u}, \bar{p}) = S(\bar{\ell}, \bar{\mathscr{L}})$*, an adjoint state* $(w, r) \in \mathscr{V} \times \mathscr{P}_{\bar{p}}$*, and a multiplier* $\mu \in \mathscr{P}$ *satisfies the system* (3.8)*, then it is B-stationary for* ($\mathrm{P}_{\mathrm{plast}}$).

Remark 3.6 It is noteworthy that, in this example in contrast to the previous one in Sect. 3.1, the space $V_{\bar{y}}$ and the operator $A(\bar{y})$ differ from the original state space Y and the "smooth part" in the VI in (3.6) associated with the control-to-state map. This effect will also appear in the two examples in the next sections.

Let us finally remark that a strong stationarity system for optimal control of static plasticity in dual, i.e., stress-based formulation, is derived in [14] under slightly more restrictive assumptions.

3.3 The Lasso Problem in Sobolev Spaces

This subsection is devoted to an optimal control problem governed by the following VI of the second kind:

$$\begin{aligned} &y \in H^1_0(\Omega), \\ &\int_\Omega \nabla y \cdot \nabla(v - y)\mathrm{d}x + \|v\|_{L^1(\Omega)} - \|y\|_{L^1(\Omega)} \geq \langle u, v - y\rangle \quad \forall\, v \in H^1_0(\Omega), \end{aligned} \tag{3.9}$$

where $\Omega \subset \mathbb{R}^d$, $d \geq 2$, is a bounded Lipschitz domain in the sense of [10, Definition 4.4]. In finite dimensions, VIs of this type arise in the context of sparse linear regression and are occasionally called lasso problem, see, e.g., [24]. By the direct method of the calculus of variations, one shows that, for every right-hand side $u \in H^{-1}(\Omega)$, there exists a unique solution $y \in H^1_0(\Omega)$ of (3.9), and the associated solution operator, denoted by $S : H^{-1}(\Omega) \to H^1_0(\Omega)$, is globally Lipschitz. Its differentiability properties however constitute a delicate issue. Via convex duality, one can transform the VI in (3.9) into an equivalent obstacle problem in $H^{-1}(\Omega)$. However, its feasible set given by

$$\Lambda := \Big\{\lambda \in H^{-1}(\Omega) : \exists\, q \in L^\infty(\Omega) \text{ such that } |q| \leq 1 \text{ a.e. in } \Omega \\ \text{and } \langle \lambda, v\rangle = \int_\Omega q\, v\, \mathrm{d}x\Big\}$$

is in general not polyhedric, as the counterexamples in [5, 6, 28] demonstrate. If one assumes that this set is polyhedric, then the analysis of [12, 19] can be adapted to prove the directional differentiability of S, see [17, 23]. In [9], comparatively restrictive conditions are established, which guarantee that this set behaves like a polyhedric set. Another approach that goes without polyhedricity is pursued in [5] and yields the following result:

Proposition 3.7 ([5], [3, Theorem 5.2.15]) *Let $u \in H^{-1}(\Omega)$ be given with associated state $y \in H_0^1(\Omega)$, and suppose that the following assumptions are fulfilled:*

- (Regularity) *It holds $y \in C^1(\Omega) \cap H_0^1(\Omega)$.*
- (Structure of the Active Set) *There exists a set $\mathcal{C} \subseteq \partial\{y \neq 0\} \cup \partial\Omega$ such that*
  1. *$\mathcal{C}$ is closed and has $H^1(\mathbb{R}^d)$-capacity zero,*
  2. *$(\partial\{y \neq 0\} \cup \partial\Omega) \setminus \mathcal{C}$ is a (strong) $(d-1)$-dimensional Lipschitz submanifold of $\mathbb{R}^d$,*
  3. *the sets*

$$\mathcal{N}_+ := \{\nabla y = 0\} \cap \partial\{y > 0\} \setminus \mathcal{C}, \quad \mathcal{N}_- := \{\nabla y = 0\} \cap \partial\{y < 0\} \setminus \mathcal{C}$$

are relatively open in $(\partial\{y \neq 0\} \cup \partial\Omega) \setminus \mathcal{C}$.

Then, S is directionally differentiable at u in every direction $h \in H^{-1}(\Omega)$, and the directional derivative $\delta = S'(u; h)$ is given by the unique solution of the following VI of the first kind:

$$\begin{aligned} &\delta \in \mathcal{K}(y), \\ &\int_\Omega \nabla\delta \cdot \nabla(v - \delta)\mathrm{d}x + 2\int_{\mathcal{M}} \frac{\tau(\delta)\,\tau(v-\delta)}{\|\nabla y\|_2}\,\mathrm{d}\mathcal{H}^{d-1} \geq \langle h, v - \delta\rangle \quad \forall\, v \in \mathcal{K}(y), \end{aligned} \tag{3.10}$$

where $\mathcal{M} := \{y = 0\} \cap \{\nabla y \neq 0\}$ and τ is the associated trace operator. Furthermore, the convex cone $\mathcal{K}(y)$ is given by

$$\begin{aligned} \mathcal{K}(y) := \Big\{ v \in H_0^1(\Omega) : \tau(v)^- = 0 \text{ a.e. on } \mathcal{N}_+,\ \tau(v)^+ = 0 \text{ a.e. on } \mathcal{N}_-, \\ |v| = \lambda\, v \text{ a.e. in } \{y = 0\},\ \int_{\mathcal{M}} \frac{\tau(v)^2}{\|\nabla y\|_2}\mathrm{d}\mathcal{H}^{d-1} < \infty \Big\}, \end{aligned} \tag{3.11}$$

where $\lambda \in L^\infty(\Omega)$ is the unique element of $\partial\|\cdot\|_{L^1(\Omega)}(y)$ that satisfies $\lambda = u + \Delta y$.

Note that the sets $\mathcal{C}$, $\mathcal{N}_\pm$, and $\mathcal{M}$ depend on the solution $y = S(u)$, but we suppress this dependency in order to simplify the notation.

Remark 3.8 It is to be noted that, in contrast to the previous examples in Sects. 3.1 and 3.2, there is—to the best of our knowledge—no result available in the literature that guarantees the directional differentiability of the solution operator to (3.9) without further assumptions. The assumptions in Proposition 3.7 are on the one hand easily verifiable, once the solution y is known (in contrast to the polyhedricity of Λ) and on the other hand substantially less restrictive compared to the assumptions in [9]. Proposition 3.7 therefore can be seen as the most rigorous differentiability result for the solution operator of (3.9).

Remark 3.9 The lack of polyhedricity of Λ is also illustrated by the integral over the set $\mathcal{M}$ in (3.10), which does not appear, if the set Λ is polyhedric, see [9, 17]. This integral is closely related to the pullback of the second distributional derivative of the absolute value function, see [3, Section 5.2.2] for details.

Similarly to the previous examples, we consider an optimal control problem of the form

$$\left.\begin{aligned} \min \quad & J(y,u) \\ \text{s.t.} \quad & (y,u) \in H_0^1(\Omega) \times L^2(\Omega), \\ & (y,u) \text{ satisfy (3.9)}, \end{aligned}\right\} \tag{$\mathrm{P}_{\mathrm{lasso}}$}$$

where $J : H_0^1(\Omega) \times L^2(\Omega) \to \mathbb{R}$ is a given Fréchet-differentiable objective. To keep the discussion concise, we again set $U_{\mathrm{ad}} = L^2(\Omega)$ in order to fulfill the constraint qualification in Assumption 2.4. As before, we consider a fixed but arbitrary local minimizer of ($\mathrm{P}_{\mathrm{lasso}}$), denoted by $(\bar{y}, \bar{u}) \in H_0^1(\Omega) \times L^2(\Omega)$. If we set

$$U := L^2(\Omega), \quad Y := H_0^1(\Omega), \quad V_{\bar{y}} := \left\{ v \in H_0^1(\Omega) : \int_{\mathcal{M}} \frac{\tau(v)^2}{\|\nabla \bar{y}\|_2} \mathrm{d}\mathcal{H}^{d-1} < \infty \right\},$$

$$\langle A(\bar{y})w, v\rangle := \int_\Omega \nabla w \cdot \nabla v \,\mathrm{d}x + 2 \int_{\mathcal{M}} \frac{\tau(w)\,\tau(v)}{\|\nabla \bar{y}\|_2} \,\mathrm{d}\mathcal{H}^{d-1}, \quad v, w \in V_{\bar{y}},$$

and $\mathcal{K}(\bar{y})$ as defined in (3.11) (with $y = \bar{y}$), then ($\mathrm{P}_{\mathrm{lasso}}$) fits into our general setting. With the obvious scalar product

$$(w, v)_{V_{\bar{y}}} := \langle A(\bar{y})w, v\rangle,$$

$V_{\bar{y}}$ becomes a Hilbert space and $A(\bar{y})$ is clearly strongly monotone in this space. Moreover, $\mathcal{K}(\bar{y})$ is closed in this space. Hence, the general theory from Sect. 2 is applicable and yields the following:

Theorem 3.10

1. *Suppose that $\bar{u} \in L^2(\Omega)$ with associated state $\bar{y} = S(\bar{u})$ is locally optimal for ($\mathrm{P}_{\mathrm{lasso}}$), and assume moreover that $\bar{y}$ is such that the assumptions of Proposition 3.7 on the regularity of $\bar{y}$ and the structure of its active set are*

fulfilled (with $y = \bar{y}$). Then, there exist an adjoint state $p \in V_{\bar{y}}$ and a multiplier $\mu \in V_{\bar{y}}^*$ such that

$$p + \partial_u J(\bar{y}, \bar{u}) = 0 \quad \text{a.e. in } \Omega, \tag{3.12a}$$

$$\left.\begin{aligned} \int_\Omega \nabla p \cdot \nabla v \, \mathrm{d}x + 2 \int_{\mathcal{M}} \frac{\tau(p)\,\tau(v)}{\|\nabla \bar{y}\|_2} \, \mathrm{d}\mathcal{H}^{d-1} \\ = \langle \partial_y J(\bar{y}, \bar{u}), v \rangle_{H_0^1(\Omega)} - \langle \mu, v \rangle_{V_{\bar{y}}}, \quad \forall\, v \in V_{\bar{y}}, \end{aligned}\right\} \tag{3.12b}$$

$$p \in \mathcal{K}(\bar{y}), \quad \langle \mu, v \rangle_{V_{\bar{y}}} \geq 0 \quad \forall\, v \in \mathcal{K}(\bar{y}). \tag{3.12c}$$

2. *Let $\bar{u} \in L^2(\Omega)$ be given such that its state $\bar{y} = S(\bar{u})$ satisfies the assumptions in Proposition 3.7 (with $y = \bar{y}$). If an adjoint state $p \in V_{\bar{y}}$ and a multiplier $\mu \in V_{\bar{y}}^*$ exist such that (3.12) holds true, then $\bar{u}$ is B-stationary for ($\mathrm{P}_{\mathrm{lasso}}$).*

Again, we observe that the space $V_{\bar{y}}$ differs from the original state space Y and that the bilinear form of the adjoint equation differs from the one in the VI defining the control-to-state map, similarly to the elastoplastic system in the previous section.

3.4 Non-Newtonian Fluids: The Mosolov Problem

Our last example arises in the modeling of non-Newtonian fluids. To be more precise, we consider the so-called Mosolov problem, which models the steady-state motion of a viscoplastic fluid in a cylindrical pipe of cross-section $\Omega \subset \mathbb{R}^2$ under no-slip boundary conditions; see [21] for details on the physical background. After setting all material parameters to one, the model is similar to the lasso problem and reads

$$\begin{aligned} &y \in H_0^1(\Omega), \\ &\int_\Omega \nabla y \cdot \nabla(v - y) \mathrm{d}x + \int_\Omega |\nabla v| \mathrm{d}x - \int_\Omega |\nabla y| \mathrm{d}x \geq \langle u, v - y \rangle \quad \forall\, v \in H_0^1(\Omega), \end{aligned} \tag{3.13}$$

where $\Omega \subset \mathbb{R}^2$ is a bounded and simply connected Lipschitz domain and $|\cdot|$ denotes the Euclidean norm of a vector. The state variable $y : \Omega \to \mathbb{R}$ describes the velocity of the fluid in direction of the pipe (i.e., perpendicular to Ω) and $u : \Omega \to \mathbb{R}$ is a volume force acting in this direction. The restriction to the two-dimensional setting is on the one hand motivated by the application background and on the other hand essential for the mathematical analysis presented in the following.

Again, the existence and uniqueness for (3.13) follow immediately from the direct method of the calculus of variations. The associated solution map S is globally Lipschitz from $H^{-1}(\Omega)$ to $H_0^1(\Omega)$. However, as in case of the lasso problem, the directional differentiability of S is a challenging issue. As it turns out, we again need additional assumptions on the respective state to obtain the existence of a directional derivative. Unfortunately, these assumptions are not as easily checked as in case of the lasso problem. In order to formulate these assumptions, we need to define the following sets. Given the solution $y \in H_0^1(\Omega)$ of (3.13) and assuming that this solution admits a continuously differentiable representative, we introduce the active, inactive, and biactive sets as follows:

$$\mathcal{I} := \{|\nabla y| > 0\}, \quad \mathcal{A} := \{|\nabla y| = 0\}, \quad \mathcal{A}^\circ := \operatorname{int}(\mathcal{A}), \quad \mathcal{B} := \partial\mathcal{A} \cup \partial\Omega,$$

$$\mathcal{B}^\circ := \Big\{ x \in \partial\mathcal{A} : \begin{array}{l} \text{there exists an open neighborhood } D \subseteq \Omega \text{ of } x \text{ such that} \\ D \cap \partial\mathcal{A} \text{ is a one-dimensional } C^1\text{-submanifold of } \mathbb{R}^2 \\ \text{and such that } D \cap \partial\mathcal{A} = D \cap \partial\{y = c\} \text{ for some } c \in \mathbb{R} \end{array} \Big\}.$$

Of course, these sets depend on the respective solution y, but we suppress this dependency for the ease of notation. Moreover, given a set $\mathcal{M} \subset \bar{\Omega}$, we denote the set of all connected components of $\mathcal{M}$ by $\{\mathcal{M}_i\}$.

Proposition 3.11 ([3, Theorem 5.1.37]) *Let $u \in H^{-1}(\Omega)$ be given, and suppose that the associated state $y = S(u)$ satisfies the following hypotheses:*

- (Regularity) *It holds $y \in C^{1,1}(\Omega) \cap H_0^1(\Omega)$ and $\Delta y + u \in L^\infty(\Omega)$.*
- (Structure of the Active and the Inactive Set)
  1. *the collections $\{\mathcal{I}_i\}$, $\{\mathcal{A}_i\}$, $\{\mathcal{A}_i^\circ\}$, $\{\mathcal{B}_i\}$ are finite,*
  2. *the components $\mathcal{A}_i^\circ$ and $\mathcal{I}_i$ are Lipschitz domains for all i,*
  3. *the components $\mathcal{A}_i$ and $\mathcal{B}_i$ are Lipschitz connected for all i,*
  4. *the set $\overline{\mathcal{B}^\circ} \setminus \mathcal{B}^\circ$ is finite and $\mathcal{B} = \overline{\mathcal{B}^\circ} \cup \partial\Omega$.*
- (Well-Behavedness of the Normalized Gradient Field) *There exist a function $\omega \in C^{0,1}(\Omega)$, a constant $C > 0$, and an open set $D \subseteq \mathbb{R}^2$ with $\mathcal{A} \cup \partial\Omega \subseteq D$ and*

$$\omega = 0 \text{ on } \mathcal{A} \cup \partial\Omega, \quad \operatorname{dist}(\cdot, \mathcal{A} \cup \partial\Omega) \le C\omega \text{ a.e. in } \mathcal{I} \cap D,$$

$$\left(\frac{\nabla y^\perp}{|\nabla y|} \cdot \frac{\nabla \omega}{|\nabla \omega|} \right)^2 \le C|\nabla y| \text{ a.e. in } \mathcal{I} \cap D,$$

where, here and in all what follows, $(a, b)^\perp = (b, -a)$ for $a, b \in \mathbb{R}$.

Under these assumptions, the solution operator S of (3.13) is directionally differentiable at u in every direction $h \in H^{-1}(\Omega)$, and the directional derivative

$\delta := S'(u;h)$ is uniquely characterized by the following VI of the first kind:

$$\delta \in \mathcal{K}(y),$$
$$\int_\Omega \nabla\delta\cdot\nabla(v-\delta)\mathrm{d}x + \int_{\mathcal{I}} \frac{(\nabla y^\perp\cdot\nabla\delta)(\nabla y^\perp\cdot\nabla(v-\delta))}{|\nabla y|^3}\,\mathrm{d}x \geq \langle h, v-\delta\rangle \quad \forall\, v\in\mathcal{K}(y),$$

where the convex cone $\mathcal{K}(y)$ is given by

$$\mathcal{K}(y) := \left\{ v\in H_0^1(\Omega) : \int_{\mathcal{I}} \frac{(\nabla y^\perp\cdot\nabla v)^2}{|\nabla y|^3}\,\mathrm{d}x < \infty,\ |\nabla v| = \lambda\cdot\nabla v \text{ a.e. in } \mathcal{A}\right\}, \tag{3.14}$$

where $\lambda\in L^\infty(\Omega;\mathbb{R}^2)$ is any element of $\partial\|.\|_{L^1(\Omega;\mathbb{R}^2)}(\nabla y)$ that satisfies $\operatorname{div}\lambda = u+\Delta y$ *in* $H^{-1}(\Omega)$.

Remark 3.12 Some words concerning the above proposition are in order. First of all, the existence of λ simply follows from the reformulation of the VI in (3.13) by means of the chain rule for convex subdifferentials, and it is easily shown that, for every such λ, the set $\mathcal{K}(y)$ is the same. Moreover, while the regularity assumptions as well as the structural assumptions on the active and inactive sets can directly be checked, if a solution y is given, the intrinsic third assumption is hard to verify in practice, see [3, Section 5.1.5] for details. Finally, for the notion of Lipschitz connected sets, we refer to [3, Definition 5.1.24].

Similarly to the previous examples, the optimal control problem associated with (3.13) reads

$$\left.\begin{array}{ll} \min & J(y,u)\\ \text{s.t.} & (y,u)\in H_0^1(\Omega)\times L^2(\Omega),\\ & (y,u) \text{ satisfy (3.13)},\end{array}\right\} \tag{$\mathrm{P}_{\mathrm{moso}}$}$$

where $J : H_0^1(\Omega)\times L^2(\Omega)\to\mathbb{R}$ is a given Fréchet-differentiable objective. Again, we set $U_{\mathrm{ad}} = L^2(\Omega)$ so that the constraint qualification in Assumption 2.4 is automatically fulfilled. As before, we consider a fixed but arbitrary local minimizer $\bar u\in L^2(\Omega)$ with associated state $\bar y\in H_0^1(\Omega)$. This time we set

$$U := L^2(\Omega),\quad Y := H_0^1(\Omega),\quad V_{\bar y} := \left\{ v\in H_0^1(\Omega) : \int_{\mathcal{I}} \frac{(\nabla\bar y^\perp\cdot\nabla v)^2}{|\nabla\bar y|^3}\,\mathrm{d}x<\infty\right\},$$
$$\langle A(\bar y)w, v\rangle := \int_\Omega \nabla w\cdot\nabla v\,\mathrm{d}x + \int_{\mathcal{I}} \frac{(\nabla\bar y^\perp\cdot\nabla w)(\nabla\bar y^\perp\cdot\nabla v)}{|\nabla\bar y|^3}\,\mathrm{d}x,\quad v,w\in V_{\bar y},$$

and $\mathcal{K}(\bar y)$ as defined in (3.14) (with $y=\bar y$ and λ associated with $\bar y$). As in case of the lasso problem, $V_{\bar y}$ becomes a Hilbert space if endowed with the scalar product

$(v, w)_{V_{\bar{y}}} := \langle A(\bar{y})v, w\rangle$ so that $A(\bar{y})$ is automatically strongly monotone and $\mathcal{K}(\bar{y})$ is closed in this space. Hence, we can again apply the general theory, which results in the following:

Theorem 3.13

1. *Let $\bar{u} \in L^2(\Omega)$ be locally optimal for (P_{moso}), and assume that the associated state $\bar{y} = S(\bar{u})$ satisfies the assumptions in Proposition 3.11 (with $y = \bar{y}$). Then, there exist an adjoint state $p \in V_{\bar{y}}$ and a multiplier $\mu \in V_{\bar{y}}^*$ such that the following optimality conditions are fulfilled:*

$$p + \partial_u J(\bar{y}, \bar{u}) = 0 \quad \text{a.e. in } \Omega, \tag{3.15a}$$

$$\left.\begin{aligned} \int_\Omega \nabla p \cdot \nabla v \,\mathrm{d}x + \int_{\mathcal{I}} \frac{(\nabla \bar{y}^\perp \cdot \nabla p)(\nabla \bar{y}^\perp \cdot \nabla v)}{|\nabla \bar{y}|^3} \,\mathrm{d}x \\ = \langle \partial_y J(\bar{y}, \bar{u}), v\rangle_{H_0^1(\Omega)} - \langle \mu, v\rangle_{V_{\bar{y}}}, \quad \forall\, v \in V_{\bar{y}}, \end{aligned}\right\} \tag{3.15b}$$

$$p \in \mathcal{K}(\bar{y}), \quad \langle \mu, v\rangle_{V_{\bar{y}}} \geq 0 \quad \forall\, v \in \mathcal{K}(\bar{y}). \tag{3.15c}$$

2. *If $\bar{u} \in L^2(\Omega)$ is such that $\bar{y} = S(\bar{u})$ fulfills the assumptions in Proposition 3.11 and there exist $p \in V_{\bar{y}}$ and $\mu \in V_{\bar{y}}^*$, with (3.15), then $\bar{u}$ is B-stationary for (P_{moso}).*

4 Conclusion

Within this chapter, we constructed a general framework for the derivation of strong stationarity conditions for optimal control problems governed by VIs. Moreover, we demonstrated by means of application-driven examples that our general analysis also applies in case of VIs of the second kind. However, as our two last examples show, sometimes additional assumptions, which may even be hard to verify, are necessary to guarantee that the control-to-state map associated with the VI under consideration is directionally differentiable and the directional derivatives possess the desired structure. Under these assumptions, though, our general framework is applicable and yields stationarity conditions, which are the most rigorous possible ones.

It is however an open question how to solve these strong stationarity systems numerically. The reason is that, in neither of our four examples, the adjoint equation together with the (generalized) sign conditions on the adjoint state and the multiplier μ forms a VI or a complementarity system, which would be amenable for numerical computations. Even worse, as the investigations on optimal control of non-smooth PDEs in [4] show, strong stationarity systems may be potentially overdetermined.

The construction of algorithms for the reliable numerical computation of strongly stationary points is therefore a field of future research.

Acknowledgments The research of this work was carried out in Project P16 (Optimal Control of Variational Inequalities of the Second Kind with Application to Yield Stress Fluids) within the DFG Priority Program SPP 1962 (Non-smooth and Complementarity-Based Distributed Parameter Systems: Simulation and Hierarchical Optimization). The support by the DFG is gratefully acknowledged.

Appendix A: Auxiliary Results

Lemma A.1 *Under Assumption 2.3, U is dense in $V_{\bar{y}}^*$.*

Proof Let us assume that U is not a dense subset of $V_{\bar{y}}^*$ so that there exists a $g \in V_{\bar{y}}^* \setminus \overline{U}^{V_{\bar{y}}^*}$. Then, the strict separation theorem in combination with the reflexivity of $V_{\bar{y}}$ implies the existence of a $v \in V_{\bar{y}}$, $v \neq 0$, such that

$$(h, v)_U = 0 < \langle g, v \rangle_{V_{\bar{y}}} \quad \forall\, h \in U.$$

Since $V_{\bar{y}} \hookrightarrow U$ and the embedding is injective, this yields $v = 0$, which is a contradiction. □

References

1. T. Betz, *Optimal control of two variational inequalities arising in solid mechanics*. PhD thesis, Technische Universität Dortmund, 2015.
2. J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*. Springer, 2000.
3. C. Christof, *Sensitivity Analysis of Elliptic Variational Inequalities of the First and the Second Kind*. PhD thesis, Technische Universität Dortmund, 2018.
4. C. Christof, C. Clason, C. Meyer, and S. Walther, *Optimal control of a non-smooth semilinear elliptic equation*. Math. Control and Related Fields **8** (2018), 247–276.
5. C. Christof and C. Meyer, *Sensitivity analysis for a class of H_0^1-elliptic variational inequalities of the second kind*. Set-Valued and Variational Analysis **27** (2019), 469–502.
6. C. Christof and G. Wachsmuth, *On the Non-Polyhedricity of Sets with Upper and Lower Bounds in Dual Spaces*. GAMM Reports **40** (2018), 339–350.
7. J. C. de los Reyes, *Optimization of mixed variational inequalities arising in flow of viscoplastic materials*. COAP **52** (2012), 757–784.
8. J. C. de los Reyes, *On the optimal control of some nonsmooth distributed parameter systems arising in mechanics*. GAMM Reports **40** (2017), 268–286.
9. J. C. de los Reyes and C. Meyer, *Strong stationarity conditions for a class of optimization problems governed by variational inequalities of the second kind*. J. Optim. Theory Appl. **168** (2016), 375–409.
10. L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*. CRC Press, 2015.

11. W. Han and B. D. Reddy, *Plasticity: Mathematical Theory and Numerical Analysis*. Springer, 1999.
12. A. Haraux, *How to differentiate the projection on a convex set in Hilbert space. Some applications to variational inequalities*. J. Math. Soc. Japan **29** (1977), 615–631.
13. R. Herzog and C. Meyer, *Optimal control of static plasticity with linear kinematic hardening*. ZAMM **91** (2011), 777–794.
14. R. Herzog, C. Meyer and G. Wachsmuth, *B- and strong stationarity for optimal control of static plasticity with hardening*. SIAM J. Optim. **23** (2013), 321–352.
15. M. Hintermüller and I. Kopacka, *Mathematical programs with complementarity constraints in function space: C- and strong stationarity and a path-following algorithm*. SIAM Journal on Optimization **20** (2009), 86–902.
16. M. Hintermüller and T. Surowiec, *First-Order Optimality Conditions for Elliptic Mathematical Programs with Equilibrium Constraints via Variational Analysis*. SIAM Journal on Optimization **21** (2011), 1561–1593.
17. M. Hintermüller and T. Surowiec, *On the Directional Differentiability of the Solution Mapping for a Class of Variational Inequalities of the Second Kind*. Set-Valued and Variational Analysis **26** (2018), 631–642.
18. C. Meyer and D. Wachsmuth, *Strong stationarity is not a necessary optimality condition for boundary control of the obstacle problem*. Preprint Nr. 327, Ergebnisberichte des Instituts für Angewandte Mathematik, Universität Würzburg, 2014.
19. F. Mignot, *Contrôle dans les inéquations variationelles elliptiques*. J. Funct. Anal. **22** (1976), 130–185.
20. F. Mignot and J. P. Puel, *Optimal control in some variational inequalities*. SIAM J. Control Optim. **22** (1984), 466–476.
21. P. P. Mosolov and V. P. Miasnikov, *Variational methods in the theory of the fluidity of a viscousplastic medium*. J. Appl. Math. Mech. **29** (1965), 545–577.
22. B. Schweizer, *Partielle Differentialgleichungen*. Springer, 2013.
23. J. Sokołowski, *Sensitivity analysis of contact problems with prescribed friction*. Appl. Math. Optim. **18** (1988), 99–117.
24. R. Tibshirani, *Regression Analysis and Selection via the Lasso*. Royal Statistical Society Series **58** (1996) 267–288.
25. G. Wachsmuth, *Strong Stationarity for Optimal Control of the Obstacle Problem with Control Constraints*. SIAM Journal on Optimization **24** (2014), 1914–1932.
26. G. Wachsmuth, *Mathematical Programs with Complementarity Constraints in Banach Spaces*. Journal of Optimization Theory and Applications **166** (2015), 480–507.
27. G. Wachsmuth, *Strong stationarity for optimization problems with complementarity constraints in absence of polyhedricity*. Set-Valued Var. Anal. **25** (2017), 133–175.
28. G. Wachsmuth, *A guided tour of polyhedric sets*. Journal of Convex Analysis **26** (2019), 153–188.
29. I. Yousept, *Hyperbolic Maxwell variational inequalities for Bean's critical-state model in type-II superconductivity*. SIAM J. Numer. Anal. **55** (2017), 2444–2464.

Optimizing Fracture Propagation Using a Phase-Field Approach

Andreas Hehl, Masoumeh Mohammadi, Ira Neitzel, and Winnifried Wollner

Abstract We consider an optimal control problem of tracking type governed by a time-discrete phase-field fracture or damage, respectively, propagation model. Pointwise inequality constraints on the phase field, which model an irreversibility condition for the fracture growth, are first regularized by a smooth regularization term, removing the inequality constraints from the lower level problem and resulting in an Euler–Lagrange equation as optimality condition for the lower level problem. We take the regularization limit in the first-order optimality conditions and prove convergence of first order necessary points of the regularized control problem to certain limits satisfying an optimality system of a limit problem governed by a variational inequality. Moreover, SQP methods for the regularized problem and its limit are analyzed with respect to solvability of the subproblems. In the case of convergence, it is proven that the limit is a first-order necessary point of the respective problem. Finally, the finite element discretization and its convergence for the linear quadratic SQP subproblems are discussed.

Keywords Optimal control · Regularized fracture model · Phase field · Regularization limit · SQP methods · Finite element discretization

Mathematics Subject Classification (2020) Primary 49M15; Secondary 49M25, 74R10

The authors gratefully acknowledge the funding by the Deutsche Forschungsgemeinschaft (DFG)—Projektnummer 314067056; Project P17 in Priority Programme 1962

A. Hehl · I. Neitzel
Institut für Numerische Simulation, Bonn, Germany
e-mail: hehl@ins.uni-bonn.de; neitzel@ins.uni-bonn.de

M. Mohammadi · W. Wollner (✉)
Technische Universität Darmstadt, Darmstadt, Germany
e-mail: mohammadi@mathematik.tu-darmstadt.de; wollner@mathematik.tu-darmstadt.de

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_13

1 Introduction

In this survey, we consider the problem investigated in [18, 19], i.e., an optimal control problem of tracking type for fracture or damage propagation, which is governed by the Euler–Lagrange equations of a regularized fracture propagation problem modeled by a phase-field approach, which has first been proposed in [3, 4, 7]. This approach, which we will explain in more detail in Sect. 2, allows to treat almost arbitrary fracture paths, as opposed to the first results on the control of fractures with prescribed path [14] or fixed length [12]. In a first step, to circumvent the difficulties posed by a pointwise irreversibility condition, this inequality constraint was regularized by a smooth penalization term. In [18], the existence of global solutions as well as first-order KKT-like necessary optimality conditions under a regularity assumption was proven for the regularized problem. Constraint violation estimates as well as convergence of solutions with respect to taking this penalization parameter to its limit were then shown in [19] for a problem formulation with viscous regularization corresponding to a time-step restriction in the spatially continuous but time-discrete model problem, cf. [13].

In Sect. 2, we will give a precise description of the model problem under consideration. Then, building on the results from [19], we will prove convergence of the optimality systems with respect to taking the limit in the penalty parameter in Sect. 3. To elaborate, let us point out that the (uncontrolled) fracture propagation problem is itself an energy minimization problem, the so-called lower level problem. Adding an outer optimal control problem leads to a bi-level optimization problem, where the lower level problem is usually replaced by its first-order necessary conditions. In case of the regularized problem, this is a system of Euler–Lagrange equations, so that the control problem resembles a PDE-constrained optimization problem without inequality constraints but a quasilinear PDE constraint. We formulate the optimality conditions for this problem and then derive a system satisfied by certain limit points when the penalization parameter tends to infinity. In addition to the convergence results for the primal variables, i.e., control, displacement, and phase field toward a solution of the unregularized problem, we now obtain a limit optimality system, which exhibits the presence of a variational inequality as constraint of the outer optimal control problem.

In Sect. 4, we formulate the method of sequential quadratic programming (SQP) for the regularized problem formulation and show that the limit point of convergent sequences produced by the SQP method actually satisfies the first-order optimality system for the regularized problem. The results are combined with convergence results for the finite element discretization of a linearized fracture control problem, such as the SQP subproblems, from [17]. A key ingredient for obtaining a priori error estimates is an improved regularity of the solutions giving a gap between the norm in which the error is calculated and the regularity of the approximated function. Such estimates have been shown only recently in [10].

Eventually, in Sect. 5, an SQP method for the unregularized problem is formulated, and convergence of the finite element method for the quadratic subproblems is derived.

2 Problem Setting

We consider the problem investigated in [18, 19], i.e., an optimal control problem of tracking type for fracture propagation, which is governed by the Euler–Lagrange equations of a regularized fracture propagation problem modeled by a phase-field approach.

Before presenting the precise control problem, let us elaborate on the fracture propagation problem, which we want to control. The model goes back to Griffith's model of brittle fracture [8] or, more precisely, a variational formulation by Francfort and Marigo in [7]. Within this model, fracture propagation occurs when the elastic energy restitution rate reaches a critical value $G_{\mathcal{C}}$, leading to a minimization problem where the total energy

$$E(u, \mathcal{C}) = \frac{1}{2} (\mathbb{C}e(u), e(u))_{\Omega \setminus \mathcal{C}} - (\tau, u)_{\partial_N \Omega} + G_{\mathcal{C}} \mathcal{H}^{d-1}(\mathcal{C})$$

is to be minimized. Here, u denotes a vector-valued displacement field, $\mathcal{C}$ denotes the crack, assumed to be compactly contained in the domain Ω without reaching the boundary, τ is a force applied to the part $\partial_N \Omega$ of the boundary, which will later be our optimization variable in the optimal control problem, and $\mathcal{H}^{d-1}$ is the $d-1$ dimensional Hausdorff measure, when $d \in \{2, 3\}$ denotes the dimension of Ω. By means of $\mathbb{C}$ and $e(u)$, a linear elasticity model is described.

This energy functional is to be minimized with respect to all kinematically admissible displacements u, and any fracture set satisfying a fracture growth condition, making sure that once a fracture has appeared it does not close again. To avoid the difficulty introduced by the Hausdorff measure, we use a regularization proposed by Bourdin et al. [3, 4]. Precisely, we introduce a time-dependent phase-field variable φ, defined on $\Omega \times (0, T)$, where $\varphi = 1$ describes non-fractured regions, and $\varphi = 0$ fractured regions, with a smooth transition. Using such an Ambrosio–Tortorelli regularization, cf. [1, 2], of the fracture function leads to a regularized energy functional to be minimized:

$$E_\varepsilon(u, \varphi) = \frac{1}{2} \Big(((1-\kappa)\varphi^2 + \kappa)\mathbb{C}e(u), e(u) \Big) - (\tau, u)_{\partial_N \Omega} + G_{\mathcal{C}} \left(\frac{1}{2\varepsilon} \|1 - \varphi\|^2 + \frac{\varepsilon}{2} \|\nabla \varphi\|^2 \right), \tag{2.1}$$

where ε is a positive parameter, which, when sent to zero, leads to the Hausdorff measure in the sense of a Γ-limit, and $\kappa = o(\varepsilon)$ is a positive parameter used to avoid degeneracy of the energy function when $\varphi = 0$.

The critical issue here is that the energy functional is not convex in both solution variables simultaneously, but only in each single variable when the other one is fixed. While the forward model problem is relatively well studied, the control of fractures remains to pose a lot of challenges. Controlling this energy functional would lead to a bi-level minimization problem in function spaces, where the lower level problem is nonconvex and subject to additional inequality constraints

$$\varphi(t_2) \leq \varphi(t_1) \quad \forall\, t_1 \leq t_2, \tag{2.2}$$

describing the irreversibility condition. Fixing the spatial dimension $d = 2$, we obtain for a control space Q, to be specified later, the following problem formulation:

$$\begin{aligned} \min_{q,\mathbf{u}} J(q,\mathbf{u}) &:= \frac{1}{2}\|u - u_d\|^2_{L^2(\Omega;\mathbb{R}^2)} + \frac{\alpha}{2}\|q\|^2_Q \\ &\text{subject to } \mathbf{u} \text{ solves (2.1) given } \tau = q \\ &\text{as well as (2.2).} \end{aligned} \tag{2.3}$$

We tackle the difficulties by the following adaptations to the model problem:

- We will consider a time-discrete but spatially continuous problem formulation.
- We regularize the inequality constraints in the lower level problem by a penalty approach introduced by Meyer, Rademacher, and Wollner, [15], i.e., adding a term $\frac{\gamma}{4}\|\max(0, \varphi_i - \varphi_{i-1})\|^4_{L^4}$ when φ_i denotes the value of the phase field at time t_i. One of our main goals is then to analyze the problem with respect to considering the limit $\gamma \to \infty$.
- We will follow standard procedure and replace the lower level problem by its Euler–Lagrange equations, leading to a PDE-constrained optimization problem with quasilinear PDE.
- For some of our results, it proved helpful to introduce a further viscous regularization of the energy functional with a regularization parameter $\eta \geq 0$, see below.

2.1 Model Problem, Notation, and Assumptions

Following the before-mentioned steps, we arrive at the following short description of the model problem. Note that while the original problem formulation is spatially continuous but time discrete, for simplicity of notation we consider only one time-step of the fracture evolution, giving us the regularized optimization problem for

finding $q_\gamma \in Q$ and $\mathbf{u}_\gamma = (u_\gamma, \varphi_\gamma) \in V$ solving

$$\begin{aligned} \min_{q_\gamma, \mathbf{u}_\gamma} \; & J(q_\gamma, \mathbf{u}_\gamma) := \frac{1}{2}\|u_\gamma - u_d\|^2_{L^2(\Omega;\mathbb{R}^2)} + \frac{\alpha}{2}\|q_\gamma\|^2_Q \\ & \text{subject to } A(\mathbf{u}_\gamma) + R(\gamma;\varphi) = B(q_\gamma). \end{aligned} \tag{NLP$^\gamma$}$$

Here, for spaces to be defined below, $A\colon W^{1,p}(\Omega;\mathbb{R}^2) \times W^{1,p}(\Omega) \subset V \to V^*$ denotes a nonlinear phase-field operator, $R\colon V_\varphi \to V_\varphi^*$ is a regularization operator penalizing deviation from an irreversibility condition for the fracture growth, and $B\colon Q \to V^*$ is the control-action operator. They are defined by

$$\begin{aligned} \langle A(\mathbf{u}), \mathbf{v})\rangle :&= \Big(g(\varphi)\mathbb{C}e(u), e(v^u)\Big) + \varepsilon(\nabla\varphi, \nabla v^\varphi) - \frac{1}{\varepsilon}(1-\varphi, v^\varphi) \\ &\quad + \eta(\varphi - \varphi^-, v^\varphi) + (1-\kappa)(\varphi\mathbb{C}e(u) : e(u), v^\varphi), \\ \langle R(\gamma;\varphi), v^\varphi\rangle :&= \gamma([(\varphi - \varphi^-)^+]^3, v^\varphi), \\ \langle Bq, (v^u, v^\varphi)\rangle :&= (q, v^u)_Q \end{aligned}$$

for any $\mathbf{v} = (v^u, v^\varphi) \in V$. Here, g is given as

$$g(x) := (1-\kappa)x^2 + \kappa,$$

$\kappa, \varepsilon, \gamma > 0$ are given parameters as explained above, and $\eta \geq 0$ is an additional parameter which can be regarded as a viscous regularization, cf. [13]. Choosing η sufficiently large serves two purposes: on the one hand, it makes the lower level energy function strictly convex and hence uniquely solvable. On the other hand, this helps to include damage problems in addition to pure fracture. It also corresponds to choosing a sufficiently small time-step in the temporal discretization of the problem. φ^- is the given initial phase field, and $\mathbb{C}$ is the rank-4 elasticity tensor with the usual properties. The problem is defined on a polygonal domain $\Omega \subset \mathbb{R}^2$ with boundary $\partial\Omega = \Gamma \dot\cup \Gamma_D$, such that the union $\Omega \cup \Gamma$ is Gröger regular [9].

Note that the term $\eta(\varphi - \varphi^-, v^\varphi)$ corresponds to a viscous regularization of the problem for $\eta > 0$ that can also be interpreted as a restriction on the time-step in the temporal discretization of the problem, cf. [13]. In [18], a setting with $\eta = 0$ was considered, requiring additional assumptions due to a lack of convexity.

We define

$$\begin{aligned} V_u &= H^1_D(\Omega;\mathbb{R}^2) := \{v \in H^1(\Omega;\mathbb{R}^2) \mid v = 0 \text{ on } \Gamma_D\}, \\ V_\varphi &= H^1(\Omega), \\ V &= V_u \times V_\varphi, \\ Q &= L^2(\Gamma) \end{aligned}$$

and denote the respective dual spaces with a superscript *, e.g., V^*. For spaces such as $W^{s,p}$ and $H^s = W^{s,2}$, we understand that they are defined on the domain Ω unless otherwise stated. We will further use the following notation for the scalar product/norm: $(\cdot,\cdot)$ denotes the usual L^2 scalar product with corresponding norm $\|\cdot\|$, and $(\cdot,\cdot)_Q$ corresponds to the scalar product of Q. In addition, $\langle\cdot,\cdot\rangle$ stands for a duality pairing where the spaces are omitted if obvious from the context. In what follows, we also define

$$W = W_u \times W_\varphi = W_D^{1,p} \cap H^{1+s} \times W^{2,q}$$

and the corresponding space for the right-hand side of the equation

$$W^\times = W^{-1,p} \cap H^{-1+s} \times L^q,$$

where $W^{-1,p} = (W^{1,p'})^*$ and $H^{-1+s} = (H^{1-s})^*$ are the respective dual spaces. As common, for any $p \in [1,\infty]$, we denote the dual exponent by p', i.e., $\frac{1}{p} + \frac{1}{p'} = 1$. Finally, we will implicitly rely on the following standing assumptions: for the parameters p, q, and s, we require

$$p > 2, \quad q = p/2 > 1, \quad \text{and} \quad s \in (0, 1/2).$$

Furthermore, we assume that p and s are chosen such that $H^{1+s} \subset W^{1,p}$.

With this notation, we point out that taking the limit $\gamma \to \infty$ in (NLP$^\gamma$) yields the MPCC

$$\min_{q,\mathbf{u}} J(q,\mathbf{u}) \qquad \text{subject to} \begin{cases} A(\mathbf{u}) + \lambda = B(q) & \text{in } V^*, \\ \lambda \geq 0 & \text{in } V_\varphi^*, \\ \varphi \leq \varphi^- & \text{a.e. in } \Omega, \\ \langle \lambda, \varphi - \varphi^- \rangle = 0, \end{cases} \tag{NLP$^{\text{VI}}$}$$

where $\lambda \in V_\varphi^*$, and we implicitly use the natural embedding $V_\varphi^* \ni \lambda \mapsto (0,\lambda) \in V^*$.

We remark that if $\varphi^- \in W_\varphi$ and $q_\gamma, q \in Q$, by [10, Section 7], the solutions $\mathbf{u}_\gamma$ for the equality constraint in (NLP$^\gamma$) and $\mathbf{u}$ for the constraints in (NLP$^{\text{VI}}$) satisfy $\mathbf{u}_\gamma, \mathbf{u} \in W$ for some $p > 2$.

2.2 The Phase-Field Equation

This section provides a short analysis of the linearized operators $A'(\mathbf{u})\colon V \to V^*$ and $R'(\gamma,\varphi)\colon V_\varphi \to V_\varphi^*$, which, for $\mathbf{u} \in V \cap W$, are defined via

$$\begin{aligned}\langle A'(\mathbf{u})\mathbf{d}^{\mathbf{u}}, \mathbf{v}\rangle := &\Big(g(\varphi)\mathbb{C}e(d^u), e(v^u)\Big) + 2(1-\kappa)(\varphi\mathbb{C}e(u) : e(d^u), v^\varphi)\\ &+ \varepsilon(\nabla d^\varphi, \nabla v^\varphi) + \frac{1}{\varepsilon}(d^\varphi, v^\varphi) + \eta(d^\varphi, v^\varphi)\\ &+ (1-\kappa)(d^\varphi\mathbb{C}e(u) : e(u), v^\varphi) + 2(1-\kappa)(\varphi\mathbb{C}e(u)d^\varphi, e(v^u)),\end{aligned} \tag{2.4}$$

$$\langle R'(\gamma;\varphi)d^\varphi, v^\varphi\rangle := 3\gamma([(\varphi-\varphi^-)^+]^2 d^\varphi, v^\varphi),$$

for any $\mathbf{v} = (v^u, v^\varphi)$, introducing the notation $\mathbf{d}^{\mathbf{u}} := (d^u, d^\varphi)$.

A quick calculation shows two properties of importance for the following calculations: firstly, coercivity of A', i.e., for $\eta \geq 0$ sufficiently large, there exists a $\beta_\eta > 0$ such that

$$\langle (A'(\mathbf{u}))\mathbf{v}, \mathbf{v}\rangle \geq \beta_\eta \|\mathbf{v}\|_V^2 \quad \forall\, \mathbf{v} \in V, \tag{2.5}$$

and, secondly, the following non-negativity statement for R' for all $v^\varphi \in V_\varphi$:

$$\langle R'(\gamma;\varphi)v^\varphi, v^\varphi\rangle = 3\gamma([(\varphi-\varphi^-)^+]^2 v^\varphi, v^\varphi) \geq 0. \tag{2.6}$$

Following the results of [10], $A'(\mathbf{u})\colon V \mapsto V^*$ is well defined and an isomorphism if $\eta \geq 0$ is sufficiently large. In particular, for $\mathbf{u} \in V \cap W$, the operator

$$\mathbf{d}^{\mathbf{u}} \mapsto A'(\mathbf{u})\mathbf{d}^{\mathbf{u}} + R'(\gamma,\varphi)d^\varphi\colon V \to V^* \tag{2.7}$$

is invertible. In a similar way to [18, Lemma 5.2], we can establish the following improved regularity result for data in $W^\times \hookrightarrow V^*$.

Proposition 2.1 *Let $\mathbf{u} \in V \cap W$, $\varphi^- \in W_\varphi$, and $b \in W^\times \hookrightarrow V^*$, recalling $p > 2$. Then, the solution $\mathbf{d}^{\mathbf{u}} = (d^u, d^\varphi) \in V$ of*

$$A'(\mathbf{u})\mathbf{d}^{\mathbf{u}} + R'(\gamma,\varphi)d^\varphi = b$$

has improved regularity $\mathbf{d}^{\mathbf{u}} \in V \cap W$.

Furthermore, for regular $\mathbf{u} \in W$, we can define the second derivative operators $A''(\mathbf{u})\colon V \times V \to (V \cap W^{1,p})^*$ and $R''(\gamma,\varphi)\colon V_\varphi \times V_\varphi \to V_\varphi^*$ by

$$\begin{aligned}\langle A''(\mathbf{u})[\mathbf{d}_1^{\mathbf{u}}, \mathbf{d}_2^{\mathbf{u}}], \mathbf{v}\rangle &= 2(1-\kappa)(d_2^\varphi \mathbb{C}e(u)d_1^\varphi, e(v^u)) \\ &\quad + 2(1-\kappa)(d_2^\varphi \mathbb{C}e(d_1^u)\varphi, e(v^u)) \\ &\quad + 2(1-\kappa)(d_2^\varphi \mathbb{C}e(u) : e(d_1^u), v^\varphi) \\ &\quad + 2(1-\kappa)(\varphi \mathbb{C}e(d_2^u)d_1^\varphi, e(v^u)) \\ &\quad + 2(1-\kappa)(d_1^\varphi \mathbb{C}e(d_2^u) : e(u), v^\varphi) \\ &\quad + 2(1-\kappa)(\varphi \mathbb{C}e(d_2^u) : e(d_1^u), v^\varphi), \\ \langle R''(\gamma;\varphi)[d_1^\varphi, d_2^\varphi], v^\varphi\rangle &= 6\gamma([(\varphi-\varphi^-)^+]d_1^\varphi d_2^\varphi, v^\varphi).\end{aligned}$$

We note that for regular data $\mathbf{u}$, $\mathbf{v}$, the second derivatives are continuous on V in the following sense:

Lemma 2.2 *Let $\mathbf{u}, \mathbf{v} \in V \cap W^{1,p}$ be given. Then, there exists a constant c depending on $\|\mathbf{u}\|_{1,p}$, $\|\mathbf{v}\|_{1,p}$ such that*

$$|\langle A''(\mathbf{u})[\mathbf{d}_1^{\mathbf{u}}, \mathbf{d}_2^{\mathbf{u}}], \mathbf{v}\rangle| \le c\|\mathbf{d}_1^{\mathbf{u}}\|_V \|\mathbf{d}_2^{\mathbf{u}}\|_V.$$

Analog estimates hold if any two of the four variables, $\mathbf{u}$, $\mathbf{v}$, $\mathbf{d}_1^{\mathbf{u}}$, and $\mathbf{d}_2^{\mathbf{u}}$ are in $V \cap W^{1,p}$.

Following the regularity results of [10], any solution $\mathbf{u}_\gamma$ to the equation in (NLP$^\gamma$) satisfies the additional regularity $\mathbf{u}_\gamma \in W$, and thus $A'(\mathbf{u})$ and $A''(\mathbf{u})$ are well defined for all points $\mathbf{u}$ of the same regularity. Furthermore, by (2.5), $A'(\mathbf{u})$ is an isomorphism if η is sufficiently large.

To simplify the following arguments, we make the following assumption:

Assumption 2.3 Let $\eta \ge 0$ be chosen such that $A'(\mathbf{u})\colon V \mapsto V^*$ is coercive, i.e., (2.5) holds.

3 The Limiting First-Order Necessary Conditions

We see that for a local minimizer $(q_\gamma, \mathbf{u}_\gamma)$ of (NLP$^\gamma$), there exists $\mathbf{z}_\gamma \in V$, $\lambda_\gamma, \mu_\gamma \in V_\varphi^*$, $\theta_\gamma \in V_\varphi$ such that the following system is satisfied:

$$\begin{aligned}A(\mathbf{u}_\gamma) + \lambda_\gamma &= Bq_\gamma && \text{in } V^*, \\ \lambda_\gamma &= R(\gamma;\varphi_\gamma) && \text{in } V_\varphi^*,\end{aligned}$$

$$\begin{aligned}
(A'(\mathbf{u}_\gamma))^*\mathbf{z}_\gamma &= u_\gamma - u_d - \mu_\gamma && \text{in } V^*,\\
B^*\mathbf{z}_\gamma + \alpha q_\gamma &= 0 && \text{in } V^*,\\
z_\gamma^\varphi - \theta_\gamma &= 0 && \text{in } V_\varphi,\\
\mu_\gamma - R'(\gamma;\varphi_\gamma)\theta_\gamma &= 0 && \text{in } V_\varphi^*.
\end{aligned} \tag{FON$^\gamma$}$$

Clearly, the variables λ_γ, μ_γ, and θ_γ can easily be eliminated, but they are useful as separate quantities as they have a meaning as multipliers for the limit, cf. [15] as well as (FON$^{\mathrm{VI}}$). Moreover, improved regularity for $\mathbf{u}_\gamma \in W$ and $\lambda_\gamma \in L^q(\Omega)$ holds as remarked at the end of Sect. 2.

We will see that certain limits $(\bar q, \bar{\mathbf{u}}, \bar\lambda, \bar\theta, \bar\mu)$ of first order necessary points of (NLP$^\gamma$) satisfy the system (C-stationarity)

$$\begin{aligned}
A(\bar{\mathbf{u}}) + \bar\lambda &= B\bar q && \text{in } V^*,\\
\bar\lambda &\geq 0 && \text{in } V_\varphi^*,\\
\bar\varphi &\leq \varphi^- && \text{a.e. in } \Omega,\\
\langle \bar\lambda, \bar\varphi - \varphi^-\rangle &= 0,\\
(A'(\bar{\mathbf{u}}))^*\bar{\mathbf{z}} &= \bar u - u_d - \bar\mu && \text{in } V^*,\\
B^*\bar{\mathbf{z}} + \alpha\bar q &= 0 && \text{in } V^*,\\
\bar z^\varphi - \bar\theta &= 0 && \text{in } V_\varphi,\\
\langle \bar\theta, \bar\lambda\rangle &= 0,\\
\langle \bar\mu, \bar\varphi - \varphi^-\rangle &= 0,\\
\langle \bar\theta, \bar\mu\rangle &\geq 0.
\end{aligned} \tag{FON$^{\mathrm{VI}}$}$$

Indeed, the following theorem holds:

Theorem 3.1 *Let $q_\gamma \to \bar q$ be a convergent sequence of local minimizers of* (NLP$^\gamma$) *for $\gamma \to \infty$. Then, up to selecting a subsequence, the following convergence*

$$\begin{aligned}
\mathbf{u}_\gamma &\to \bar{\mathbf{u}} && \text{in } V,\\
u_\gamma &\to \bar u && \text{in } W_u,\\
\varphi_\gamma &\rightharpoonup \bar\varphi && \text{in } W_\varphi,\\
\lambda_\gamma &\to \bar\lambda && \text{in } V_\varphi^*,\\
\mathbf{z}_\gamma &\rightharpoonup \bar{\mathbf{z}} && \text{in } V,\\
\mu_\gamma &\rightharpoonup \bar\mu && \text{in } V_\varphi^*,\\
\theta_\gamma &\rightharpoonup \bar\theta && \text{in } V_\varphi
\end{aligned}$$

holds. Furthermore, any such limit satisfies (FON$^{\mathrm{VI}}$).

Proof By [19, Corollary 3.10], we obtain the first three convergence claims as well as the satisfaction of the first four lines of (FON$^{\mathrm{VI}}$). Furthermore, from this convergence, the convergence of λ_γ in V_φ^* follows from the first equation.

It remains to show the convergence of the dual variables and the limits in the adjoint equation, the gradient equation, and the complementary slackness conditions. We start with the weak convergence of $\mathbf{z}_\gamma$ in V. To this end, we replace μ_γ and θ_γ in the equation for $\mathbf{z}_\gamma = (z_\gamma^u, z_\gamma^\varphi)$ in (FON$^\gamma$) and obtain

$$(A'(\mathbf{u}_\gamma))^* \mathbf{z}_\gamma + R'(\gamma, \varphi_\gamma) z_\gamma^\varphi = u_\gamma - u_d.$$

Testing with $\mathbf{z}_\gamma$, we arrive at

$$\langle (A'(\mathbf{u}_\gamma))^* \mathbf{z}_\gamma, \mathbf{z}_\gamma \rangle + \langle R'(\gamma; \varphi_\gamma) z_\gamma^\varphi, z_\gamma^\varphi \rangle = \langle u_\gamma - u_d, z_\gamma^u \rangle, \tag{3.1}$$

and using (2.5) and (2.6), from (3.1), we receive

$$\|\mathbf{z}_\gamma\|_V \leq \frac{1}{\beta_\eta} \|u_\gamma - u_d\|_{H^{-1}}. \tag{3.2}$$

Since u_γ is bounded independently of γ in $W_u \hookrightarrow V_u \hookrightarrow H^{-1}$ as proven in [19, Lemma 3.1], $\mathbf{z}_\gamma$ is bounded in V, and we deduce the existence of a subsequence that converges weakly to some $\bar{\mathbf{z}}$ in V.

Next, we want to take the limit in the adjoint equation, so we have to establish weak convergence of $(A'(\mathbf{u}_\gamma))^* \mathbf{z}_\gamma$ in V^* first. Using the already shown convergence $u_\gamma \to \bar{u}$ in W_u and $\varphi_\gamma \rightharpoonup \bar{\varphi}$ in W_φ, we obtain $A'(\mathbf{u}_\gamma) - A'(\mathbf{u}) \to 0$ in $L(V, V^*)$ showing

$$(A'(\mathbf{u}_\gamma))^* \mathbf{z}_\gamma \rightharpoonup (A'(\bar{\mathbf{u}}))^* \bar{\mathbf{z}} \text{ in } V^*.$$

Since u_γ converges strongly in $W_u \hookrightarrow V_\varphi^*$, the convergence of μ_γ and the limit in the adjoint equation $(A'(\bar{\mathbf{u}}))^* \bar{\mathbf{z}} = \bar{u} - u_d - \bar{\mu}$ in (FON$^{\mathrm{VI}}$) follow by

$$\mu_\gamma = u_\gamma - u_d - (A'(\mathbf{u}_\gamma))^* \mathbf{z}_\gamma \rightharpoonup \bar{u} - u_d - (A'(\bar{\mathbf{u}}))^* \bar{\mathbf{z}} =: \bar{\mu} \text{ in } V_\varphi^*.$$

Since $\theta_\gamma = z_\gamma^\varphi$, the weak convergence of θ_γ to $\bar{\theta} = \bar{z}^\varphi$ is an immediate consequence.

Next, we can pass to the limit in the gradient equation $B^* \mathbf{z}_\gamma + \alpha q_\gamma = 0$ in (FON$^\gamma$), to obtain

$$B^* \bar{\mathbf{z}} + \alpha \bar{q} = 0.$$

We have shown the convergence results for all functions as stated in the theorem and established the limits in the first seven lines of (FON$^{\mathrm{VI}}$). We can now verify

the complementary slackness conditions given by the last three lines of (FON$^{\mathrm{VI}}$). By definition of $\lambda_\gamma := R(\gamma, \varphi_\gamma) = \gamma[(\varphi_\gamma - \varphi^-)^+]^3$ and introducing the set $\mathcal{A} := \{x \in \Omega \,|\, \varphi_\gamma > \varphi^-\}$, we find that

$$|\langle \lambda_\gamma, \theta_\gamma \rangle| = \left| \int_{\mathcal{A}} \gamma[(\varphi_\gamma - \varphi^-)^+]^3 \theta_\gamma \, dx \right| \leq \|\lambda_\gamma\|_{L^q(\mathcal{A})} \|\theta_\gamma\|_{L^{q'}(\mathcal{A})} \tag{3.3}$$

is true. By [19, Lemma 3.8], we know that $\|\lambda_\gamma\|_q \leq C$ holds independently of γ, which implies a uniform bound for $\|\lambda_\gamma\|_{L^q(\mathcal{A})}$. For the second term of (3.3), exploiting (3.2), we receive a uniform bound on the subsequence θ_γ in $H^1 \hookrightarrow L^{q'}$ noting that $q' \in (1, \infty)$. By the convergence result for the primal variables, it has already been proven that $\varphi_\gamma \to \bar{\varphi}$ in V_φ as well as $\bar{\varphi} \leq \varphi^-$, and hence for $\gamma \to \infty$ it holds $|\mathcal{A}| \to 0$ and thus

$$\|\theta_\gamma\|_{L^{q'}(\mathcal{A})} \to 0 \text{ for } \gamma \to 0.$$

So overall, from (3.3), we obtain

$$|\langle \lambda_\gamma, \theta_\gamma \rangle| \leq C \|\theta_\gamma\|_{L^{q'}(\mathcal{A})} \to 0 \text{ for } \gamma \to \infty. \tag{3.4}$$

We already know that λ_γ converges strongly in V_φ^*. In combination with the weak convergence of θ_γ in V_φ, (3.4) yields

$$\langle \bar{\lambda}, \bar{\theta} \rangle = \lim_{\gamma \to \infty} \langle \lambda_\gamma, \theta_\gamma \rangle = 0,$$

which is the third-to-last line of (FON$^{\mathrm{VI}}$).

Next, by definition of $\mu_\gamma := R'(\gamma; \varphi_\gamma)\theta_\gamma = 3\gamma[(\varphi_\gamma - \varphi^-)^+]^2\theta_\gamma$, using (3.4), it holds

$$\begin{aligned} \langle \mu_\gamma, \varphi_\gamma - \varphi^- \rangle &= 3 \int_\Omega \gamma[(\varphi_\gamma - \varphi^-)^+]^2 \theta_\gamma (\varphi_\gamma - \varphi^-) \, dx \\ &= 3\gamma \int_\Omega [(\varphi_\gamma - \varphi^-)^+]^3 \theta_\gamma \, dx \\ &= 3\langle \lambda_\gamma, \theta_\gamma \rangle \to 0. \end{aligned}$$

Since $\mu_\gamma \rightharpoonup \bar{\mu}$ in V_φ^* and $\varphi_\gamma \to \bar{\varphi}$ in V_φ, this proves $\langle \bar{\mu}, \bar{\varphi} - \varphi^- \rangle = 0$.

Finally, it remains to show the last line in (FON$^{\mathrm{VI}}$), $\langle \bar{\theta}, \bar{\mu} \rangle \geq 0$. We test both

$(A'(\mathbf{u}_\gamma))^*\mathbf{z}_\gamma + \mu_\gamma = u_\gamma - u_d$ and $(A'(\bar{\mathbf{u}}))^*\bar{\mathbf{z}} + \bar{\mu} = \bar{u} - u_d$ with $\mathbf{z}_\gamma - \bar{\mathbf{z}}$ and subtract the equations to arrive at

$$\begin{aligned}\langle \mu_\gamma - \bar{\mu}, \theta_\gamma - \bar{\theta}\rangle &= \langle \mu_\gamma - \bar{\mu}, z_\gamma^\varphi - \bar{z}^\varphi\rangle \\ &= \langle u_\gamma - \bar{u}, z_\gamma^u - \bar{z}^u\rangle - \langle (A'(\mathbf{u}_\gamma))^*\mathbf{z}_\gamma - (A'(\bar{\mathbf{u}}))^*\bar{\mathbf{z}}, \mathbf{z}_\gamma - \bar{\mathbf{z}}\rangle \\ &= \langle u_\gamma - \bar{u}, z_\gamma^u - \bar{z}^u\rangle - \langle (A'(\mathbf{u}_\gamma))^*(\mathbf{z}_\gamma - \bar{\mathbf{z}}), \mathbf{z}_\gamma - \bar{\mathbf{z}}\rangle \\ &\quad - \langle (A'(\mathbf{u}_\gamma) - A'(\bar{\mathbf{u}}))^*\bar{\mathbf{z}}, \mathbf{z}_\gamma - \bar{\mathbf{z}}\rangle \\ &\le \langle u_\gamma - \bar{u}, z_\gamma^u - \bar{z}^u\rangle - \langle (A'(\mathbf{u}_\gamma) - A'(\bar{\mathbf{u}}))^*\bar{\mathbf{z}}, \mathbf{z}_\gamma - \bar{\mathbf{z}}\rangle,\end{aligned} \tag{3.5}$$

where the last inequality follows from coercivity of A', i.e., Assumption 2.3. As before, convergence of $\mathbf{u}_\gamma$ provides

$$\langle (A'(\mathbf{u}_\gamma) - A'(\bar{\mathbf{u}}))^*\bar{\mathbf{z}}, \mathbf{z}_\gamma - \bar{\mathbf{z}}\rangle \to 0 \text{ for } \gamma \to \infty. \tag{3.6}$$

By definition of μ_γ, we also find that

$$\langle \mu_\gamma, \theta_\gamma\rangle = 3\int_\Omega \gamma[(\varphi_\gamma - \varphi^-)^+]^2\theta_\gamma^2\, dx \ge 0. \tag{3.7}$$

We thus arrive at

$$\begin{aligned}\langle \bar{\mu}, \bar{\theta}\rangle &= \langle \bar{\mu}, \theta_\gamma\rangle + \langle \mu_\gamma, \bar{\theta}\rangle - \langle \mu_\gamma, \theta_\gamma\rangle + \langle \mu_\gamma - \bar{\mu}, \theta_\gamma - \bar{\theta}\rangle \\ &\le \langle \bar{\mu}, \theta_\gamma\rangle + \langle \mu_\gamma, \bar{\theta}\rangle + \langle u_\gamma - \bar{u}, z_\gamma^u - \bar{z}^u\rangle - \langle (A'(\mathbf{u}_\gamma))^*\bar{\mathbf{z}} - (A'(\bar{\mathbf{u}}))^*\bar{\mathbf{z}}, \mathbf{z}_\gamma - \bar{\mathbf{z}}\rangle,\end{aligned}$$

where we used the non-negativity of $\langle \mu_\gamma, \theta_\gamma\rangle$ from (3.7) in the second line as well as (3.5). Because $u_\gamma \to \bar{u}$ in V_u, $\theta_\gamma \rightharpoonup \bar{\theta}$ in V_φ and $\mu_\gamma \rightharpoonup \bar{\mu}$ in V_φ^*, the first two terms of the right-hand side converge to $2\langle \bar{\mu}, \bar{\theta}\rangle$ and the third term converges to zero. By (3.6), also the last term converges to zero, and the desired sign condition in the last line of (FON$^{\mathrm{VI}}$) follows.

□

4 An SQP Method for (NLP$^\gamma$)

In this section, we introduce the sequential quadratic programming method for the regularized problem (NLP$^\gamma$). Toward a complete convergence analysis of this algorithm, we are interested in the following tasks: after introducing the algorithm, we discuss solvability of the SQP subproblem (QP$^\gamma$) in the spaces provided by the regularity of the prior iterates, relying on a typical coercivity condition on the second

derivative of the Lagrangian, which is expected to be used when deriving second-order sufficient optimality conditions. Typically, conditions like that allow to prove local convergence of the algorithm. For the purpose of this chapter, we assume the existence of a convergent sequence and show that the limit satisfies (FON$^\gamma$), i.e., in case of convergence, the limit is in fact a critical point of the problem under consideration. Last, we are interested in the convergence behavior of finite element discretizations of the SQP subproblems.

4.1 The SQP Algorithm

Let us start by defining the Lagrangian $\mathcal{L}$ corresponding to (NLP$^\gamma$) via

$$\mathcal{L}(q, \mathbf{u}, \mathbf{z}) := J(q, \mathbf{u}) - \langle A(\mathbf{u}) + R(\gamma, \varphi) - B(q), \mathbf{z} \rangle.$$

Let $(q^k, \mathbf{u}^k) = (q^k, u^k, \varphi^k) \in Q \times V \cap W$ with associated $z^k \in V \cap W$ denote a given iterate of the solution algorithm, and define the notation

$$\mathbf{d} := (d^q, \mathbf{d}^\mathbf{u}) := (d^q, d^u, d^\varphi) = (q - q^k, \mathbf{u} - \mathbf{u}^k) = (q - q^k, u - u^k, \varphi - \varphi^k)$$

for the update directions. We note first that the second derivative of the Lagrangian, twice with respect to (q, u) for directions $[\mathbf{d}, \mathbf{d}]$ at the current iterate as linearization point, is given and denoted by

$$\begin{aligned}\mathcal{L}''_{(q,\mathbf{u}),(q,\mathbf{u})}(q^k, \mathbf{u}^k, \mathbf{z}^k)[\mathbf{d}, \mathbf{d}] = \|d^u\|^2 + \alpha \|d^q\|_Q^2 - \langle A''(\mathbf{u}^k)[\mathbf{d}^\mathbf{u}, \mathbf{d}^\mathbf{u}], \mathbf{z}^k \rangle \\ - \langle R''(\gamma, \varphi^k)[d^\varphi, d^\varphi], z^{\varphi,k} \rangle.\end{aligned}$$

Together with the first-order derivative of the objective function J with respect to (q, u) at point $(q^k, \mathbf{u}^k)$ in direction $\mathbf{d}$, denoted by

$$J'_{(q,\mathbf{u})}(q^k, \mathbf{u}^k)\mathbf{d},$$

we formulate for given $\varphi^- \in W_\varphi$ the linear quadratic subproblem (QP$^\gamma$) as follows:
Find the solution $\mathbf{d} = (d^q, \mathbf{d}^\mathbf{u}) \in Q \times V \cap W$ of

$$\min_{\mathbf{d}} J'_{(q,\mathbf{u})}(q^k, \mathbf{u}^k)\mathbf{d} + \frac{1}{2}\mathcal{L}''_{(q,\mathbf{u}),(q,\mathbf{u})}(q^k, \mathbf{u}^k, \mathbf{z}^k)[\mathbf{d}, \mathbf{d}] \tag{QP$^\gamma$}$$

$$\text{s. t. } A'(\mathbf{u}^k)\mathbf{d}^\mathbf{u} + R'(\gamma; \varphi^k)d^\varphi = B(d^q) + B(q^k) - A(\mathbf{u}^k) - R(\gamma; \varphi^k). \tag{4.1}$$

The local SQP algorithm for solving (NLP$^\gamma$) then reads as follows:

Algorithm 4.1 Sequential quadratic programming method for (NLP$^\gamma$):

0. Choose $(q^0, \mathbf{u}^0, \mathbf{z}^0) \in Q \times V \cap W \times V \cap W$, sufficiently close to the optimal triple $(\bar{q}, \bar{\mathbf{u}}, \bar{\mathbf{z}})$, and set $k = 0$.
1. STOP, if $(q^k, \mathbf{u}^k, \mathbf{z}^k)$ is a KKT point of (NLP$^\gamma$), i.e., satisfies (FON$^\gamma$).
2. Solve (QP$^\gamma$) to receive $\mathbf{d}$ with associated adjoint $\mathbf{z}$.
3. Set $(q^{k+1}, \mathbf{u}^{k+1}) := (q^k, \mathbf{u}^k) + \mathbf{d}$, $\mathbf{z}^{k+1} = \mathbf{z}$, $k := k + 1$, and go to Step 1.

Solvability of (QP$^\gamma$) will be shown in Proposition 4.3, and the optimality conditions to be satisfied in the second step of Algorithm 4.1 are stated in (4.3). To derive these results, we need the solvability of (4.1), which follows from Proposition 2.1 and the properties of $A'(\mathbf{u})$ in Sect. 2.2.

Corollary 4.2 *Assuming $(q^k, \mathbf{u}^k, \mathbf{z}^k) \in Q \times W \times W$, the linearized partial differential equation given in (4.1) has a solution $\mathbf{d}^\mathbf{u} \in V \cap W$ for data $d^q \in Q$.*

Proof By the regularity assumptions on $(q^k, \mathbf{u}^k) \in Q \times W$, one can see that all terms of $A(\mathbf{u}^k) - R(\gamma, \varphi^k)$ are at least in $W^\times$. Also, by assumption on the data, it holds $d^q, q^k \in Q \hookrightarrow W^\times$. Thus, the right-hand side of (4.1) is an element of $W^\times$, and the desired result follows by Proposition 2.1. □

Next, we discuss solvability of the quadratic subproblem (QP$^\gamma$).

Proposition 4.3 *Let $(\bar{q}, \bar{\mathbf{u}})$ be a locally optimal solution to (NLP$^\gamma$), and let the linearization triple $(q^k, \mathbf{u}^k, \mathbf{z}^k) \in Q \times V \cap W \times V \cap W$ be given. Assume there exists an $\alpha'' > 0$ such that*

$$\mathcal{L}''_{(q,\mathbf{u}),(q,\mathbf{u})}(q^k, \mathbf{u}^k, \mathbf{z}^k)[(d^q, \mathbf{d}^\mathbf{u}), (d^q, \mathbf{d}^\mathbf{u})] \geq \alpha'' \|(d^q, \mathbf{d}^\mathbf{u})\|^2_{Q \times L^2(\Omega;\mathbb{R}^3)} \tag{4.2}$$

holds for all $(d^q, \mathbf{d}^\mathbf{u}) \in Q \times V \cap W$ that satisfy

$$A'(\mathbf{u}^k)\mathbf{d}^\mathbf{u} + R'(\gamma, \varphi^k)d^\varphi = B(d^q).$$

Then, there exists a unique global solution $(d^q, \mathbf{d}^\mathbf{u}) \in Q \times V \cap W$ to (QP$^\gamma$).

Proof For every $d^q \in Q$, the linearized partial differential equation (4.1) in (QP$^\gamma$) has a unique solution $\mathbf{d}^\mathbf{u} \in V \cap W$ by Corollary 4.2. Let M_{feas} denote the feasible set, i.e.,

$$M_{\text{feas}} := \{(d^q, \mathbf{d}^\mathbf{u}) \in Q \times V \cap W \text{ satisfying (4.1)}\}.$$

It is immediate that M_{feas} is nonempty, closed, and convex. Due to (4.2), the cost functional of (QP$^\gamma$) is strictly convex and continuous, hence weakly lower semicontinuous, as well as radially unbounded, so (QP$^\gamma$) is uniquely solvable in $Q \times V$. Due to Corollary 4.2, $\mathbf{d}^\mathbf{u}$ is an element of W. □

4.2 First-Order Optimality Conditions for (QP^γ) and Its Limit

In order to prove that any limit point of sequences generated by Algorithm 4.1 is in fact a first-order necessary point for (NLP^γ), let us point out that for $(q^k, \mathbf{u}^k, \mathbf{z}^k)$ be given as in Algorithm 4.1, in the $(k+1)$st-step, the functions

$$\mathbf{d}^{\mathbf{u}} = (d^u, d^\varphi) = (u^{k+1} - u^k, \varphi^{k+1} - \varphi^k)$$

with associated adjoint $\mathbf{z}^{k+1}$ satisfy the first-order optimality conditions of (QP^γ), which are given by

$$A'(\mathbf{u}^k)\mathbf{d}^{\mathbf{u}} + R'(\gamma;\varphi^k)d^\varphi = B(d^q) + B(q^k) - A(\mathbf{u}^k) - R(\gamma;\varphi^k), \tag{4.3a}$$

$$(A'(\mathbf{u}^k))^*\mathbf{z}^{k+1} + R'(\gamma,\varphi^k)z^{\varphi,k+1} = -A''(\mathbf{u}^k)[\mathbf{d}^{\mathbf{u}},\,\cdot\,]^*\mathbf{z}^k - R''(\gamma,\varphi^k)[d^\varphi,\,\cdot\,]z^{\varphi,k} + \mathbf{d}^{\mathbf{u}} + u^k - u_d, \tag{4.3b}$$

$$B^*\mathbf{z}^{k+1} + \alpha(d^q + q^k) = 0. \tag{4.3c}$$

These optimality conditions are necessary and sufficient, since (QP^γ) is a convex linear quadratic problem due to (4.2). These properties allow to prove our desired convergence result.

Theorem 4.4 *Assume that Algorithm 4.1 generates an infinite sequence* $(q^k, \mathbf{u}^k, \lambda^k, \mathbf{z}^k, \theta^k, \mu^k)$ *with a limit point* $(\hat{q}, \hat{\mathbf{u}}, \hat{\lambda}, \hat{\mathbf{z}}, \hat{\theta}, \hat{\mu})$ *in the sense that*

$$\begin{aligned} q^k &\to \hat{q} \quad \text{in } Q, & \mathbf{u}^k &\to \hat{\mathbf{u}} \quad \text{in } V,\\ u^k &\to \hat{u} \quad \text{in } W_u, & \varphi^k &\rightharpoonup \hat{\varphi} \quad \text{in } W_\varphi,\\ \lambda^k &\to \hat{\lambda} \quad \text{in } V_\varphi^*, & \mathbf{z}^k &\rightharpoonup \hat{\mathbf{z}} \quad \text{in } V,\\ \mu^k &\rightharpoonup \hat{\mu} \quad \text{in } V_\varphi^*, & \theta^k &\rightharpoonup \hat{\theta} \quad \text{in } V_\varphi. \end{aligned}$$

Then, the limit satisfies (FON^γ).

Proof We examine the limit for $k \to \infty$ in the Eqs. (4.3a), (4.3b), and (4.3c) separately, starting with (4.3a).

By definition, we have $\mathbf{d}^{\mathbf{u}} = \mathbf{u}^{k+1} - \mathbf{u}^k = (u^{k+1} - u^k, \varphi^{k+1} - \varphi^k)$. Thus, by the given convergence and regularity assumptions, there exists a limit point $\hat{\mathbf{d}}^{\mathbf{u}} = (\hat{d}^u, \hat{d}^\varphi) = 0$ in V and $\hat{d}^q = 0$ in Q.

Analogous to the proof of Theorem 3.1, convergence of $\mathbf{u}^k$ in (4.3a) shows that this limit solves

$$0 = A'(\hat{\mathbf{u}})\hat{\mathbf{d}}^{\mathbf{u}} + R'(\gamma;\hat{\varphi})\hat{d}^\varphi - B(\hat{d}^q) = B(\hat{q}) - A(\hat{\mathbf{u}}) - R(\gamma;\hat{\varphi}).$$

Defining $\hat{\lambda} = R(\gamma; \hat{\varphi})$ gives the first two lines of (FON$^\gamma$).

Taking the limit in (4.3b), and defining $\hat{\theta} = \hat{z}^\varphi$ and $\hat{\mu} = R'(\gamma; \hat{\varphi})\hat{\theta}$, shows the third, fifth, and sixth lines of (FON$^\gamma$).

Finally, convergence in (4.3c) gives the fourth line of (FON$^\gamma$). □

4.3 Approximation of (QP$^\gamma$) by Finite Elements

For a practical implementation of Algorithm 4.1, the QP step cannot be performed exactly. Instead, an approximate solution of (QP$^\gamma$) is needed, where the PDE (4.1) is discretized by finite elements. To this end, let $\mathcal{T}_h$ be a sequence of shape regular and quasi-uniform meshes with element diameter $h_T \le h \to 0$ for all $T \in \mathcal{T}_h$. We assume, for simplicity, that the elements T are open triangles, pairwise disjoint, and provide a decomposition of the domain Ω, i.e., $\overline{\Omega} = \bigcup_{T\in\mathcal{T}_h} \overline{T}$. Furthermore, we assume that the elements match the splitting of the boundary into Γ and Γ_D.

Now, we define the finite element space of piecewise linear finite elements

$$V_h = \{v \in V \mid v\big|_T \in P_1(T), \quad \forall T \in \mathcal{T}_h\}.$$

Then, for $\mathbf{u}^k \in V \cap H^{1+s}$ and $q^k \in Q$, we can define the discretized QP subproblem

$$\begin{aligned} \min_{\mathbf{d}\in Q\times V_h} \; & J'_{(q,\mathbf{u})}(q^k, \mathbf{u}^k)\mathbf{d} + \frac{1}{2}\mathcal{L}''_{(q,\mathbf{u}),(q,\mathbf{u})}(q^k, \mathbf{u}^k, \mathbf{z}^k)[\mathbf{d}, \mathbf{d}] \\ \text{s. t. } & \langle A'(\mathbf{u}^k)\mathbf{d}^{\mathbf{u}}, \mathbf{v}_h\rangle + \langle R'(\gamma; \varphi^k)d^\varphi, v_h^\varphi\rangle \\ & = \langle B(d^q) + B(q^k) - A(\mathbf{u}^k), \mathbf{v}_h\rangle \\ & \quad - \langle R(\gamma; \varphi^k), v_h^\varphi\rangle \quad \forall \mathbf{v}_h \in V_h. \end{aligned} \tag{QP$_h^\gamma$}$$

Note that, although no discretization is enforced for d^q, the optimality conditions immediately induce a natural discretization, see, e.g., [11].

Under the growth condition (4.2), the analysis of [17, Theorem 3.3, Corollary 3.8] can be transferred to this situation and yields the following:

Proposition 4.5 *Given $(q^k, \mathbf{u}^k, \mathbf{z}^k)$ satisfying (4.2), and let η be such that Assumption 2.3 holds. Then, there exists $h_0 > 0$, depending on $\|q^k\|_Q$, $\|\mathbf{u}^k\|_{1+s}$ only, such that for any $h \le h_0$, problem (QP$_h^\gamma$) has a unique solution $(d_h^q, \mathbf{d}_h^u)$, and for the solution $(d^q, \mathbf{d}^u)$ of (QP$^\gamma$), it holds the error estimate*

$$\alpha''\Big(\|d^q - d_h^q\|_Q^2 + \|\mathbf{d}^u - \mathbf{d}_h^u\|^2\Big) + \|\mathbf{d}^u - \mathbf{d}_h^u\|_V^2 + \|\mathbf{z} - \mathbf{z}_h\|_V^2 \le ch^{2s}.$$

The constant c depends on $\|\mathbf{u}^k\|_{1+s}$ and $R'(\gamma; \varphi^k)$.

Combining these estimates with the convergence in Theorem 4.4, we see that convergence can be asserted as long as $h \to 0$ sufficiently fast for $k \to \infty$. Of course, to assert global convergence of the sequence, suitable globalization strategies are needed. In these, additional requirements on the accuracy of the iterates need to be required to reliably evaluate sufficient descent conditions, cf. [20, 21].

However, our results [19] only show that it is reasonable to assert bounds on $R(\gamma; \varphi^k)$ but not on $R'(\gamma; \varphi^k)$. Hence, it is not clear whether a uniform bound on the constant in Proposition 4.5 can be proven throughout a globalized SQP-type method. Thus, we will also discuss an alternative SQP-like algorithm in which the regularization is not used when building subproblems. However, a detailed analysis of this alternative is beyond the scope of this chapter.

5 An SQP Method for ($\mathrm{NLP}^{\mathrm{VI}}$)

Along the lines of the last section, we would now like to consider an SQP algorithm for the problem ($\mathrm{NLP}^{\mathrm{VI}}$) and briefly discuss solvability of the SQP subproblems. Instead of investigating the convergence analysis, we place special emphasis on the finite element discretization of the quadratic problem governed by complementarity conditions.

It should be noted that in this setting, the quadratic subproblems contain the linearized operator, while the feasible set is not linearized, similar to the way Newton's method is utilized for generalized equations, e.g., in [6]. This means that our resulting QP still is an MPEC in function space. However, in contrast to Proposition 4.5, we will be able to provide uniform discretization error estimates for the resulting QP problems. Note, again, that in Proposition 4.5, uniform finite element estimates are only true under the assumption that $R'(\gamma; \varphi^k)$ remains bounded, a property which up to now is not even proven for the central path where φ_γ solves (NLP^γ).

5.1 SQP Algorithm for ($\mathrm{NLP}^{\mathrm{VI}}$)

Similar as before, we consider the Lagrangian

$$\mathcal{L}(q, \mathbf{u}, \lambda, \mathbf{z}) := J(q, \mathbf{u}) - \langle A(\mathbf{u}) + \lambda - B(q), \mathbf{z} \rangle$$

and point out that we have

$$\mathcal{L}''_{(q,\mathbf{u})(q,\mathbf{u})}(q, \mathbf{u}, \lambda, \mathbf{z})[\mathbf{d}, \mathbf{d}] = \|d^u\|^2 + \alpha \|d^q\|_Q^2 - \langle A''(\mathbf{u})[\mathbf{d}^{\mathbf{u}}, \mathbf{d}^{\mathbf{u}}]; \mathbf{z} \rangle.$$

Based on this, we can define the QP approximation to (NLP$^{\mathrm{VI}}$) in a given point $(q^k, \mathbf{u}^k, \lambda^k, \mathbf{z}^k)$ for $\mathbf{d} = (d^q, \mathbf{d}^{\mathbf{u}}) \in Q \times V$ as

$$\min_{\mathbf{d}\in Q\times V} J'_{(q,\mathbf{u})}(q^k, \mathbf{u}^k)(d^q, \mathbf{d}^{\mathbf{u}}) + \frac{1}{2}\mathcal{L}''_{(q,\mathbf{u})(q,\mathbf{u})}(q^k, \mathbf{u}^k, \lambda^k, \mathbf{z}^k)[\mathbf{d}, \mathbf{d}]$$

$$\text{s. t.} \begin{cases} A(\mathbf{u}^k) + A'(\mathbf{u}^k)\mathbf{d}^{\mathbf{u}} + \lambda^{k+1} - B(q^k) - B(d^q) = 0, \\ \lambda^{k+1} \geq 0, \\ \varphi^- - \varphi^k - d^\varphi \geq 0, \\ \langle \lambda^{k+1}, \varphi^k - \varphi^- + d^\varphi \rangle = 0. \end{cases} \tag{QP$^{\mathrm{VI}}$}$$

Indeed, this problem corresponds to the linearization of the PDE operator in (NLP$^{\mathrm{VI}}$) while keeping the inequality constraint. Thus, the step d^φ needs to be found in

$$K^k := \{\mathbf{v} = (v^u, v^\varphi) \in V \mid v^\varphi \leq \varphi^- - \varphi^k \text{ a.e. in } \Omega\}$$

to assert $\varphi^{k+1} = \varphi^k + d^\varphi \leq \varphi^-$. Thus, the constraint in (QP$^{\mathrm{VI}}$) can equivalently be written as

$$\langle A(\mathbf{u}^k) + A'(\mathbf{u}^k)\,\mathbf{d}^{\mathbf{u}}, \mathbf{v} - \mathbf{d}^{\mathbf{u}} \rangle \geq \langle B(q^k + d^q), v^u - d^u \rangle, \quad \forall \mathbf{v} \in K^k, \tag{5.1}$$

and λ^{k+1} is the corresponding Lagrange multiplier.

With this, we obtain the following local SQP-type iteration:

Algorithm 5.1 Sequential quadratic programming method for (NLP$^{\mathrm{VI}}$):

0. Choose $(q^0, \mathbf{u}^0, \lambda^0, \mathbf{z}^k) \in Q \times V \times V_\varphi^* \times V$, and set $k = 0$.
1. If $(q^k, \mathbf{u}^k, \lambda^k, \mathbf{z}^k)$ is a KKT point of (NLP$^{\mathrm{VI}}$), STOP.
2. Derive a KKT point $(\mathbf{d}, \lambda^{k+1}, \mathbf{z}^{k+1})$ of the problem (QP$^{\mathrm{VI}}$).
3. Set $(q^{k+1}, \mathbf{u}^{k+1}) = (q^k, \mathbf{u}^k) + \mathbf{d}$, $k := k + 1$, and go to step 1.

Similar as in the previous section, we need to assume a growth condition to have well-posedness of the QP subproblem; the analog to (4.2) is now:

Assumption 5.2 Let us assume that for given $(q^k, \mathbf{u}^k, \lambda^k, \mathbf{z}^k)$, there exists α'' such that for all $\mathbf{d} = (d^q, \mathbf{d}^{\mathbf{u}}) \in Q \times V \cap W$ satisfying (5.1), it holds

$$\mathcal{L}''_{(q,\mathbf{u}),(q,\mathbf{u})}(q^k, \mathbf{u}^k, \lambda^k, \mathbf{z}^k)[\mathbf{d}, \mathbf{d}] \geq \alpha'' \|\mathbf{d}\|^2_{Q\times L^2(\Omega;\mathbb{R}^3)}.$$

Once this assumption holds, it follows by standard arguments that (QP$^{\mathrm{VI}}$) has a global solution, noting that coercivity of $A'(\mathbf{u})$ implies that the variational inequality (5.1) is the necessary and sufficient optimality condition for a strictly convex energy minimization.

5.2 Convergence of FE Approximation to (QP^{VI})

In fact, the subproblem (QP^{VI}) is a quadratic minimization problem with inequality constraints. To analyze its approximation by finite elements, we can proceed similarly as in [16], with the slight complication that the linear second-order operator in the obstacle problem (5.1) is not H^2-regular.

We abbreviate the objective function of the QP problem by

$$J^k(\mathbf{d}) := J'_{(q,\mathbf{u})}(q^k, \mathbf{u}^k)(d^q, \mathbf{d}^\mathbf{u}) + \frac{1}{2}\mathcal{L}''_{(q,\mathbf{u})(q,\mathbf{u})}(q^k, \mathbf{u}^k, \lambda^k, \mathbf{z}^k)[\mathbf{d}, \mathbf{d}].$$

By Assumption 5.2, for any $d^q \in Q$, there exists a unique solution $\mathbf{d}^\mathbf{u} \in K^k$ of the constraint (5.1). Therefore, we can define the solution operator $S : Q \to K^k$, which maps d^q to $\mathbf{d}^\mathbf{u}$, and thus we can define the reduced objective function

$$\begin{aligned} j^k &: Q \to \mathbb{R} \\ j^k(d^q) &:= J^k(d^q, S(d^q)) \end{aligned}$$

with which we can equivalently write (QP^{VI}) as

$$\min_{d^q \in Q} j^k(d^q). \tag{5.2}$$

Moreover, if $\varphi^- \in W_\varphi$ and $\mathbf{u}^k \in W$, then (5.1) implies the additional regularity $\mathbf{d}^\mathbf{u} = S(d^q) \in W$, cf., [10, Remark 7]. As a consequence, the corresponding multiplier λ^{k+1} satisfies $\lambda^{k+1} \in H^{-1+s}$.

For the discretization, we proceed as in Sect. 4.3; except now solutions need to be found in the set

$$K_h^k := \{\mathbf{v}_h \in V_h : v_h^\varphi \le I_h(\varphi^- - \varphi^k) \text{ in } \Omega\}.$$

Where $I_h : C(\overline{\Omega}) \mapsto V_h$ is the nodal interpolation operator satisfying

$$\|w - I_h w\|_V \le C_I h^s \|w\|_{1+s}, \quad \|w - I_h w\|_{1-s} \le C h^{2s} \|w\|_{1+s}, \tag{5.3}$$

for any $w \in H^{1+s}$.

From the discrete analog of (5.1), we get the linearized solution operator $S_h : Q \to K_h^k \subset V_h, d^q \mapsto \mathbf{d}_h^\mathbf{u}$ and the discretized reduced objective

$$\begin{aligned} j_h^k &: Q \to \mathbb{R} \\ j_h^k(d^q) &:= J^k(d^q, S_h(d^q)) \end{aligned}$$

and the discretized problem

$$\min_{d_h^q \in Q} j_h^k(d_h^q). \tag{5.4}$$

Lemma 5.3 *Let Assumption 2.3 be satisfied. Let* $(q^k, \mathbf{u}^k, \lambda^k, \mathbf{z}^k) \in Q \times W \times V_\varphi^* \times W$ *and* $d^q \in Q$ *be given.*

Then, there exists $c > 0$ *such that* $\mathbf{d}^{\mathbf{u}} = S(d^q)$ *and* $\mathbf{d}_h^{\mathbf{u}} = S_h(d^q)$ *satisfy*

$$\|\mathbf{d}^{\mathbf{u}} - \mathbf{d}_h^{\mathbf{u}}\|_V \leq ch^s(\|d^q\|_Q + 1),$$

where $c = c(\|\lambda^{k+1}\|_{-1+s})$ *depends on the* H^{-1+s}*-norm of the multiplier for the variational inequality (5.1).*

Proof We follow [5] and derive the best approximation result

$$\begin{aligned}\langle A'(\mathbf{u}^k)(\mathbf{d}^u - \mathbf{d}_h^u), \mathbf{d}^u - \mathbf{d}_h^u\rangle \leq \langle A'(\mathbf{u}^k)(\mathbf{d}^u - \mathbf{d}_h^u), \mathbf{d}^u - \mathbf{v}_h\rangle \\ - \langle \lambda^{k+1}, \mathbf{v}_h - \mathbf{d}_h^u\rangle \quad \forall \mathbf{v}_h \in K_h^k.\end{aligned} \tag{5.5}$$

Indeed, the result shows the claim using continuity and coercivity of $A'(\mathbf{u}^k)$ as well as the following simple calculation using the complementarity and sign relations of ($\mathrm{QP}^{\mathrm{VI}}$):

$$\begin{aligned}-\langle \lambda^{k+1}, \mathbf{v}_h - \mathbf{d}_h^u\rangle &= -\langle \lambda^{k+1}, \mathbf{v}_h - I_h(\varphi^- - \varphi^k) - \mathbf{d}^u + \varphi^- - \varphi^+\rangle \\ &\quad - \langle \lambda^{k+1}, \mathbf{d}^u - \varphi^- + \varphi^k\rangle \\ &\quad - \langle \lambda^{k+1}, I_h(\varphi^- - \varphi^k) - \mathbf{d}_h^u\rangle \\ &\leq -\langle \lambda^{k+1}, I_h(\mathbf{v}_h - \varphi^- + \varphi^k) - (\mathbf{d}^u - \varphi^- + \varphi^+)\rangle \\ &\leq \|\lambda^{k+1}\|_{-1+s}\|I_h(\mathbf{v}_h - \varphi^- + \varphi^k) - (\mathbf{d}^u - \varphi^- + \varphi^+)\|_{1-s}.\end{aligned}$$

Taking $\mathbf{v}_h = I_h\mathbf{d}^u$ in (5.5) thus yields

$$\begin{aligned}\beta_\eta\|\mathbf{d}^u - \mathbf{d}_h^u\|_V^2 \leq C\|\mathbf{d}^u - \mathbf{d}_h^u\|_V\|\mathbf{d}^u - I_h\mathbf{d}^u\|_V \\ + \|\lambda^{k+1}\|_{-1+s}\|I_h(\mathbf{d}_h^u - \varphi^- + \varphi^k) - (\mathbf{d}^u - \varphi^- + \varphi^+)\|_{1-s},\end{aligned}$$

and the interpolation error estimate (5.3) yields the assertion.

To show (5.5), we calculate for arbitrary $\mathbf{v}_h \in K_h^k$

$$\begin{aligned}\langle A'(\mathbf{u}^k)(\mathbf{d}^u - \mathbf{d}_h^u), \mathbf{d}^u - \mathbf{d}_h^u\rangle &= \langle A'(\mathbf{u}^k)(\mathbf{d}^u - \mathbf{d}_h^u), \mathbf{d}^u - \mathbf{v}_h\rangle \\ &\quad + \langle A'(\mathbf{u}^k)(\mathbf{d}^u - \mathbf{d}_h^u), \mathbf{v}_h - \mathbf{d}_h^u\rangle \\ &= \langle A'(\mathbf{u}^k)(\mathbf{d}^u - \mathbf{d}_h^u), \mathbf{d}^u - \mathbf{v}_h\rangle\end{aligned}$$

$$
\begin{aligned}
&- \langle \lambda^{k+1} + A(\mathbf{u}^k) - B(q^k + d^q), \mathbf{v}_h - \mathbf{d}_h^u \rangle \\
&- \langle A'(\mathbf{u}^k)\mathbf{d}_h^u, \mathbf{v}_h - \mathbf{d}_h^u \rangle \\
&\leq \langle A'(\mathbf{u}^k)(\mathbf{d}^u - \mathbf{d}_h^u), \mathbf{d}^u - \mathbf{v}_h \rangle - \langle \lambda^{k+1}, \mathbf{v}_h - \mathbf{d}_h^u \rangle,
\end{aligned}
$$

where we utilized the Lagrange multiplier λ^{k+1} for the variational inequality (5.1) for the second equation and the discretized variational inequality for (5.1) in the last step, showing (5.5). □

With this, we obtain a discretization error estimate for the reduced cost functional.

Lemma 5.4 *Let $(q^k, \mathbf{u}^k, \lambda^k, \mathbf{z}^k) \in Q \times W \times V_\varphi^* \times V$ be given. Then, it holds for any $d^q \in Q$*

$$
|j^k(d^q) - j_h^k(d^q)| \leq ch^s \|d^q\|_Q.
$$

Proof By definition, it holds

$$
\begin{aligned}
j^k(d^q) - j_h^k(d^q) &= J'(q^k, \mathbf{u}^k)(0, (S - S_h)d^q) \\
&\quad + \frac{1}{2}\Big(\mathcal{L}''_{(q,\mathbf{u}),(q,\mathbf{u})}(q^k, \mathbf{u}^k, \lambda^k, z^k)[(d^q, Sd^q), (d^q, Sd^q)] \\
&\qquad - \mathcal{L}''_{(q,\mathbf{u}),(q,\mathbf{u})}(q^k, \mathbf{u}^k, \lambda^k, z^k)[(d^q, S_h d^q), (d^q, S_h d^q)] \Big) \\
&= J'(q^k, \mathbf{u}^k)(0, (S - S_h)d^q) + \frac{1}{2}\Big(\|Sd^q\|^2 - \|S_h d^q\|^2 \\
&\quad - \langle A''(\mathbf{u}^k)[Sd^q, Sd^q], \mathbf{z}^k \rangle + \langle A''(\mathbf{u}^k)[S_h d^q, S_h d^q], \mathbf{z}^k \rangle \Big) \\
&= J'(q^k, \mathbf{u}^k)(0, (S - S_h)d^q) + \frac{1}{2}\Big(\|Sd^q\|^2 - \|S_h d^q\|^2 \\
&\quad + \langle A''(\mathbf{u}^k)[(S + S_h)d^q, (S - S_h)d^q], \mathbf{z}^k \rangle \Big).
\end{aligned}
$$

Using that $(S + S_h)d^q \in W^{1,p}$, we get from Lemma 2.2

$$
|j^k(d^q) - j_h^k(d^q)| \leq c\|(S - S_h)d^q\|_V,
$$

and Lemma 5.3 shows the assertion. □

In order to derive error estimates for the optimal arguments, we need to rely on the following quadratic growth condition. Let $\bar{d}^q \in Q$ be a local solution to (5.2). We assume the following:

Assumption 5.5 (Quadratic Growth Condition) There exists $\delta > 0$ such that

$$
j^k(\bar{d}^q) \leq j^k(d^q) - \delta \|d^q - \bar{d}^q\|_Q^2, \quad \forall\, d^q \in Q. \tag{5.6}
$$

In many cases, it can be shown that such a condition is a direct consequence of Assumption 5.2. Whether this holds in the given situation is currently being investigated.

From the quadratic growth condition, a standard argument gives the following convergence estimate:

Theorem 5.6 *Let $\bar{d}^q$ and $\bar{d}_h^q$ be the optimal solutions to the problems (5.2) and (5.4), respectively. Then, there exists a constant $c > 0$, independent of the mesh size h, such that the following holds:*

$$\|\bar{d}^q - \bar{d}_h^q\|_Q^2 \leq ch^s.$$

Proof From Assumption 5.5, we get, using the optimality of $\bar{d}_h^q$,

$$\begin{aligned}
\delta \|\bar{d}_h^q - \bar{d}^q\|_Q^2 &\leq j^k(\bar{d}_h^q) - j^k(\bar{d}^q) \\
&= j^k(\bar{d}_h^q) - j_h^k(\bar{d}_h^q) + j_h^k(\bar{d}^q) - j^k(\bar{d}^q) + j_h^k(\bar{d}_h^q) - j_h^k(\bar{d}^q) \\
&\leq |j^k(\bar{d}_h^q) - j_h^k(\bar{d}_h^q)| + |j^k(\bar{d}^q) - j_h^k(\bar{d}^q)| \\
&\leq ch^s(\|\bar{d}^q\|_Q + \|\bar{d}_h^q\|_Q),
\end{aligned}$$

where the last inequality follows from Lemma 5.4. □

References

1. L. Ambrosio and V. Tortorelli, *Approximation of functionals depending on jumps by elliptic functionals via γ-convergence*, Comm. Pure Appl. Math., 2 (1990), pp. 999–1036.
2. ——, *On the approximation of free discontinuity problems*, Boll. Un. Mat. Ital. B, (1992), pp. 105–123.
3. B. Bourdin, G. A. Francfort, and J.-J. Marigo, *Numerical experiments in revisited brittle fracture*, J. Mech. Phys. Solids, 48 (2000), pp. 797–826.
4. ——, *The variational approach to fracture*, J. Elasticity, 91 (2008), pp. 1–148.
5. F. Brezzi, W. W. Hager, and P. A. Raviart, *Error estimates for the finite element solution of variational inequalities. Part I. Primal theory*, Numer. Math., 28 (1977), pp. 431–443.
6. A. L. Dontchev and R. T. Rockafellar, *Newton's method for generalized equations: a sequential implicit function theorem*, Math. Program., 123 (2010), pp. 139–159.
7. G. A. Francfort and J.-J. Marigo, *Revisiting brittle fracture as an energy minimization problem*, J. Mech. Phys. Solids, 46 (1998), pp. 1319–1342.
8. A. Griffith, *The phenomena of rupture and flow in solids*, Philos. Trans. R. Soc. Lond., 221 (1921), pp. 163–198.
9. K. Gröger, *A $W^{1,p}$-estimate for solutions to mixed boundary value problems for second order elliptic differential equations*, Math. Ann., 283 (1989), pp. 679–687.
10. R. Haller-Dintelmann, H. Meinlschmidt, and W. Wollner, *Higher regularity for solutions to elliptic systems in divergence form subject to mixed boundary conditions*, Ann. Mat. Pura Appl., 198(4) (2019), pp. 1227–1241.
11. M. Hinze, *A variational discretization concept in control constrained optimization: The linear-quadratic case*, Comput. Optim. Appl., 30 (2005), pp. 45–61.

12. A. M. KHLUDNEV AND G. LEUGERING, *Optimal control of cracks in elastic bodies with thin rigid inclusions*, ZAMM Z. Angew. Math. Mech., 91 (2011), pp. 125–137.
13. D. KNEES, R. ROSSI, AND C. ZANINI, *A vanishing viscosity approach to a rate-independent damage model*, Math. Models Methods Appl. Sci., 23 (2013), pp. 565–616.
14. G. LEUGERING, J. SOKOŁOWSKI, AND A. ŻOCHOWSKI, *Control of crack propagation by shape-topological optimization*, Discrete Contin. Dyn. Syst., 35 (2015), pp. 2625–2657.
15. C. MEYER, A. RADEMACHER, AND W. WOLLNER, *Adaptive optimal control of the obstacle problem*, SIAM J. Sci. Comput., 37 (2015), pp. A918–A945.
16. C. MEYER AND O. THOMA, *A priori finite element error analysis for optimal control of the obstacle problem*, SIAM J. Numer. Anal., 51 (2013), pp. 605–628.
17. M. MOHAMMADI AND W. WOLLNER, *A priori error estimates for a linearized fracture control problem*, Optim. Eng., 22 (2021), pp. 2127–2149.
18. I. NEITZEL, T. WICK, AND W. WOLLNER, *An optimal control problem governed by a regularized phase-field fracture propagation model*, SIAM J. Control Optim., 55 (2017), pp. 2271–2288.
19. I. NEITZEL, T. WICK, AND W. WOLLNER, *An optimal control problem governed by a regularized phase-field fracture propagation model. part II the regularization limit*, SIAM J. Control Optim., 3 (2019), pp. 1672–1690.
20. J. ZIEMS, *Adaptive multilevel inexact SQP-methods for PDE-constrained optimization with control constraints*, SIAM J. Optim., 23 (2013), pp. 1257–1283.
21. J. C. ZIEMS AND S. ULBRICH, *Adaptive multilevel inexact SQP methods for PDE-constrained optimization*, SIAM J. Optim., 21 (2011), pp. 1–40.

Algorithms for Optimal Control of Elastic Contact Problems with Finite Strain

Anton Schiela and Matthias Stöcklein

Abstract Optimal control of hyperelastic contact problems in the regime of finite strains combines various severe theoretical and algorithmic difficulties. Apart from being large scale, the main source of difficulties is the high nonlinearity and non-convexity of the elastic energy functional, which precludes uniqueness of solutions and simple local sensitivity results. In addition, the contact conditions add non-smoothness to the overall problem.

In this chapter, we discuss algorithmic approaches to address these issues. In particular, the non-smoothness is tackled by a path-following approach, whose theoretical properties are reviewed. The subproblems are highly nonlinear optimal control problems, which can be solved by an affine invariant composite step method. For increased robustness and efficiency, this method has to be adapted to the particular problem, taking into account its large-scale nature, its function space structure and its non-convexity.

Keywords Nonlinear elasticity · Optimal control · Contact problem

Mathematics Subject Classification (2020) Primary 49M37; Secondary 90C55

1 Introduction

The analysis and simulation of elastic contact problems, in particular for small deformations, are a classical subject of applied mathematics. Already in their simplest form, the Signorini problem [6, 15, 28], they are intrinsically non-smooth and lead to a variational inequality on an appropriate Sobolev space. Nevertheless, even in this simple convex setting, the simulation of linearly elastic contact may be challenging [15], depending on the geometric configuration.

A. Schiela (✉) · M. Stöcklein
Mathematisches Institut, Universität Bayreuth, Bayreuth, Germany
e-mail: anton.schiela@uni-bayreuth.de; matthias.stoecklein@uni-bayreuth.de

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_14

Nonlinearly elastic contact problems combine these difficulties with those that arise in nonlinear hyperelasticity. Solutions of hyperelastic problems can be modelled as energy minimizers [5]. Due to non-convexity of the energy, minimizers do not have to be unique. Also, owing to the high nonlinearity of the problem, local minimizers do not have to satisfy the weak form of the equilibrium equation in general.

Here, we consider optimization problems in the context of nonlinear elasticity and contact. This connects to the works [16, 17], where first steps into the topic were taken. In particular, the existence of optimal solutions to such kinds of problems was shown. A composite step method [18] was developed for the numerical solution of these problems. A similar topic was considered in [11], where optimal control problems in the context of biological models were investigated and solved by a quasi-Newton approach. Optimal control of linear contact problems has been considered in [3, 20, 30].

In the last few years, additional progress has been made for this class of problems, and the aim of this chapter is to report on this progress, both concerning theoretical results and algorithmic concepts. We build upon and extend the results from [18, 26, 27], where additional details can be found.

This work consists of two main parts: the first part is a concise recapitulation of the available theoretical results for optimal control of finite strain contact problems. It is mainly based on [27] and, after fixing the framework, describes analytic results on a path-following approach for the solution of these problems. The second part deals with the algorithmic development, which has taken place recently. Here, a number of inherent numerical and practical challenges are described, and algorithmic ways to deal with them are presented. First, elastic problems in three spatial dimensions yield large-scale systems after discretization. Hence, efficient iterative solvers for the computation of steps have to be used. To this end, in [26], an algorithmic framework for inexact step computations was developed, but our class of problems demands further advance in this direction. Second, an appropriate choice of functional analytic framework is discussed. It turns out that the choice of norms has a decisive impact on the performance of the algorithms. Finally, we consider the treatment of the inherent non-convexity in the problem, both in the objective and in the energy functionals. Our discussions are illustrated by numerical examples.

2 Contact Problems in Hyperelasticity

Generally speaking, we study the deformation of a nonlinear elastic body made of a hyperelastic material. The body is considered to be under stress from an external boundary force which causes the deformation. Additionally, deformations are constrained by an obstacle which the body cannot penetrate. In the context of hyperelasticity, computing such deformations corresponds to solving an energy minimization problem.

In this section, we introduce the setting and review the central results for nonlinear elastic contact problems. In particular, we give an overview of the existence theory in nonlinear elasticity, and we address a suitable regularization approach for the contact constraints.

Nonlinear Elasticity In our setting, the nonlinear elastic body is represented by a domain $\Omega \subset \mathbb{R}^3$, which is required to be Lipschitz continuous. In addition, its boundary is divided into three subsets as follows:

$$\Gamma = \overline{\Gamma_D \cup \Gamma_N \cup \Gamma_C},$$

where each subset has non-zero boundary measure. Here, Γ_D and Γ_N denote the parts where the Dirichlet and Neumann boundary conditions hold, respectively. Furthermore, $\Gamma_U \subset \Gamma_N$ and Γ_C denote the parts where the boundary force acts and where the contact constraints are enforced, respectively.

Next, we denote by

$$y : \overline{\Omega} \to \mathbb{R}^3 \text{ and } u : \Gamma_U \to \mathbb{R}^3$$

the deformation of the body and the boundary forces, respectively.

For simplicity, we consider the following contact constraint:

$$y_3 \geq 0 \text{ a.e. on } \Gamma_C.$$

This describes a setting where the body has to stay above the plane that is spanned by the first two canonical basis vectors.

As deformation space, we choose the Sobolev space $W^{1,p}(\Omega; \mathbb{R}^3)$ with $p \geq 2$. Correspondingly, as space for the boundary force, we choose $L^2(\Gamma_U, \mathbb{R}^3)$. If there is no risk of ambiguity, we will skip the notation for the image space in all vector-valued spaces. In the setting of optimal control, deformations will act as the state, and boundary forces will act as the control. Accordingly, we introduce the notation $Y = W^{1,p}(\Omega)$ and $U = L^2(\Gamma_U)$. Furthermore, let $id : \overline{\Omega} \to \overline{\Omega}$ be the identity mapping, and let $\mathbb{M}^3_+$ denote the space of invertible 3×3 matrices with positive determinant. Lastly, if not stated otherwise, we define the matrix norm $\|M\| := \sqrt{\operatorname{tr} M^T M}$, and we denote by $\operatorname{Cof} M := \det(M) M^{-T}$ the cofactor matrix for $M \in \mathbb{M}^3_+$.

For hyperelastic materials, computing the deformation of a body subjected to an external load is equivalent to finding a respective energy minimizer. The corresponding total energy functional $I : Y \times U \to \mathbb{R}$ can be defined by

$$I(y, u) := \int_\Omega \hat{W}(\omega, \nabla y(\omega))\, d\omega - \int_{\Gamma_U} yu\, ds,$$

with the common splitting

$$I_{\text{strain}}(y) = \int_\Omega \hat{W}(\omega, \nabla y(\omega))\, d\omega \quad \text{and} \quad I_{\text{out}}(y,u) = \int_{\Gamma_U} yu\ ds.$$

Here, $\hat{W} : \Omega \times \mathbb{M}^3_+ \to \mathbb{R}$ denotes the stored energy function which depends on the material. For detailed discussion of the specific choice of $\hat{W}$, we refer to [4]. For the further analysis, we require the following assumptions, which are standard in hyperelasticity.

Assumption 2.1 *Let $\hat{W} : \Omega \times \mathbb{M}^3_+ \to \mathbb{R}$ be the stored energy function. We assume that the following properties hold:*

1. *Polyconvexity: For almost all $\omega \in \Omega$, there is a convex function $\mathbb{W}(\omega, \cdot, \cdot, \cdot) : \mathbb{M}^3 \times \mathbb{M}^3 \times]0, +\infty[\to \mathbb{R}$ such that*

$$\hat{W}(\omega, M) = \mathbb{W}(\omega, M, \operatorname{Cof} M, \det M), \quad \text{for all } M \in \mathbb{M}^3_+.$$

   The function $\mathbb{W}(\cdot, M, \operatorname{Cof} M, \det M) : \Omega \to \mathbb{R}$ is measurable for all $M \in \mathbb{M}^3_+$.
2. *For almost all $\omega \in \Omega$, the implication $\det M \to 0^+ \Rightarrow \hat{W}(\omega, M) \to \infty$ holds.*
3. *The sets of admissible deformations defined by*

$$\begin{aligned}\mathcal{A} := \{y \in W^{1,p}(\Omega),\ & Cof\, \nabla y \in L^s(\Omega),\ \det \nabla y \in L^r(\Omega),\\ & y = id \text{ a.e. on } \Gamma_D,\ \det \nabla y > 0 \text{ a.e. in } \Omega\},\\ \mathcal{A}_c := \{y \in \mathcal{A} : y_3 \geq 0 \text{ a.e. on } \Gamma_c\},\end{aligned}$$

   for $p \geq 2$, $s \geq \frac{p}{p-1}$, $r > 1$ are non-empty.
4. *Coercivity: There exist $a \in \mathbb{R}$ and $b > 0$, such that*

$$\hat{W}(\omega, M) \geq a + b(\|M\|^p + \|\operatorname{Cof} M\ \|^s + |\det M|^r).$$

5. *The identity $id : \overline{\Omega} \to \overline{\Omega}$ satisfies*

$$id \in \operatorname*{argmin}_{v \in \mathcal{A}_c} I(v, 0) \text{ and } id_3 \geq 0 \text{ a.e. on } \Gamma_C.$$

With this at hand, computing the deformation of a body constrained by an obstacle can be described by the optimization problem

$$y \in \operatorname*{argmin}_{v \in \mathcal{A}_c} I(v, u). \tag{1}$$

The existence of energy minimizers has been established in [5, Theorem 4.2], extending techniques from [1].

Regularization of Contact Constraints Contact constraints add non-smoothness to an already highly nonlinear and non-convex problem. Therefore, we will introduce a suitable regularization approach.

Here, we apply the normal compliance regularization used in [19, 21]. In this context, we introduce the penalty functional $P : Y \to \mathbb{R}_0^+$ defined by

$$P(v) := \frac{1}{k} \int_{\Gamma_C} [-v_3]_+^k \, ds, \quad k \in \mathbb{N}, \ k > 1, \quad v \in Y,$$

which measures the violation of the constraints. We add the scaled penalty function P to the total energy functional I

$$I_\gamma(y, u) := I(y, u) + \gamma P(y) \qquad \gamma > 0.$$

This approach allows us to drop the contact constraints. As a result, we obtain the regularized minimization problem:

$$y \in \operatorname*{argmin}_{v \in \mathcal{A}} I_\gamma(v, u). \tag{2}$$

The well-posedness of the regularized problem (2) and a convergence result that links (1)–(2) have been proven in [27, Theorem 2.3, Proposition 2.1].

Theorem 2.1 *Let $\gamma > 0$ be a fixed penalty parameter and $u \in U$ be some fixed boundary force. Then, under Assumption 2.1, the regularized total energy functional $I_\gamma(\cdot, u)$ has at least one global minimizer in $\mathcal{A}$.*

3 Optimal Control of Nonlinear Elastic Contact Problems

In the optimal control setting, we aim at minimizing an objective functional

$$J : Y \times U \to \mathbb{R},$$

subject to the constraint that an optimal state y_* is a minimizer of the total energy functional, i.e.,

$$y_* \in \operatorname*{argmin}_{v \in \mathcal{A}_c} I(v, u_*),$$

where u_* is the corresponding optimal control. We restrict ourselves here to a tracking type functional of the form

$$J(y, u) := \frac{1}{2} \|y - y_\mathrm{d}\|^2_{L^2(\Omega)} + \frac{\alpha}{2} \|u\|^2_{L^2(\Gamma_U)},$$

for $y_\mathrm{d} \in L^2(\Omega)$, $\alpha > 0$ and $U = L^2(\Gamma_U)$. Accordingly, the optimal control problem reads as follows:

$$\min_{(y,u)\in Y\times U} J(y,u) \quad s.t. \quad y \in \operatorname*{argmin}_{v\in\mathcal{A}_c} I(v,u). \tag{3}$$

Based on the analysis in [17], the following existence result was derived in [27, Theorem 4.1]:

Theorem 3.1 *Problem* (3) *has at least one optimal solution.*

Solving those kinds of optimal control problems numerically is already a challenging task, even without contact constraints. Therefore, we are going to apply the previously introduced normal compliance regularization in order to avoid dealing with the contact constraints numerically. As a result, we obtain the regularized optimal control problem:

$$\min_{(y,u)\in Y\times U} J(y,u) \qquad s.t. \quad y \in \operatorname*{argmin}_{v\in\mathcal{A}} I_\gamma(v,u), \tag{4}$$

for some fixed parameter $\gamma > 0$.

Analogously to above, we can show the existence of optimal solutions.

Theorem 3.2 *For each $\gamma > 0$, problem* (4) *has at least one optimal solution.*

Proof See [27, Proof of Theorem 4.2]. □

Convergence of Solutions of the Regularized Problem With the regularized optimal control problem at hand, we now have to verify that solutions of the regularized problem (4) approach solutions of the original control problem (3).

However, concerning the regularization (4), one corner case precludes the desired result: it may happen that the energy minimization problems in (3) admit a larger set of energy minimizers than can be approximated by solutions of (4). So the desired convergence theory for (4) can only be established under an assumption that rules this case out, as discussed in [27]. However, a modified regularization can solve this problem.

We introduce the following alternative regularized problem:

$$\mathcal{E}_\gamma(y,u) := I_\gamma(y,u) + \varphi(\gamma)\frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)}, \tag{5}$$

where $\varphi : [0,\infty[\to]0,\infty[$ is a positive function in γ, which is monotonically decreasing, such that

$$\lim_{\gamma\to\infty} \varphi(\gamma) = 0.$$

Again, the well-posedness and a convergence result for this new regularization were established in [27].

For the convergence analysis, it has to be ensured that the regularization function φ does not approach zero too quickly. This is necessary to guarantee that the minimization of a fraction of the objective functional J is sufficiently weighted at all times. This property is specified in the following assumption.

Assumption 3.1 *Let $u \in U$ be fixed. Assume that*

$$\lim_{\gamma \to \infty} \frac{\min_{v \in \mathcal{A}_c} I(v,u) - \min_{v \in \mathcal{A}} I_\gamma(v,u)}{\varphi(\gamma)} = 0.$$

With this in mind, we can show convergence for this approach.

Theorem 3.3 *Let $\gamma_n \to \infty$ be a positive and monotonically increasing sequence of penalty parameters. Furthermore, let (y_*, u_*) denote an optimal solution to problem (3). In addition, let $(y_n, u_n) \subset \mathcal{A} \times U$ be a sequence of optimal solutions to the corresponding regularized problems, where the regularization function φ satisfies Assumption 3.1 w.r.t. u_*. Then,*

$$\lim_{n \to \infty} J(y_n, u_n) = \min_{\mathcal{S}} J.$$

Furthermore, there exist a subsequence (y_{n_k}, u_{n_k}) and a pair $(\overline{y}, \overline{u}) \in \mathcal{A}_c \times U$ such that we obtain the weak convergence $y_{n_k} \rightharpoonup \overline{y}$ in Y and the strong convergence $u_{n_k} \to \overline{u}$ in $L^2(\Gamma_U)$. Additionally, $(\overline{y}, \overline{u})$ solves the original problem (3).

Proof See [27, Proof Theorem 5.4]. □

In [27], additional results are established that allow an a priori choice of φ, depending on the regularity of the geometric configuration.

Formal KKT Conditions We recall that problem (2) admits multiple solution, which rules out the application of efficient algorithms. Thus, we replace the minimization problem (2) by its formal first-order optimality condition. We have to keep in mind that this approach creates a different problem and we have to interpret results carefully, e.g., optimal solution of this new problem might not satisfy (2) but may be just stationary points of the energy.

Let us define

$$c_\gamma(y,u)v = \partial_y \mathcal{E}_\gamma(y,u)v \quad \forall v \in P,$$

where P is a reflexive space of test functions. This mapping corresponds to equilibrium conditions of our hyperelastic problem. We thus have a nonlinear mapping:

$$c_\gamma : Y \times U \to P^*,$$

which can be split additively as follows:

$$c_\gamma(y,u) = A_\gamma(y) - Bu$$

into a nonlinear operator $A_\gamma : Y \to P^*$ and a linear operator $B : U \to P^*$. Then, formally, the KKT conditions at a minimizer x_* state the existence of an adjoint state p such that

$$\begin{aligned} J'(y_*,u_*) + c'_\gamma(y_*,u_*)^* p &= 0 \\ c_\gamma(y_*,u_*) &= 0. \end{aligned}$$

We stress that a rigorous derivation of these conditions seems to be out of reach at the moment. The main reason is the lack of local sensitivity results of solutions of hyperelasticity with respect to perturbations of u.

4 Numerical Optimization Algorithms

In order to algorithmically approach this problem, we formally replace the energy minimizing constraint by its first-order optimality condition. Then, the reformulated problem reads as follows:

$$\min_{(y,u)\in Y\times U} J(y,u) \quad s.t. \quad c_\gamma(y,u) = 0. \tag{6}$$

As a result, we obtain an equality-constrained optimization problem for each parameter $\gamma > 0$. This formulation allows the application of solution algorithms for equality constraints. Nevertheless, a couple of intrinsic difficulties have to be considered. To overcome them, measures have to be taken that go beyond standard equality-constrained optimization:

- The problem is posed in function space, and even after discretization (which is done here by a displacement formulation), this inherent structure should be taken into account by the algorithm.
- In three-dimensional elasticity, the use of direct solvers severely limits the resolution of discretizations. Thus, all arising linear systems have to be solved by iterative methods, preferably of conjugate gradient type. In this context, the issue of finding appropriate preconditioners and termination criteria arises.
- Although the elastic energy minimization problem has been replaced by its equilibrium conditions, the goal remains to compute stable solutions, i.e., energy minimizers. Hence, our algorithm should have built-in preference towards energy decreasing search directions. This issue arises due to non-convexity, in particular, if the problem of linearized elasticity yields a Hessian matrix that is not positive

definite. This case also precludes the direct application of a conjugate gradient method.
- The high nonlinearity of the energy functional also includes a singularity near $\det \nabla y = 0$, while local self-penetration, i.e., $\det \nabla y < 0$ is infeasible.

While we concentrated [27] on the construction of a path-following method for the regularization of the contact constraints, the aim of this section is to propose algorithmic measures to tackle the above problems, which are sometimes specific to nonlinear elasticity, and to illustrate their numerical performance.

For the following numerical computations, we always employ the tracking type functional, described above and a nonlinear material that is used for modelling soft biological tissue. They differ in terms of geometric configuration and in terms of the desired deformed state y_d.

4.1 An Affine Covariant Composite Step Method

For brevity of notation, we set $x := (y, u)$ and $X = Y \times U$, which yields the formulation

$$\min_{x \in X} J(x) \quad s.t. \quad c_\gamma(x) = 0. \tag{7}$$

The space of iterates is equipped with an appropriately chosen scalar product $\langle \cdot, \cdot \rangle$ that gives rise to a Riesz isomorphism $M : X \to X^*$, i.e., $\langle v, w \rangle = (Mv)(w)$. As usual, we define the Lagrangian function $L : X \times P \to \mathbb{R}$ via $L(x, p) := J(x) + pc_\gamma(x) = J(x) + p \circ c_\gamma(x)$.

For the solution of the minimization problem (7), we apply a composite step algorithm based on the preceding work [18], from which we recapitulate the main ideas.

The idea of composite step methods is to split the Newton update δx into a normal step δn and a tangential step δt for a precise treatment of optimality and feasibility. A normal step δn satisfies $\delta n \in \ker c'(x)^\perp$ and aims for feasibility. It can be computed via the augmented system:

$$\begin{pmatrix} M & c'_\gamma(x)^* \\ c'_\gamma(x) & 0 \end{pmatrix} \begin{pmatrix} \delta n \\ q \end{pmatrix} + \begin{pmatrix} 0 \\ c_\gamma(x) \end{pmatrix} = 0, \tag{8}$$

which corresponds to the minimization problem:

$$\min_v \frac{1}{2} \langle v, v \rangle_M \quad s.t. \quad c'_\gamma(x)v + c_\gamma(x) = 0.$$

If necessary, a damping factor $\nu \in]0, 1]$ is applied. The tangential step δt satisfies $\delta t \in \ker c'(x)$ and aims for a decrease of the functional value. It can be computed by solving the problem:

$$\begin{pmatrix} L_{xx}(x, p) & c'_\gamma(x)^* \\ c'_\gamma(x) & 0 \end{pmatrix} \begin{pmatrix} \delta t \\ q \end{pmatrix} + \begin{pmatrix} L_x(x, p) + L_{xx}(x, p)\nu\delta n \\ 0 \end{pmatrix} = 0. \tag{9}$$

This problem corresponds to the following minimization problem for $\delta x = \nu\delta n + \delta t$ with fixed $\nu\delta n$:

$$\min_{\delta t} f'(x)\delta x + \frac{1}{2} L_{xx}(x, p)(\delta x, \delta x) \quad \text{s.t.} \quad c'(x)\delta t = 0.$$

The Lagrange multiplier p, which is needed for this step, is computed before as follows:

$$\begin{pmatrix} M & c'(x)^* \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} g \\ p \end{pmatrix} + \begin{pmatrix} f'(x) \\ 0 \end{pmatrix} = 0. \tag{10}$$

In contrast to (8), the upper left block in (9) need not be positive definite. The best we can hope for is that $L_{xx}(x, p)$ is positive definite on $\ker c'_\gamma(x)$ if x is close to a strict minimizer of the problem.

Adding δn and δt directly results in a full Lagrange–Newton step, while damping can be used to construct a globalization procedure, e.g., or the form:

$$\delta x := \nu\delta n + \tau\delta t,$$

where $\nu \in]0, 1]$ and $\tau > 0$ are damping and step size parameters. This class of methods is popular in equality-constrained optimization and optimal control, and there are various realizations of this principal idea available [12, 22, 25, 29, 32].

The motivation of the approach in [18] is due to functional analytic considerations. Most methods for equality-constrained optimization use residual norms of the form $\|c_\gamma(x)\|_{P^*}$, e.g., within a merit function or a filter. If $c_\gamma(x)$ models a partial differential equation, appropriate norms are dual norms and thus not easy to evaluate. The use of a simple norm at this position may degrade the performance of the globalization procedure considerably.

The concept of affine covariance [8] allows to dispense with the evaluation of residuum norms. Instead, a simplified Newton step δs is used, which is defined as a minimum norm solution of a simplified Newton equation:

$$\begin{pmatrix} M & c'_\gamma(x)^* \\ c'_\gamma(x) & 0 \end{pmatrix} \begin{pmatrix} \delta s \\ q \end{pmatrix} + \begin{pmatrix} 0 \\ c_\gamma(x + \delta x) - c_\gamma(x) - c'_\gamma(x)\delta x \end{pmatrix} = 0. \tag{11}$$

By computing the ratio $\Theta := \|\delta s\|_X / \|\delta x\|_X$, it is possible to estimate the Newton contraction towards the feasible manifold. If $\Theta \ll 1$, we are in the region of fast local convergence of Newton's method for the solution of the underdetermined problem $c_\gamma(x) = 0$. This globalization idea is combined with an appropriate decrease criterion. Details are elaborated in [18]. There it is also shown that δs helps to overcome the Maratos effect, since $J(x + \delta x + \delta s)$ is approximated better by the quadratic model of the Lagrange–Newton step than $J(x + \delta x)$ is. Thus, δs plays the role of a second-order correction and thus is beneficial in two ways.

To enforce the non-self-penetration condition $\det \nabla y > 0$, additional damping is applied if the trial iterate violates this condition.

4.2 Computation of Steps by Iterative Solvers

Solving the systems (8)–(11) by direct factorization is only possible for small problems. Consequently, as the degrees of freedom increase, iterative solvers have to be deployed. Here, we build upon the analysis and techniques discussed in [16, Chapter 4] and extend these to a framework where can deal with possible non-convexities in the constraints. A general overview concerning numerical approaches for saddle point problems can be found in [2]. For projected conjugate gradient methods, we refer here to the summaries in [9, 23].

We observe that the systems (8)–(11) all have a common structure, which, after splitting of $X = Y \times U$, can be written as follows:

$$\begin{pmatrix} H & C^* \\ C & 0 \end{pmatrix} \begin{pmatrix} v \\ q \end{pmatrix} + \begin{pmatrix} c_{1,2} \\ c_3 \end{pmatrix} = 0 \;\Leftrightarrow\; \begin{pmatrix} H_y & 0 & A^* \\ 0 & H_u & -B^* \\ A & -B & 0 \end{pmatrix} \begin{pmatrix} v_y \\ v_u \\ q \end{pmatrix} + \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = 0. \tag{12}$$

Here, we used the observation that $J(y, u) = J_1(y) + J_2(u)$ so that H is block diagonal and set $Cx = Ay - Bu$. Since $J_2(u) = \alpha/2\|u\|^2$, we know that H_u is positive definite. However, H may not always be positive definite on $\ker C$ in case of the tangential step (9). This can only be expected close to a minimizer. Moreover, in our context of hyperelasticity, the block $A = A'_\gamma(y) = \partial_{yy}\mathcal{E}_\gamma(y, u)$ is symmetric, but not always positive definite (and may be singular) due to non-convexity of the elastic energy. However, as we will discuss below, it is possible to modify A and H, such that the modified operators are positive definite, so that, in particular, A is invertible. Finally, we observe that $c_3 = 0$ in (10) and (9), while by solving the system $Av_{y,0} + c_3 = 0$, we can reduce (8) and (11) to a problem, where $c_3 = 0$ holds as well.

In this setting, a conjugate gradient method on $\ker C$ can be applied to (12). For its implementation, a constraint preconditioner is needed that guarantees that iterates remain in $\ker C$, as long as $c_3 = 0$. We obtain a projected conjugate gradient

method, cf., e.g., [10, 24]. Taking into account the block structure of (12), we use the following block lower triangular preconditioner P:

$$P := \begin{pmatrix} 0 & 0 & A^* \\ 0 & \tilde{H}_u & -B^* \\ A & -B & 0 \end{pmatrix},$$

dropping H_y and replacing H_u by a preconditioner $\tilde{H}_u$, e.g., if H_u is a mass matrix, its diagonal. This decouples (12) into three equations, which can be solved sequentially:

$$A^* q = -c_1 \qquad \to q \tag{13}$$

$$\tilde{H}_u v_u = B^* q - c_2 \qquad \to v_u \tag{14}$$

$$A v_y = B v_u \qquad \to v_y. \tag{15}$$

The main computational effort is spent solving (13) and (15), which are problems of linearized elasticity in 3D. For coarse discretizations, a sparse direct solver can be used to factorize $A = A^*$ and solve (13) and (15). In that case, the preconditioned cg-method can be applied directly to (12) despite its saddle point structure. Preconditioners of this form have been considered in [23, Chapter 5].

4.3 Inexact Constraint Preconditioning

For fine discretizations, A cannot be factorized directly, and we have to resort to preconditioned conjugate gradients. Here, we use a multigrid preconditioner of BPX-type, equipped with a block Jacobi smoother that uses the diagonal of 3×3 blocks of A, respecting the vector-valued nature of the problem.

Some care has to be taken, when this method is implemented. If (13) and (15) are solved only inexactly, the iterates are not contained in $\ker C$ any longer. Instead of (13) and (15), one actually solves nearby problems

$$\widetilde{A^*} q = -c_1 \tag{16}$$

$$\widetilde{A} v_y = B v_u \tag{17}$$

with $\widetilde{A} \approx \widetilde{A^*} \approx A = A^*$, changing slightly from step to step. If the cg-method applies the operator in (12) as the forward operator, this may lead to spurious occurrence of directions of negative curvature, unless the accuracy of solution of (13) and (15) is very high.

Hence, we have to avoid application of the A and A^* blocks in (12). This can be done by a simple auxiliary recursion within the projected cg-method.

Although it is known that the convergence theory of conjugate gradients requires that preconditioners are linear mappings that do not change during the iteration, the cg-method then tolerates small errors in the application of the preconditioners. It can be observed, however, that loose tolerances in the solution of (16) and (17) are detrimental for the speed of convergence of the outer cg-iteration.

Remark 4.1 It is desirable to reduce the accuracy requirement for (16) and (17) even further. Then, a linear iteration scheme, such as a preconditioned Chebyshev semi-iteration with a fixed number of steps, has to be employed, and we end up with a solution of the perturbed problem:

$$\begin{pmatrix} H_y & 0 & \tilde{A}^* \\ 0 & H_u & -B^* \\ \tilde{A} & -B & 0 \end{pmatrix} \begin{pmatrix} \tilde{v}_y \\ \tilde{v}_u \\ q \end{pmatrix} + \begin{pmatrix} c_1 \\ c_2 \\ 0 \end{pmatrix} = 0, \tag{18}$$

where in contrast to before, $\tilde{A}$ is a linear operator, so that the outer cg-iteration really solves a well-defined linear problem.

Solving (15) with $\tilde{v}_u$ on the right-hand side (by conjugate gradients) yields solutions in $\ker C$ again. A linear solver, based on this idea, has been tested with promising results for optimal control of linear elliptic problems. Application to optimal control of nonlinear elasticity is under current investigation.

Remark 4.2 Within the approach of Byrd–Omojokun composite step methods, inexact system solvers were considered in [12, 13, 25]. Here, an alternative route is taken. A GMRES method is used to solve normal steps inexactly, allowing for loose tolerances in the evaluation of A and A^*. The solver for this problem also serves as a preconditioner for the tangential step, for which a projected cg with re-orthogonalization is used.

4.4 Accuracy Matching

For the efficiency of the overall method, it is important that the steps are computed neither with too tight tolerances, which renders each step too expensive, nor with too loose tolerances, which may lead to loss of robustness and increase of the number of outer iterations. Setting fixed tolerances is usually not the best way to cope with this problem, since each step of the outer iteration has a different characteristic. For example, if the tangential step is dominant (which often happens close to the optimal solution), then normal and simplified normal step can be computed with low relative accuracy. A strategy for accuracy matching has been proposed and tested in [26], where in particular the impact of inexact normal and simplified normal steps on the outer iteration was considered, and adaptive termination criteria were derived.

4.5 Choice of Functional Analytic Framework

A very important issue for our problem is a good choice of a Hilbert space norm on $X = Y \times U$ that measures the step lengths and also defines what *normal* means, when normal steps are computed. Our numerical results show that this has considerable impact on the performance of our algorithm. Recall that the Riesz operator $M : X \to X^*$ enters the definition of normal and simplified normal step, as well as the computation of the Lagrange multiplier in (8), (11), and (10).

For certain classes of optimal control problems with mildly nonlinear PDEs (e.g., semilinear equations), it is possible to find a Hilbert space norm on X that allows a rigorous (local) convergence theory in function space, based on the corresponding functional analytic setting. For an overview of this topic, we refer to [14, Chapter 1–2]. To obtain analogous results for nonlinear elasticity is illusory. Even solution algorithms for the forward problem suffer from a two-norm discrepancy: differentiability of the energy functional cannot be expected in a space less regular than $W^{1,\infty}(\Omega)$, while the energy space is only $W^{1,2}(\Omega)$. We thus have a norm-gap that is hard to bridge.

As a consequence, mesh-dependent behaviour of solvers has to be expected, at least, if difficult problems with large strains are solved. If the problem is not too hard, however, additional regularity of the steps can usually be observed, which alleviates the difficulty in practice. This favourable effect depends on the concrete configuration and is hard to grasp a priori in a mathematical theory.

Our numerical observations confirm our considerations. In the following, consider two alternative norms:

$$\|(y,u)\|_{M_0}^2 := \frac{1}{2}\|y\|_{L_2(\Omega)}^2 + \frac{\alpha}{2}\|u\|_{L_2(\Gamma)}^2,$$
$$\|(y,u)\|_{M_1}^2 := \frac{1}{2}\|y\|_{H^1(\Omega)}^2 + \frac{\alpha}{2}\|u\|_{L_2(\Gamma)}^2.$$

Clearly, $\|\cdot\|_{M_0}$ is in close correspondence with the objective functional to be minimized but does not take into account the regularity requirements of the nonlinearity, at all. In contrast, $\|\cdot\|_{M_1}$ promotes smoother states, so, although not guaranteeing $W^{1,\infty}(\Omega)$ regularity, is certainly considerably closer to the ideal situation.

We test these two alternatives at a problem, described in Fig. 1. The results of a numerical comparison are depicted in Figs. 2 and 3. We observe a marked difference, although the computed final solutions, as inspection showed, are the same. Equipped with $\|\cdot\|_{M_0}$, our algorithm takes about 200 steps and shows quite irregular behaviour. If $\|\cdot\|_{M_1}$ is used, we observe from Fig. 3 that the behaviour of our algorithm is much faster and also much more regular, concerning choice of damping factors.

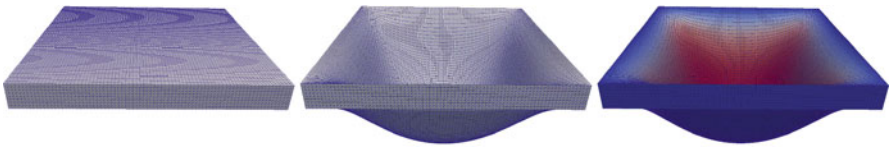

Fig. 1 Problem of pushing down a plate. Left: undeformed domain. Middle: desired deformation. Right: optimal deformation. Upper horizontal boundary: Γ_U, colour codes intensity of force. Lower horizontal boundary: $\Gamma_N \setminus \Gamma_U$. Vertical boundaries: Γ_D

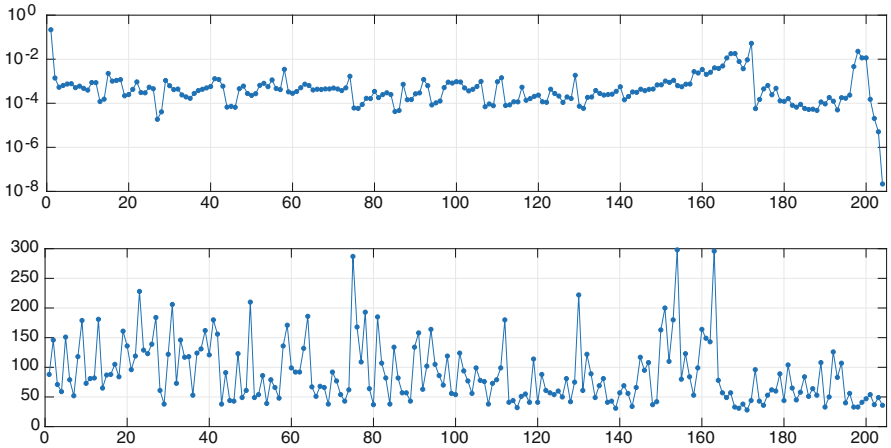

Fig. 2 Iteration history for problem from Fig. 1, using M_0. Top: norms of steps, taken by the algorithm. Bottom: the total number of outer cg-iterations in each step

4.6 Non-convexity of Objective and Energy

A major difficulty in the considered class of problems is the occurrence of non-convexity, not only in the objective functional but also in the energy functional. The difficulties, introduced by non-convexity, are well known: we usually obtain non-unique local minimizers and additional stationary points. Furthermore, quadratic models used in SQP methods are not positive definite anymore. As a consequence, Lagrange–Newton methods, even if equipped with some damping, are usually not appropriate for finding minimizers of non-convex problems. Various algorithmic techniques for non-convex optimization have been developed in the last few decades (cf., e.g., [7]).

Two popular techniques for large-scale problems are truncated conjugate gradients, where the cg-method is performed, until a direction of negative curvature is detected and Hessian modification, where a positive definite term is added to the Hessian, such that the sum is positive definite. In our context, the latter strategy yields markedly more robust behaviour, if applied appropriately. For this, it is

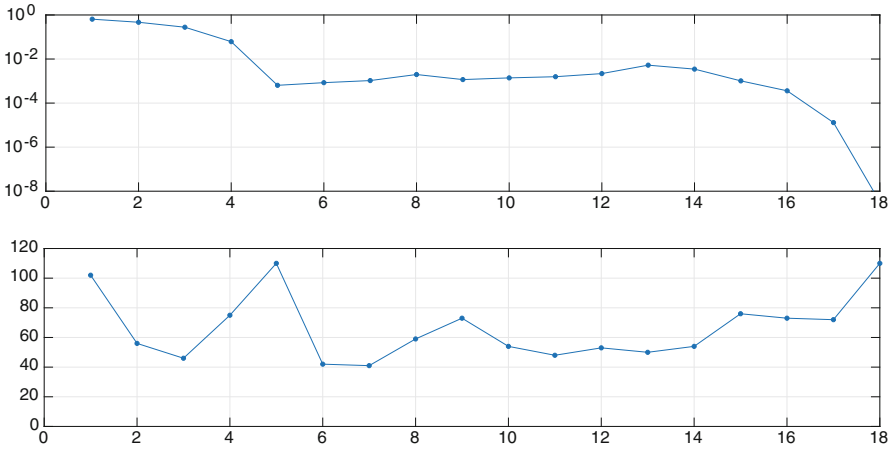

Fig. 3 Iteration history for problem from Fig. 1, using M_1. Top: norms of steps, taken by the algorithm. Bottom: the total number of outer cg-iterations in each step

important that the regularization term is chosen adequately, taking into account the underlying functional analytic structure.

While truncated cg-methods typically yield cheaper steps, the computed search directions are often very irregular (an effect that is not present in $\mathbb{R}^n$) and thus yield very small damping parameters and many outer iterations. The reason is that steps result from an algebraic computation that is just terminated at a point where things became particularly difficult with often irregular cg-iterates.

On the contrary, an appropriate Hessian modification yields a well-defined problem in function space. For example, a regularized elastic problem is still an elastic problem, however, with a different stiffer material. Hence, solutions of regularized problems typically have better regularity properties than an arbitrary element of the energy space. This is in agreement with the above observation that additional regularity of solutions helps to bridge the problem inherent norm-gap.

Non-convexity of the Objective If $L_{xx}(x, p)$ is not positive definite on $\ker c'_\gamma(x)$, then (9) does not correspond to a quadratic minimization problem, and descent of tangential steps is not guaranteed, unless appropriate modifications are made. Just as described above, we use a Hessian modification approach. Instead of solving (9), we solve the modified problem:

$$\begin{pmatrix} L_{xx}(x, p) + \lambda M & c'_\gamma(x)^* \\ c'_\gamma(x) & 0 \end{pmatrix} \begin{pmatrix} \delta t \\ q \end{pmatrix} + \begin{pmatrix} L_x(x, p) + L_{xx}(x, p)\delta n \\ 0 \end{pmatrix} = 0, \qquad (19)$$

where M is the Riesz isomorphism that is used to define normal steps, and $\lambda \geq 0$ is an algorithmic parameter that is chosen large enough to render $L_{xx}(x, p) + \lambda M$ positive definite on $\ker c'_\gamma(x)$.

This system corresponds to the following minimization problem:

$$\min_{\delta t} f'(x)\delta x + \frac{1}{2}(L_{xx}(x, p) + \lambda \tilde{M})(\delta x, \delta x) \quad \text{s.t.} \quad c'(x)\delta t = 0,$$

where $\tilde{M} = M$ on $\ker c'_\gamma(x)$, but $\tilde{M} = 0$ on $(\ker c'_\gamma(x))^\perp$. Our quadratic model is thus mainly modified on $\ker c'_\gamma(x)$ but not on its orthogonal complement.

In our current implementation, each step is started with $\lambda = 0$. If non-convexity is encountered, $\lambda > 0$ is chosen and a new attempt to compute a step is made. Subsequently, λ is increased, until directions of negative curvature are no longer encountered.

4.7 Non-convexity of the Energy

If models for elastic materials are intended to be realistic for large deformations, they have to be non-convex. A classical and practically relevant example, caused by non-convexity, is buckling. It can be observed, if compressive forces act on the opposite ends of a slim body. For small forces, the body is compressed in the direction of force, but as forces increase, this state becomes unstable and energy is decreased if the body is bent in some direction. Depending on the symmetries of the body, the new energy minimizers may be non-unique, and they may differ dramatically from the previous solution. Seemingly, stable structures collapse suddenly, if a certain critical force is exceeded (Fig. 5).

If such a behaviour is encountered during the course of solution of an optimal control problem, which is the case for the problem, described in Fig. 4, a couple of numerical difficulties arise. First of all, the operator A in (15) is likely to be

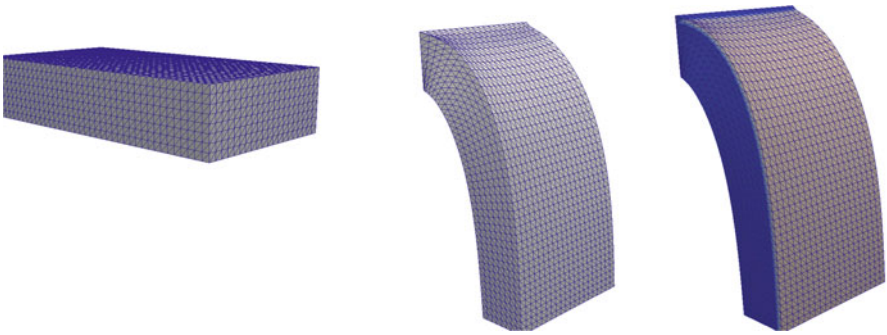

Fig. 4 Problem of bending down a horizontal cantilever. Left: undeformed domain. Middle: desired deformation y_d. Right: optimal deformation. Upper/front horizontal boundary: Γ_U, colour codes intensity of applied forces. Rear vertical boundary: Γ_D. All other boundaries: $\Gamma_N \setminus \Gamma_U$

indefinite, so (15) does not correspond to a quadratic energy minimization problem and also cannot be solved by a cg-method. Furthermore, solutions of the nonlinear equation $c_\gamma(y,u) = 0$ cease to be energy minimizers or change rapidly, if u is perturbed.

In analogy to the regularization of the objective function, we apply a regularization term to the energy functional. This has to be done in such a way that δn, δt and δs can be computed in a consistent way.

Assume that at the iterate (y_k, u_k), the operator $A = A'_\gamma(y_k)$ is not positive definite. This can be detected during the attempt to solve (13) or (15) by a cg-method. In that case, we choose a regularization factor $\lambda > 0$ and define a regularized energy functional as follows:

$$\hat{\mathcal{E}}_\gamma(y,u) := \mathcal{E}_\gamma(y,u) + \frac{\lambda}{2} q(y - y_k),$$

where q is a quadratic positive definite energy. For our computations, we have chosen $q(v) = \langle \nabla v, \nabla v \rangle_{L_2}$. For this modification, we compute

$$\partial_y \hat{\mathcal{E}}_\gamma(y,u) = \partial_y \mathcal{E}_\gamma(y,u) + \lambda q'(y - y_k) \;\Rightarrow\; \partial_y \hat{\mathcal{E}}_\gamma(y_k,u_k) = \partial_y \mathcal{E}_\gamma(y_k,u_k),$$

$$\hat{A} := \partial^2_{yy} \hat{\mathcal{E}}_\gamma(y_k,u_k) = \partial^2_{yy} \mathcal{E}_\gamma(y_k,u_k) + \lambda q''(0).$$

The effect of this regularization is threefold: first of all, if λ is sufficiently large to render $\hat{A}$ is positive definite, then the solution of (13) or (15) with A replaced by $\hat{A}$ is a minimization problem. Thus, conjugate gradients can be applied. Second, due to ellipticity of $\hat{A}$, normal steps are shifted towards descent for $\mathcal{E}_\gamma(y,u)$ at (y_k,u_k), because the linearized constraint imposed on δn reads

$$\hat{A}\delta n_y - B\delta n_u + \partial_y \mathcal{E}_\gamma(y_k,u_k) = 0.$$

Since $\mathcal{E}_\gamma(y_k,u_k)$ is linear in u, we conclude

$$\partial_y \mathcal{E}_\gamma(y_k, u_k + \delta n_u)\delta n_y = (\partial_y \mathcal{E}_\gamma(y_k,u_k) - B\delta n_u)\delta n_y = -(\hat{A}\delta n_y)\delta n_y < 0.$$

So, δn_y is a descent direction for the total energy at the point $(y_k, u_k + \delta n_u)$. Thus, finding energy minimizers is promoted. Third, long steps are penalized, which results in a more stable behaviour of the optimization algorithm in the presence of an instability of the elastic problem.

As a numerical example, we consider the problem, described in Fig. 4. In Fig. 5 and Fig. 6, we see an iteration history and some of the iterates taken by our optimization algorithm for a problem, where non-convex behaviour of the energy functional is encountered. In the beginning of the iteration, a buckling type non-convexity is encountered. We observe that the applied forces in the early phase of

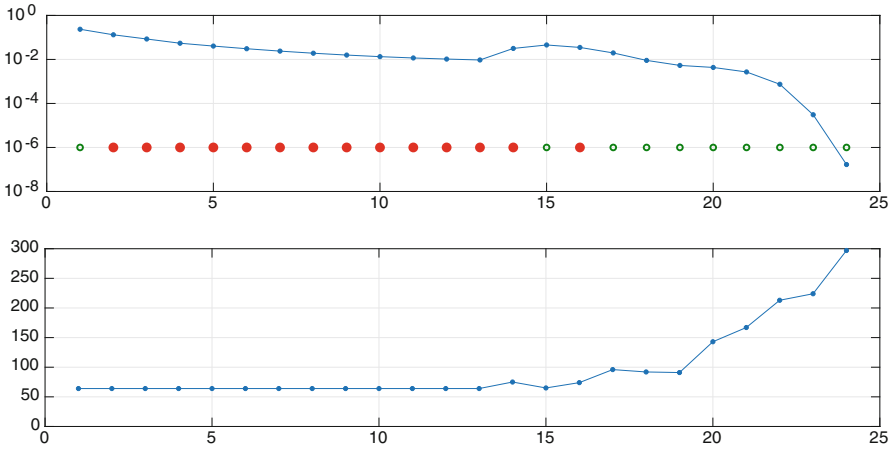

Fig. 5 Iteration history for problem from Fig. 4. Top: norms of steps, taken by the algorithm. Green circles: convex energy, red dots: non-convex energy. Bottom: the total number of outer cg-iterations in each step

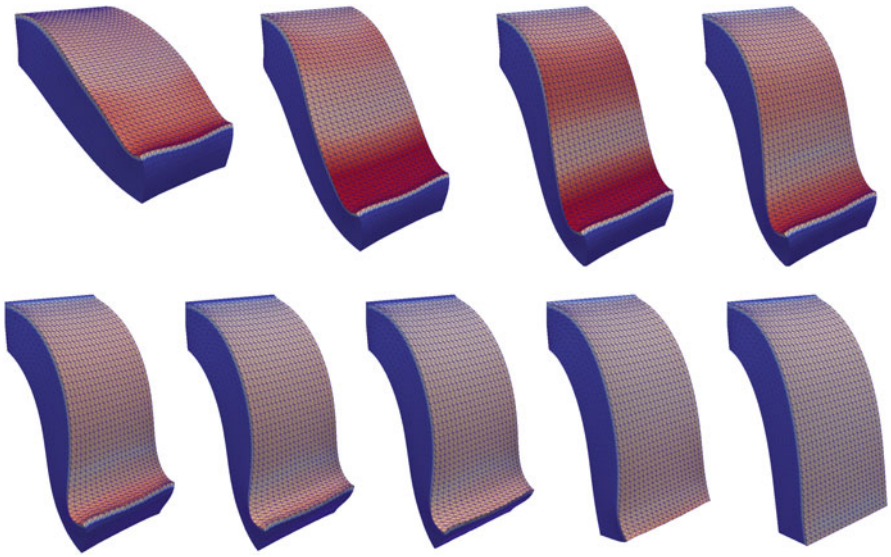

Fig. 6 Iterates taken by the composite step method for problem from Fig. 4 without contact. The colour codes the intensity of the forces. Top row: iterates 1,2,3,4. Bottom row: iterate 7,10,13,16,19

the algorithm are rather intense, since the material resists the applied compressive forces. Also, the necessary regularization of the energy adds some artificial stiffness to the material. After some bending has taken place, the algorithm finds the optimal solution using comparably small forces.

An interesting effect is the S-shape shown by some intermediate deformations. This is a consequence of the nonlinearity of the problem, which occurs in particular for boundary forces. A tangential traction force tries to push the body towards the desired deformation. Due to the nonlinearity of the problem, however, the body is bent in upward direction, due to the moment introduced by this force.

Remark 4.3 An alternative to the presented idea was also tested. If non-convexity of the energy was encountered during the iteration at (y_k, u_k), the control u_k was kept fixed and an energy minimization algorithm, based on the ideas of [31], was applied to compute a minimizer of $\mathcal{E}_\gamma(\cdot, u_k)$ was computed. So, we temporarily switched to a black-box method. At least for our test case, the performance of this variant was not satisfactory. It showed a rather unstable and erratic behaviour. The reason for this seems to be that buckling can occur during such an algorithm.

4.8 Path Following

With a robust solver for (6) at hand, we can now use a path-following method to approximate solutions of the original optimal control problem with contact. We use a simple approach, where after (6) has been solved for some γ_k, the regularization parameter is multiplied by some fixed factor $s > 1$. A choice of $s = 10$ has proven quite appropriate.

We added contact constraints to the problem, described in Fig. 4. An illustration of the path-following procedure is given in Fig. 7. Obviously, the regularization procedure works as intended. For moderate γ, the contact constraint is clearly violated, but if γ becomes larger, this violation gradually vanishes. We also observe that our method is well capable to deal with large deformations and strains.

A detailed discussion of the problem from Fig. 1 with added contact constraints, including convergence plots, can be found in [27].

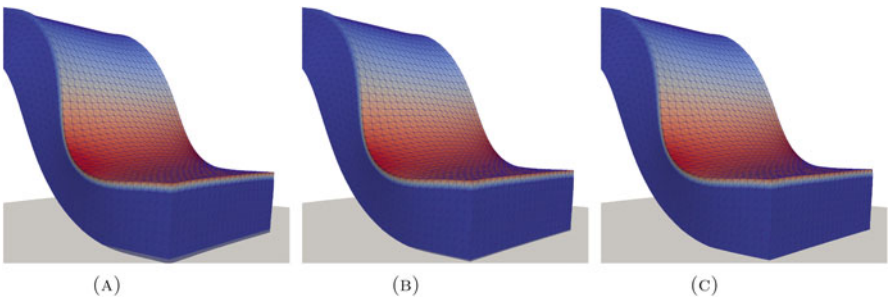

Fig. 7 Optimal deformations with penalty parameter γ for problem from Fig. 4 with contact. Colour codes intensity of the forces. (**a**) $\gamma = 10^3$. (**b**) $\gamma = 10^4$. (**c**) $\gamma = 10^6$

5 Conclusion and Outlook

We conclude that optimal control problems with finite strain hyperelastic materials and contact offer a broad range of challenges, concerning both theoretical and algorithmic aspects.

The main theoretical challenge is that not much analytic structure is available to build a theory upon. The main result of polyconvexity, a weak lower semi-continuity property of the energy functional, could be exploited to conclude results about the path of regularized solutions. However, satisfactory stronger results were only possible by employing refined techniques. Still there are many questions that remain open, most importantly a local sensitivity result that could permit a rigorous derivation of optimality conditions for the optimal control problem.

From a numerical view point, this class of problems combines high nonlinearity and non-convexity with large scale. To obtain efficient and robust solution algorithms, significant advances had to be made, compared to generic optimization methods. Decisive ingredients are a good choice of functional analytic framework, a sound concept for inexact computation of steps by iterative solvers, and a proper treatment of non-convexities, both in the objective and in the energy. In this chapter, we concentrated on these computation aspects. An observed key ingredient to efficient algorithmic behaviour is to produce regular steps where possible.

A couple of algorithmic concepts are subject to current work. First, as pointed out in Remark 4.1, new ideas, concerning the use of iterative solvers for the A-block in (12) are currently under investigation. Second, the nonlinearity of finite deformation problems exhibits some very interesting geometrical structure. The analysis hints on using nonlinear updates, instead of the usual linear ones. Currently, promising numerical results that go into this direction are available for solving the energy minimization problem. The application of this concept to optimal control problems is still subject to current research and is planned to be published in a forthcoming paper.

As a future perspective, the algorithmic solution of our class of problems has to be extended to real world applications. A particular example is inverse problems in the context of biomechanics, where elastic contact problems occur in joints. In addition to the described difficulties, an envisioned solution algorithm will have to deal with complicated contact geometries.

Acknowledgments This work was supported by the DFG grant SCHI 1379/2-1 within the priority programme SPP 1962 (Non-smooth and Complementarity-Based Distributed Parameter Systems: Simulation and Hierarchical Optimization).

References

1. John M. Ball. Convexity conditions and existence theorems in nonlinear elasticity. *Archive for Rational Mechanics and Analysis*, 63(4):337–403, 1977.

2. Michele Benzi, Gene H. Golub, and Jörg Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14:1–137, 2005.
3. Thomas Betz. *Optimal control of two variational inequalities arising in solid mechanics*. PhD thesis, 2015.
4. P.G. Ciarlet. *Mathematical Elasticity: Three-dimensional elasticity*. Number Bd. 1. North-Holland, 1994.
5. Philippe G. Ciarlet and Jindřich Nečas. Unilateral problems in nonlinear, three-dimensional elasticity. *Archive for Rational Mechanics and Analysis*, 87(4):319–338, 1985.
6. Marius Cocu. Existence of solutions of Signorini problems with friction. *International journal of engineering science*, 22(5):567–575, 1984.
7. A.R. Conn, N.I.M. Gould, and P.L. Toint. *Trust-Region Methods*. SIAM, 2000.
8. Peter Deuflhard. *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms*. Springer Publishing Company, Incorporated, 2011.
9. Hilary Dollar. *Iterative linear algebra for constrained optimization*. PhD thesis, University of Oxford, 2005.
10. N.I.M. Gould, M.E. Hribar, and J. Nocedal. On the solution of equality constrained quadratic programming problems arising in optimization. *SIAM Journal on Scientific Computing*, 23(4):1376–1395, 2001.
11. Andreas Günnel and Roland Herzog. Optimal control problems in finite-strain elasticity by inner pressure and fiber tension. *Frontiers in Applied Mathematics and Statistics*, 2:4, 2016.
12. M. Heinkenschloss and D. Ridzal. A matrix-free trust-region SQP method for equality constrained optimization. *SIAM J. Optim.*, 24(3):1507–1541, 2014.
13. M. Heinkenschloss and L.N. Vicente. Analysis of inexact trust-region SQP algorithms. *SIAM J. Optim.*, 12(2):283–302, 2001/02.
14. Michael Hinze, René Pinnau, Michael Ulbrich, and Stefan Ulbrich. *Optimization with PDE constraints*, volume 23. Springer Science & Business Media, 2008.
15. Noboru Kikuchi and John Tinsley Oden. *Contact problems in elasticity: a study of variational inequalities and finite element methods*, volume 8. SIAM, 1988.
16. Lars Lubkoll. *An Optimal Control Approach to Implant Shape Design : Modeling, Analysis and Numerics*. PhD thesis, Bayreuth, 2015.
17. Lars Lubkoll, Anton Schiela, and Martin Weiser. An optimal control problem in polyconvex hyperelasticity. *SIAM J. Control Opt.*, 52(3):1403–1422, 2014.
18. Lars Lubkoll, Anton Schiela, and Martin Weiser. An affine covariant composite step method for optimization with PDEs as equality constraints. *Optimization Methods and Software*, 32:1132–1161, 2017.
19. J.A.C. Martins and J.T. Oden. Existence and uniqueness results for dynamic contact problems with nonlinear normal and friction interface laws. *Nonlinear Analysis: Theory, Methods and Applications*, 11(3):407–428, 1987.
20. Georg Müller and Anton Schiela. On the control of time discretized dynamic contact problems. *Computational Optimization and Applications*, 68(2):243–287, Nov 2017.
21. J.T. Oden and J.A.C. Martins. Models and computational methods for dynamic friction phenomena. *Computer Methods in Applied Mechanics and Engineering*, 52(1):527–634, 1985.
22. E. O. Omojokun. *Trust Region Algorithms for Optimization with Nonlinear Equality and Inequality Constraints*. PhD thesis, Boulder, CO, USA, 1989. UMI Order No: GAX89-23520.
23. Tyrone Rees. *Preconditioning iterative methods for PDE constrained optimization*. PhD thesis, Oxford University, 2010.
24. Tyrone Rees, H. Dollar, and Andrew Wathen. Optimal solvers for PDE-constrained optimization. *SIAM J. Scientific Computing*, 32:271–298, 01 2010.
25. D. Ridzal. *Trust-region SQP methods with inexact linear system solves for large-scale optimization*. ProQuest LLC, Ann Arbor, MI, 2006. Thesis (Ph.D.)–Rice University.
26. Manuel Schaller, Anton Schiela, and Matthias Stöcklein. A composite step method with inexact step computations for PDE constrained optimization. Preprint SPP1962-098, 10 2018.
27. Anton Schiela and Matthias Stöcklein. Optimal control of static contact in finite strain elasticity. Preprint SPP1962-097, 10 2018.

28. Antonio Signorini. Sopra alcune questioni di elastostatica. *Atti della Societa Italiana per il Progresso delle Scienze*, 1933.
29. A. Vardi. A trust region algorithm for equality constrained minimization: convergence properties and implementation. *SIAM J. Numer. Anal.*, 22(3):575–591, 1985.
30. Jan Christoph Wehrstedt. *Formoptimierung mit Variationsungleichungen als Nebenbedingung und eine Anwendung in der Kieferchirurgie*. PhD thesis, 2007.
31. M. Weiser, P. Deuflhard, and B. Erdmann. Affine conjugate adaptive Newton methods for nonlinear elastomechanics. *Opt. Meth. Softw.*, 22(3):414–431, 2007.
32. J.C. Ziems and S. Ulbrich. Adaptive multilevel inexact SQP methods for PDE-constrained optimization. *SIAM J. Optim.*, 21(1):1–40, 2011.

Algorithms Based on Abs-Linearization for Non-smooth Optimization with PDE Constraints

Olga Weiß, Andrea Walther, and Stephan Schmidt

Abstract This chapter presents two optimization algorithms to solve non-smooth optimization problems subject to PDE constraints. Throughout, all non-differentiabilities are assumed to be caused by the Lipschitz-continuous operator abs() as well as the related min() and max() operators. The two approaches are based on a special treatment of the absolute value operator called abs-linearization. They do not require any regularization for the non-smoothness but instead allow to explicitly exploit the structure caused by the non-smoothness.

Keywords Non-smooth optimization · Abs-Linearization · SALMIN · SCALi

1 Motivation and Introduction

The design of efficient solution methods for non-smooth, infinite dimensional optimization problems still forms a challenging task. This is due to the fact that one is interested in an effective as well as efficient handling of the non-smoothness in addition to a desirable degree of applicability. Although problems with specific structures of non-smoothness, in particular those of the popular L^1 or total variation regularization, can be solved via splitting or ADMM-type schemes, [10, 11], almost all established approaches for general non-smoothness are based on appropriate regularization techniques to avoid the explicit treatment of the non-smoothness such that one can apply for example semi-smooth Newton-type methods to solve the regularized and modified optimization problem, see, e.g. [4].

In contrast to smoothing and regularization approaches, we aim for an explicit exploitation of the structure caused by the non-smoothness. For this purpose, we employ a special treatment of the absolute value operator depending on the level where the non-smoothness occurs. First, one may consider the situation, where the

O. Weiß (✉) · A. Walther · S. Schmidt
Institut für Mathematik, Humboldt-Universität zu Berlin, Berlin, Germany
e-mail: olga.weiss@hu-berlin.de; andrea.walther@math.hu-berlin.de; s.schmidt@hu-berlin.de

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_15

PDE constraints are such that still a Fréchet-differentiable control-to-state operator exists. This is for example the case, when a non-smooth regularization is added to the target function, but the PDE constraint is smooth as considered for example in [3, 20]. Then, the so-called abs-linearization as established already in finite dimensions see, e.g., [9, 12, 13], can be extended to the infinite case as a special handling of the absolute value operator. Second, the non-smoothness may occur at the PDE level such that the control-to-state operator is also non-smooth. Then, a special reformulation of the absolute value evaluation yields a cascade of smooth optimization problems that can be solved by standard smooth approaches. Doing so in an appropriate way allows to calculate also a solution of the original optimization problem with the non-smooth state constraint.

These observations with respect to the character of non-smooth optimization problems with PDE constraints motivate the following structure of this chapter. The abs-linearization is introduced in Sect. 2. This includes also an analysis of the properties of the resulting model. In Sect. 3, we present and discuss the resulting SALMIN algorithm for optimization problems with a Fréchet-differentiable control-to-state operator. For non-smooth PDE constraints, we introduce the SCALi algorithm in Sect. 4. Furthermore, numerical results illustrate the efficiency of the presented algorithm. Finally, we draw conclusions and provide an outlook in the final Sect. 5.

2 The Abs-Linearization

To explicitly exploit the structure caused by the kind of non-smoothness considered here, we use the abs-linearization as a special handling of the absolute value operator. The abs-linearization itself was developed in the finite dimensional setting in [12] and is already well analysed for unconstrained optimization, the solution of non-smooth equations systems and the integration of dynamical systems with non-smooth right-hand sides.

To extend this approach to the infinite dimensional setting, we consider throughout a function space V over a bounded domain $\Omega \subset \mathbb{R}^n$ that is either an L^p space with $1 < p < \infty$ or a Hilbert space such that the absolute value operator is Lipschitz-continuous. Note, that due to our choice of p, the considered function space V is a reflexive Banach space. Furthermore, we assume that the non-smoothness is only caused by the operator

$$\begin{aligned} &\mathrm{abs} : V \to V, \\ &[\mathrm{abs}(v)](x) = |v(x)| \quad \text{for every } v \in V \text{ and for almost all } x \in \Omega \end{aligned} \tag{2.1}$$

as the Nemytskii or superposition operator induced by the absolute value function. For more information on superposition operators, we refer the reader to [22]. For better readability we will sometimes omit the local argument x and thus consider

Table 1 Structured Evaluation of $\varphi(v)$

$v_0 = v$
for $i = 1, \ldots, s$ do
$\quad z_i = \psi_i((v_j)_{j<i})$
$\quad \sigma_i = \text{sign}(z_i)$
$\quad v_i = \sigma_i z_i = \text{abs}(z_i)$
end for
$w = \psi_{s+1}(v_j)_{j<s+1} = \varphi(v)$

abs(.) directly as an operator on the function space. The abs(.) operator can enforce sparsity if included appropriately in the target function, see, e.g., [3, 20]. Furthermore, it can be used to describe a class of partial differential equations involving non-smooth but Lipschitz-continuous and directionally differentiable nonlinearities such as those appearing in the two-phase Stefan problem [4].

In general Banach spaces, it is not clear whether the absolute value operator is Lipschitz-continuous, see, e.g., [7]. However in the function spaces considered here, the absolute value operator abs : $V \to V$, $\text{abs}(v) := |v|$, is Lipschitz continuous and even nonexpansive [24, Prop. 2.1].

The class of operators considered here is denoted by $C^1_{abs}(V)$ and defined as follows.

Definition 2.1 (Operator Class $C^1_{abs}(V)$) Let V as well as $V_i, \tilde{V}_i$ for $1 \leqslant i \leqslant s \in \mathbb{N}$ be reflexive Banach spaces, which preserve the Lipschitz-continuity of the Nemitzkii operator induced by the absolute value operator abs as defined in Eq. (2.1). The class $C^1_{abs}(V)$ contains all operators $\varphi : V \to \mathbb{R}$ such that φ can be represented as a composition of Lipschitz-continuously Fréchet-differentiable operators $\psi_i : V_i \to \tilde{V}_i$ and the absolute value operator.

Depending on the specific situation, the Lipschitz-continuously Fréchet differentiable operators ψ_i are mappings between various Banach spaces, which preserve the Lipschitz-continuity. However, for our purpose, only the overall mapping from V to $\mathbb{R}$ is important. Using the well-known reformulations

$$\begin{aligned} \min(v, u) &= (v + u - \text{abs}(v - u))/2 \qquad \text{and} \\ \max(v, u) &= (v + u + \text{abs}(v - u))/2\,, \end{aligned} \tag{2.2}$$

a large class of non-smooth operators is contained in $C^1_{abs}(V)$.

Following the idea in the finite dimensional setting, we assume that the considered non-smooth function, given here by $\varphi \in C^1_{abs}$, can be described as a composition of elemental operators that are either Lipschitz-continuously Fréchet differentiable or the absolute value operator. Subsequently, consecutive continuously Fréchet-differentiable elemental operators can be conceptually combined to obtain a representation, where all applications of the absolute value operator can be clearly identified and exploited, see Table 1.

In the finite dimensional case $V = \mathbb{R}^n$, one has $z_i \in \mathbb{R}$ and therefore $\sigma_i \in \{-1, 0, 1\}$. For the function space scenario considered here, it follows that $z_i \in V_i$ and the functions σ_i are also Nemytskii operators defined by

$$\sigma_i : V_i \to V_i, \qquad [\sigma_i(z_i)](x) = \operatorname{sign}(z_i(x)) \cdot z_i(x) \quad \text{for almost all } x \in \Omega$$

as a function of z_i. This choice ensures that $v_i = \sigma_i z_i = \operatorname{abs}(z_i) \in V_i$ holds. Furthermore, it follows from the representation in Table 1 that φ is locally Lipschitz continuous. Hence, φ is also continuous due to the assumed smoothness of ψ_i, $i = 1, \ldots, s$, [16, Thm. 3.15] and [26, Chap. 1].

The abs-linearization, applied to the class of non-smooth operators considered here, makes use of the structured evaluation and extends the propagation of derivative information in a suitable way to cover also the absolute value operator. For given elements $v, u, \Delta v, \Delta u \in V$ and a continuously Fréchet differentiable ψ, we may use the linearizations

$$\Delta w = \Delta v \pm \Delta u \qquad \text{for} \quad w = v \pm u \ , \tag{2.3}$$

$$\Delta w = \psi'(v)(\Delta v) \qquad \text{for} \quad w = \psi(v) \neq \operatorname{abs}(v) \ , \tag{2.4}$$

where $\psi'(v)$ denotes the Fréchet derivative of ψ.

For linear operators A, the linearizations are simply given by

$$\Delta w = A\, \Delta v \qquad \text{for} \quad w = A\, v \ . \tag{2.5}$$

Thus we observe the fact that Fréchet differentiation is equivalent to linearizing all elemental operators. Now the question arises which linearization to choose for the absolute value operator. Our method of choice is the so-called abs-linearization given by

$$\Delta w = \operatorname{abs}(v + \Delta v) - w \qquad \text{for} \quad w = \operatorname{abs}(v) \ . \tag{2.6}$$

As can be seen, the linearized values Δw depend on both the argument v itself and the direction Δv. If required, we will denote this dependency by $\Delta w(v; \Delta v)$. However, most of the time we will drop these arguments v and Δv for notational simplicity. Similarly, the dependence of the intermediates v_i occurring during the evaluation of φ as described in Table 1 on the argument v is denoted by $v_i(v)$.

The formal definition of the abs-linearization is therefore given by:

Definition 2.2 (Abs-Linearization) Suppose $\varphi : V \to \mathbb{R}$ is an element of the operator class $C^1_{abs}(V)$ as defined in Definition 2.1. For a fixed argument $v \in V$ and $w = \varphi(v)$ the abs-linearization $\Delta w(v; .) : V \to \mathbb{R}$ based on the linearizations Eqs. (2.3)–(2.6) is constructed in the following way:

$$
\begin{aligned}
&v_0 = v,\ \Delta v_0 = \Delta v \\
&\text{for } i = 1, \ldots, s \text{ do} \\
&\qquad z_i \quad = \psi_i((v_j)_{j<i}) \\
&\qquad \Delta z_i \ = \psi_i'((v_j)_{j<i})((\Delta v_j)_{j<i}) \\
&\qquad \sigma_i \quad = \operatorname{sign}(z_i) \\
&\qquad v_i \quad = \sigma_i z_i = \operatorname{abs}(z_i) \\
&\qquad \Delta v_i \ = \operatorname{abs}(z_i + \Delta z_i) - \operatorname{abs}(z_i) \\
&\text{end for} \\
&w = \psi_{s+1}(v_j)_{j<s+1} = \varphi(v),\ \Delta w = \psi_{s+1}'((v_j)_{j<s+1})((\Delta v_j)_{j<s+1})
\end{aligned}
$$

This approach was first introduced in [12], where the term abs-linearization was coined and is still used here in the function space setting. Moreover, the abs-linearization generates a locally linear model for the class of non-smooth functions and operators considered here, which justifies the term linearization in "abs-linearization".

Considering the overall aim, the minimization of the objective functional φ, we restate first-order necessary conditions as well as introduce Clarke's concept of generalized derivatives, see, e.g. [6, Sec 1.2].

Definition 2.3 (Clarke Generalized Gradient) Suppose, $\varphi \in C^1_{\text{abs}}(V)$, i.e., φ is also locally Lipschitz-continuous. Let $\bar{v}, h \in V$ be given. Then the limit superior

$$\limsup_{\substack{v \to \bar{v} \\ \lambda \to 0_+}} \frac{1}{\lambda} (\varphi(v + \lambda h) - \varphi(v)) \equiv \varphi^C(\bar{v}, h)$$

exists and is called *Clarke derivative* of φ at $\bar{v}$ in direction h. Since this limit superior exists for all $h \in V$, the function φ is called Clarke differentiable at $\bar{v}$. The set

$$\partial_C \varphi(\bar{v}) \equiv \{\xi \in V^* : \varphi^C(\bar{v}, h) \geqslant \xi(h)\ \forall h \in V\} \subset V^*$$

denotes the Clarke *generalized gradient* or *subdifferential* of φ at $\bar{v}$, where V^* refers to the dual space of V.

Since one has for a function $\varphi : V \to \mathbb{R}$ that is Fréchet differentiable at $\bar{v}$ the inclusion $\varphi^C(\bar{v}, .) \in \partial_C \varphi(\bar{v})$ [6, Prop. 2.2.2], the concept of Clarke derivatives fits well for the non-smooth case analysed in this chapter. As a necessary optimality condition, one has for φ being an element of the considered non-smooth function class the following result: If v^* is a minimal point of φ then the functional 0_{V^*} is an element of $\partial_C \varphi(v^*)$, see e.g. [5, Prop. 6] and [16, Theo. 3.46].

The abs-linearization provides a local model that satisfies the following approximation properties.

Proposition 2.4 (Approximation Properties) *Suppose, $\varphi \in C^1_{abs}(V)$. For all $\bar{v} \in W$ with $W \subset V$ a closed convex subset there exists a Lipschitz-continuous local model $\varphi_{loc}(\bar{v}; .) : V \to \mathbb{R}$ with $\varphi_{loc}(\bar{v}; .) \in C^1_{abs}(V)$, given by a finite composition of linear functions and the absolute value operator. There exists a constant $q > 0$ such that for all pairs $\bar{v}, v \in W$ one has*

$$\varphi(\bar{v}) = \varphi_{loc}(\bar{v}; 0), \qquad |\varphi(v) - \varphi_{loc}(\bar{v}; v - \bar{v})| \leqslant q\|v - \bar{v}\|_V^2 \,. \tag{2.7}$$

Proof See [24, Prop. 4.3]. □

The quadratic model corresponding to such a local model is then defined by

$$\varphi_Q(\bar{v}; .) \equiv \varphi_{loc}(\bar{v}; .) + q\|.\|_V^2 \tag{2.8}$$

Hence, the approach of abs-linearization can be explicitly transferred to the infinite dimensional setting while preserving the good approximation property of the generated local model. However, it should be noted that in contrast to the finite dimensional setting, the local model $\varphi_{loc}(\bar{v}; .)$ is no longer piecewise linear as this concept does not transfer to the infinite dimensional setting.

In the following sections, we will demonstrate how the abs-linearization can be used in solution algorithms to solve non-smooth optimization problems in reflexive function spaces by explicitly exploiting the non-smooth structure of the given problem.

3 The SALMIN Algorithm

In this section we present the algorithm SALMIN (Successive Abs-Linear MINimization) that targets non-smooth optimization problems, where the non-smoothness appears only in the objective functional such that there still exists a Fréchet differentiable control-to-state operator. Hence, objective functions belonging to the class of L^1-regularized problems as well as PDE-constrained optimization problems incorporating the L^1-penalty term in the objective functional fit into the considered class of non-smooth optimization problems.

The SALMIN algorithm for infinite dimensional optimization problems, as presented in [24], can be interpreted as a quadratic overestimation method due to the approximation property of the abs-linearization, see Proposition 2.7. Hence, this approach is similar to proximal-point methods as analysed for the infinite dimensional setting, for example, in [8, 14, 19]. However, it is not possible to transfer the available results directly to the situation considered here. This is due to the fact that, in contrast to the results presented in these publications, SALMIN uses a local model of the function to be minimized in the current iteration rather than the original function.

In contrast to many other approaches, we aim at first-order minimality that is defined as follows.

Definition 3.1 (First-Order Minimality) Suppose, $\varphi \in C^1_{abs}(V)$. The operator φ is called *first-order minimal* at $v_* \in V$ if one has

$$0 \leqslant \varphi'(v_*, h) \qquad \text{for all} \qquad h \in V \,.$$

Then, v_* is called *first-order minimal point*.

Algorithm 1 SALMIN

Require: Let $v_0 \in V$ be such that $\varphi(.)$ is bounded on the bounded level set $\mathcal{N}_0$, $q^0 > 0$, $\tau > 0$.
for $k = 0, 1, 2, \ldots$ **do**
Compute

$$\Delta v_k = \arg\min_{\Delta v \in V} \varphi_{loc}(v_k; \Delta v) + \tfrac{1}{2}(1+\tau) q^k \|\Delta v\|_V^2$$

if $\Delta v = 0$ **then**
STOP
end if
if $\varphi(v_k + \Delta v_k) < \varphi(v_k)$ **then**
$v_{k+1} = v_k + \Delta v_k$
Compute $q^{k+1} = \max\{q^k, \hat{q}(v_k, \Delta v_k)\}$
else
Compute $q^k = \max\{(1+\tau) q^k, \hat{q}(v_k, \Delta v_k)\}$
end if
end for

It is important to note that this property is stronger than the frequently used concept of Clarke stationary. Often, first-order minimality is also called criticality as defined in [1] and [2], where $0 \in \mathbb{R}^n$ must be a Fréchet subgradient.

For the local model given by the abs-linearization, one obtains the following result:

Lemma 3.2 *Suppose for $\varphi \in C^1_{abs}$ and $v_* \in V$ that Assumption 2.4 holds for the local model $\varphi_{loc}(v_*, .)$ in a neighbourhood of v_*. Then one has:*

1. *If the quadratic model $\varphi_Q(v_*; .)$ is Clarke stationary at $\Delta v = 0$ for one $q \geqslant 0$, then φ is Clarke stationary at v_*.*
2. *If φ is first-order minimal at v_*, then the quadratic model $\varphi_Q(v_*; .)$ is first-order minimal at the argument $\Delta v = 0$ for all $q \in \mathbb{R}$, $q \geqslant 0$.*
3. *If the quadratic model is first-order minimal at $\Delta v = 0$ for one $q \geqslant 0$, then φ is first-order minimal at v_*.*

Proof [24, Lem. 2.8] □

Hence, once more it is possible to directly transfer the result for the finite dimensional case to the infinite dimensional setting. For the solution of a class of elliptic MPECs, a similar relation of the full model and a different local model with respect to Clarke stationarity was presented in [15, Cor. 2.2].

Lemma 3.2 motivates the termination criterion given for the SALMIN algorithm as stated in Algorithm 1 aiming for first-order minimal points.

The convergence theory for this direct transfer of the SALMIN algorithm from the finite dimensional setting to a function space formulation is rather involved. Especially, additional assumptions are required to obtain strong convergence for a subsequence of the generated iterates, see [24] for a first convergence analysis. Moreover, the handling of more general situations where the control-to-state

operator is not Fréchet differentiable is so far not covered by the convergence theory. This motivates the design of a second algorithm that employs the idea of abs-linearization in a different fashion as described in the next section, such that a broader class of non-smooth PDE-constrained optimization problems can be handled.

4 The SCALi Algorithm

In contrast to the class of model problems considered before, we will now focus on optimization problems constrained by a non-smooth partial differential equation. To illustrate the approach, we consider the following class of PDE-constrained optimization problems:

$$\begin{aligned} \min_{(y,u)\in H_0^1(\Omega)\times L^2(\Omega)} \quad & \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2 \\ \text{s.t.} \quad & -\Delta y + \ell(y) - u = 0 \ \text{ in } \Omega \end{aligned} \tag{4.1}$$

with a convex and twice continuously Fréchet-differentiable objective functional and a semi-linear elliptic PDE constraint.

The non-smoothness in this model problem is given by the non-smooth operator $\ell : H_0^1(\Omega) \to L^2(\Omega)$. Throughout this section, we assume that the model problem (4.1) has the following properties:

Assumption 4.1

(i) The domain $\Omega \subset \mathbb{R}^n$, $n \in \mathbb{N}$, is a Lipschitz domain.
(ii) The operator $\ell : H_0^1(\Omega) \to L^2(\Omega)$ denotes the Nemytskii operator induced by an operator, which is bounded and measurable in $x \in \Omega$ for every fixed y, monotone in y for almost every $x \in \Omega$ and locally Lipschitz-continuous.
(iii) The operator ℓ can be expressed as composition of the absolute value function and Fréchet-differentiable operators similar to the structured evaluation as given in Table 1.

In addition to these assumptions on the non-smooth PDE, it can easily be observed that the objective functional $\mathcal{J} : H_0^1(\Omega) \times L^2(\Omega) \to \mathbb{R}$ in Eq. (4.1) is weakly lower semi-continuous and twice continuously Fréchet differentiable.

Applying standard arguments for nonlinear monotone operators [27], it can be shown that for any given control $u \in L^2(\Omega)$ the PDE of the optimization problem (4.1) is well posed and has a unique solution y. Further analysis reveals that the optimal control problem admits a solution under the given assumptions. Such non-smooth optimization problems with a PDE as constraint, which involves the non-differentiable functions abs(.), min(.) and max(.) arise in many applications. For example, a corresponding semi-linear elliptic partial differential equation

Table 2 Structured evaluation of $\ell(y)$

for $i = 1, \ldots, s$ do
$z_i = \psi_i(y, (\sigma_j z_j)_{j<i})$
$\sigma_i = \text{sign}(z_i)$
$\hat{\ell}(y, \sigma z) = \psi_{s+1}(y, (\sigma_i z_i)_{1 \leqslant i \leqslant s})$ with $\sigma z = (\sigma_1 z_1, \ldots, \sigma_s z_s)$

describes the deflection of a stretched thin membrane partially covered by water, see [17]. Furthermore, a similar non-smooth partial differential equation arises in free boundary problems for a confined plasma, see, e.g., [17, 21].

The non-smoothness in the governing PDE constraint causes the control-to-state operator to be non-smooth as well. Hence the classical chain rule is no longer valid, which makes it challenging to consider a general reduced unconstrained problem formulation which is also well defined and unique. This is the reason why standard optimal control techniques for obtaining first-order optimal points and also the algorithm SALMIN as proposed in the last section cannot be applied. Therefore, we employ here a penalty-based approach to treat the PDE constraint explicitly, where we follow the key idea for the finite dimensional case in that stationary points are determined by an appropriate decomposition of the original problem into several smooth so-called branch problems. Each of these smooth branch problems can be solved by classical methods for smooth PDE-constrained optimization. Then, the exploitation of standard optimality conditions for the smooth case determines the next branch problem and ensures the reduction of the target function value. In deriving necessary optimality conditions, the difficulty lies in the fact that, while the solution domain of the PDE is compact, the number and location of the solutions is unknown. For this reason, a direct approach, i.e., first-discretize-then-optimize, is presented for the numerical solution of the optimization problems.

Similar to the previous two sections, we assume that the non-smooth operator ℓ can be described as a composition of elemental functions that are either continuously Fréchet differentiable or the absolute value operator. The structured evaluation procedure from Table 1 adapted to the operator ℓ is shown in Table 2 and results in an equivalent reformulation for ℓ denoted by $\hat{\ell}$.

It should be noted that $(\sigma_j z_j)_{j<i}$ indicates that ψ_i might depend also implicitly on the previously defined switching functions z_j with $j < i$. Hence, the switching function z_1 is defined as the argument of the first absolute value evaluation, i.e. as $\psi_1(y)$.

We use the notation $\hat{\ell}(y, \sigma z) = \ell(y)$ for $\sigma z = (\sigma_1 z_1, \ldots, \sigma_s z_s)$ to refer explicitly to this particular representation of the non-smooth part $\ell(y)$ based on the auxiliary variables z_i, the so-called switching functions, and σ_i, $1 \leqslant i \leqslant s$.

Clearly, the operator $\hat{\ell}(.,.)$ is not smooth in z since σ depends non-Fréchet differentiably on z. However, it is important to note, that the new function $\hat{\ell}(.,.)$ is smooth i.e., Fréchet differentiable, in its two arguments y and σz, due to the chosen formulation. This fact will be exploited later to define the smooth branch problems.

Table 3 Structured evaluation for $\ell(y) = \min(y,\ y|y|)$

z_1	$= \psi_1(y)$	$= y$
σ_1	$= \operatorname{sign}(z_1)$	
z_2	$= \psi_2(y, \sigma_1 z_1)$	$= y - y\sigma_1 z_1$
σ_2	$= \operatorname{sign}(z_2)$	
$\hat{\ell}(y, \sigma z)$	$= \psi_3(y, \sigma z)$	$= \frac{1}{2}\Big(y + y\sigma_1 z_1 - \sigma_2 z_2\Big)$

Example 1 Consider the non-smooth operator $\ell(y) = \min(y,\ y|y|)$. Exploiting the identities (2.2), we can reformulate ℓ as a function in terms of the absolute value operator and smooth elemental functions in the following way:

$$\ell(y) = \min(y,\ y|y|) = \tfrac{1}{2}\Big(y + y|y| - \big|y - y|y|\big|\Big)\,.$$

The corresponding structured evaluation is shown in Table 3.

Inserting the formulation $\hat{\ell}(y, \sigma z)$ with the auxiliary functions σ_i and z_i of ℓ into the original optimal control problem (4.1), one obtains for the functions $(y, z, u) \in H_0^1(\Omega) \times [H^1(\Omega)]^s \times L^2(\Omega)$ a smooth optimization problem with state constraints

$$\begin{aligned} \min_{y,z,u,\sigma} \quad & \tfrac{1}{2}\|y - y_d\|_{L^2}^2 + \tfrac{\alpha}{2}\|u\|_{L^2}^2 \\ \text{s.t} \quad & -\Delta y + \hat{\ell}(y, \sigma z) - u = 0 \\ & \left.\begin{aligned} \psi_i(y, (\sigma_j z_j)_{j<i}) - z_i \quad &= 0 \\ \sigma_i z_i \quad &\geqslant 0 \\ \sigma_i : \Omega \to \{-1, 0, 1\} \quad & \end{aligned}\right\} \ \forall\, i = 1, \ldots, s\,. \end{aligned} \tag{4.2}$$

Here, $[H^1(\Omega)]^s$ denotes the product $H^1(\Omega) \times \cdots \times H^1(\Omega)$ of the Hilbert spaces the switching function $z = (z_1, \ldots, z_s)$ lives on.

Assume that u^* and the corresponding $y^* := y^*(u^*)$ are solutions of the original optimization problem (4.1). Defining the auxiliary functions z_i^* and σ_i^* by

$$z_i^* = \psi_i(y, (\sigma_j^* z_j^*)_{j<i}), \qquad \sigma_i^* = \operatorname{sign}(z_i^*) \qquad \forall \quad i = 1, \ldots, s\,,$$

it follows that (y^*, z^*, u^*) is a solution of the optimization problem (4.2) if $\sigma_i = \sigma_i^*$ holds. Here, the additional equality and inequality constraints for the definitions of the additional functions z_i^* and σ_i^*, $1 \leqslant i \leqslant s$, ensure that $\sigma_i^*(z_i^*) = \operatorname{abs}(z_i^*) \in L^2(\Omega)$ is valid for $1 \leqslant i \leqslant s$. This observation motivates the optimization algorithm SCALi, i.e., a solution of a sequence of smooth subproblems of the form Eq. (4.2) to solve the original non-smooth optimization problem (4.1).

Defining and Solving the Branch Problems

The so-called branch problem corresponding to the problem formulation (4.2) for fixed functions $\bar{\sigma}_i \in L^2(\Omega)$, $\bar{\sigma}_i : \Omega \to \{-1, 1\}$ for $1 \leqslant i \leqslant s$ is defined as follows:

$$\min_{y,z,u} \quad \mathcal{J}(y,u) \tag{4.3}$$

$$\text{s.t.} \quad -\Delta y + \hat{\ell}(y, \bar{\sigma} z) - u = 0 \tag{4.4}$$

$$\psi_i(y, (\bar{\sigma}_j z_j)_{j<i}) - z_i = 0 \quad \forall\, i = 1, \ldots, s \tag{4.5}$$

$$\bar{\sigma}_i z_i \geqslant 0 \quad \forall\, i = 1, \ldots, s\,. \tag{4.6}$$

All functions occurring in this branch problem are smooth in the variables y, u and z because the function $\hat{\ell}(.,.)$ is smooth in its arguments as mentioned before. Therefore, standard smooth optimization methods can be used to solve the branch problem (4.3)–(4.6). Here we use a penalty-based approach to solve the optimization problem (4.3)–(4.6), where the constraints (4.4) and (4.5) are handled explicitly. From a formal point of view, we treat the inequality constraints (4.6) with a penalty approach such that the target function (4.3) is modified to

$$\min_{y,z,u} \quad \mathcal{J}(y,u) + \mu \int_\Omega \sum_{i=1}^{s} \Big(\max(-\bar{\sigma}_i z_i, 0) \Big)^4 \, d\Omega \tag{4.7}$$

with a penalty factor $\mu > 0$. We chose the exponent 4 to ensure that the target function is twice continuously differentiable despite the max function that is used for the formulation of the penalty function. Here, it is important to note that the penalty approach is used only to handle the inequality constraint (4.6). It is not introduced to regularize the non-smoothness.

The modified target function (4.7) coupled with the equality constraints by means of Lagrange multipliers yields the Lagrangian

$$\begin{aligned} \mathcal{L}^p(y, z, u, \lambda_{PDE}, \lambda_1, \ldots, \lambda_s) = \mathcal{J}(y,u) &+ \big(\nabla \lambda_{PDE}, \nabla y\big)_{L^2(\Omega)} \\ &+ \big(\lambda_{PDE}, \hat{\ell}(y, \bar{\sigma} z) - u\big)_{L^2(\Omega)} \\ &+ \sum_{i=1}^{s} \big(\lambda_i, \psi_i(y, (\bar{\sigma}_j z_j)_{j<i}) - z_i\big)_{L^2(\Omega)} \\ &+ \mu \int_\Omega \sum_{i=1}^{s} \Big(\max(-\bar{\sigma}_i z_i, 0) \Big)^4 \, d\Omega\,. \end{aligned} \tag{4.8}$$

A similar penalty approach was studied in [23], where the logarithm was used as barrier function. Here, we use the max function since we have to evaluate the penalty function also at 0. For a branch problem with fixed functions $\bar{\sigma}_i \in L^2(\Omega)$,

$\bar{\sigma}_i : \Omega \to \{-1, 1\}$ for $1 \leqslant i \leqslant s$, the first-order necessary optimality conditions can now be derived from the Lagrangian (4.8) using standard KKT theory for smooth PDE-constrained optimization problems. Furthermore, the optimality conditions for problem (4.2) coincide with the optimality conditions for problem (4.3)–(4.6), except for the conditions arising from the derivative with respect to the switching functions. Hence, if one computes a solution of the slightly modified branch problem, there is a very strong relation to the original non-smooth problem (4.1). This is analysed in detail in [25].

Successive Constant Abs-Linearization (SCALi)
In the following, we derive a heuristic for the switching strategy based on a discretized version of the original non-smooth problem (4.1) and the also discretized branch problems.

Applying Farkas Lemma for the discretized systems, the corresponding discrete Lagrange multiplier λ_k identifies the regions where the sign of $\bar{\sigma}_k$ has to be changed to obtain a reduction in the function value. For this purpose, the Lagrange multipliers λ_i corresponding to the solution of the current discretized branch problem are projected to the adequate function space and their max-norm is computed in order to determine the Lagrange multiplier λ_k with maximum influence. If this maximum value (almost) vanishes, the stationary point is already reached and the algorithm stops. Otherwise the sign of the corresponding discretized $\bar{\sigma}_k$ is switched at those mesh points where $|\lambda_k|$ is large and exceeds a certain threshold. If no switching occurs the algorithm stops. Otherwise the branch problem is updated accordingly and a new solution is computed by once again solving the nonlinear variational Lagrange problem by applying Newton's method.

Since the proposed algorithm is essentially motivated by the special handling of the absolute value operator, i.e., the abs-linearization, we call the resulting optimization algorithm presented in Algorithm 2 SCALi for Successive Constant Abs-Linearization.

Algorithm 2 SCALi

Input: Initial values: $\bar{\sigma}^0 = (\bar{\sigma}_1^0, \ldots, \bar{\sigma}_s^0)$, $y^0, z^0 = (z_1^0, \ldots, z_s^0), u^0$
Parameter: $\alpha, \mu, i = 0$
for $i = 0, 1, \ldots$ **do**
Solve branch problem (4.7) with constraints (4.4)–(4.5) to obtain $y^i, z^i, u^i, \lambda_{PDE}^i, \lambda^i$
Identify index κ and use λ_κ to either stop or to define $\bar{\sigma}_\kappa^{i+1}$
Set $\bar{\sigma}_k^{i+1} = \bar{\sigma}_k^i$ for $k = 1, \ldots, s, k \neq \kappa$
$i+ = i$
end for

Despite the fact that the switching from one branch problem to the other is currently based on a heuristic, we observed a successive reduction in the objective function value for the numerous examples that we considered indicating that the proposed strategy works well in practice. To illustrate this fact, the convergence

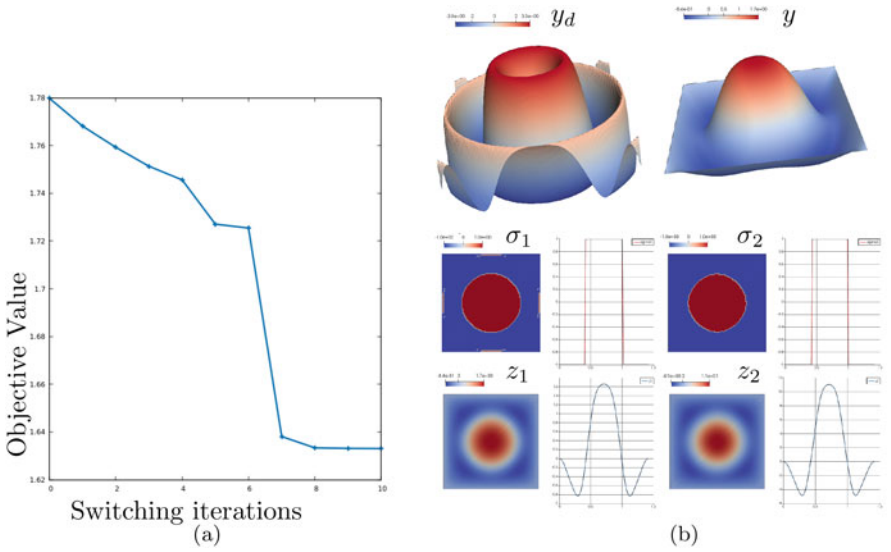

Fig. 1 (**a**) History of the objective function value with respect to the branch problem switches corresponding to the parameters given in the first row in Table 6. (**b**) Final iteration step with final branch problem and resulting solution for y, z_1 and z_2

history for one specific example is shown in Fig. 1. To some extent the convergence obtained for the sequel of discretized branch problems by the successive reduction of the objective value is not that surprising. In the discretized version, the original problem was decomposed into finitely many discretized branch problems and the objective function value decreases with each iteration step. Therefore a minimal solution must be reached after finitely many steps. This observation also leads the way to a convergence analysis in infinite dimensions by analysing the limit case when the step size defining the discretization approaches zero.

It should be noted that the algorithm proposed in this section is not limited to the considered class of semi-linear PDE or this kind of objective functionals. The arguments can easily be adapted to more general cases with, for example, a general linear elliptic differential operator of second order instead of the Laplacian operator.

Obviously, other strategies to choose the index k as alternatives to the greedy approach described here might be applied as well. Despite the fact that our heuristic for the finite dimensional setting obtained after discretization works well in practice, we will continue to develop our existing approach further and adapt it for a related systematic switching strategy of the branch problems. Furthermore, we plan to investigate the infinite dimensional case.

Numerical Results

For the actual discretizations of the branch problems, we applied a standard finite element method with piecewise linear and continuous ansatz functions for the functions y and z_i, $i = 1, \ldots, s$, and piecewise constant ansatz functions for the

control u. The resulting problem is solved by the Galerkin method combined with a Newton method for the solution of the smooth modified branch problems within the open source simulation tool FEniCS [18].

The nonlinear variational Lagrange problem is solved by Newton's method using the derivatives calculated within FEniCS. The computed solution is examined according to the switching rule and the branch problem is modified by updating the corresponding $\bar{\sigma}_i$ using the discrete Lagrange multipliers as indicators which and where to change the functions $\bar{\sigma}_i \in L^2(\Omega)$, $\bar{\sigma}_i : \Omega \to \{-1, 1\}$ for $1 \leqslant i \leqslant s$.

Example 2 For the numerical tests, we considered three different two dimensional examples:

(a)

$$\begin{aligned} &\min_{(y,u)} \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2 \\ &\text{s.t.} \quad -\Delta y + \max(0, y) - u = f \ \text{ in } \Omega = (0,1)^2 , \\ &\text{with } \ y_d(x_1, x_2) = \begin{cases} ((x_1 - \frac{1}{2})^4 + \frac{1}{2}(x_1 - \frac{1}{2})^3)\sin(\pi x_2), & \text{if } x \leqslant \frac{1}{2} \\ 0, & \text{otherwise} . \end{cases} \end{aligned} \tag{4.9}$$

(b)

$$\begin{aligned} &\min_{(y,u)} \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2 \\ &\text{s.t.} \quad -\Delta y + \min(y, y|y|) - u = 0 \ \text{ in } \Omega , \\ &\text{with } \ y_d(x_1, x_2) = (x_1 - \tfrac{1}{2})^3 \cos(\pi x_2) . \end{aligned} \tag{4.10}$$

(c)

$$\begin{aligned} &\min_{(y,u)} \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2 \\ &\text{s.t.} \quad -\Delta y + \max(5y, y|y|) - u = 0 \ \text{ in } \Omega , \\ &\text{with } \ y_d(x_1, x_2) = \frac{\sin\left(10\pi((x_1 - \frac{1}{2})^2 + (x_2 - \frac{1}{2})^2)\right)}{\sqrt{\frac{1}{100} + (x_1 - \frac{1}{2})^2 + (x_2 - \frac{1}{2})^2}} - 1 . \end{aligned} \tag{4.11}$$

Analysing these examples, one finds that the desired state is reachable for the examples (a) and (b), whereas the desired state is not reachable for example (c). The resulting effects in the convergence behaviour are discussed below.

For all three examples, the domain Ω was chosen to be the unit square and we take as an initial guess $y \equiv 0, u \equiv 0, z_1 \equiv 0, z_2 \equiv 0$. Furthermore, $\bar{\sigma}_1$ and $\bar{\sigma}_2$ are

Table 4 Numerical results for (4.9) compared with [4]

			SCALi			[4]
h	α	μ	$\frac{\|y_d - y_h\|_{L^2}}{\|y_d\|_{L^2}}$	# Switches	# Newton	# Newton
3.009e − 02	1e − 4	50	5.765e − 04	0	1	4
1.537e − 02	1e − 4	50	1.514e − 04	0	1	5
7.728e − 03	1e − 4	50	3.790e − 05	0	1	3
3.885e − 03	1e − 4	50	9.664e − 06	0	1	3
3.009e − 02	1e − 4	100	5.764e − 04	0	1	4
1.537e − 02	1e − 4	100	1.514e − 04	0	1	5
7.728e − 03	1e − 4	100	3.790e − 05	0	1	3
3.885e − 03	1e − 4	100	9.664e − 06	0	1	3
7.728e − 03	1e − 4	500	3.790e − 05	0	1	3
7.728e − 03	1e − 2	100	8.106e − 05	0	1	2
7.728e − 03	1e − 3	100	6.609e − 05	0	1	2
7.728e − 03	1e − 5	100	1.237e − 05	0	1	5
7.728e − 03	1e − 6	100	3.056e − 06	0	1	no conv.

chosen such that they fit the ones defined by the desired state y_d. We terminate the iteration if either the L^∞-Norm of the Lagrange multipliers λ_i becomes less than 10^{-9} and therefore no further switching between branch problems is done, or if the difference between the Lagrange function value which includes the bi-quadratic penalty terms and the original objective functional becomes less than 10^{-12}. The latter implicitly ensures that the sign condition $\bar{\sigma}_i z_i \geqslant 0$ is correctly adhered to.

We would like to emphasize that the vanishing Lagrange multiplier λ_k corresponds to the equality constraint Eq. (4.5) for the definition of the switching function z_k. The termination condition due to this vanishing Lagrange multiplier is based on the requirement that the associated equality constraint Eq. (4.5) is satisfied naturally at the solution.

The numerical results, considering different values of the mesh size denoted by h, the penalty parameter α for the control in the objective functional, and the penalty parameter μ in the bi-quadratic penalty term, are presented in the tables Tables 4, 5, and 6. These tables show also the quality of the resulting approximation which is given by the relative error $\|y_d - y_h\|_{L^2} / \|y_d\|_{L^2}$.

Example (a) was taken from [4]. A commonly used method for solving such non-smooth problems are semi-smooth Newton-like methods. Therefore, we also provide a comparison with results obtained with a semi-smooth Newton approach used in [4] to compute a solution of the non-smooth PDE-constrained optimization problem. It can be observed that in the more involved example, according to [4], the approach presented here requires only one single Newton step and no switches between branch problems to compute the optimal solution. Here, the fact that the desired state is reachable allows one to find a good choice for the initial σ_i functions. That is, no switching of the branch problem was required. On the other hand, the semi-smooth Newton method requires on average three to five steps to obtain the

Table 5 Numerical results for (4.10)

h	α	Objective	$\frac{\|y-y_h\|_{L^2}}{\|y\|_{L^2}}$	$\max_{i=1,2}\{\|\sigma_i z_i - \|z_i\|\|_{L^2}\}$	#Switches	#Newt.
2.8e − 02	1e − 02	5.572e − 04	9.97e − 01	1.1e − 11	0	2
2.8e − 02	1e − 03	5.434e − 04	9.73e − 01	3.8e − 10	0	2
2.8e − 02	1e − 04	4.750e − 04	8.76e − 01	5.0e − 09	0	2
2.8e − 02	1e − 06	2.183e − 04	5.54e − 01	1.5e − 08	0	2
1.4e − 02	1e − 02	5.565e − 04	9.97e − 01	9.7e − 12	0	2
1.4e − 02	1e − 03	5.426e − 04	9.73e − 01	7.7e − 11	0	2
1.4e − 02	1e − 04	4.737e − 04	8.75e − 01	6.2e − 10	0	2
1.4e − 02	1e − 06	2.143e − 04	5.47e − 01	1.8e − 09	0	2
7.7e − 03	1e − 02	5.564e − 04	9.97e − 01	1.2e − 11	0	2
7.7e − 03	1e − 03	5.425e − 04	9.73e − 01	2.9e − 11	0	2
7.7e − 03	1e − 04	4.735e − 04	8.75e − 01	4.8e − 12	0	2
7.7e − 03	1e − 06	2.133e − 04	5.45e − 01	1.2e − 11	0	2

Table 6 Numerical results for (4.11) with smooth but non-reachable y_d

h	α	μ	Objective	$\frac{\|y-y_h\|_{L^2}}{\|y\|_{L^2}}$	# Switches	# Newton
1.537e − 02	1e − 4	100	1.633	7.996e − 01	10	63
1.159e − 02	1e − 4	100	1.640	8.000e − 01	11	65
7.071e − 03	1e − 4	100	1.645	8.005e − 01	21	72
1.159e − 02	1e − 4	500	1.640	8.002e − 01	15	89
7.071e − 03	1e − 4	500	1.646	8.009e − 01	13	90
1.159e − 02	1e − 6	100	0.361	2.920e − 01	3	23
7.071e − 03	1e − 6	100	0.363	2.925e − 01	4	28

optimal solution for the considered problem. Hence, the SCALi approach reduced the numerical complexity of the solution process considerably.

We observed the no-switching behaviour for all examples that we considered if the desired state y_d is reachable. Hence, our reformulation of the non-smoothness offers one approach to solve the non-smooth PDE-constrained optimization problem with classical means, i.e., smooth optimization algorithms. To illustrate this with a more complex problem, the absolute value operators are nested in example (b), where we set $\mu = 500$. In addition to the information given already, Table 5 displays also the compliance with the absolute value of the product $\bar{\sigma}_i z_i$ which is given by $\max\{\|\sigma_i z_i - \text{abs}(z_i)\|_{L^2}\}$. Hence, the resulting inequalities are fulfilled up to a very high degree. Furthermore, it can be observed that in almost all cases only a few Newton iterations are needed to solve the problem and to compute the minimal solution. The increase in the number of Newton iterations may be caused by the nested absolute value operators yielding a more complicated optimization problem.

Once more, the fact, that SCALi does not require any switches between branch problems is mainly due to the fact that the reformulation described in Table 1 allows to exploit as much information as possible given by the optimization

problem and in particular by the given desired state y_d. The initial choice of the $\bar{\sigma}_i$ motivated by the desired state already provides the perfect guess of the σ_i. Since the desired state is reachable by the given state equation, no switches between branch problems are required and the optimal solution can be computed by solving the initial branch problem, which is already the final one. As additional observation, Table 5 suggests a further special property of the SCALi algorithms namely mesh independence. Regardless of the mesh size, the behaviour for the relative error $\|y_d - y\|_{L^2} / \|y_d\|_{L^2}$ with respect to different parameters α remains the same. Moreover, it is clearly evident that in each parameter setting the desired condition $\bar{\sigma}_i z_i = \mathrm{abs}(z_i)$ for $i \in \{1, \ldots, s\}$ is met.

Finally, example (c) considers a non-reachable desired state such that switches from one branch problem to another are required. Consequently, also the number of Newton steps required to solve each branch problem increases as illustrated in Table 6. Once more, the results obtained for the different mesh sizes indicate a mesh independent behaviour.

5 Conclusion and Outlook

We presented two approaches based on the abs-linearization technique for the solution of non-smooth optimization problems in reflexive Banach spaces. These methods enable optimization for the considered class of genuinely non-smooth problems without any substitute assumptions and regularizations for the non-smoothness.

The first method, SALMIN, represents a quadratic overestimation approach based on a local model with approximation properties of second order, constructed with the abs-linearization. It is applicable to non-smooth optimization problems in function spaces with Fréchet-differentiable control-to-state operator. It can be shown that this method results in convergence to first-order minimal points and hence a stronger stationarity concept than Clarke stationarity. The presented theory can easily be extended to more general reflexive Banach spaces where the absolute value function is Lipschitz-continuous.

For infinite dimensional optimization problems with non-smooth PDE constraints, and hence non-Fréchet-differentiable control-to-state operators the second method, SCALi, was introduced. The key idea is to appropriately decompose the non-smooth problem into smooth branch problems, which can be solved by classical smooth optimization problems. An indicator strategy is used to determine the sequence of branch problems to be solved. Several non-smooth PDE-constrained problems that fit into the considered setting were discussed and corresponding numerical results emphasize the beneficial properties and application range of the presented algorithm. The convergence in the considered discretized version is ensured by the already available convergence theory. An extension of this convergence theory to the infinite setting is subject of ongoing work.

Acknowledgments This research was supported by the DFG Priority Program SPP1962 within project 19 "Shape Optimization for Maxwell's Equations Including Hysteresis Effects in the Material Laws".

References

1. P.-A. Absil, R. Mahony, and B. Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005.
2. H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming. Series B*, 116(1–2):5–16, 2009.
3. E. Casas, R. Herzog, and G. Wachsmuth. Optimality conditions and error analysis of semilinear elliptic control problems with L^1 cost functional. *SIAM Journal on Optimization*, 22(3):795–820, 2012.
4. C. Christof, C. Clason, C. Meyer, and S. Walther. Optimal control of a non-smooth semilinear elliptic equation. *Mathematical Control and Related Fields*, pages 247–276, 2018.
5. F. Clarke. A new approach to Lagrange multipliers. *Mathematics of Operations Research*, 1:167–174, 1976.
6. F. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, 1990.
7. P.G. Dodds, T.K. Dodds, B. de Pagter, and F.A. Sukochev. Lipschitz continuity of the absolute value and Riesz projections in symmetric operator spaces. *Journal of Functional Analysis*, 148(1):28–69, 1997.
8. J. Eckstein and D. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(3 (A)):293–318, 1992.
9. S. Fiege, A. Walther, and A. Griewank. An algorithm for nonsmooth optimization by successive piecewise linearization. *Mathematical Programming, Series A*, 2018. DOI: 10.1007/s10107-018-1273-5.
10. T. Goldstein, B. O'Donoghue, S. Setzer, and R. Baraniuk. Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences*, 7(3):1588–1623, 2014.
11. T. Goldstein and S. Osher. The split Bregman method for $L1$-regularized problems. *SIAM Journal on Imaging Sciences*, 2(2):323–343, 2009.
12. A. Griewank. On stable piecewise linearization and generalized algorithmic differentiation. *Optimization Methods and Software*, 28(6):1139–1178, 2013.
13. A. Griewank and A. Walther. Relaxing kink qualifications and proving convergence rates in piecewise smooth optimization. *SIAM Journal on Optimization*, 29(1):262–289, 2019.
14. O. Guler. On the convergence of the proximal point algorithm for convex minimization. *SIAM Journal on Control and Optimization*, 29(2):403–419, 1991.
15. M. Hintermüller and T. Surowiec. A bundle-free implicit programming approach for a class of elliptic MPECs in function space. *Mathematical Programming*, 160(1–2 (A)):271–305, 2016.
16. J. Jahn. *Introduction to the theory of nonlinear optimization*. Springer, 2007.
17. F. Kikuchi, K. Nakazato, and T. Ushijima. Finite element approximation of a nonlinear eigenvalue problem related to MHD equilibria. *Japan Journal of Applied Mathematics*, 1(2):369–403, 1984.
18. A. Logg, K.-A. Mardal, and G. N. Wells. *Automated Solution of Differential Equations by the Finite Element Method*. Springer, 2012.
19. R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14:877–898, 1976.
20. G. Stadler. Elliptic optimal control problems with L^1-control cost and applications for the placement of control devices. *Computational Optimization and Applications*, 44(2), Nov 2007.

21. R. Temam. A non-linear eigenvalue problem: the shape at equilibrium of a confined plasma. *Archive for Rational Mechanics and Analysis*, 60:51–73, 1975.
22. M. Ulbrich. Nonsmooth Newton-like methods for variational inequalities and constrained optimization problems in function spaces. Habilitation, 2002. Technische Universität München.
23. M. Ulbrich and S. Ulbrich. Primal-dual interior-point methods for PDE-constrained optimization. *Mathematical Programming*, 117(1):435–485, 2009.
24. A. Walther, O. Weiß, A. Griewank, and S. Schmidt. Nonsmooth optimization by successive abs-linearisation in function spaces. Applicable Analysis. 2020. https://doi.org/10.1080/00036811.2020.1738397.
25. O. Weiß and A. Walther. A structure exploiting algorithm for non-smooth semi-linear elliptic optimal control problems. Submitted
26. E. Zeidler. *Applied functional analysis. Applications to mathematical physics.* Springer, 1995.
27. E. Zeidler and L.F. Boron. *Nonlinear Functional Analysis and its Applications: II/B: Nonlinear Monotone Operators*. Springer New York, 2013.

Shape Optimization for Variational Inequalities of Obstacle Type: Regularized and Unregularized Computational Approaches

Volker H. Schulz and Kathrin Welker

Abstract Two approaches to the solution of variational inequalities of obstacle type are discussed: quadratic regularization enabling a standard non-linear PDE constrained versus a novel approach avoiding regularization. The novel approach avoiding regularization is shown to provide analytic insight as well as superior numerical techniques.

Keywords Semi-smooth optimization · Variational inequality · Obstacle problem · Shape optimization · Numerical methods · Adjoint methods

Mathematics Subject Classification (2020) 65K15, 49Q10, 49M29, 35Q93, 35J86, 49J40.

1 Introduction

Shape optimization is an active field of research. It is of particular importance in a model based optimization context. For the case of process models in the form of classical partial differential equations (PDE), many questions are already answered and a methodological basis for the numerical solution of such shape optimization problems has been established. Thus, generalizations of shape optimization techniques to other types of model equations are in reach. This research is devoted to model equations in the form of variational inequalities (VI), which are ubiquitous in industrial applications. Nevertheless, the literature in this area is very scarce.

V. H. Schulz (✉)
Trier University, Trier, Germany
e-mail: volker.schulz@uni-trier.de

K. Welker
Helmut-Schmidt-University, University of the Federal Armed Forces Hamburg, Hamburg, Germany
e-mail: welker@hsu-hh.de

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_16

There are only very few approaches in the literature to the problem class of VI constrained shape optimization problems. Shape optimization of 2D elasto-plastic bodies is studied in [14] and shape optimization for 2D graph-like domains is investigated in [21]. In [27, Chap. 4], shape derivatives of elliptic VI problems are presented in the form of solutions to again VIs. Also [17] presents existence results for shape optimization problems which can be reformulated as optimal control problems, whereas [5, 8] show existence of solutions in a more general set-up. In [21], level-set methods are proposed and applied to graph-like two-dimensional problems. Moreover, [11] presents a regularization approach to the computation of shape and topological derivatives in the context of elliptic VIs and, thus, circumventing the numerical problems in [27, Chap. 4]. However, all these mentioned problems have in common that one cannot expect for an arbitrary shape functional depending on solutions to VIs to obtain the shape derivative as a linear mapping (cf. [27, Example in Chap. 1]). In order to circumvent the numerical problems related to the non-linearity of the shape derivative (cf., e.g., [27, Chap. 4]), [11] presents a regularization approach to the computation of shape and topological derivatives in the context of elliptic VIs. In one part of this paper, we also consider a regularization strategy, leading to novel possibilities to numerically exploit structures. In the other part, we avoid a regularization technique.

This paper is structured as follows. It has two major parts. In Sect. 4, we essentially circumvent the challenges of variational inequalities by employing a very smooth regularization technique leading to additional non-linearities, which are remedied by a linearization technique. Analytical and numerical investigations are performed and the viability of the approach is stated. In contrast, Sect. 5 discusses a novel approach, which avoids regularization and is found to possess superior analytic as well as numerical properties. In particular, the performance of the unregularized approach does not deteriorate, when the obstacle of a Signorini-type problem is more and more tight.

2 Model Problem and Its Challenges

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain equipped with a sufficiently smooth boundary $\partial\Omega$. This domain is assumed to be partitioned in a subdomain $\Omega_{\text{out}} \subset \Omega$ and an interior domain $\Omega_{\text{int}} \subset \Omega$ with boundary $\Gamma_{\text{int}} := \partial\Omega_{\text{int}}$ such that $\Omega_{\text{out}} \sqcup \Gamma_{\text{int}} \sqcup \Omega_{\text{int}} = \Omega$, where $\sqcup$ denotes the disjoint union. We consider Ω depending on Γ_{int}, i.e., $\Omega = \Omega(\Gamma_{\text{int}})$. In the following, we write only Ω instead of $\Omega(\Gamma_{\text{int}})$ for readability. Figure 1 illustrates this situation. In the following, the boundary Γ_{int} of the interior domain is called the interface and an element of an appropriate shape space $\mathcal{X}$. In contrast to the outer boundary $\partial\Omega$, which is assumed to be fixed, the inner boundary Γ_{int} is variable. If Γ_{int} changes, then the subdomains $\Omega_{\text{int}}, \Omega_{\text{out}} \subset \Omega$ change in a natural manner.

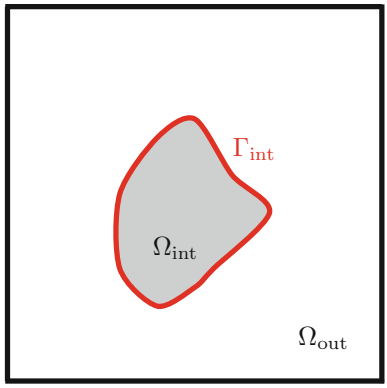

Fig. 1 Example of a domain $\Omega = \Omega_{\text{out}} \sqcup \Gamma_{\text{int}} \sqcup \Omega_{\text{int}}$.

Let $\nu > 0$ be an arbitrary constant. For the objective function

$$J(y, \Gamma_{\text{int}}) := j(y, \Gamma_{\text{int}}) + j^{\text{reg}}(\Gamma_{\text{int}}) := \frac{1}{2}\int_{\Omega} |y - \bar{y}|^2 \, dx + \nu \int_{\Gamma_{\text{int}}} 1 \, ds \tag{2.1}$$

we consider the following shape optimization problem:

$$\min_{\Gamma_{\text{int}} \in \mathcal{X}} J(y, \Gamma_{\text{int}}) \tag{2.2}$$

constrained by the following obstacle type variational inequality:

$$a(y, v - y) \geq \langle f, v - y \rangle \quad \forall v \in K := \{\theta \in H_0^1(\Omega) \colon \theta(x) \leq \psi(x) \text{ in } \Omega\}, \tag{2.3}$$

where $f \in L^2(\Omega)$ is explicitly dependent on the shape, and $\langle \cdot, \cdot \rangle$ denotes the duality pairing. Please note that the solution of (2.3) depends on the shape Γ_{int}. Of course, the formulation (2.3) is to be understood only formally because of the dependence of f from the shape. In the most general case considered in [18], $a(\cdot, \cdot)$ is a general elliptic bilinear form

$$\begin{aligned} a \colon H_0^1(\Omega) \times H_0^1(\Omega) &\to \mathbb{R} \\ (y, v) &\mapsto \sum_{i,j} a_{i,j} \partial_i y \partial_j v + \sum_i d_i (\partial_i y v + y \partial_i v) + b y v \end{aligned} \tag{2.4}$$

defined by coefficient functions $a_{i,j} \in C^1(\bar{\Omega})$, $d_j, b \in L^\infty(\Omega)$. However, for ease of presentation, we assume here that $a_{i,j} = \delta_{i,j}$ (Kronecker delta), $d_i = 0$, and $b = 0$.

With the tracking-type objective j the model is fitted to data measurements $\bar{y} \in L^2(\Omega)$. The second term j^{reg} in the objective function J is a perimeter regularization, which is frequently used to overcome ill-posedness of shape optimization problems. In (2.3), ψ denotes an obstacle which needs to be an element of $L^1_{\text{loc}}(\Omega)$ such that the set of admissible functions K is non-empty (cf. [27]). If additionally $\partial\Omega$ is $C^{1,1}$ or a polyhedron and $\psi \in H^2(\Omega)$, then the solution to (2.3) satisfies $y \in H^1_0(\Omega)$, given that the assumptions from above hold (cf. [4, 12, 28]). Further, (2.3) can be equivalently expressed as

$$a(y, v) + (\lambda, v)_{L^2(\Omega)} = (f, v)_{L^2(\Omega)} \quad \forall v \in H^1_0(\Omega) \tag{2.5}$$

$$\begin{aligned} \lambda &\geq 0 \quad \text{in } \Omega \\ y &\leq \psi \quad \text{in } \Omega \\ \lambda(y - \psi) &= 0 \quad \text{in } \Omega \end{aligned} \tag{2.6}$$

with $(\cdot, \cdot)_{L^2(\Omega)}$ denoting the L^2-scalar product and $\lambda \in L^2(\Omega)$.

It is well-known, e.g., from [4], that under these assumptions there exists a unique solution y to the obstacle type variational inequality (2.3) and an associated Lagrange multiplier λ.

The direct handling of obstacle type variational inequalities formulated as in (2.5)–(2.6) poses several problems. One problem is that in general the multiplier λ is only an element of $H^{-1}(\Omega)$, leading to severe numerical challenges. Under the assumptions above, which are also found in [13], we have $\lambda \in L^2(\Omega)$, meaning that we have a representation of the distribution as a L^2-function. It can be easily verified that this in turn gives the possibility to summarize the conditions (2.6) equivalently into a single condition of the form

$$\lambda = \max\big(0, \lambda + c(y - \psi)\big) \quad \text{for any } c > 0. \tag{2.7}$$

This formulation still leaves us with the difficulty of finding such a multiplier. For this reason, a regularization is often employed by substitution of $\lambda \in L^2(\Omega)$ by an independent $\bar{\lambda} \in L^2(\Omega)$. This results in the equation

$$a(y_c, v) + (\max\big(0, \bar{\lambda} + c(y_c - \psi)\big), v)_{L^2(\Omega)} = (f, v)_{L^2(\Omega)} \qquad \forall v \in H^1_0(\Omega). \tag{2.8}$$

Explicit dependence on λ is avoided, making the resulting semi-linear elliptic equation tractable, for example by semi-smooth Newton methods, see, e.g., [13]. Moreover, the authors of [13] prove L^2-convergence of the regularized multiplier $\max\big(0, \bar{\lambda} + c \cdot (y_c - \psi)\big)$ to the original λ for their method.

However, with problem (2.8) we are still left to solve a non-linear, semi-smooth problem, giving rise to problems concerning existence of adjoints. Hence, standard

smoothing strategies can be applied to render this problem smooth enough to show existence of adjoints. In general, it is not guaranteed that an adjoint state can be introduced (cf. [27, example in chap. 1, chap. 4]). Moreover, note that shape optimization problems constrained by VIs are especially challenging because, in general, the shape derivative of VI constrained shape optimization problems is not linear (cf. [11, 27]). This potential non-linearity of the shape derivative complicates its use in algorithms.

Before we address the question how to solve a regularized version of (2.2) constrained by (2.5)–(2.6) and the original problem (2.2) constrained by (2.5)–(2.6), i.e., the unregularized version, we give a brief overview of optimization approaches based on the Steklov–Poincaré metric in the next section, which is needed to solve the regularized and unregularized problem.

3 Optimization Based on the Steklov–Poincaré Metric

The solution techniques for the unregularized and regularized problem base on an optimization algorithm arising from the Steklov–Poincaré metric. Thus, this section focuses on this techniques and presents a way to solve the VI constrained minimization problem computationally in a suitable shape space $\mathcal{X}$. Please note that there exists no common shape space suitable for all applications. The modeling of a shape space is a challenging task and different approaches lead to diverse models. There is a multitude of shape spaces in the literature like landmark vectors, plane curves, surfaces, multiphase objects, characteristic functions of measurable sets, morphologies of images, etc. In this paper, we concentrate on the well-investigated manifold of smooth shapes in $\mathbb{R}^2$. In [20], the *set of all one-dimensional smooth shapes* is characterized by $B_e = B_e(S^1, \mathbb{R}^2) := \mathrm{Emb}(S^1, \mathbb{R}^2)/\mathrm{Diff}(S^1)$, where $\mathrm{Emb}(S^1, \mathbb{R}^2)$ denotes the set of all embeddings from the unit circle S^1 into $\mathbb{R}^2$, which contains all simple closed smooth curves in $\mathbb{R}^2$, and $\mathrm{Diff}(S^1)$ is the set of all diffeomorphisms from S^1 into itself, which characterize all smooth reparametrizations. In the following, we choose $\mathcal{X} = B_e$.

Remark 3.1 From a computational point of view, one has to deal with polygonal shape representations arising in the setting of constrained shape optimization. This is owed to the fact that finite element methods usually discretize the models. Of course, one can ask how to relax the C^∞-assumptions of shapes. Recently, the space of $H^{1/2}$-shapes, denoted by $\mathcal{H}^{1/2}$, is introduced and investigated in [30]. In [26], it is outlined that the combination of this shape space with the so-called Steklov–Poincaré metric (defined below) is an essential step toward applying efficient FE solvers. Of course, it is possible to choose this or other shape space models than B_e. However, in order to work with these weaker shape spaces, new optimization approaches need to be investigated which is beyond the scope of this paper.

If we want to optimize on a Riemannian shape manifold, we have to find a representation of the shape derivative with respect to the Riemannian metric under

consideration, called the *Riemannian shape gradient*. In [24], the authors present a metric based on the Steklov–Poincaré operator, which allows for the computation of the Riemannian shape gradient as a representative of the shape derivative in volume form. Besides saving analytical effort during the calculation process of the shape derivative, this technique is computationally more efficient than using an approach which needs the surface shape derivative form. For example, the volume form allows us to optimize directly over the hold-all domain Ω containing one or more elements $\Gamma_{\text{int}} \in B_e$, whereas the surface formulation would give us descent directions (in normal directions) for the boundary Γ_{int} only, which would not help us to move mesh elements around the shape. Additionally, when we are working with a surface shape derivative, we need to solve another PDE in order to get a mesh deformation in the hold-all domain Ω as outlined for example in [29]. The volumetric formulation of the shape derivative has also been used in a number of other publications [9, 15, 16].

Remark 3.2 In this paper, the shape derivative of a volume shape functional j at Ω in direction of a sufficiently smooth vector field V is denoted by $Dj(\Gamma_{\text{int}})[V]$. Shape derivatives can always be expressed as boundary integrals due to the Hadamard structure theorem [27, Theorem 2.27]. Note that the shape derivative arises in two equivalent notational forms:

$$Dj(\Gamma_{\text{int}})[V] := \int_{\Omega} R(x)V(x)\,dx \qquad \text{(volume/weak formulation)}$$

$$Dj(\Gamma_{\text{int}})[V] := \int_{\Gamma_{\text{int}}} r(s)\,\langle V(s), n(s)\rangle\,ds \qquad \text{(surface/strong formulation)}$$

Here, R is a differential operator acting linearly on the vector field V and $r \in L^1(\Gamma_{\text{int}})$.

Following the ideas presented in [24], we choose the Steklov–Poincaré metric defined by

$$G^S\colon H^{1/2}(\Gamma_{\text{int}}) \times H^{1/2}(\Gamma_{\text{int}}) \to \mathbb{R},\ (v,u) \mapsto \int_{\Gamma_{\text{int}}} v(s)[(S^{\text{pr}})^{-1}u](s)\,ds,$$

where $S^{\text{pr}}\colon H^{-1/2}(\Gamma_{\text{int}}) \to H^{1/2}(\Gamma_{\text{int}})$, $v \mapsto \operatorname{tr}(U)^T n$ denotes the projected Poincaré-Steklov operator with $\operatorname{tr}\colon H_0^1(\Omega,\mathbb{R}^2) \to H^{1/2}(\Gamma_{\text{int}},\mathbb{R}^2)$ denoting the trace operator on Sobolev spaces for vector-valued functions and $U \in H_0^1(\Omega,\mathbb{R}^2)$ solving the Neumann problem

$$a^{\text{deform}}(V,U) = \int_{\Gamma_{\text{int}}} v(\operatorname{tr}(V))^T n\,ds \quad \forall V \in H_0^1(\Omega,\mathbb{R}^2), \tag{3.1}$$

where $a^{\text{deform}}\colon H_0^1(\Omega,\mathbb{R}^2)\times H_0^1(\Omega,\mathbb{R}^2)\to\mathbb{R}$ is a symmetric and coercive bilinear form. In the setting of the shape space B_e, the mesh deformation vector $U\in H_0^1(\Omega,\mathbb{R}^2)$ can be viewed as an extension of a Riemannian shape gradient to the hold-all domain Ω because of the identities

$$G^S(v,u)=DJ(y,\Gamma_{\text{int}})[V]=a^{\text{deform}}(V,U)\quad\forall V\in H_0^1(\Omega,\mathbb{R}^2),\tag{3.2}$$

where $v=(\text{tr}(V))^Tn, u=(\text{tr}(U))^Tn\in T_{\Gamma_{\text{int}}}B_e$ with $T_{\Gamma_{\text{int}}}B_e\cong\{\delta n\colon\delta\in C^\infty(S^1)\}$and $DJ(y,\Gamma_{\text{int}})[V]$ denotes the shape derivative of J at Ω in direction U with y denoting the solution of the regularized or unregularized state equation. One option for the operator $a^{\text{deform}}(\cdot,\cdot)$ is chosen to be the bilinear form associated with the linear elasticity problem. To summarize, we need to solve the following so-called *deformation equation*: find $U\in H_0^1(\Omega,\mathbb{R}^2)$ s.t.

$$a^{\text{elas}}(V,U)=DJ(y,\Gamma_{\text{int}})[V]\quad\forall V\in H_0^1(\Omega,\mathbb{R}^2).\tag{3.3}$$

In this equation, we need the solution y of the regularized or unregularized state equation. Section 4 considers the regularized case, in which a *linearized adapted primal-dual active set (laPDAS) algorithm* can be applied to solve the regularized state equation. In contrast, a solution technique for the unregularized minimization problem is given in Sect. 5. The main advantage of the Steklov–Poincaré metric approach is that the identity (3.2) holds, meaning that the Riemannian metric $G^S(\cdot,\cdot)$, which is naturally defined over the interfaces, can be equivalently reformulated in terms of the bilinear form $a(\cdot,\cdot)$ over the whole domain. This last observation is the main approach we will use in the numerical solution of our model problem.

Remark 3.3 In general, $u=(\text{tr}(U))^Tn$ is not necessarily an element of $T_{\Gamma_{\text{int}}}B_e$ because it is not ensured that $U\in H_0^1(\Omega,\mathbb{R}^2)$ is $\mathcal{C}^\infty$. Under special assumptions depending on the coefficients of a second-order partial differential operator and the right-hand side of a PDE, a weak solution U which is at least H_0^1-regular is $\mathcal{C}^\infty$ by the regularity theorem of infinite differentiability (cf. [6]).

The right-hand side of the deformation equation is given by the shape derivative which can be of mixed expressions, i.e.,

$$DJ(y,\Gamma_{\text{int}})[V]:=DJ_{\text{vol}}(\Gamma_{\text{int}})[V]+DJ_{\text{surf}}(\Gamma_{\text{int}})[V].$$

Here $J_{\text{surf}}(\Gamma_{\text{int}})$ denotes parts of the objective function leading to surface shape derivative expressions—in our setting above, the perimeter regularization j^{reg}. The shape derivative $DJ_{\text{surf}}(\Gamma_{\text{int}})[V]$ of these terms are incorporated as Neumann boundary conditions. Parts of the objective function leading to volume shape derivative expressions are denoted by $J_{\text{vol}}(\Gamma_{\text{int}})$—in our setting above, the objective function j. However, note that from a theoretical point of view the volume and

surface shape derivative formulations have to be equal to each other for all test functions. Thus, $DJ_{\text{vol}}[V]$ is assembled only for test functions V whose support includes Γ_{int}, i.e.,

$$DJ_{\text{vol}}(\Gamma_{\text{int}})[V] = 0 \quad \forall V \text{ with } \operatorname{supp}(V) \cap \Gamma_{\text{int}} = \emptyset.$$

The entire optimization algorithm is given in Fig. 1 and explained in detail in the next sections for the regularized and unregularized case.

Algorithm 1 Optimization algorithm based on the Steklov–Poincaré metric

(1) Evaluate objective

(2) Solve the state and adjoint equation

(3) Assemble the right hand-side of the deformation equation:

(i) Assemble $DJ_{\text{vol}}(\Gamma_{\text{int}})[V]$ for V with $\Gamma_{\text{int}} \cap \operatorname{supp}(V) \neq \emptyset$ as source term
(ii) Assemble $DJ_{\text{surf}}(\Gamma_{\text{int}})[V]$ in form of Neumann boundary conditions

(5) Solve the deformation equation 3.3

(6) Apply the resulting deformation $U \in H_0^1(\Omega, \mathbb{R}^2)$ to the finite element mesh

(7) Stop or go to (1)

Remark 3.4 An unmodified right hand-side of the deformation equation (3.3) leads to wrong meshes due to discretization errors. This is outlined and illustrated in [24].

Remark 3.5 In general, we need the concept of the exponential map and vector transports in order to formulate optimization methods on a shape manifold. The calculations of optimization methods have to be performed in tangent spaces because manifolds are not necessarily linear spaces. This means points from a tangent space have to be mapped to the manifold in order to get a new shape-iterate, which can be realized with the help of the exponential map. However, the computation of the exponential map is prohibitively expensive in the most applications because a calculus of variations problem must be solved or the Christoffel symbols need to be known. It is much easier and much faster to use a first-order approximation of the exponential map. In [1], it is shown that a so-called *retraction* is such a first-order approximation and sufficient in most applications. We refer to [25], where a suitable retraction on B_e is given.

4 Solution Techniques Based on the Regularized Problem

In this section, we consider the optimization Algorithm 1 in detail for the regularized model problem of (2.2) constrained by (2.5)–(2.6). In particular, we explain how the regularized state equation can be solved (Sect. 4.1) and how the deformation equation (3.3) looks like (Sect. 4.2).

As already mentioned, it is not guaranteed that the shape derivative of VI constrained shape optimization problems is linear or exists. In order to circumvent the problems of non-linearity, in [11], a regularized version of (2.2) constrained by (2.5)–(2.6) is considered, on which we focus in this section. Moreover, for convenience, we focus on a special bilinear form: We assume the bilinear form $a(\cdot,\cdot)$ of the state equation to correspond to the Laplacian $-\Delta$. In this setting, a regularized version of (2.2) constrained by (2.5)–(2.6) is given by Hintermüller and Laurain (cf. [11]):

$$\min_{\Gamma_{\text{int}} \in \mathcal{X}} \; J(y_c, \Omega) \tag{4.1}$$

subject to

$$-\Delta y_c + \lambda_c = f \quad \text{in } \Omega \tag{4.2}$$

$$y_c = 0 \quad \text{on } \partial\Omega \tag{4.3}$$

with $\lambda_c = \max(0, \overline{\lambda} + c(y_c - \psi))^2$, where $c > 0$ and $0 \leq \overline{\lambda} \in L^4(\Omega)$ fixed. In the following, we call (4.1)–(4.3) *regularized state equation* or *regularized obstacle problem*. In [11], it is mentioned that for a large parameter c the associated solution of the regularized state Eqs. (4.2)–(4.3) is an excellent approximation of the solution to the unregularized VI. Moreover, it is shown in [11] that the shape derivative for the regularized problem converges to the solution of a linear problem which depends linearly on a perturbation vector field. Numerical tests in [11] show the efficiency of the approach to introduce a regularization of the VI, which allows to apply the usual theory for obtaining shape derivatives.

4.1 Linearized Adapted Primal-dual Active Set Algorithm

In the optimization algorithm based on the Steklov–Poincaré metric (cf. Algorithm 1), we have to solve the state equation—i.e., in the setting of the regularized problem, the regularized state Eqs. (4.2)–(4.3). To solve this problem, we adapt the primal-dual active set (PDAS) algorithm given in [13], where a similar regularized problem is solved. The solution of the problem described in [13] and the solution of our model problem converge to the same result (cf. [11]). In Algorithm 2, the PDAS algorithm of [13] is adapted to our problem.

Note that Algorithm 2 involves a non-linear equation (c.f. (4.4) is not linear in y_{k+1}). In order to avoid solving a non-linear equation, we compute the increment $\Delta y := y_{k+1} - y_k$ instead of y_{k+1}, i.e., the new iterate is given by $y_{k+1} = y_k + \Delta y$. Computing Δy instead of y_{k+1} leads to the following equation:

$$a(y_k + \Delta y, v) + \left([\lambda_k + c(y_k + \Delta y - \psi)]^2 , \chi_{\mathcal{A}_{k+1}} v \right) = (f, v) \qquad \forall v \in H_0^1(\Omega). \tag{4.5}$$

Algorithm 2 Adapted PDAS (mPDAS) algorithm

(1) Choose y_0, $k = 0$ and $\lambda_0 = 0$

(2) $\mathcal{A}_{k+1} := \{x\colon\ [\lambda_k + c(y - \psi)]\,(x) > 0\}$ and $\mathcal{I}_{k+1} := \Omega \backslash \mathcal{A}_{k+1}$

(3) Compute $y_{k+1} \in H_0^1(\Omega)$ as solution of

$$a(y_{k+1}, v) + \left([\lambda_k + c(y_{k+1} - \psi)]^2\,, \chi_{\mathcal{A}_{k+1}} v\right) = (f, v) \qquad \forall v \in H_0^1(\Omega) \tag{4.4}$$

(4) $\lambda_{k+1} := \begin{cases} 0 & \text{if } x \in \mathcal{I}_{k+1} \\ \lambda_k + c(y_{k+1} - \psi) & \text{if } x \in \mathcal{A}_{k+1} \end{cases}$

(5) Stop or $k := k + 1$ and go to (2)

The second term in (4.5) is still non-linear due to the square, but it is possible to linearize this equation. By putting the linearization

$$(\lambda_k + c(y_k + \Delta y - \psi))^2 \doteq (\lambda_k + c(y_k - \psi))^2 + 2c\Delta y(\lambda_k + c(y_k - \psi))$$

in (4.5), we get

$$\begin{aligned} &a(\Delta y, v) + (2c\Delta y\,[\lambda_k + c(y_k - \psi)]\,, \chi_{\mathcal{A}_{k+1}} v) \\ &= (f, v) - a(y_k, v) - \left([\lambda_k + c(y_k - \psi)]^2, \chi_{\mathcal{A}_{k+1}} v\right) \qquad \forall v \in H_0^1(\Omega), \end{aligned} \tag{4.6}$$

which is linear in Δy. The idea of this linearization is inspired by the concept of internal numerical differentiation, which is due to Hans Georg Bock [2] and his legacy. Due to linearized version (4.6) of (4.4), we can formulate the linearized mPDAS algorithm (cf. Algorithm 3). Here the third step can be iterated several times. This should be done as soon as y_{k+1} changes significantly.

Algorithm 3 Linearized aPDAS (laPDAS) algorithm

(1) Choose y_0, $k = 0$ and $\lambda_0 = 0$

(2) $\mathcal{A}_{k+1} := \{x\colon\ [\lambda_k + c(y - \psi)]\,(x) > 0\}$ and $\mathcal{I}_{k+1} := \Omega \backslash \mathcal{A}_{k+1}$

(3) (i) Compute Δy as solution of

$$\begin{aligned} &a(\Delta y, v) + \left(2c\Delta y\,[\lambda_k + c(y_k - \psi)]\,, \chi_{\mathcal{A}_{k+1}} v\right) \\ &= (f, v) - a(y_k, v) - \left([\lambda_k + c(y_k - \psi)]^2\,, \chi_{\mathcal{A}_{k+1}} v\right) \qquad \forall v \in H_0^1(\Omega) \end{aligned}$$

(ii) $y_{k+1} := y_k + \Delta y$

(4) $\lambda_{k+1} := \begin{cases} 0 & \text{if } x \in \mathcal{I}_{k+1} \\ \lambda_k + c(y_{k+1} - \psi) & \text{if } x \in \mathcal{A}_{k+1} \end{cases}$

(5) Stop or $k := k + 1$ and go to (2)

4.2 Deformation Equation

An essential part of the shape optimization techniques outlined in Algorithm 1 is to update the finite element mesh after each iteration. For this purpose, we use a solution of the deformation Eq. (3.3). The right hand-side of this equation is given by the shape derivative and the left hand-side can be chosen for example to be the bilinear form associated with the linear elasticity problem. In the setting of linear elasticity, the deformation equation—for the model problem above—is given by

$$\int_\Omega \sigma(U) : \epsilon(V)\, dx = Dj(y_c, \Omega)[V] + Dj^{\text{reg}}(\Omega)[V] \quad \forall V \in H_0^1(\Omega, \mathbb{R}^2), \tag{4.7}$$

where $\sigma(U) := \lambda^{\text{elas}} \text{tr}(\epsilon(U))I + 2\mu^{\text{elas}}\epsilon(U)$ and $\epsilon(U) := \frac{1}{2}(\nabla U + \nabla U^T)$ are the strain and stress tensor, respectively. Here λ^{elas} and μ^{elas} denote the Lamé parameters, which can be expressed in terms of Young's modulus E and Poisson's ratio ν^{elas} as

$$\lambda^{\text{elas}} = \frac{\nu^{\text{elas}} E}{(1+\nu^{\text{elas}})(1-2\nu^{\text{elas}})}, \quad \mu^{\text{elas}} u = \frac{E}{2(1+\nu^{\text{elas}})}. \tag{4.8}$$

The right hand-side of (4.7) is given by the shape derivative. The shape derivative of j in direction V is given by

$$\begin{aligned} Dj(y_c, \Omega)[V] = \int_\Omega & -\nabla y_c^T \left(\nabla V + \nabla V^T\right) \nabla p_c - V^T \nabla f p_c - (y_c - \bar{y}) V^T \nabla \bar{y} \\ & + \text{div}(V) \left(\frac{1}{2}(y_c - \bar{y})^2 + \nabla y_c^T \nabla p_c + \lambda_c p_c - f p_c\right) dx, \end{aligned} \tag{4.9}$$

where $\lambda_c = \max(0, \overline{\lambda} + c(y_c - \psi))^2$ with $c > 0$, $0 \leq \overline{\lambda} \in L^4(\Omega)$, $\psi \in H^4(\Omega)$ with $0 < \psi \leq M$ for some $M > 0$, and $p_c \in H_0^1(\Omega)$ is the weak solution of the adjoint problem to (4.1) constrained by (4.2)–(4.3) given in strong form by the following equation (cf. [7]):

$$-\triangle p_c + 2c\sqrt{\lambda_c}\, p = -(y_c - \bar{y}) \quad \text{in } \Omega \tag{4.10}$$

$$p_c = 0 \quad \text{on } \partial\Omega. \tag{4.11}$$

The shape derivative of the perimeter regularization is given by

$$Dj^{\text{reg}}(\Gamma_{\text{int}})[V] = \nu \int_{\Gamma_{\text{int}}} \text{div}(V) - \langle V, n\rangle\, n\, ds. \tag{4.12}$$

Remark 4.1 In the literature, for the shape derivative of a perimeter regularization, the formulation

$$Dj^{\mathrm{reg}}(\Gamma_{\mathrm{int}})[V] = \nu \int_{\Gamma_{\mathrm{int}}} \kappa \, \langle V, n \rangle \, ds \qquad (\kappa := \mathrm{div}_{\Gamma_{\mathrm{int}}}(n) \text{ mean curvature of } \Gamma_{\mathrm{int}})$$

is much more known instead of (4.12). However, as outlined in [26], formulation (4.12) is attractive from a computational point of view, since the evaluation of κ in each iteration is a surface-only operation.

The solution $U : \Omega \to \mathbb{R}^2$ of (4.7) is added to the coordinates of the finite element nodes. The Lamé parameters do not need to have a physical meaning here. It is rather essential to understand their effect on the mesh deformation. Here E states the stiffness of the material, which enables to control the step size for the shape update, and λ^{elas} gives the ratio how much the mesh expands in the remaining coordinate directions when compressed in one particular direction. The Lamé parameters should not be chosen constant. In [23], it is observed that locally varying Lamé parameters have a good influence on the mesh. For example, a good strategy is to choose $\lambda^{\mathrm{elas}} = 0$ and μ^{elas} as the solution of the following Laplace equation:

$$\begin{aligned} -\Delta \mu^{\mathrm{elas}} &= 0 \quad \text{in } \Omega \\ \mu^{\mathrm{elas}} &= \mu^{\mathrm{elas}}_{\mathrm{max}} \quad \text{on } \Gamma_{\mathrm{int}} \\ \mu^{\mathrm{elas}} &= \mu^{\mathrm{elas}}_{\mathrm{min}} \quad \text{on } \partial\Omega. \end{aligned} \tag{4.13}$$

Here $\mu^{\mathrm{elas}}_{\mathrm{min}}, \mu^{\mathrm{elas}}_{\mathrm{max}} \in \mathbb{R}$ influence the step size of the optimization algorithm. A small step is achieved by the choice of a large $\mu^{\mathrm{elas}}_{\mathrm{max}}$.

4.3 Summary

In consideration of Algorithm 3 and the discussion about the deformation equation, we can formulate a technique to solve our model problem. Such a technique is outlined in Algorithm 4. In Sect. 6, we apply this technique and give the corresponding numerical results.

5 Toward the Unregularized Problem

We consider the optimization Algorithm 1 in detail for the original model problem (2.2) constrained by (2.5)–(2.6).

Algorithm 4 Optimization algorithm to solve the regularized model problem

(1) Evaluate objective

(2) Solve the regularized equation (4.2)–(4.3) with the laPDAS Algorithm 3

(3) Solve the adjoint equation (4.10)–(4.11)

(4) Assemble the linear elasticity equation (4.7):

(i) Compute μ^{elas} by solving (4.13)
(ii) Assemble the right hand-side (cf. step (3) in Algorithm 1)

(5) Solve the linear elasticity equation

(6) Apply the resulting deformation $U \in H_0^1(\Omega, \mathbb{R}^2)$ to the finite element mesh

(7) Stop or go to (1)

Besides the problem of non-linearity of the shape derivatives, we also have the problem that there is no guarantee that an adjoint to VI constrained shape optimization problem exists. As already mentioned above, standard smoothing strategies can be applied to render this problem smooth enough to show existence of adjoints. In [18], analytical investigations are done to formulate an efficient optimization algorithm for the unregularized problem. To be more precise, existence of adjoints for smoothed problems and convergence to adjoints of the unregularized problem are proved. Moreover, shape derivatives for the smoothed problem are derived and convergence to a limit object is proved.

In this section, we summarize the main analytical results (without proofs) of this work. Based on this results, we formulate an efficient optimization algorithm for the unregularized problem (5.10).

5.1 Analytical Investigations: Existence of Adjoints and Shape Derivatives

In light of [3, 22], we pose the following assumptions on the smoothed max-function, which from now on is called $\max_\gamma : \mathbb{R} \to [0, \infty)$, with $\gamma > 0$ being the smoothing parameter:

(A1) $\max_\gamma \in C^1(\Omega)$ for all $\gamma > 0$;
(A2) there exists a function $g : (0, \infty) \to [0, \infty)$ with $g(\gamma) \to 0$ as $\gamma \to \infty$, s.t. $|\max_\gamma(x) - \max(0, x)| \leq g(\gamma)$ for all $x \in \mathbb{R}$ and for all $\gamma > 0$;
(A3) $\max'_\gamma(x) \in [0, 1]$ for all $x \in \mathbb{R}$ and all $\gamma > 0$;
(A4) $\max'_\gamma$ converges uniformly to 0 on $(-\infty, -\delta)$ and 1 on (δ, ∞) for all $\delta > 0$ for $\gamma \to \infty$.

Applying $\max_\gamma$ instead of max in (2.8) gives the following equation, which we call *smooth* or *fully regularized state equation*:

$$a(y_{\gamma,c}, v) + \left(\max\nolimits_\gamma\left(\bar{\lambda} + c(y_{\gamma,c} - \psi)\right), v\right)_{L^2(\Omega)} = (f, v)_{L^2(\Omega)} \quad \forall v \in H_0^1(\Omega). \tag{5.1}$$

So linearizing the corresponding Lagrangian with respect to $y_{\gamma,c}$ results in the typical adjoint equation

$$\begin{aligned} &a(p_{\gamma,c}, v) + c \cdot \left(\operatorname{sign}_\gamma(\bar{\lambda} + c(y_{\gamma,c} - \psi)) \cdot p_{\gamma,c}, v\right)_{L^2(\Omega)} \\ &= -(y_{\gamma,c} - \bar{y}, v)_{L^2(\Omega)} \quad \forall v \in H_0^1(\Omega) \end{aligned} \tag{5.2}$$

with $\operatorname{sign}_\gamma$ being the derivative of $\max_\gamma$ (see, e.g., [22] or [10] in the context of optimal control). As in [22], smoothness of the state Eq. (5.1) in $y_{\gamma,c}$ guarantees existence of solutions to the linearized equation (5.2) for a given $L^2(\Omega)$ right-hand side and, thus, existence of adjoints in the case of the considered tracking-type objective functional (2.1).

In [18], it is shown that solutions of (5.1) converge strongly in H^1 to solutions of (2.5)–(2.6) for $\gamma, c \to \infty$ if $a(\cdot, \cdot)$ is chosen by an elliptic bilinear form as in (2.4) on a compact domain Ω with polyhedric or $C^{1,1}$–boundary, $f \in L^2(\Omega)$, $\gamma, c > 0$, $\psi \in H^2(\Omega)$, $\bar{\lambda} \in L^2(\Omega)$, and $\max_\gamma : \mathbb{R} \to \mathbb{R}$ satisfy the assumptions (A1)–(A4).

The following definition is needed to state the convergence of adjoints.

Definition 5.1 Let $\Omega \subset \mathbb{R}^n$ be a bounded, open domain with Lipschitz boundary. A set $A \subseteq \Omega$ is called regularly decomposable, if there exists an $N \in \mathbb{N}$ and disjoint, path-connected, and closed $A_i \subset \Omega$ with non-empty interior and Lipschitz boundaries ∂A_i such that $A = \bigsqcup_{i=1}^{N} A_i$.

With this definition, it is possible to state the convergence of adjoints corresponding to the fully regularized problems and to characterize the limit object.

Theorem 5.2 (Convergence of the Adjoints) *Let $\Omega \subset \mathbb{R}^2$ be a bounded, open domain with C^2-boundary. Moreover, let the following assumptions be satisfied:*

(i) *$\psi \in H^2(\Omega)$, $f \in L^2(\Omega)$, and coefficient functions $a_{i,j}, d_j, b \in L^\infty(\Omega)$ in (2.5)–(2.6);*

(ii) *The active set $A = \{x \in \Omega \mid y - \psi \geq 0\}$ corresponding to (2.5)–(2.6) is regularly decomposable;*

(iii) *$A_c := \{x \in \Omega \mid \bar{\lambda} + c \cdot (y_c - \psi) \geq 0\}$ is regularly decomposable and*

$$A_c \subseteq A \quad \forall c > 0, \tag{5.3}$$

where y_c solves the regularized state equation (2.8);

(iv) *The following convergence holds:*

$$\|\mathrm{sign}_\gamma(\bar{\lambda} + c \cdot (y_{\gamma,c} - \psi)) - \mathrm{sign}(\bar{\lambda} + c \cdot (y_c - \psi))\|_{L^1(\Omega)} \to 0 \quad \text{for } \gamma \to \infty. \tag{5.4}$$

Then the adjoints $p_{\gamma,c} \to p_c$ *in* $H_0^1(\Omega)$ *for* $\gamma \to \infty$ *for all* $c > 0$*, where* p_c *is the solution to*

$$a(p_c, v) + c \cdot \int_\Omega \mathbb{1}_{A_c} \cdot p_c \cdot v \, dx = -\int_\Omega (y_c - \bar{y}) \cdot v \, dx \quad \forall v \in H_0^1(\Omega). \tag{5.5}$$

Moreover, there exists $p \in H^{-1}(\Omega)$ *to(2.5)–(2.6) and* p *is representable as an* H_0^1*-function given by the extension of* $\tilde{p} \in H_0^1(\Omega \setminus A)$ *to* $\bar{\Omega}$*, i.e.,*

$$p = \begin{cases} \tilde{p} & \text{in } \Omega \setminus A \\ 0 & \text{in } A \end{cases}, \tag{5.6}$$

where $\tilde{p} \in H_0^1(\Omega \setminus A)$ *is the solution of the elliptic problem*

$$a_{\Omega\setminus A}(\tilde{p}, v) = -\int_{\Omega\setminus A} (y - \bar{y}) v \, dx \qquad \forall v \in H_0^1(\Omega \setminus A) \tag{5.7}$$

with $a_{\Omega\setminus A}$ *being the restriction of the bilinear form* $a(\cdot,\cdot)$ *to* $\Omega \setminus A$*. Further, the solutions* p_c *of (5.5) converge strongly in* $H_0^1(\Omega)$ *to the* H_0^1*-representation of* p*.*

Proof See [18]. □

There are a few non-trivial assumptions in Theorem 5.2: assumption (iv) and (v). To the first one: It is possible to fulfill Assumption (5.3) on inclusion of the active sets $A_c \subset A$ by choosing a sufficient $\bar{\lambda} \in L^2(\Omega)$. To be more precise, since we assume $\psi \in H^2(\Omega)$, we can choose $\bar{\lambda} := \max\{0, f - S\psi\}$ with S being the differential operator corresponding to the elliptic bilinear form $a(\cdot,\cdot)$ in (2.5), guaranteeing feasibility $y_{c_1} \le y_{c_2} \le y \le \psi$ for all $0 < c_1 \le c_2$. For the proof of this, we refer to [13, Section 3.2]. To the second one: Assumption (5.4) ensures that convergence of sign_γ is compatible with convergence of $y_{\gamma,c}$ for $\gamma \to \infty$. We refer to [18] for a working example.

Remark 5.3 The limit object $p \in H_0^1(\Omega)$ of the adjoints $p_{\gamma,c}$ as defined in (5.6) is the solution of an elliptic problem (5.7) on a domain $\Omega \setminus A$ with topological dimension greater than 0. This can be exploited in numerical computations, for instance by a fat boundary method for finite elements on domains with holes as proposed by the authors of [19].

Next, we formulate similar convergence results for the shape derivatives of the shape optimization problem constrained by the fully regularized state equation (5.1). Please remember, shape derivatives of the unregularized VI constrained shape opti-

mization problems do not exist. Nevertheless, it is possible—with the convergence results above—to show existence of an object behaving as a shape derivative as well as convergence of the shape derivatives of the fully regularized problem to the latter. We split the main results into two theorems, the first one being the shape derivative for the fully regularized equation, the second one being convergence of the former for $\gamma, c \to \infty$. For convenience, we only consider the shape functional J defined in (2.1) without regularization term j^{reg}, i.e., we focus only on j. The shape derivative of J is given by the sum of the shape derivative of j and j^{reg}, where $Dj^{\mathrm{reg}}(\Gamma_{\mathrm{int}})[V]$ is given in (4.12). Please note that the objective functional and the shape derivative in correlation with the regularized VI (5.1) depends on the parameters γ and c. In order to denote this dependency, we use the notation $j_{\gamma,c}$ and $Dj_{\gamma,c}(\Gamma_{\mathrm{int}})[V]$ for the objective functional and its shape derivative, respectively.

Theorem 5.4 *Assume the setting of the shape optimization problem above. Let the assumptions of Theorem 5.2 hold. Moreover, let $M := (a_{i,j})_{i,j=1,2}$ be the matrix of coefficient functions to the leading order terms in (2.4). Furthermore, assume $D_m(y_{\gamma,c}), D_m(p_{\gamma,c}) \in H_0^1(\Omega)$ for all $\gamma, c > 0$, where $D_m(\cdot)$ denotes the material derivative. Then the shape derivatives of j, as defined in (2.1), constrained by a fully regularized VI (5.1) in direction of a vector field $V \in H_0^1(\Omega)$ is given by*

$$
\begin{aligned}
&Dj_{\gamma,c}(\Gamma_{int})[V] \\
&= \int_\Omega -(y_{\gamma,c} - \bar{y})\nabla\bar{y}^T V - \nabla y_{\gamma,c}^T(\nabla V^T M - \nabla M \cdot V + M^T \nabla V)\nabla p_{\gamma,c} \\
&\quad + (\nabla b^T V) y_{\gamma,c} p_{\gamma,c} + y_{\gamma,c} \cdot ((\nabla d^T V)^T \nabla p_{\gamma,c} - d^T(\nabla V \nabla p_{\gamma,c})) \\
&\quad + p_{\gamma,c} \cdot ((\nabla d^T V)^T \nabla y_{\gamma,c} - d^T(\nabla V \nabla y_{\gamma,c})) \\
&\quad + \mathrm{sign}_\gamma(\bar{\lambda} + c \cdot (y_{\gamma,c} - \psi)) \cdot (\nabla\bar{\lambda} - c \cdot \nabla\psi)^T V \cdot p_{\gamma,c} - \nabla f^T V p_{\gamma,c} \\
&\quad + \mathrm{div}(V)\Big(\frac{1}{2}(y_{\gamma,c} - \bar{y})^2 + b y_{\gamma,c} p_{\gamma,c} + \sum_{i,j} a_{i,j}\partial_i y_{\gamma,c}\partial_j p_{\gamma,c} \\
&\qquad\qquad + \sum_i d_i(\partial_i y_{\gamma,c} p_{\gamma,c} + y_{\gamma,c}\partial_i p_{\gamma,c}) \\
&\qquad\qquad + \max_\gamma\big(\bar{\lambda} + c \cdot (y_{\gamma,c} - \psi)\big) p_{\gamma,c} - f p_{\gamma,c}\Big)\, dx\,.
\end{aligned}
\tag{5.8}
$$

Proof See [18]. □

Theorem 5.5 *Assume the setting of the shape optimization problem above and let the assumptions of Theorem 5.2 hold. Moreover, let $M := (a_{i,j})_{i,j=1,2}$ be the matrix of coefficient functions to the leading order terms in (2.4). Then, for all $V \in H_0^1(\Omega)$,*

the shape derivatives $Dj_{\gamma,c}(\Gamma_{int})[V]$ *in (5.8) converge to* $Dj(\Gamma_{int})[V]$ *for* $\gamma, c \to \infty$, *where*

$$\begin{aligned} &Dj(\Gamma_{int})[V] \\ &:= \int_\Omega -(y-\bar{y})\nabla\bar{y}^T V - \nabla y^T(\nabla V^T M - \nabla M \cdot V + M^T \nabla V)\nabla p \\ &\qquad + y\cdot((\nabla d^T V)^T \nabla p - d^T(\nabla V \nabla p)) + p\cdot((\nabla d^T V)^T \nabla y - d^T(\nabla V \nabla y)) \\ &\qquad + (\nabla b^T V) yp - \nabla f^T V p \\ &\qquad + \operatorname{div}(V)\Big(\frac{1}{2}(y_{\gamma,c}-\bar{y})^2 + \sum_{i,j} a_{i,j}\partial_i y \partial_j p \\ &\qquad + \sum_i d_i(\partial_i yp + y\partial_i p) + byp - fp\Big)\, dx \\ &\quad + \int_A (\psi - \bar{y})\nabla\psi^T V\, dx. \end{aligned} \tag{5.9}$$

Proof See [18]. □

Remark 5.6 If $f \in L^2(\Omega)$ or $\psi \in H^2(\Omega)$ depend explicitly on the shape Ω with shape derivatives $f', \psi' \in H_0^1(\Omega)$, then the shape derivatives need to be modified accordingly by replacing terms including $\nabla f^T V$ and $\nabla\psi^T V$ by $\nabla f^T V + f'$ and $\nabla\psi^T V + \psi'$.

5.2 Optimization Algorithm

Based on the results in the previous subsection, we can formulate an optimization algorithm to solve the unregularized problem. For convenience—as in (4)—we focus on the special bilinear form $a(\cdot,\cdot)$ corresponding to the Laplacian $-\triangle$. In this setting, (2.5)–(2.6) are given by the following VI:

$$\begin{aligned} \int_\Omega \nabla y^T \nabla v\, dx + \langle\lambda, v\rangle &= \int_\Omega f v\, dx \quad \forall v \in H_0^1(\Omega) \\ \lambda &\geq 0 \quad \text{in } \Omega \\ y &\leq \psi \quad \text{in } \Omega \\ \lambda(y-\psi) &= 0 \quad \text{in } \Omega. \end{aligned} \tag{5.10}$$

This VI can be solved, e.g., by the semi-smooth Newton method proposed in [13]. Calculating the limit p of the adjoints $p_{\gamma,c}$ as in (5.6) and (5.7) is performed in

several steps. First, a linear system corresponding to

$$\begin{aligned} -\Delta p &= -(y - \bar{y}) \quad \text{in } \Omega \\ p &= 0 \qquad\qquad\quad \text{on } \partial\Omega \end{aligned} \tag{5.11}$$

is assembled without incorporation of information from the active set A. Afterwards, the vertex indices corresponding to the points in the active set $A = \{x \in \Omega \mid y - \psi \geq 0\}$ are collected by checking the condition

$$y(x) - \psi(x) \geq -\varepsilon_{\text{adj}} \tag{5.12}$$

for some error bound $\varepsilon_{\text{adj}} > 0$. The error bound ε_{adj} is incorporated since y is feasibly approximated by y_i with the semi-smooth Newton method from [13], i.e., $y_i \leq \psi$ for all $i \in \mathbb{N}$. After this, the collected vertex indices are used to incorporate the Dirichlet boundary conditions $p = 0 \quad$ in A into the linear system corresponding to (5.11). The resulting systems can be solved with a conjugate gradient solver.

As in Sect. 4, to calculate gradients $U \in H_0^1(\Omega, \mathbb{R}^2)$, we choose the linear elasticity equation as left hand-side of the deformation equation and assemble the shape derivative $DJ(\Gamma_{\text{int}})[V] = Dj(\Gamma_{\text{int}})[V] + Dj^{\text{reg}}(\Gamma_{\text{int}})[V]$ given in (4.12) and (5.9) as the right-hand side, which ends up in the deformation equation:

$$\begin{aligned} \int_\Omega \sigma(U) : \epsilon(V)\, dx &= DJ(\Gamma_{\text{int}})[V] \qquad \forall V \in H_0^1(\Omega, \mathbb{R}^2) \\ \sigma(U) &:= \lambda^{\text{elas}} \text{tr}(U) I + 2\mu^{\text{elas}} \epsilon(U) \\ \epsilon(U) &:= \frac{1}{2}(\nabla U^T + \nabla U),\ \epsilon(V) := \frac{1}{2}(\nabla V^T + \nabla V) \end{aligned} \tag{5.13}$$

with the Lamé parameters λ^{elas} and μ^{elas}. Here, we choose $\lambda^{\text{elas}} = 0$ and μ^{elas} as the solution of the Poisson problem (4.13). All this ends up in Algorithm 5.

Remark 5.7 In order to improve the convergence, a linesearch technique can be employed. For example, one can use Armijo linesearch techniques. However, one can also use a simple backtracking linesearch with sufficient descent criterion, where U_k denotes the shape derivative calculated at the corresponding interface in Ω_k in step number k, $\mathcal{T}_{\tilde{U}}(\Omega_k) := \{y \in \mathbb{R}^2 : y = x + \tilde{U}(x) \text{ for some } x \in \Omega_k\}$ the linearized vector transport by $\tilde{U}$ and $y_{\tilde{U}}$, the state solution in $\mathcal{T}_{\tilde{U}}(\Omega_k)$.

6 Numerical Results

In this section, we implement Algorithms 4 and 5 and analyze the results of these methods for a numerical experiment, the deformation of a circle into a broken donut like shape (cf. shape with dotted lines in Fig. 2).

Algorithm 5 Optimization algorithm to solve the unregularized model problem

(1) Evaluate objective

(2) Solve the unregularized state equation (5.11)

(3) Solve the adjoint equation (5.6)–(5.7):

- (i) Assemble adjoint system (5.11) neglecting active set
- (ii) Collect vertex indices of active set by (5.12)
- (ii) Implement Dirichlet conditions of active set
- (iv) Solve modified adjoint linear system

(4) Assemble the linear elasticity Eq. (5.13):

- (i) Compute μ^{elas} by solving (4.13)
- (ii) Assemble the right hand-side (cf. step (3) in Algorithm 1)

(5) Solve the linear elasticity equation

(6) Apply the resulting deformation $U \in H_0^1(\Omega, \mathbb{R}^2)$ to the finite element mesh

(7) Stop or go to (1)

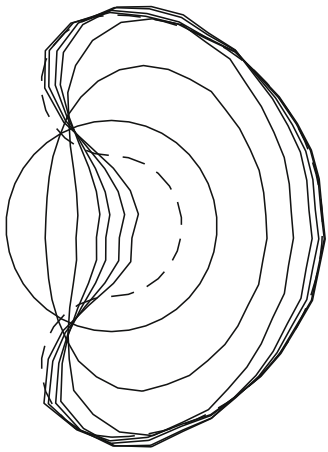

Fig. 2 Shape iterates, where the target shape is represented with dotted lines; initial shape is the circle

As already considered in the previous sections, we specialize the more general constraint (2.5)–(2.6) to a Laplacian version (cf. (5.10)). In this setting, the shape derivative 5.9 for the unregularized approach simplifies to

$$\begin{aligned} &Dj(\Gamma_{\text{int}})[V] \\ &= \int_\Omega -(y-\bar{y})\nabla\bar{y}^T V - \nabla y^T(\nabla V^T + \nabla V)\nabla p \\ &\qquad + \operatorname{div}(V)\Big(\frac{1}{2}(y-\bar{y})^2 + \nabla y^T \nabla p - fp\Big)\,dx + \int_A (\psi - \bar{y})\nabla\psi^T V\,dx. \end{aligned} \tag{6.1}$$

We use test cases within the domain $\Omega = (0, 1)^2$, which contains a compact and closed subset Ω_{int} with variable boundary Γ_{int} (cf. Fig. 1). The parameter f_{int} is valid in the interior Ω_{int} and chosen as $f_{\text{int}} = 100$, and the parameter f_{ext} is valid in the exterior $\Omega_{\text{ext}} = \Omega \backslash \overline{\Omega}_{\text{int}}$ and chosen as $f_{\text{ext}} = -10$. Further, the perimeter regularization in Eq. (2.1) is weighted by $\nu = 10^{-5}$. Moreover, the obstacle is chosen to be the following parabola:

$$\psi : \Omega \to \mathbb{R},\ (x, y) \mapsto 80 \left(\left(x - \frac{1}{2} \right)^2 + \left(y - \frac{1}{2} \right)^2 \right). \tag{6.2}$$

The calculations are performed with Python using the finite element package FEniCS. As initial shape we choose a circle with radius 0.15, illustrated in Fig. 2. The computational grid of the initial shape, which is embedded in the hold-all-domain $(0, 1)^2 \subset \mathbb{R}^2$, consists of 2184 vertices with 4206 cells, having a maximum cell diameter of 0.0359 and a minimum cell diameter of 0.018.

The target data $\bar{y} \in L^2(\Omega)$ is computed by using the mesh of the target interface to calculate a corresponding state solution of (5.10) by the semi-smooth Newton method proposed in [13]. Then, we add noise to the measurements $\bar{y}$, which is distributed according to $\mathcal{N}(0.0, 0.5)$. The state solution for the target shape is visualized in Fig. 3b for the obstacle (parabola defined in Eq. (6.2)). In contrast, Fig. 3a shows the state solution without obstacle. One can observe that the parabola bores a hole into the solution such that we lost any shape information there. We apply the same method for calculating state variables y in the unregularized optimization approach. In contrast, the regularized state equation is solved with laPDAS Algorithm 3. The adjoint p_c to the regularized equation is calculated by solving Eq. (5.5) with first-order elements by using the FEniCS standard linear algebra back end solver PETSc. In contrast, the adjoint p of the unregularized equation is calculated in several steps as outlined in Sect. 5 (cf. step (3) in Algorithm 5). To solve the resulting system, we use the standard PETSc back end conjugate gradient solver.

In order to update the finite element mesh after each iteration, we use the solution of the deformation equation, which is chosen in our experiments by (4.7) and (5.13) in the regularized and unregularized case, respectively. The solution $U : \Omega \to \mathbb{R}^2$ is then added to the coordinates of the finite element nodes. We choose the Lamé parameters in the linear elasticity equation as described above, i.e., $\lambda^{\text{elas}} = 0$ and μ^{elas} as solution of the Poisson equation (4.13), where we set $\mu^{\text{elas}}_{\text{min}} = 1$ and $\mu^{\text{elas}}_{\text{max}} = 20$, in both approaches.

The initial and final shape geometry together with the shape iterates are plotted in Fig. 2 for the unregularized approach using $\varepsilon_{\text{adj}} = 10^{-9}$. We plotted only this case because we could observe that there is a vanishing difference between approaches using regularized calculation with high c and the unregularized one. One can see that the expected shape (dashed shape) cannot be achieved. This is due to some

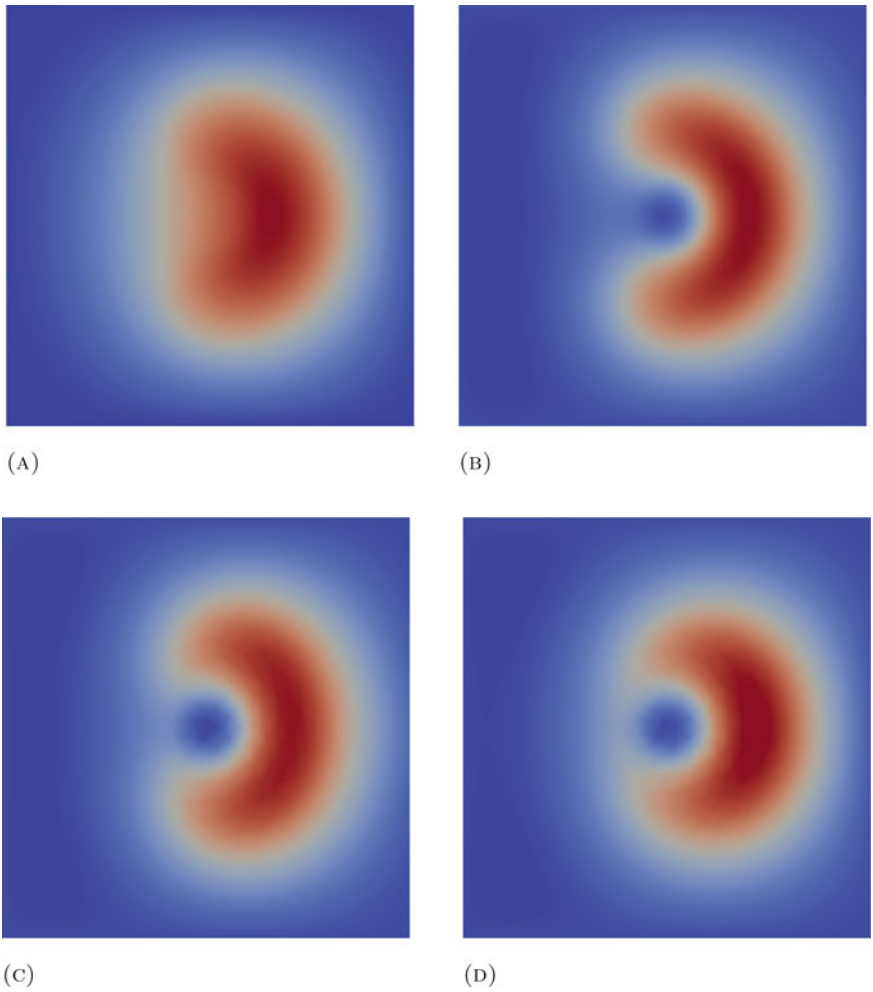

Fig. 3 Solutions of the regularized and unregularized state equation. (**a**) State solution without obstacle for target shape. (**b**) State solution with obstacle for target shape. (**c**) Unregularized state solution with obstacle for optimal shape. (**d**) Regularized state with obstacle for optimal shape

loss of shape information in active regions of the variational inequality. If we look on the state solutions for the achieved optimal shape (Fig. 3c, d), we see that these are very close to the state solution for the target shape (Fig. 3b). However, it is worth to mention that for small regularization parameters c, the solved state and adjoint equations begin to differ from the original problem and, thus, slowing down convergence, and for very low c no convergence at all.

We conclude this section with some convergence observations. The values of the objective function and the mesh distance in each iteration are given in a plot

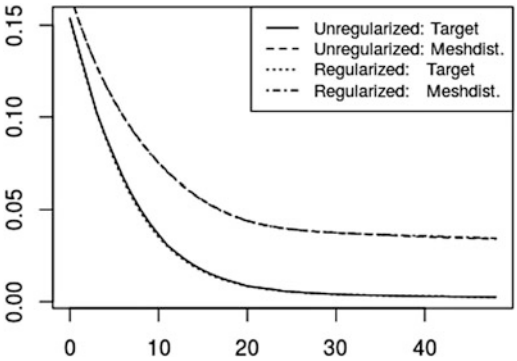

Fig. 4 Convergence rates: Mesh distances and objective values

in Fig. 4. In both approaches, the shape distance between two shapes $\Gamma^1_{\text{int}}, \Gamma^2_{\text{int}}$ is approximated by the integral

$$\int_{x\in\Gamma^1_{\text{int}}} \max_{y\in\Gamma^2_{\text{int}}} \|x-y\|_2\, dx,$$

where $\|\cdot\|_2$ denotes the Euclidean norm. One can observe that we get equal convergence rates for the regularized and unregularized approach. However, it is worth to mention that the convergence behavior of the unregularized method strongly depends on the selection of the active set. When the state solution y is not calculated with sufficient precision, the numerical errors lead to misclassification of vertex indices. Hence wrong Dirichlet conditions are incorporated in the adjoint system, creating errors in the adjoint. This makes the gradient sensitive to error for smaller ε_{adj}. In order to compensate this, the condition for checking active set indices (5.12) can be relaxed by increasing ε_{adj}. This increases likelihood of correctly classifying the true active indices, while also increasing likelihood of misclassification of inactive indices. Such a relaxation can lead to errors in the adjoint increasing with ε_{adj} and, thus, trading convergence speed for robustness. Moreover, it is worth to mention that implementing the unregularized state and adjoint becomes especially numerically exploitable with higher resolution meshes and more strongly binding obstacles ψ, i.e., larger active sets A. This is possible by sparse solvers due to the incorporation of Dirichlet conditions on the active set, as we have proposed, or by a fat boundary method as in [19]. So in contrast to the regularized method, where performance slows down for more active obstacle ψ, we do not notice unusual slowdown in performance with the unregularized method, and even offer possibility to actually benefit numerically from more binding obstacle ψ.

Acknowledgments The authors are indebted to Daniel Luft for computational support. This work has been supported by the German Research Foundation within the priority program SPP 1962 “Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization” under contract number Schu804/15-1.

References

1. P.A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.
2. H. Bock. Numerical treatment of inverse problems in chemical reaction kinetics. In K. Ebert, P. Deuflhard, and W. Jäger, editors, *Modelling of Chemical Reaction Systems*, volume 18 of *Springer Series Chemical Physics*, pages 102–125. Springer, 1981.
3. C. Christof, C. Clason, C. Meyer, and S. Walther. Optimal Control of a Non-Smooth Semilinear Elliptic Equation. *arXiv:1705.00939*, 2017.
4. G. Stampacchia D. Kinderlehrer. *An Introduction to Variational Inequalities and Their Applications*, volume 31. SIAM, 1980.
5. Z. Denkowski and S. Migorski. Optimal Shape Design for Hemivariational Inequalities. *Universitatis Iagellonicae Acta Matematica*, 36:81–88, 1998.
6. L.C. Evans. *Partial Differential Equations*. American Mathematical Society, 1993.
7. B. Führ, V.H. Schulz, and K. Welker. Shape Optimization for Interface Identification with Obstacle Problems. *Vietnam Journal of Mathematics*, 2018. DOI: 10.1007/s10013-018-0312-0.
8. L. Gasiński. Mapping Method in Optimal Shape Design Problems Governed by Hemivariational Inequalities. In J. Cagnol, M. Polis, and J.-P. Zolésio, editors, *Shape Optimization And Optimal Design*, number 216, pages 277–288. New York; Marcel Dekker, 2001.
9. Johannes Haubner, Michael Ulbrich, and Stefan Ulbrich. Analysis of shape optimization problems for unsteady fluid-structure interaction. Technical report, ZU Munich, 2019.
10. M. Hintermüller. An Active-Set Equality Constrained Newton Solver with Feasibility Restoration for Inverse Coefficient Problems in Elliptic Variational Inequalities. *Inverse Problems*, 24(3):034017, 2008.
11. M. Hintermüller and L. Laurain. Optimal Shape Design Subject to Elliptic Variational Inequalities. *SIAM Journal on Control and Optimization*, 49(3):1015–1047, 2011.
12. K. Ito and K. Kunisch. Optimal Control of Elliptic Variational Inequalities. *Applied Mathematics and Optimization*, 41(3):343–364, 2000.
13. K. Ito and K. Kunisch. Semi-Smooth Newton Methods for Variational Inequalities of the First Kind. *ESIAM: Mathematical Modelling and Numerical Analysis*, 37:41–62, 2003.
14. M. Kocvara and J. Outrata. Shape Optimization of Elasto-Plastic Bodies Governed by Variational Inequalities. In J.-P. Zolésio, editor, *Boundary Control and Variation*, number 163 in Lecture Notes in Pure and Applied Mathematics, pages 261–271. Marcel Dekker, 1994.
15. Antoine Laurain and Kevin Sturm. Distributed shape derivative via averaged adjoint method and applications, 2015.
16. C. Leithäuser, R. Feßler, and R. Pinnau. Shape optimization for stokes flows using conformal metrics. *PAMM*, pages 581–582, 2010.
17. W.B. Liu and J.E. Rubio. Optimal shape design for systems governed by variational inequalities, part 1: Existence theory for the elliptic case. *Journal of Optimization Theory and Applications*, 69(2):351–371, 1991.
18. D. Luft, V.H. Schulz, and K. Welker. Efficient techniques for shape optimization with variational inequalities using adjoints. *SIAM Journal on Optimization*, 30(3):1922–1953, 2020.
19. B. Maury. A Fat Boundary Method for the Poisson Problem in a Domain with Holes. *Journal of Scientific Computing*, 16(3):319–339, 2001.
20. P.W. Michor and D. Mumford. Riemannian geometries on spaces of plane curves. *Journal of the European Mathematical Society*, 8(1):1–48, 2006.
21. A. Myśliński. Level Set Method for Shape and Topology Optimization of Contact Problems. In *IFIP Conference on System Modeling and Optimization*, pages 397–410. Springer, 2007.
22. A. Schiela and D. Wachsmuth. Convergence Analysis of Smoothing Methods for Optimal Control of Stationary Variational Inequalities. *ESAIM: Mathematical Modelling and Numerical Analysis*, 47(3):771–787, 2013.

23. V.H. Schulz and M. Siebenborn. Computational comparison of surface metrics for PDE constrained shape optimization. *Computational Methods in Applied Mathematics*, 16(3):485–496, 2016.
24. V.H. Schulz, M. Siebenborn, and K. Welker. Efficient PDE Constrained Shape Optimization based on Steklov-Poincaré Type Metrics. *SIAM Journal on Optimization*, 26(4):2800–2819, 2016.
25. V.H. Schulz and K. Welker. On optimization transfer operators in shape spaces. In *Shape Optimization, Homogenization and Optimal Control*, pages 259–275. Springer, 2018.
26. M. Siebenborn and K. Welker. Algorithmic Aspects of Multigrid Methods for Optimization in Shape Spaces. *SIAM Journal on Scientific Computing*, 39(6):B1156–B1177, 2017.
27. J. Sokolowski and J.-P. Zolésio. *Introduction to Shape Optimization*, volume 16 of *Computational Mathematics*. Springer, 1992.
28. G.M. Troianiello. *Elliptic Differential Equations and Obstacle Problems*. Springer Science & Business Media, 2013.
29. K. Welker. Optimization in the space of smooth shapes. In F. Nielsen and F. Barbaresco, editors, *Geometric Science of Information*, volume 10589 of *Lecture Notes in Computer Science*, pages 65–72. Springer, 2017.
30. K. Welker. Suitable Spaces for Shape Optimization. *arXiv:1702.07579*, 2017. https://arxiv.org/abs/1702.07579.

Extensions of Nash Games in Finite and Infinite Dimensions with Applications

Jan Becker, Alexandra Schwartz, Sonja Steffensen, and Anna Thünen

Abstract Over the past several years, many applications have emerged, in which several agents can be accurately modeled as multi-leader-multi-follower game, where the agents are influencing each other as well as possibly some state coupled to their decisions via a differential equation. When all players decide simultaneously, much is known about the resulting finite dimensional, often convex games and there are also some analogous results for the infinite dimensional case. However, only very little is available for multi-level games in infinite or even in finite dimensions. Thus, our goal within this article is to extend the existing knowledge to be able to tackle the more general class of multi-level games. To this end, we formalize the multi-level games needed to model applications, describe which classes of games can already be solved, and provide first results to close the gap between the two. We close the article by providing some insight into possible next steps on the way to general multi-level games in function space.

Keywords Multi-leader–follower games · Nash equilibrium problems · Control problems · Optimality conditions · Existence of Nash equilibria

Mathematics Subject Classification (2020) Primary 91A35; Secondary 91A80

J. Becker · A. Schwartz (✉)
Technische Universität Darmstadt, Darmstadt, Germany
e-mail: janbecker@mathematik.tu-darmstadt.de; schwartz@gsc.tu-darmstadt.de

S. Steffensen · A. Thünen
Institut für Geometrie und Praktische Mathematik, RWTH Aachen, Aachen, Germany
e-mail: steffensen@igpm.rwth-aachen.de; thuenen@igpm.rwth-aachen.de

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_17

1 Introduction and State-of-the-Art

Models that mathematically describe and simulate the behavior of several agents, whose decisions influence each other, appear in many applications from economics, operations research, computer science, and robotics, see e.g. [1, 17, 20, 30, 32]. Often, some of the players have a temporal advantage over their rivals, i.e. there are several stages, on which decisions are made. One possibility to model such situations is a *multi-leader–follower game (MLFG)*. Here, we divide the agents into leaders and followers. The leaders ν choose their strategies u_ν first, not knowing which strategies v_i the followers i will choose later. Thus, the leaders $\nu = 1, \dots, N$ have to solve the following kind of problem:

$$\min_{u_\nu} F_\nu(u_\nu, u_{-\nu}, v) \quad \text{s.t.} \quad u_\nu \in U_\nu(u_{-\nu}), \quad v \in S(u)$$

where we use $u_{-\nu}$ as a shorthand for the strategies of all other leaders $\eta \neq \nu$, and $S(u)$ denotes the solution set of the followers' problems. The followers $i = 1, \dots, M$, already knowing the leaders' strategies u_ν, then solve the problem

$$\min_{v_i} f_i(v_i, v_{-i}, u) \quad \text{s.t.} \quad v_i \in V_i(v_{-i}, u),$$

where v_{-i} is a shorthand for the other followers' strategies v_j with $j \neq i$. Here, the individual agents' problems are allowed to be coupled both via the objective functions and the feasible sets. (Note, that if $S(u)$ is not single-valued for all u, we use an optimistic formulation for the leaders' bilevel optimization problem.) Furthermore, in the case of dynamic (i.e. time-dependent) games, the agents control a joint or individual state, such that the strategies are control functions. This leads to infinite dimensional problems. Two important special situations are the case without followers, which is called a *generalized Nash equilibrium problem (GNEP)* and the case with only a single leader, which is called a *Stackelberg game*, see Fig. 1. The most common solution concept for this kind of problem is the *Nash equilibrium*, which is a combination of feasible strategies $(u^*, v^*) = (u_1^*, \dots, u_N^*, v_1^*, \dots, v_M^*)$ such that v_i^* solves

$$\min_{v_i} f_i(v_i, v_{-i}^*, u^*) \quad \text{s.t.} \quad v_i \in V_i(v_{-i}^*, u^*),$$

for all followers $i = 1, \dots, M$ and u_ν^* solves

$$\min_{u_\nu} F_\nu(u_\nu, u_{-\nu}^*, v) \quad \text{s.t.} \quad u_\nu \in U_\nu(u_{-\nu}^*), \quad v \in S(u_\nu, u_{-\nu}^*)$$

for all leaders $\nu = 1, \dots, N$.

For convex finite dimensional GNEPs, many results on the existence of Nash equilibria and solution algorithms are known, see e.g. [11, 14] and the references therein. However, less is available for nonconvex, nondifferentiable, or infinite

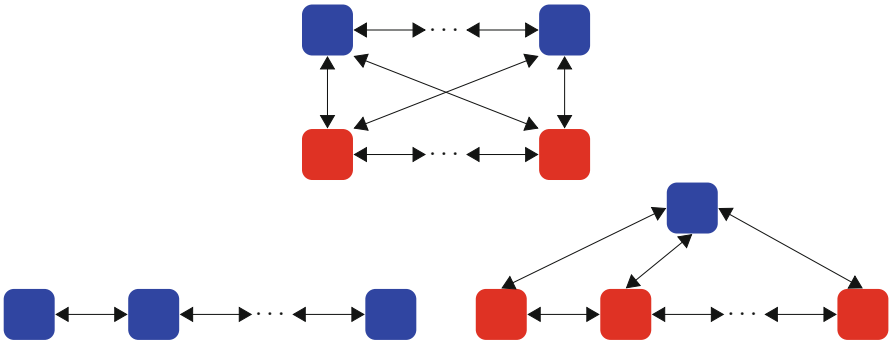

Fig. 1 A multi-leader–follower game (*top*), a generalized Nash equilibrium problem (*left*) and a Stackelberg game (*right*), where one box signifies the problem of one player

dimensional problems, see [3, 10, 13, 27, 42] for some exceptions. When a Stackelberg game is considered, especially in its reformulation as a mathematical program with complementarity constraints (MPCC), due to the arising nonconvexity the focus so far has mostly been on various kinds of stationarity and algorithms to compute stationary points, see [8, 16, 53]. Some generalizations of these ideas to infinite dimensional problems can be found in [25, 26, 52].

The generalization from a GNEP or a Stackelberg game to a MLFG generates several complications: The first difficulty appears when the followers' reaction to the leaders' strategies is not unique, because then the leaders can have different expectations of the followers' behavior. Even if the followers' response is unique and one can thus reformulate the MLFG into a single-level game, the resulting GNEP is usually nonsmooth and nonconvex. This is not surprising, since computing a Nash equilibrium of the MLFG requires a global solution of each leader's Stackelberg problem, which is usually a nonconvex problem, if one replaces the followers' optimization problems by the corresponding KKT conditions. Some results on stationarity conditions and special MLFGs can be found e.g. in [2, 28, 29, 33, 36, 42, 45, 49].

Within this paper, we provide some new steps toward MLFGs in finite and infinite dimensions. In Sect. 2, we consider finite dimensional problems, starting with a nonconvex GNEP motivated by computation offloading for mobile devices in Sect. 2.1, in which the players problems are coupled by vanishing constraints, and provide an explicit formula for the unique Nash equilibrium. Then we analyze a special class of quadratic MLFGs in Sect. 2.2, for which we can show existence and uniqueness of Nash equilibria and provide a smoothing-based solution algorithm. Afterwards, we move on to infinite dimensional games in Sect. 3 and develop some stationarity conditions for MLFGs in Banach spaces in Sect. 3.1, which are then applied to some classes of MLFGs with a single quadratic lower level problem in Sect. 3.2. A Stackelberg game with an infinite number of followers solving an optimal control problem is the topic of Sect. 3.3. Finally, we show existence and

uniqueness for a GNEP motivated by gas pipelines, where each player solves an optimal control problem with a partial differential equation in Sect. 3.4. We close the paper by pointing out some future research topics needed on the way from convex Nash games to MLFGs.

2 Games in Finite Dimensions

2.1 *A GNEP with Vanishing Constraints for Computation Offloading*

Many people nowadays use smartphones and other mobile devices for a multitude of tasks. However, even though the computational capability of these devices is steadily increasing, their limited battery capacity makes executing computationally expensive tasks such as augmented reality or video processing on a mobile device a challenge. For this reason, computation offloading, which allows mobile users to offload expensive tasks at least partially to a remote location, e.g. a cloudlet service, has become an active field of research, see for example [5, 35, 39]. We introduce a generalized Nash game to model the interaction between mobile users sharing the same cloudlet service, where the computation tasks are splittable, i.e. they can also be partially offloaded. The results presented below are an excerpt from [44], which also includes all proofs.

We consider the following model: Every user $\nu = 1, \ldots, N$ wants to complete a computational task of a certain size as fast as possible. To this end, the user offloads a percentage $u_\nu \in [0, 1]$ of this task and completes the remaining $1 - u_\nu$ locally on the mobile device. The time needed to complete the local part of the computations is given by

$$T_\nu^{\text{local}} = \alpha_\nu (1 - u_\nu),$$

where the constant $\alpha_\nu > 0$ depends on the size of the task and the local computation power. The time needed for all offloaded computations to be finished on the cloudlet is given by

$$T^{\text{offload}} = \sum_{\eta=1}^{N} \beta_\eta u_\eta + C,$$

where $\beta_\eta > 0$ depends on the computational power of the cloudlet, the size of the offloaded tasks, and possibly also the transmission rates. After all computations on the cloudlet are finished, the results are transmitted back to the users. The constant $C > 0$ allows to model situations, where some tasks are already running on the cloudlet.

All mobile users want to minimize their total completion time, i.e.

$$F_\nu(u_\nu, u_{-\nu}) = \begin{cases} T_\nu^{\text{local}} & \text{if } u_\nu = 0, \\ \max\{T_\nu^{\text{local}}, T^{\text{offload}}\} & \text{if } u_\nu > 0. \end{cases}$$

In order to eliminate the distinction of cases in the objective function, we introduce an additional variable τ_ν for the total completion time of each user ν. Then we can rewrite the optimization problem of user ν as

$$\min_{u_\nu, \tau_\nu} \ \tau_\nu \text{ s.t. } \alpha_\nu(1 - u_\nu) \le \tau_\nu, \quad \Big(\sum_{\eta=1}^N \beta_\eta u_\eta + C\Big) u_\nu \le \tau_\nu u_\nu, \quad u_\nu \in [0, 1]. \tag{2.1}$$

This is a GNEP, where the individual problems are coupled via vanishing constraints. Those vanishing constraints are coupled but not shared by all players. Since this constraint is nonconvex, most standard theory using fixed point theorems or potential game reformulations for generalized Nash games cannot be applied.

Nevertheless, a direct analysis of the game allowed us to obtain the following result including an explicit formula for the Nash equilibrium:

Theorem 2.1 *The game (2.1) has exactly one Nash equilibrium (u^*, τ^*). In this Nash equilibrium, the set of all users offloading a part of their computation to the cloudlet is given by*

$$O := \{\nu \in \{1, \dots, N\} \mid u_\nu^* > 0\} = \left\{ \nu \in \{1, \dots, N\} \mid \alpha_\nu > \frac{C + \sum_{\eta \in O} \beta_\eta}{1 + \sum_{\eta \in O} \frac{\beta_\eta}{\alpha_\eta}} \right\},$$

the equilibrium strategies for users $\nu \in O$ are given by

$$u_\nu^* = 1 - \frac{1}{\alpha_\nu} \frac{C + \sum_{\eta \in O} \beta_\eta}{1 + \sum_{\eta \in O} \frac{\beta_\eta}{\alpha_\eta}}$$

and $u_\nu^* = 0$ for $\nu \notin O$. In both cases, the equilibrium completion time is given by

$$\tau_\nu^* = \alpha_\nu(1 - u_\nu^*).$$

In this result, the set of offloading users is given implicitly. If the users are ordered such that $\alpha_1 \ge \dots \ge \alpha_N$, then the set can also be explicitly described by

$$O = \left\{ \nu \in \{1, \dots, N\} \mid \alpha_\nu > \frac{C + \sum_{\eta=1}^{\nu} \beta_\eta}{1 + \sum_{\eta=1}^{\nu} \frac{\beta_\eta}{\alpha_\eta}} \right\}.$$

This game has the interesting property that the Nash equilibrium can also be computed as the unique solution of the following centralized optimization problem:

$$\min_{u_1,\dots,u_n,\tau} \tau \text{ s.t. } \alpha_\nu(1-u_\nu) \le \tau, \quad u_\nu \in [0,1] \quad \forall \nu = 1,\dots,N,$$
$$\sum_{\nu=1}^{N} \beta_\nu u_\nu + C \le \tau,$$

although it is not a potential game in the usual sense, see e.g. [13, 33, 41].

Finally, we can also extend the game to a hierarchical setting, where "premium" users can decide first how much they want to offload. Afterwards, the "regular" users decide on their offloading strategy. The resulting game then has the following form for the premium users/leaders:

$$\min_{u_\nu,\tau_\nu} \tau_\nu \text{ s.t. } \alpha_\nu(1-u_\nu) \le \tau_\nu, \quad \Big(\sum_{\eta=1}^{N} \beta_\eta u_\eta + \sum_{i=1}^{M} b_i v_i^*\Big) u_\nu \le \tau_\nu u_\nu, \quad u_\nu \in [0,1],$$

where (v^*, t^*) is the unique Nash equilibrium of the regular users/followers game

$$\min_{v_i,t_i} t_i \text{ s.t. } a_i(1-v_i) \le t_i, \quad \Big(\sum_{j=1}^{M} b_j v_j + \sum_{\nu=1}^{N} \beta_\nu u_\nu\Big) v_i \le t_i v_i, \quad v_i \in [0,1].$$

The constant C now has the value $C(u) = \sum_{\nu=1}^{N} \beta_\nu u_\nu$. As it turns out, this MLFG has a unique Nash equilibrium, which coincides with the unique Nash equilibrium of the one-level game, where all users decide on their strategy simultaneously.

So besides having found the solution of a nonconvex GNEP with vanishing constraints, the analysis of this game raises some interesting questions: The structure of the Nash equilibria derived here is very similar to the Nash equilibria of Cournot games and all-pay auctions, which appear as lower level in MLFGs describing strategic booking and nomination decisions of gas suppliers and in Stackelberg games describing contest design problems respectively, see e.g. [15, 17]. Thus, it would be interesting to derive a closed form solution for a more general class of such GNEPs, which can then be used to eliminate the lower level from corresponding MLFGs. To obtain a description of these Nash equilibria, it could be useful to generalize the notion of potential games, looking for games which can be replaced by a joint optimization problem. Finally, we have seen that in the MLFG version, we can ignore the hierarchy and instead solve the corresponding GNEP, which is of course much easier. An interesting direction of future research would be to identify conditions under which this is possible for general MLFGs, see e.g. [34] for some results on linking bilevel problems to NEPs.

2.2 Quadratic Multi-Leader–Follower Game

The following quadratic MLFG is related to competitive interactions which appear in e.g. tolling [20, 32] and energy markets [1, 21, 22, 30].

We consider a single follower problem given by the convex optimization problem

$$\min_{v\in\mathbb{R}^m} f(v,u) = \frac{1}{2}v^\top Q_v v - b(u)^\top v \quad \text{s.t.} \quad v \geq l(u), \tag{2.2}$$

where v denotes the follower's strategy and $u = (u_1, \ldots, u_N)$ the leaders' strategies. Here, we assume that $Q_v \in \mathbb{R}^{m\times m}$ is a positive definite diagonal matrix and $b_i, l_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, M$ are convex and smooth functions describing the coupling of the player variables.

The single follower can be also interpreted as multiple followers playing a potential game; for details on potential games, the reader is referred to [41]. Due to the structure of potential games, we can assume that the number of follower players equals the total number of follower variables $M = m$, and discuss a single follower without loss of generality.

Since problem (2.2) has a strictly convex objective $f(v,u)$ and a convex strategy set, the follower admits a unique optimal solution,

$$v^*(u) = \max\left\{Q_v^{-1}b(u), l(u)\right\}, \tag{2.3}$$

also called best response, which may be derived by the KKT conditions to (2.2).

For $\nu = 1, \ldots, N$, the leader problems are given by

$$\min_{u_\nu\in\mathbb{R}^{n_\nu}} F_\nu(u_\nu, u_{-\nu}) = \frac{1}{2}u_\nu^\top Q_\nu u_\nu + c_\nu^\top u_\nu + a^\top v \quad \text{s.t.} \quad u_\nu \in U_\nu, \tag{2.4}$$

with nonempty, closed, and convex strategy sets U_ν. The quadratic objective F_ν is assumed to be strictly convex with $Q_\nu \in \mathbb{R}^{n_\nu\times n_\nu}$ symmetric positive definite, $c_\nu \in \mathbb{R}^{n_\nu}$, and $a \in \mathbb{R}^m_+$.

The MLFG (2.2) and (2.4) is reformulated by plugging the best response (2.3) in the leader game, yielding the nonsmooth Nash game for $\nu = 1, \ldots, N$

$$\min_{u_\nu\in\mathbb{R}^{n_\nu}} \frac{1}{2}u_\nu^\top Q_\nu u_\nu + c_\nu^\top u_\nu + \sum_{i=1}^{M} a_i \max\left\{\left(Q_v^{-1}b(u)\right)_i, l_i(u)\right\} \quad \text{s.t.} \quad u_\nu \in U_\nu. \tag{2.5}$$

Each optimization problem has a nonsmooth but convex objective and a convex strategy set. For compact strategy sets, we can prove the existence of Nash equilibria in the following theorem:

Theorem 2.2 (Existence of Nash Equilibria for Compact Strategy Sets) *Assume that the nonsmooth Nash equilibrium problem in (2.5) has a convex and compact joint strategy set $U = U_1 \times \cdots \times U_N$, where all U_ν are nonempty.*

Then there exists at least one Nash equilibrium. Therefore, also the quadratic MLFG given by (2.2) and (2.4) has at least one Nash equilibrium.

Proof The objectives in (2.5) are continuous in $(u_\nu, u_{-\nu})$ and convex in u_ν as a sum of a strictly convex quadratic term and the maximum of two convex functions. Furthermore, the admissible strategy sets U_ν are nonempty, convex, and compact. Thus, the conditions of [43, Theorem 3.1] are fulfilled, which guarantees the existence of at least one Nash equilibrium. □

For further studies, the data is assumed to be linear:

$$b(u) = B^\top u \quad \text{and} \quad l(u) = L^\top u,$$

where $B, L \in \mathbb{R}^{n\times m}$ and $n = n_1 + \ldots + n_N$. In the following, we analyze an approximate problem to (2.5), which is derived by smoothing the best response function (2.3). Let the smoothed best response of the follower be of the structure

$$v_\varepsilon(u) = \frac{1}{2}\left[\left(L^\top + Q_\nu^{-1}B^\top\right)u + \tilde{\phi}_\varepsilon\left(\left(L^\top - Q_\nu^{-1}B^\top\right)u\right)\right],$$

where $\tilde{\phi}_\varepsilon$ coincides with the absolute value function if the smoothing parameter ε vanishes, i.e. $\varepsilon = 0$. An example for a smooth and convex function $\tilde{\phi}_\varepsilon$ is $\tilde{\phi}_\varepsilon^{\mathrm{MIN}}(z) = \sqrt{z^2 + 4\varepsilon^2}$, which corresponds to the smooth minimum function. The smoothed best response function v_ε plugged into the leader's objectives yields a smooth Nash equilibrium problem, where for $\nu = 1, \ldots, N$ we have

$$\begin{aligned}\min_{u_\nu \in \mathbb{R}^{n_\nu}} F_\nu^\varepsilon(u_\nu, u_{-\nu}) &= \frac{1}{2}u_\nu^\top Q_\nu u_\nu + c_\nu^\top u_\nu \\ &\quad + \frac{1}{2}\sum_{i=1}^{M} a_i\left[\left(L^\top + Q_\nu^{-1}B^\top\right)u + \tilde{\phi}_\varepsilon\left(\left(L^\top - Q_\nu^{-1}B^\top\right)u\right)\right]_i \\ \text{s.t. } u_\nu &\in U_\nu.\end{aligned} \tag{2.6}$$

For this game, we state an existence and uniqueness theorem.

Theorem 2.3 (Existence and Uniqueness) *Assume that the Nash equilibrium problem (2.6) has a convex and closed strategy set $U = U_1 \times \cdots \times U_N$, where all U_ν are nonempty, and $\tilde{\phi}_\varepsilon$ is convex. Then the Nash equilibrium problem has a unique equilibrium for every smoothing parameter $\varepsilon > 0$.*

Proof (Sketch) Formulate (2.6) as variational inequality (VI) and show that the concatenated gradients of the strictly convex objective are uniformly monotone. Then, the VI has a unique solution by Facchinei and Pang [12, Theorem 2.3.3], which in turn is the unique Nash equilibrium using [12, Proposition 1.4.2]. □

This guarantees a unique Nash equilibrium of the approximate problem for every smoothing parameter $\varepsilon > 0$. It can be proven that a sequence of smoothing parameters $\varepsilon_k \to 0$ yields a sequence of Nash equilibria $u^*(\varepsilon_k)$, which has at least one accumulation point for compact U. We call that accumulation point a *limiting Nash equilibrium* and denote it by $u^*(0)$.

In order to analyze the relation of the limiting Nash equilibrium to the original MLFG, we remark that each leader's problem can be formulated as an MPCC, which together provide a GNEP formulation of the MLFG. In the following theorem, it is verified that the limiting Nash strategy $u_\nu^*(0)$ is an S-stationary point for every leader.

Theorem 2.4 *For all $\nu = 1, \ldots, N$, the limiting Nash strategy $u_\nu^*(0)$ is an S-stationary point of the leader's MPCC.*

Proof (Sketch) The limiting Nash strategy $u_\nu^*(0)$ satisfies the Fritz-John conditions of Clarke for every leader $\nu = 1, \ldots, N$. Under Slater's condition, the multipliers of MPCC strong stationarity can be constructed using the Fritz-John multipliers. □

Besides nice analytical properties, the approximating problems are also advantageous for computations since classical derivatives of the objectives are available. To compute the limiting Nash equilibrium, the smooth Nash equilibrium problem (2.6) is solved for a sequence of decreasing smoothing parameters ε_k by a globalized semismooth Newton method on the joint optimality system. In Fig. 2(left), the convergence behavior of the Nash equilibria of the approximating problems and

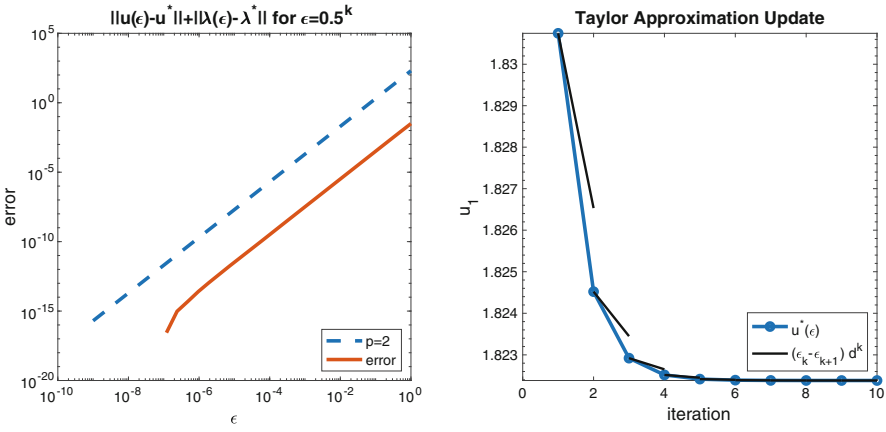

Fig. 2 Quadratic convergence to reference solution $x^*(0)$ (*left*) and Taylor expansion based update on the first entry of u, in blue the Nash equilibria for decreasing ε, in black the update (*right*)

associated multipliers $\lambda(\varepsilon_k)$ is illustrated. Usually, the Nash equilibrium to the precedent (larger) smoothing parameter initializes the current iteration, i.e. $u^0(\varepsilon_k) = u^*(\varepsilon_{k-1})$. More efficiently, the previous Nash equilibrium may be updated by

$$u^0(\varepsilon_k) = u^*(\varepsilon_{k-1}) - (\varepsilon_k - \varepsilon_{k+1})d^k \quad \text{with} \quad d^k = \frac{\partial u^*}{\partial \varepsilon}(\varepsilon_k)$$

which is based on a formal Taylor expansion of the map $\varepsilon \mapsto u^*(\varepsilon)$, c.f. Fig. 2(right). The reader is referred to [23] for more details on theory and computations for the MLFG (2.2) and (2.4).

3 Games in Infinite Dimensions

3.1 Stationarity Concepts for MLFGs in Banach Spaces

This section serves as an introduction to several stationarity concepts motivated by the work of Mehlitz and Wachsmuth (see [40, 52]) on MPCCs in Banach spaces.

We consider the following class of MLFGs: For all $\nu = 1, \ldots, N$, leader ν solves the parametric bilevel optimization problem

$$\min_{u_\nu, v} F_\nu(u_\nu, u_{-\nu}, v) \quad \text{s.t.} \quad u_\nu \in U_{\text{ad}}^\nu, \ v \in S(u_\nu, u_{-\nu}), \tag{3.1}$$

where $S(u)$ is the solution set mapping of the single follower's problem

$$\min_v f(v, u) \quad \text{s.t.} \quad v \in V_{\text{ad}}, \ g(v, u) \in K. \tag{3.2}$$

We assume that $f : U \times V \to \mathbb{R}$ and $g : U \times V \to Z$ are twice continuously Fréchet differentiable mappings where $U = U^1 \times \cdots \times U^N$ is an arbitrary Banach space and V and Z are reflexive Banach spaces. The feasible sets $V_{\text{ad}} \subseteq V$ and $K \subset Z$ are nonempty, closed, and convex cones. Additionally, we assume that f is strictly convex in v, g is $-K$-convex in v, and for all $(u, v) \in U_{\text{ad}} \times V$ with $g(u, v) \in K$ the constraint qualification (CQ)

$$\nabla_v g(v, u) \left[R_{V_{\text{ad}}}(v) \right] = Z \tag{3.3}$$

holds, where $R_{V_{\text{ad}}}(v)$ denotes the radial cone of V_{ad} with respect to v.

Then, $\bar{v} \in S(u)$ is equivalent to the following necessary and sufficient first-order optimality condition:

$$\nabla_v f(\bar{v}, u) + \nabla_v g(\bar{v}, u)^* \lambda + \mu = 0, \quad (\mu, \lambda) \in N_{V_{ad}}(\bar{v}) \times N_K\big(g(\bar{v}, u)\big)$$

where $N_C(x)$ defines the classical convex normal cone of a convex set C and A^* is the adjoint of some operator A.

So instead of analyzing the MLFG directly, we study the following *equilibrium problem with complementarity constraints* (EPCC): For all $\nu = 1, \ldots, N$, leader ν solves the parametric MPCC

$$\begin{aligned} &\min F_\nu(u_\nu, u_{-\nu}, v) \quad \text{over} (u_\nu, v, \lambda, \mu) \in U^\nu \times V \times Z^* \times V^* \\ &\text{s.t. } u_\nu \in U_{\text{ad}}^\nu, \\ &\quad \nabla_v f(u_\nu, u_{-\nu}, v) + \nabla_v g(u_\nu, u_{-\nu}, v)^* \lambda + \mu = 0, \\ &\quad v \in V_{\text{ad}},\ \mu \in V_{\text{ad}}^\circ,\ \langle \mu, v \rangle_V = 0, \\ &\quad g(u_\nu, u_{-\nu}, v) \in K,\ \lambda \in K^\circ,\ \langle \lambda, g(u_\nu, u_{-\nu}, v) \rangle_Z = 0. \end{aligned} \tag{3.4}$$

Here, we have applied the identity $N_X(x) = X^\circ \cap \{x\}^\perp$ for conic sets X with polar cone X° and annihilator $\{x\}^\perp$.

Remark 3.1 The lower level CQ (3.3) ensures that the corresponding set of Lagrange multipliers

$$\Lambda(u, \bar{v}) := \left\{ (\mu, \lambda) \in N_{V_{ad}}(\bar{v}) \times N_K\big(g(\bar{v}, u)\big) \;\middle|\; \nabla_v f(\bar{v}, u) + \nabla_v g(\bar{v}, u)^* \lambda + \mu = 0 \right\}$$

is at most single-valued for feasible (u, v). Now, let $\nu \in \{1, \ldots, N\}$ be fixed. Then it can be shown that for any $u_{-\nu} \in U_{\text{ad}}^{-\nu}$, (u_ν^*, v^*) is a solution of (3.1) if and only if there exist multipliers $(\mu, \lambda) \in \Lambda(u_\nu^*, u_{-\nu}, v^*)$ such that $(u_\nu^*, v^*, \mu, \lambda)$ is a solution of (3.4).

This motivates the following result.

Lemma 3.2 *A point (u^*, v^*) is a leader–follower equilibrium of the MLFG if and only if there exists multipliers $(\mu^*, \lambda^*) \in \Lambda(u^*, v^*)$ such that $(u_\nu^*, v^*, \mu^*, \lambda^*)$ is a solution of (3.4) with respect to the optimal opponent strategy vector $u_{-\nu}^*$ for all $\nu = 1, \ldots, N$.*

Proof (Sketch) The lemma can easily be verified by using the definition of a leader–follower equilibrium and the observation made in Remark 3.1. □

Besides the question whether a leader–follower equilibrium does exist, we cannot expect that KKT-type conditions are satisfied at solutions of MPCCs, since CQs of suitable strength for nonlinear programming as *Kurcyusz–Robinson–Zowe-Constraint Qualification* fail to hold. As a consequence, we have to deal with weaker stationarity conditions. Similar to EPCCs in finite dimensions (see [49, Definition 3.1]), we thus define stationarity concepts in Banach spaces.

Definition 3.3 A point (u^*, v^*) is an *S-stationary (C-stationary, M-stationary, W-stationary) equilibrium* of the MLFG, if there exists $(\mu^*, \lambda^*) \in \Lambda(u^*, v^*)$ such that

$(u_\nu^*, v^*, \mu^*, \lambda^*)$ is an S-stationary (C-stationary, M-stationary, W-stationary) point of (3.4) w.r.t. $u_{-\nu}^*$ for all $\nu = 1, \ldots, N$.

We assume that the mapping $F_\nu : U \times V \to \mathbb{R}$ is continuously Fréchet differentiable and $U_{\text{ad}}^\nu \subseteq U^\nu$ is a nonempty, closed, and convex set for all $\nu = 1, \ldots, N$. Moreover, we only consider the case $V_{\text{ad}} = V$. Then we can adopt [40, Definition 3.3] and obtain an extended definition of W- and S-stationarity.

Definition 3.4 Let $(u^*, v^*) \in U \times V$ be a feasible point of the MLFG and assume there exist multipliers $\lambda \in Z^*$ and $\kappa_\nu = \left(\kappa_\nu^u, \kappa_\nu^F, \kappa_\nu^g\right) \in (U^\nu)^* \times V \times Z^*$ $(\nu = 1, \ldots, N)$. Then, (u^*, v^*) is called a

(a) *W-stationary equilibrium*, if

$$\begin{aligned} 0 &= \nabla_{u_\nu} F_\nu(u^*, v^*) + \nabla_{u_\nu} g(u^*, v^*)^*[\kappa_\nu^g] + \kappa_\nu^u \\ &\quad + \left(\nabla_{vu_\nu}^2 f(u^*, v^*)^* + \langle \lambda, \nabla_{vu_\nu}^2 g(u^*, v^*)\rangle_Z^*\right)[\kappa_\nu^F], \end{aligned} \tag{3.5}$$

$$\begin{aligned} 0 &= \nabla_v F_\nu(u^*, v^*) + \nabla_v g(u^*, v^*)^*[\kappa_\nu^g] \\ &\quad + \left(\nabla_{vv}^2 f(u^*, v^*)^* + \langle \lambda, \nabla_{vv}^2 g(u^*, v^*)\rangle_Z^*\right)[\kappa_\nu^F], \end{aligned} \tag{3.6}$$

$$\lambda \in \Lambda(u^*, v^*), \ \kappa_\nu^u \in N_{U_{\text{ad}}^\nu}(u_\nu^*), \tag{3.7}$$

$$\kappa_\nu^g \in \text{cl}\left(K^\circ - K^\circ \cap \{g(u^*, v^*)\}^\perp\right) \cap \{g(u^*, v^*)\}^\perp,$$

$$-\nabla_v g(u^*, v^*)[\kappa_\nu^F] \in \text{cl}\left(K - K \cap \{\lambda\}^\perp\right) \cap \{\lambda\}^\perp$$

(b) *S-stationary equilibrium*, if (3.5)–(3.7) and

$$\kappa_\nu^g \in \mathcal{K}_{K^\circ}\left(\lambda, g(u^*, v^*)\right), \quad -\nabla_v g(u^*, v^*)[\kappa_\nu^F] \in \mathcal{K}_K\left(g(u^*, v^*), \lambda\right)$$

are satisfied for $\nu = 1, \ldots, N$, where $\mathcal{K}_K(z, z^*) = N_K(z)^\circ \cap \{z^*\}^\perp$ is the critical cone of K w.r.t. (z, z^*).

After introducing stationarity concepts in a quite general setting of MLFGs, the ongoing focus is on exploiting more concrete structures, i.e. polyhedric cones, in this abstract framework, in order to derive necessary conditions for leader–follower (stationary) equilibria.

3.2 An MLFG with Quadratic Lower Level Problem

Modeling physical or economical phenomena e.g. elasticity, elastoplasticity, and mathematical finances (see e.g. [24, 50]) using (convex) optimization problems with bound constraints naturally leads to variational inequalities (VI). Therefore,

many efforts have been made to analyze the corresponding MPEC/MPCC (see e.g. [25, 26]). Moreover, many real-world problems consist of several decision makers which compete in a non-cooperative manner, e.g. autonomous driving, predator–prey games, and economic markets [10, 27]. In our context, we consider a hierarchical extension of an equilibrium problem, where the follower's best response can be described by a VI.

Throughout this section, we consider a special class of MLFG, introduced in Sect. 3.1, in which the νth leader considers the following problem:

$$\min_{u_\nu, v} F_\nu(u_\nu, v) := F_\nu^1(v) + F_\nu^2(u_\nu) \quad \text{s.t. } u_\nu \in U_{\text{ad}}^\nu,\ v \in S(u_\nu, u_{-\nu}), \tag{3.8}$$

where

$$S(u) = \arg\min_v \left\{ \frac{1}{2}\langle Av, v\rangle_{V^*,V} - \langle Bu, v\rangle_{V^*,V} \;\middle|\; v \geq 0 \right\}. \tag{3.9}$$

We assume that V and U^ν are Hilbert spaces such that $V \hookrightarrow U^\nu \hookrightarrow V^*$ for all $\nu = 1, \ldots, N$, where the embedding $U^\nu \hookrightarrow V^*$ is compact. The operators $A : V \to V^*$ and $B : U \to V^*$ with $Bu := \sum B_\nu u_\nu$ are bounded and linear where A is additionally self-adjoint and coercive. Then the lower level problem (3.9) admits a unique solution v for all $u \in U$. Hence, $S(u)$ is single-valued and we obtain an EPEC, where leader ν considers

$$\begin{aligned} &\min_{u_\nu, v} F_\nu^1(v) + F_\nu^2(u_\nu) \\ &\text{s.t. } u_\nu \in U_{\text{ad}}^\nu, \\ &\qquad v \geq 0,\ \langle Av - Bu, w - v\rangle_{V^*,V} \geq 0 \quad \forall w \geq 0. \end{aligned} \tag{3.10}$$

Introducing a slack variable $\xi \in V^*$, the VI can equivalently be written as the linear complementarity constraints

$$\begin{aligned} Av - \xi - Bu = 0, \\ \xi \geq 0 \text{in} V^*,\ v \geq 0 \text{in} V,\ \langle \xi, v\rangle_{V^*,V} = 0 \end{aligned}$$

and we obtain the EPCC representation of (3.10) with an additional variable ξ. Subsequently, we discuss two special cases of EPCCs, which have favorable properties, and then return to the general case.

First let us assume that the lower level problem (3.9) is unconstrained. Then the optimal solution is given by $v = A^{-1}Bu$, and the EPCC is equivalent to the following NEP: For all $\nu = 1, \ldots, N$

$$\min_{u_\nu} F_\nu^1(A^{-1}Bu) + F_\nu^2(u_\nu) \quad \text{s.t. } u_\nu \in U_{\text{ad}}^\nu.$$

Assuming that the objective function $F_\nu : U^\nu \times V \to \mathbb{R}$ is continuously Fréchet differentiable with $F_\nu^1 : V \to \mathbb{R}$ convex and $F_\nu^2 : U^\nu \to \mathbb{R}$ strictly convex for all $\nu = 1, \ldots, N$, the existence of a unique Nash equilibrium is guaranteed by the unique solvability (see [31]) of the resulting strongly monotone variational inequality

$$\text{Find} u_\nu^* \text{s.t.} \left(B_\nu^* A^{-1} \nabla_{u_\nu} F_\nu^1 (A^{-1} B u^*) + \nabla_{u_\nu} F_\nu^2 (u_\nu^*), u_\nu - u_\nu^* \right) \geq 0 \quad \forall u_\nu \in U_{\text{ad}}^\nu$$

for all $\nu = 1, \ldots, N$. Consequently, the MLFG admits a unique leader–follower equilibrium $(u^*, A^{-1} B u^*) \in U \times V$.

In the second setting, we assume that $V = H_0^1(\Omega)$, $U^\nu = L^2(\Omega)$ and $A = c \cdot \iota$ with positive scalar $c > 0$ and embedding operator $\iota \in \mathcal{L}(V, L^2(\Omega))$. Moreover, we require that $B \in \mathcal{B}(U, L^2(\Omega))$ and leader ν's objective is linear with respect to the follower's strategy, i.e. $F_\nu^1(v) = \langle \gamma_\nu, v \rangle_{V^*, V}$ where $\gamma_\nu \in V^+$ is an element of the dual cone of V. This case can be seen as an infinite dimensional analogon of the model considered in Sect. 2.2. Due to the special structure, it is known that $\xi \in L^2(\Omega)$ and we can write the complementarity constraints equivalently as

$$v - \frac{1}{c} B u \geq 0 \text{in} L^2(\Omega), \; v \geq 0 \text{in} L^2(\Omega), \; \left(v - \frac{1}{c} B u, v \right)_{L^2} = 0. \tag{3.11}$$

Then, it has been shown for instance in [51] that the complementarity system (3.11) is equivalent to

$$v = \max \left\{ c^{-1} B u, 0 \right\} \quad \text{in} L^2(\Omega).$$

Hence, the leader ν's parametric MPCC (3.8) can be written as

$$\min_{u_\nu} \left\langle \gamma_\nu, \max \left\{ c^{-1} B u, 0 \right\} \right\rangle_{H^{-1}, H_0^1} + F_\nu^2(u_\nu) \quad \text{s.t.} \, u_\nu \in U_{\text{ad}}^\nu.$$

If $F_\nu^2(u_\nu)$ is convex and continuous for all $\nu = 1, \ldots, N$, the aforementioned equilibrium problem admits a Nash equilibrium u^* by the fixed point theorem of Kakutani. Hence, $(u^*, v^*) = (u^*, \max\{c^{-1} B u^*, 0\})$ is a leader–follower equilibrium of the MLFG.

In contrast to the aforementioned two special cases, we cannot expect to obtain existence of equilibria in the general framework of (3.10). Motivated by the contributions on MPECs/MPCCs referenced in the beginning of this section, we therefore focus on several types of stationarity conditions. Based on the work of EPCCs in finite dimensions [28], our present focus is on the analysis of the nonsmooth and nonconvex NEP

$$\min_{u_\nu} F_\nu^1(v(u_\nu, u_{-\nu})) + F_\nu^2(u_\nu) \quad \text{s.t.} \, u_\nu \in U_{\text{ad}}^\nu \quad (\forall \nu = 1, \ldots, N),$$

which we obtain by plugging in the solution $v(u) \in S(u) = \{v(u)\}$ of the lower level problem into the upper level (3.8). In the context of equilibrium problems, C-stationarity seems to be an appropriate concept, since stronger concepts such as S-stationarity require additional assumptions as the absence of control constraints, i.e. $U_{\text{ad}}^{v} = U^{v}$ (see [25]). On the other hand, the well-known Lipschitz continuity of the solution operator $v(\cdot)$ motivates an approach based on Clarke's nonsmooth analysis.

3.3 Stackelberg Game with Infinitely Many Followers

We consider a Stackelberg game with infinitely many followers, whose optimality conditions are reformulated as a PDE of a distribution of followers. The followers' optimal control problems are given for $i = 1, \dots, M$ by

$$\min_{x_i, v_i} f_i(x_i, v_i, u) = \int_0^T \left[\phi(x_i(t), u(t)) + \frac{\alpha}{2} v_i^2(t) \right] \mathrm{d}t \quad \text{s.t. } \dot{x}_i(t) = v_i(t), x_i(0) = x_{0,i},$$

i.e. every follower controls its state $x_i \in X = C^1([0,T], \mathbb{R})$ by its control $v_i \in V = C^0([0,T], \mathbb{R})$. The smooth functional $\phi : X \times U \to \mathbb{R}$ models the influence of the leader, whose control is denoted by $u \in U = C^0([0,T], \mathbb{R})$. Similar to the game presented in Sect. 2.2, the followers are playing a potential game. Therefore, they can be gathered into a single optimization problem yielding a Stackelberg game of the following structure:

$$\begin{aligned} \min_{u} \; & F(u, \mathbf{v}, \mathbf{x}) = \int_0^T J\,(u(t), p(\mathbf{x}))\, \mathrm{d}t \\ \text{s.t. } & \min_{\mathbf{x},\mathbf{v}} \; f(\mathbf{x}, \mathbf{v}, u) = \frac{1}{M} \sum_{i=1}^{M} \int_0^T \left[\phi(x_i(t), u(t)) + \frac{\alpha}{2} v_i^2(t) \right] \mathrm{d}t \\ & \text{s.t.} \dot{x}_i(t) = v_i(t), \; x_i(0) = x_{0,i}, \; \text{for } i = 1, \dots, M, \end{aligned} \tag{3.12}$$

where $\mathbf{x} = (x_1, \dots, x_M)$ and $\mathbf{v} = (v_1, \dots, v_M)$. The functional $J : U \times \mathbb{R}^m \to \mathbb{R}$ describes the coupling between the leader control and the moment of the follower states, which is assumed to have the structure $p(\mathbf{x}) = \frac{1}{M} \sum_{i=1}^{M} \tilde{p}(x_i(t))$, where $\tilde{p} : X \to X$. This assumption guarantees symmetry of the cost functional which, in addition to modest assumptions on continuity and boundedness on ϕ, allows us to compute the mean field limit characterizing the asymptotic behavior as the number of players grow $N \to \infty$, c.f. [6, Ch.1]. In order to keep notation simple, time dependence of variables is no longer explicitly indicated from now on.

The Pontryagin Maximum Principle [46, Ch.1] is necessary and sufficient for suitable ϕ. Therefore, the follower problem in (3.12) can be replaced by its optimality conditions and we have the optimal control problem

$$\min_{u,\mathbf{x},\mathbf{\Lambda}} F(u, \mathbf{v}, \mathbf{x}) = \int_0^T J\,(u,\, p(\mathbf{x}))\ \mathrm{d}t$$

$$\text{s.t. for } i = 1, \dots, M$$

$$\begin{cases} \dot{x}_i = -\frac{1}{\alpha}\lambda_i, & x_i(0) = x_{i,0} \\ \dot{\lambda}_i = -\partial_{x_i}\phi(x_i, u), & \lambda_i(T) = 0, \end{cases}$$

where $\lambda \in X^*$ is the costate and $\mathbf{\Lambda} = (\lambda_1, \dots, \lambda_M)$. The formal application of the Lagrange multiplier theorem, e.g. [38, Ch.9], yields for $i = 1, \dots, M$:

$$\begin{aligned} \partial_u J\,(u,\, p(\mathbf{x})) + \frac{1}{M}\sum_{i=1}^{M} \xi_i^{(2)} \partial_u \partial_{x_i}\phi(x_i, u) &= 0, \\ \partial_p J\,(u,\, p(\mathbf{x}))\, \partial_{x_i}\tilde{p}(x_i) + \partial_{x_i}^2 \phi(x_i, u)\xi_i^{(2)} - \dot{\xi}_i^{(1)} &= 0, \\ \frac{1}{\alpha}\xi_i^{(1)} - \dot{\xi}_i^{(2)} &= 0, \\ \dot{x}_i + \frac{1}{\alpha}\lambda_i &= 0, \\ \dot{\lambda}_i + \partial_{x_i}\phi(x_i, u) &= 0, \\ \xi_i^{(1)}(T) = 0, \quad \xi_i^{(2)}(0) = 0, \quad x_i(0) = x_{i,0}, \quad \lambda_i(T) &= 0, \end{aligned} \tag{3.13}$$

where $\xi_i^{(1)} \in X^*$ is the costate to the dynamic of x_i and $\xi_i^{(2)} \in X$ to λ_i, respectively.

Since we are interested in the limit problem for infinitely many followers, we introduce the empirical measure:

$$\mu^M\left(t, x, \lambda, \xi^{(1)}, \xi^{(2)}\right) = \frac{1}{M}\sum_{i=1}^{M} \delta(x - x_i)\delta(\lambda - \lambda_i)\delta(\xi^{(1)} - \xi_i^{(1)})\delta(\xi^{(2)} - \xi_i^{(2)}).$$

The empirical measure $\mu^M(t, \cdot) \in \mathcal{P}(\mathbb{R}^4)$ is a Borel probability measure, which describes the probability to find a particle in a certain position $(x, \lambda, \xi^{(1)}, \xi^{(2)})$ at

time $t \in [0, T]$. Formal reformulations of (3.13) yield the macroscopic optimality conditions

$$\begin{aligned} 0 =& \partial_t \mu - \frac{1}{\alpha}\lambda \cdot \partial_x \mu - \partial_x \phi(x, u) \cdot \partial_\lambda \mu \\ &+ \left[\partial_p J_\mu \left(u, p_\mu(t)\right) \cdot \partial_x \tilde{p}(x) + \partial_x^2 \phi(x, u) \cdot \xi^{(2)}\right] \cdot \partial_{\xi^{(1)}} \mu + \frac{1}{\alpha}\xi^{(1)} \cdot \partial_{\xi^{(2)}} \mu, \\ 0 =& \partial_u J_\mu \left(u, p_\mu(t)\right) + \int_{\mathbb{R}} \int_{\mathbb{R}} \int_{\mathbb{R}} \int_{\mathbb{R}} \xi^{(2)} \partial_y \partial_x \phi(x, u) \mu \, \mathrm{d}\xi^{(2)} \, \mathrm{d}\xi^{(1)} \, \mathrm{d}\lambda \, \mathrm{d}x, \end{aligned} \tag{3.14}$$

where the moment is

$$p_\mu(t) = \frac{1}{M} \sum_{i=1}^{M} \tilde{p}(x_i) = \int_{\mathbb{R}} \int_{\mathbb{R}} \int_{\mathbb{R}} \int_{\mathbb{R}} \tilde{p}(x) \mu \left(t, x, \lambda, \xi^{(1)}, \xi^{(2)}\right) \mathrm{d}\xi^{(2)} \, \mathrm{d}\xi^{(1)} \, \mathrm{d}\lambda \, \mathrm{d}x,$$

and the integrand of the leader's objective is $J(u, p(\mathbf{x})) = J_\mu(u, p_\mu(t))$. The initial/terminal conditions in (3.13) are assumed to be satisfied by the empirical measure μ^M, but are omitted here.

In contrast to (3.13), the PDE in (3.14) describes the optimal control and state trajectories for a density of players. It resembles the limit for infinitely many players. Rigorous limits are known for linear problems, e.g. [9]. So far, it is unknown if the mean field limit Eq. (3.14) poses a unique solution. Provided there exists a weak solution to the mean field equation, the empirical measure fulfills conditions (3.13).

This alternative formulation has advantages if a large number of similar players are studied. However, it is arbitrary to derive the mean field of the full optimality conditions (3.13). Present focus of research is the investigation of the relationship between the conditions in (3.14) and the optimality conditions if the mean field limit is e.g. already derived in (3.12).

3.4 Dynamic Boundary Control Games with Networks of Strings

In contrast to NEPs governed by ordinary differential equations, NEPs governed by partial differential equations got an increasing interest among researchers more recently. In [4], the existence of Nash equilibria for networked hyperbolic systems of partial differential equations (scalar conservation laws) for traffic flow models on networks have been studied.

Here, we discuss a game for a star-shaped network of strings with N rays and N players, where each player influences the system state in the network through a Dirichlet boundary control at one end of the rays and the objective functionals are given by sums of squared L^2-norms with one term for the control cost and

two tracking-type terms. The motivating application we have in mind here is a gas pipeline network, where the players represent the gas market participants, i.e. producers or consumers. A simple linear model for one pipe ($\nu \in \{1, \dots, N\}$) for the dynamics in a network of N horizontal pipelines (without friction) is given by

$$\begin{cases} \rho_t^{(\nu)} + q_x^{(\nu)} = 0, \\ q_t^{(\nu)} + a^2\, \rho_x^{(\nu)} = 0, \end{cases} \tag{3.15}$$

where $\rho^{(\nu)}$ denotes the gas density, $q^{(\nu)}$ the flow rate of the gas, and $a > 0$ corresponds to the sound speed. This system implies that $\rho^{(\nu)}$ satisfies the wave equation

$$\rho_{tt}^{(\nu)} = a^2\, \rho_{xx}^{(\nu)}$$

along pipe ν. Moreover, we assume the following coupling conditions to model the flow through a junction of N pipes, where for all adjacent pipes $x = 0$ denotes the end of the junction, respectively (cf. [47]): the continuity of the density, i.e.

$$\rho^{(\nu)}(t,\, 0) = \rho^{(j)}(t,\, 0)$$

for all $\nu, j = 1, \dots, N$ and the conservation of mass, which leads similar to Kirchhoff's law to the equation $\sum_{k=1}^{N} q^{(k)}(t,\, 0) = 0$. In terms of the densities $\rho^{(\nu)}(t,\, x)$, these assumptions yield the following node conditions:

$$\rho^{(\nu)}(t,\, 0) = \rho^{(j)}(t,\, 0), \text{ for } \nu, j = 1, \dots, N \text{ and } \sum_{k=1}^{N} \rho_x^{(k)}(t,\, 0) = 0.$$

Similarly to the objective functional used in [37] for an optimal Dirichlet boundary control problem (for a single player), let the objective functional here be given by

$$J_\nu(u) = \frac{\gamma_\nu}{2} \int_0^T u_\nu(\tau)^2 \,\mathrm{d}\tau \; + \; + \frac{1}{2} \sum_{j=1}^{N} \left[\int_0^1 \left(\rho^{(j)}(T,\, x) - \rho_{D,\nu}^{(j)}(x) \right)^2 \right.$$
$$\left. + \left(q^{(j)}(T,\, x) - q_{D,\nu}^{(j)}(x) \right)^2 \mathrm{d}x \right],$$

where the lengths of the pipelines are assumed to be normalized to one and $\rho_{D,\,\nu}^{(j)}(x)$ denotes the density profile that is desired by player ν in pipe j at terminal time T. Moreover, $q_{D,\nu}^{(j)}(x)$ here denotes the profile of the flow rate in pipe j that is desired by player ν at terminal time T. The term $q(T, \cdot)$ is determined by (3.15) and in such a way that $J_\nu(u)$ is minimized. In this model $\rho_{D,\,\nu}^{(j)}(x)$ and $q_{D,\,\nu}^{(j)}(x)$ might also be the desired initial states for the next time period to consider.

In the following, we consider a particular setting: Let a terminal time $T \geq 4$ be given and assume we have a finite number $N \geq 3$ of strings of length 1 and $\Omega = (0,\ T) \times (0,\ 1)$. Then for the initial states

$$\left(y_0^{(\nu)}, y_1^{(\nu)}\right)_{\nu=1}^{N} \in \left\{ (y_0^{(\nu)}, y_1^{(\nu)})_{\nu=1}^{N} \,\middle|\, y_0^{(\nu)} \in L^2(0,\ 1),\ y_1^{(\nu)} \in H^{-1}(0,\ 1),\ \nu = 1, \ldots, N \right\},$$

where $H^{-1}(0, 1) = \{Y \in \mathcal{D}'((0,\ 1)) \mid \text{there is } f \in L^2(0, 1) \text{ such that } f' = Y\}$ and $\mathcal{D}'((0,\ 1))$ denotes the set of distributions on the interval $(0, 1)$, and $u_\nu \in L^2(0,\ T)$ for $\nu = 1, \ldots, N$, we consider the following system (**S**) that is defined by

$$w^{(\nu)}(0, x) = y_0^{(\nu)}(x), \qquad x \in (0,\ 1), \qquad \nu = 1, \ldots, N, \tag{3.16a}$$

$$w_t^{(\nu)}(0, x) = y_1^{(\nu)}(x), \qquad x \in (0,\ 1), \qquad \nu = 1, \ldots, N, \tag{3.16b}$$

$$w_{tt}^{(\nu)}(t, x) = w_{xx}^{(\nu)}(t, x), \qquad (t, x) \in \Omega, \qquad \nu = 1, \ldots, N, \tag{3.16c}$$

$$w^{(\nu)}(t, 0) = w^{(j)}(t, 0), \qquad t \in (0,\ T), \qquad \nu, j = 1, \ldots, N, \tag{3.16d}$$

$$0 = \sum_{\nu=1}^{N} w_x^{(\nu)}(t, 0), \qquad t \in (0,\ T), \tag{3.16e}$$

$$w^{(\nu)}(t,\ 1) = u_\nu(t), \qquad t \in (0,\ T), \qquad \nu = 1, \ldots, N. \tag{3.16f}$$

The system (**S**) is a star-shaped network of vibrating strings with Dirichlet boundary control action at the boundary nodes. An overview on the control of networks of vibrating strings can be found in [7] and the exact controllability of networks of vibrating strings is studied in [48].

We consider a dynamic Nash game with N players, who control the system (**S**) by their strategies i.e. control functions $u_\nu \in L^2(0,\ T)$ $(\nu = 1, \ldots, N)$, where each player's goal it is to minimize his/her own cost functional J_ν given by

$$J_\nu(w, u) = \frac{\gamma_\nu}{2} \int_0^T u_\nu(\tau)^2 \, \mathrm{d}\tau + \frac{1}{2} \sum_{j=1}^{N} \left[\int_0^1 \left(w^{(j)}(T,\ x) - g_{D,\nu}^{(j)}(x) \right)^2 \right.$$
$$\left. + \left(V_\nu^{(j)}(x) - h_{D,\nu}^{(j)}(x) \right)^2 \mathrm{d}x \right],$$

where the constants $\gamma_\nu > 0$ are given weighting factors of the control costs in J_ν, $g_{D,\nu}^{(j)}(\cdot) \in L^2(0,\ 1)$ denotes the position of the j-th string at the terminal time T desired by player ν, and $h_{D,\nu}^{(j)}(\cdot) \in L^2(0,\ 1)$ denotes the antiderivatives of the

desired velocity, where for $\nu = 1, \ldots, N$ the $V_\nu^{(j)}(x) \in L^2(0,\ 1)$ are antiderivatives of $w_t^{(j)}(T,\ x)$ (see [18]). Hence, the associated Nash game is given by

$$\begin{aligned} &\min_{u_\nu} J_\nu(w, u_\nu, u_{-\nu}) \\ &\text{s.t. } (w, u) \quad \text{solve} \quad (\mathbf{S}) \end{aligned} \qquad \nu = 1, \ldots, N. \tag{3.17}$$

In [19] it has been shown, that this boundary control game with the wave equation, i.e. (3.16a)–(3.16f), admits a unique Nash equilibrium, where the associated optimal strategies u_ν are 4-periodic and can explicitly be determined in terms of the given data.

Theorem 3.5 *Assume that $\gamma_\nu = \gamma_j$ for all $\nu,\ j = 1, \ldots, N$. Then there exists a unique Nash equilibrium for 3.17. Moreover, the optimal strategies are* 4*-periodic.*

The explicit representation of the linear operator, which maps the initial state and the desired states to the corresponding Nash equilibrium, given in [19] implies the boundedness of the operator as a map from the corresponding function spaces to the control space $(L^2(0,\ T))^N$ and thus the stability of the Nash equilibria with respect to perturbations of the initial and the desired states.

4 Outlook

The previously discussed results represent some examples of the latest achievements in understanding extensions of standard Nash games in view of our intention to analyze and solve multi-leader–follower games in function space. This research will be continued in the future by considering further extensions inspired by applications and also by combinations of the types of games, that we discussed here. Another future task concerns the improvement of the current numerical algorithms in that field. The presented algorithm for finite dimensional MLFGs has to be developed further and combined with suitable numerical methods for differential equations in order to solve the dynamic counterparts of MLFG.

Acknowledgments This work has been supported by the German Research Foundation (DFG) in the framework the SPP 1962 “Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization”, project STE 2063/2-1 and SCHW 1831/1-1.

References

1. E. Allevi, D. Aussel, and R. Riccardi, *On an equilibrium problem with complementarity constraints formulation of pay-as-clear electricity market with demand elasticity.* J. Global Optim. **70** (2018) 329–346.

2. D. Aussel and A. Svensson, *Some remarks about existence of equilibria, and the validity of the EPCC reformulation for multi-leader-follower games.* J. Nonlinear Convex Anal. **19** (2018) 1141–1162.
3. A. Borzi and C. Kanzow, *Formulation and numerical solution of Nash equilibrium multiobjective elliptic control problems.* SIAM J. Control Optim. **51** (2013) 718–744.
4. A. Bressan and K. Han, *Existence of optima and equilibria for traffic flow on networks.* Games Econom. Behav. **8** (2013) 627–648.
5. V. Cardellini, V. D. N. Personé, V. Di Valerio, F. Facchinei, V. Grassi, F. L. Presti, and V. Piccialli, *A game-theoretic approach to computation offloading in mobile cloud computing.* Math. Program. **157** (2016) 421–449.
6. R. Carmona and F. Delarue *Probabilistic Theory of Mean Field Games with Applications I*, Springer International Publishing, (2018).
7. R. Dáger and E. Zuazua, *Wave propagation, observation and control in 1-d flexible multi-structures*, Vol. 50. Springer Science & Business Media, 2006.
8. S. Dempe, *Annotated bibliography on bilevel programming and mathematical programs with equilibrium constraints.* Optimization **52** (2003) 333–359.
9. R. L. Dobrushin, *Vlasov equations.* Funct. Anal. Appl. **13** (1979) 115–123.
10. A. Dreves and J. Gwinner, *Jointly convex generalized Nash equilibria and elliptic multiobjective optimal control.* J. Optim. Theory Appl. **168** (2016) 1065–1086.
11. F. Facchinei and C. Kanzow, *Generalized Nash equilibrium problems.* Ann. Oper. Res. **175** (2010) 177–211.
12. F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007.
13. F. Facchinei, V. Piccialli, and M. Sciandrone, *Decomposition algorithms for generalized potential games.* Comput. Optim. Appl. **50** (2011) 237–262.
14. A. Fischer, M. Herrich, and K. Schönefeld, *Generalized Nash equilibrium problems-recent advances and challenges.* Pesquisa Operacional **34** (2014) 521–558.
15. J. Franke, C. Kanzow, W. Leininger, and A. Schwartz, *Effort maximization in asymmetric contest games with heterogeneous contestants.* Econom. Theory **52** (2013) 589–630.
16. H. Gfrerer and J. J. Ye, *New constraint qualifications for mathematical programs with equilibrium constraints via variational analysis.* SIAM J. Optim. **27** (2017) 842–865.
17. V. Grimm, L. Schewe, M. Schmidt, and G. Zöttl, *A multilevel model of the European entry-exit gas market.* Math. Methods Oper. Res. **89** (2019) 223–255.
18. M. Gugat, *Penalty techniques for state constrained optimal control problems with the wave equation.* SIAM J. Control Optim. **48** (2009) 3026–3051.
19. M. Gugat and S. Steffensen, *Dynamic boundary control games with networks of strings.* ESAIM: Control, Optimisation and Calculus of Variations **24** (2018) 1789–1813.
20. T. Harks, M. Schröder, and D. Vermeulen, *Toll caps in privatized road networks.* European J. Oper. Res. **276** (2019) 947–956.
21. R. Henrion and W. Römisch, *On M-stationary points for a stochastic equilibrium problem under equilibrium constraints in electricity spot market modeling.* Applications of Mathematics **52**, (2007) 473–494.
22. R. Henrion and J. Outrata, and T. Surowiec, *Analysis of M-stationary points to an EPEC modeling oligopolistic competition in an electricity spot market.* ESAIM **18**, (2012) 295–317.
23. M. Herty, S. Steffensen, and A. Thünen, *Solving quadratic multi-leader-follower games by smoothing the follower's best response.* arXiv preprint arXiv:1808.07941, 2018.
24. R. Herzog, C. Meyer, and G. Wachsmuth, *B-and strong stationarity for optimal control of static plasticity with hardening.* SIAM J. Optim. **23** (2013) 321–352.
25. M. Hintermüller and I. Kopacka, *Mathematical programs with complementarity constraints in function space: C-and strong stationarity and a path-following algorithm.* SIAM J. Optim. **20** (2009) 868–902.
26. M. Hintermüller and T. M. Surowiec, *First-order optimality conditions for elliptic mathematical programs with equilibrium constraints via variational analysis.* SIAM J. Optim. **21** (2011) 1561–1593.

27. M. Hintermüller, T. M. Surowiec, and A. Kämmler, *Generalized Nash equilibrium problems in Banach spaces: Theory, Nikaido–Isoda-based path-following methods, and applications.* SIAM J. Optim. **25** (2015) 1826–1856.
28. M. Hu and M. Fukushima, *Smoothing approach to Nash equilibrium formulations for a class of equilibrium problems with shared complementarity constraints.* Comput. Optim. Appl. **52** (2012) 415–437.
29. M. Hu and M. Fukushima, *Existence, uniqueness, and computation of robust Nash equilibria in a class of multi-leader-follower games.* SIAM J. Optim. **23** (2013) 894–916.
30. X. Hu and D. Ralph, *Using EPECs to model bilevel games in restructured electricity markets with locational prices.* Oper. Res. **55** (2007) 809–827.
31. D. Kinderlehrer and G. Stampacchia, *An introduction to variational inequalities and their applications.* SIAM (1980).
32. A. Koh and S. Shepherd, *Tolling, collusion and equilibrium problems with equilibrium constraints.* European Transport/Trasporti Europei **44** (2010) 3–22.
33. A. A. Kulkarni and U. V. Shanbhag *A shared-constraint approach to multi-leader multi-follower games.* Set-Valued Var. Anal. **22** (2014) 691–720.
34. L. Lampariello and S. Sagratella, *A bridge between bilevel programs and Nash games.* J. Optim. Theory Appl. **174** (2017) 613–635.
35. Q. H. Le, H. Al-Shatri, and A. Klein, *Optimal joint power allocation and task splitting in wireless distributed computing.* In *International ITG Conference on Systems, Communication and Coding*, VDE Verlag, 2017.
36. S. Leyffer and T. Munson, *Solving multi-leader–common-follower games.* Optim. Methods Softw. **25** (2010) 601–623.
37. J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations (Grundlehren der Mathematischen Wissenschaften)*, Vol. 170. Springer Berlin, 1971.
38. D. G. Luenberger, *Optimization by vector space methods*. John Wiley & Sons, 1997.
39. Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, *A survey on mobile edge computing: the communication perspective* IEEE Communications Surveys & Tutorials **19** (2017) 2322–2358,
40. P. Mehlitz and G. Wachsmuth, *Weak and strong stationarity in generalized bilevel programming and bilevel optimal control.* Optimization **65** (2016) 907–935.
41. D. Monderer and L. S. Shapley, *Potential games.* Games Econom. Behav. **14** (1996) 124–143.
42. B. S. Mordukhovich, *Optimization and equilibrium problems with equilibrium constraints in infinite-dimensional spaces.* Optimization **57** (2008) 715–741.
43. H. Nikaido and K. Isoda, *Note on non-cooperative convex games.* Pacific J. Math. **5** (1955) 807–815.
44. D. Nowak, T. Mahn, H. Al-Shatri, A. Schwartz, and A. Klein, *A generalized Nash game for mobile edge computation offloading.* In *6th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering*, IEEE, 2018.
45. J.-S. Pang and M. Fukushima, *Quasi-variational inequalities, generalized Nash equilibria, and multi-leader-follower games.* Comput. Manag. Sci. **2** (2005) 21–56.
46. L. Pontryagin, V. Boltyanski, R. Gamkrelidze, and E. Miscenko, *The Mathematical Theory of Optimal Processes*. John Wiley & Sons, 1962.
47. D. F. N. Rodriguez, E. Witrant, and O. Sename, *Control-oriented modeling of fluid networks: a time-delay approach.* In *Recent Results on Nonlinear Delay Control Systems*,Springer, 2016.
48. E. G. Schmidt, *On the modelling and exact controllability of networks of vibrating strings.* SIAM J. Control Optim. **30** (1992) 229–245.
49. C.-L. Su, *Equilibrium problems with equilibrium constraints: Stationarities, algorithms, and applications*. PhD thesis, Stanford University, 2005.
50. J. Sun, *Optimal Control Problem for American Put Option*. PhD thesis, Washington State University, 2011.
51. M. Ulbrich, *Semismooth Newton methods for operator equations in function spaces*. In *SIAM Journal on Optimization* **13** (2002) 805–841.

52. G. Wachsmuth, *Mathematical programs with complementarity constraints in Banach spaces.* J. Optim. Theory Appl. **166** (2015) 480–507.
53. B. Yu, J. E. Mitchell, and J.-S. Pang, *Solving linear programs with complementarity constraints using branch-and-cut.* Math. Program. Comput. **11** (2019) 267–310.

Stress-Based Methods for Quasi-Variational Inequalities Associated with Frictional Contact

Bernhard Kober, Gerhard Starke, Rolf Krause, and Gabriele Rovi

Abstract The stress-based formulation of elastic contact with Coulomb friction in the form of a quasi-variational inequality is investigated. Weakly symmetric stress approximations are constructed using a finite element combination on the basis of Raviart–Thomas spaces of next-to-lowest order. An error estimator is derived based on a displacement reconstruction and proved to be reliable under certain assumptions on the solution formulated in terms of a norm equivalence in the trace space $H^{1/2}(\Gamma)$. Numerical results illustrate the effectiveness of the adaptive refinement strategy for a Hertzian frictional contact problem in the compressible as well as in the incompressible case.

Keywords Quasi-variational inequality · Coulomb friction · A posteriori error estimation

Mathematics Subject Classification (2020) Primary 65N30; Secondary 74M15

The authors gratefully acknowledge support by DFG in the Priority Programme SPP 1962 "Non-smooth and Complementarity-based Distributed Parameter Systems" under grant number STA 402/13-1 and by SNF under grant number 200021E-167012/1.

B. Kober · G. Starke (✉)
Fakultät für Mathematik, Universität Duisburg-Essen, Essen, Germany
e-mail: bernhard.kober@uni-due.de; gerhard.starke@uni-due.de

R. Krause · G. Rovi
Institute of Computational Science, Università della Svizzera Italiana, Lugano, Switzerland
e-mail: rolf.krause@usi.ch; gabriele.rovi@usi.ch

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_18

1 Introduction

The mathematical modeling of elastic contact problems with a Coulomb friction law gives rise to a quasi-variational inequality. Formulated in terms of stresses, its difficulty lies in the fact that the admissible space depends on the solution itself, see e.g. [9, Sect. 8.5] for the details of the derivation. Our main motivation for studying the stress-based formulation is that the numerical treatment simplifies considerably due to the less involved nature of the constraints. We use a weakly symmetric approximation of the stresses using the finite element triple (for stress, displacement, and rotation) proposed in [4]. Using a displacement reconstruction approach, we derive an upper bound of the stress approximation error for sufficiently small friction parameter. This is based on the estimation of the duality gap for variational inequalities going back to [13] (cf. also [14]).

The error estimator resulting from the displacement reconstruction is used for adaptive refinement resulting in a high resolution of the stress and displacement components in the contact zone. Other adaptive approaches for frictional contact are usually based on residual error estimation, see [5, 6] for Tresca friction and [11] for Coulomb friction.

The next section introduces the stress-based formulation for the elastic contact problem with Coulomb friction. The a posteriori error estimator based on displacement reconstruction is derived in Sect. 3. The reliability estimate for sufficiently small friction relies on a norm equivalence in the trace space $H^{1/2}(\Gamma)$ which is the content of Sect. 4. Section 5 presents the numerical results for a Hertzian frictional contact problem in the compressible and incompressible case.

2 The Dual Stress-Based Formulation of Contact with Coulomb Friction

Throughout this paper, $(\,\cdot\,,\,\cdot\,)$ stands for the inner product in $L^2(\Omega)$, $L^2(\Omega)^d$, or $L^2(\Omega)^{d\times d}$, respectively, and $\langle\,\cdot\,,\,\cdot\,\rangle_\Gamma$ stands for the duality pairing of $H^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$ for $\Gamma \subseteq \partial\Omega$. The underlying material model will be linearly elastic such that the strain tensor $\boldsymbol{\varepsilon} : \Omega \to \mathbb{R}^{d\times d}$ depends on the stress $\boldsymbol{\sigma} : \Omega \to \mathbb{R}^{d\times d}$ by means of

$$\boldsymbol{\varepsilon} = \mathcal{A}\boldsymbol{\sigma} := \frac{1}{2\mu}\left(\boldsymbol{\sigma} - \frac{\lambda}{d\lambda + 2\mu}(\operatorname{tr}\boldsymbol{\sigma})\mathbf{I}\right), \tag{2.1}$$

where μ and λ denote the usual Lamé parameters. For $\lambda < \infty$, i.e., away from the incompressible limit, this relation is invertible and leads to the familiar stress–strain relation $\boldsymbol{\sigma} = 2\mu\boldsymbol{\varepsilon} + \lambda(\operatorname{tr}\boldsymbol{\varepsilon})\mathbf{I}$. Throughout this paper, μ is assumed to be on the order of one while λ may be arbitrarily big. The boundary of our domain $\Omega \subset \mathbb{R}^d$ is assumed to consist of disjoint segments Γ_D, Γ_N, and Γ_C, all of them nonempty, such

that Γ_N separates Γ_C from Γ_D. As a suitable space for the stress $\boldsymbol{\sigma}$, we introduce

$$H_{\Gamma_N}(\mathrm{div},\Omega) = \{\boldsymbol{\tau} \in L^2(\Omega)^{d\times d} : \mathrm{div}\,\boldsymbol{\tau} \in L^2(\Omega)^d\ ,\ \boldsymbol{\tau}\cdot\mathbf{n} = \mathbf{0} \text{ on } \Gamma_N\}. \tag{2.2}$$

The symmetry of the stress tensor will be enforced by the constraint $\mathbf{as}\ \boldsymbol{\sigma} = \mathbf{0}$, where **as** denotes the skew-symmetric part. Following [9, Sect. 8.5], the stress-based formulation of frictional contact then consists in finding $\boldsymbol{\sigma} \in H_{\Gamma_N}(\mathrm{div},\Omega)^d$ such that it solves, among all $\boldsymbol{\tau} \in H_{\Gamma_N}(\mathrm{div},\Omega)^d$, the constrained minimization problem

$$\begin{aligned} &\mathcal{J}(\boldsymbol{\tau}) := \frac{1}{2}(\mathcal{A}\boldsymbol{\tau},\boldsymbol{\tau}) - \langle \mathbf{u}^D, \boldsymbol{\tau}\cdot\mathbf{n}\rangle_{\Gamma_D} - \langle g, \mathbf{n}\cdot(\boldsymbol{\tau}\cdot\mathbf{n})\rangle_{\Gamma_C} \to \min! \\ &\text{subject to } \mathrm{div}\,\boldsymbol{\tau} + \mathbf{f} = \mathbf{0}\ ,\ \mathbf{as}\,\boldsymbol{\tau} = \mathbf{0} \text{ in } \Omega\ , \\ &\qquad \text{and } \mathbf{n}\cdot(\boldsymbol{\tau}\cdot\mathbf{n}) \le 0\ ,\ |\mathbf{n}\times(\boldsymbol{\tau}\cdot\mathbf{n})| \le -\nu_F(\mathbf{n}\cdot(\boldsymbol{\sigma}\cdot\mathbf{n})) \text{ on } \Gamma_C\ . \end{aligned} \tag{2.3}$$

The given functions $\mathbf{u}^D \in H^{1/2}(\Gamma_D)^d$ and $g \in H^{1/2}(\Gamma_C)$ represent the prescribed boundary displacements and the gap function, respectively. For simplicity, the volume force $\mathbf{f}$ is assumed to be piecewise affine and $\nu_F \ge 0$ stands for the friction parameter. It is important to note that the last constraint in (2.3) depends on the solution itself. This is the nature of the quasi-variational inequality in the dual formulation. With the definition of the admissible set

$$\begin{aligned} \boldsymbol{\Sigma}(\boldsymbol{\sigma}) = \{&\boldsymbol{\tau} \in H_{\Gamma_N}(\mathrm{div},\Omega)^d : \mathrm{div}\,\boldsymbol{\tau} + \mathbf{f} = \mathbf{0}\ ,\ \mathbf{as}\,\boldsymbol{\tau} = \mathbf{0} \text{ in } \Omega\ , \\ &\mathbf{n}\cdot(\boldsymbol{\tau}\cdot\mathbf{n}) \le 0\ ,\ |\mathbf{n}\times(\boldsymbol{\tau}\cdot\mathbf{n})| \le -\nu_F(\mathbf{n}\cdot(\boldsymbol{\sigma}\cdot\mathbf{n})) \text{ on } \Gamma_C\} \end{aligned} \tag{2.4}$$

the problem consists in finding $\boldsymbol{\sigma} \in \boldsymbol{\Sigma}(\boldsymbol{\sigma})$ such that

$$(\mathcal{A}\boldsymbol{\sigma}, \boldsymbol{\tau}-\boldsymbol{\sigma}) - \langle \mathbf{u}^D, (\boldsymbol{\tau}-\boldsymbol{\sigma})\cdot\mathbf{n}\rangle_{\Gamma_D} - \langle g, \mathbf{n}\cdot((\boldsymbol{\tau}-\boldsymbol{\sigma})\cdot\mathbf{n})\rangle_{\Gamma_C} \ge 0 \tag{2.5}$$

holds for all $\boldsymbol{\tau} \in \boldsymbol{\Sigma}(\boldsymbol{\sigma})$. The constraints on Γ_C require a careful interpretation and cannot be regarded pointwise due to the non-local nature of the space $H^{-1/2}(\partial\Omega)$, where these traces live. The correct interpretation and mathematically precise formulation of these constraints can be found in [9, Sect. 8.5].

If we restrict ourselves to a finite element space $\boldsymbol{\Sigma}_h \subset H_{\Gamma_N}(\mathrm{div},\Omega)^d$ based on a shape-regular family of triangulations $\mathcal{T}_h$, then the formulation of the constraints on Γ_C in (2.6) may be taken literally. The corresponding discrete admissible set is thus given by

$$\begin{aligned} \boldsymbol{\Sigma}_h(\boldsymbol{\sigma}_h) = \{&\boldsymbol{\tau}_h \in \boldsymbol{\Sigma}_h : (\mathrm{div}\,\boldsymbol{\tau}_h + \mathbf{f}, \mathbf{z}_h) = 0 \text{ for all } \mathbf{z}_h \in \mathbf{Z}_h\ , \\ &(\mathbf{as}\,\boldsymbol{\tau}_h, \boldsymbol{\gamma}_h) = 0 \text{ for all } \boldsymbol{\gamma}_h \in \mathbf{X}_h\ , \\ &\mathbf{n}\cdot(\boldsymbol{\tau}_h\cdot\mathbf{n}) \le 0\ ,\ |\mathbf{n}\times(\boldsymbol{\tau}_h\cdot\mathbf{n})| \le -\nu_F(\mathbf{n}\cdot(\boldsymbol{\sigma}_h\cdot\mathbf{n})) \text{ on } \Gamma_C\} \end{aligned} \tag{2.6}$$

and the discrete problem consists in finding $\boldsymbol{\sigma}_h \in \boldsymbol{\Sigma}_h(\boldsymbol{\sigma}_h)$ such that

$$(\mathcal{A}\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h - \boldsymbol{\sigma}_h) - \langle \mathbf{u}^D, (\boldsymbol{\tau}_h - \boldsymbol{\sigma}_h)\cdot \mathbf{n}\rangle_{\Gamma_D} - \langle g, \mathbf{n}\cdot((\boldsymbol{\tau}_h - \boldsymbol{\sigma}_h)\cdot \mathbf{n})\rangle_{\Gamma_C} \geq 0 \tag{2.7}$$

holds for all $\boldsymbol{\tau}_h \in \boldsymbol{\Sigma}_h(\boldsymbol{\sigma}_h)$. As discrete spaces, Raviart–Thomas elements of degree 1 are combined with piecewise affine (possibly discontinuous) functions for $\mathbf{Z}_h$ and piecewise affine continuous functions for $\mathbf{X}_h$. With this choice of $\mathbf{Z}_h$, the first constraint in the definition of $\boldsymbol{\Sigma}_h(\boldsymbol{\sigma}_h)$ in (2.6) is equivalent to $\operatorname{div} \boldsymbol{\tau}_h + \mathbf{f} = \mathbf{0}$ due to our assumption of $\mathbf{f}$ being piecewise affine if the triangulation $\mathcal{T}_h$ is assumed to resolve the discontinuities. For boundary value problems associated with linear elasticity, these finite element spaces constitute an inf-sup stable combination (cf. [4]). The quasi-variational structure caused by the last constraint does, however, lead to situations where the uniqueness of the solution is not guaranteed, in general. It is known that uniqueness does not hold for sufficiently large ν_F by the counterexample given in [10]. In [16], uniqueness of the solutions of the quasi-variational inequality associated with Coulomb friction is shown under additional assumptions which are physically reasonable. We will come back to this issue in Sect. 3 in the context of error estimation and present a special situation in two dimensions where it is known that there exists a unique solution.

3 A Posteriori Error Estimation by Displacement Reconstruction

The method of Lagrange multipliers applied to (2.5) leads to the equation

$$\begin{aligned}(\mathcal{A}\boldsymbol{\sigma}, \boldsymbol{\tau}) - \langle \mathbf{u}^D, \boldsymbol{\tau}\cdot\mathbf{n}\rangle_{\Gamma_D} - \langle g, \mathbf{n}\cdot(\boldsymbol{\tau}\cdot\mathbf{n})\rangle_{\Gamma_C} \\ + (\mathbf{u}, \operatorname{div}\boldsymbol{\tau}) + (\boldsymbol{\theta}, \mathbf{as}\,\boldsymbol{\tau}) + \langle \rho, \mathbf{n}\cdot(\boldsymbol{\tau}\cdot\mathbf{n})\rangle_{\Gamma_C} + \langle \boldsymbol{\eta}, \mathbf{n}\times(\boldsymbol{\tau}\cdot\mathbf{n})\rangle_{\Gamma_C} = 0\end{aligned} \tag{3.1}$$

for all $\boldsymbol{\tau} \in H_{\Gamma_N}(\operatorname{div}, \Omega)^d$, where $\mathbf{u}$, $\boldsymbol{\theta}$, ρ, and $\boldsymbol{\eta}$ constitute multipliers corresponding to the constraints in (2.6). Inserting appropriate test functions which vanish on Γ_C and integrating by parts lead to $\mathcal{A}\boldsymbol{\sigma} = \boldsymbol{\varepsilon}(\mathbf{u})$ which implies that $\mathbf{u} \in H^1(\Omega)^d$ does indeed coincide with the displacement field and satisfies $\mathbf{u} = \mathbf{u}^D$ on Γ_D. Similarly, $\boldsymbol{\theta} = \mathbf{as}\,\nabla\mathbf{u}$ stands for the rotations. With this, we obtain

$$-\langle g, \mathbf{n}\cdot(\boldsymbol{\tau}\cdot\mathbf{n})\rangle_{\Gamma_C} + \langle \mathbf{u}, \boldsymbol{\tau}\cdot\mathbf{n}\rangle_{\Gamma_C} + \langle \rho, \mathbf{n}\cdot(\boldsymbol{\tau}\cdot\mathbf{n})\rangle_{\Gamma_C} + \langle \boldsymbol{\eta}, \mathbf{n}\times(\boldsymbol{\tau}\cdot\mathbf{n})\rangle_{\Gamma_C} = 0 \tag{3.2}$$

which may be split into its normal and its tangential components. The remaining Lagrange multipliers are identified as

$$\rho = g - \mathbf{n}\cdot\mathbf{u}\,, \quad \boldsymbol{\eta} = -\mathbf{u}\times\mathbf{n}\,, \tag{3.3}$$

and we end up with the complementarity conditions

$$\begin{aligned} \langle g - \mathbf{n}\cdot\mathbf{u}, \mathbf{n}\cdot(\boldsymbol{\sigma}\cdot\mathbf{n})\rangle_{\Gamma_C} &= 0 \\ -\langle \mathbf{n}\times\mathbf{u}, \mathbf{n}\times(\boldsymbol{\sigma}\cdot\mathbf{n})\rangle_{\Gamma_C} + \nu_F\langle|\mathbf{n}\times\mathbf{u}|, \mathbf{n}\cdot(\boldsymbol{\sigma}\cdot\mathbf{n})\rangle_{\Gamma_C} &= 0\,. \end{aligned} \tag{3.4}$$

This motivates the derivation of an a posteriori error estimator using the terms

$$\begin{aligned} &\|\mathcal{A}\boldsymbol{\sigma}_h - \boldsymbol{\varepsilon}(\mathbf{u}_h^R)\|^2\,,\ \langle \mathbf{n}\cdot\mathbf{u}_h^R - g, \mathbf{n}\cdot(\boldsymbol{\sigma}_h\cdot\mathbf{n})\rangle_{\Gamma_C} \text{ and} \\ &\qquad \langle \mathbf{n}\times\mathbf{u}_h^R, \mathbf{n}\times(\boldsymbol{\sigma}_h\cdot\mathbf{n})\rangle_{\Gamma_C} - \nu_F\langle|\mathbf{n}\times\mathbf{u}_h^R|, \mathbf{n}\cdot(\boldsymbol{\sigma}_h\cdot\mathbf{n})\rangle_{\Gamma_C} \end{aligned} \tag{3.5}$$

as building blocks. The reconstruction of $\mathbf{u}_h^R \in \mathbf{u}^D + H^1_{\Gamma_D}(\Omega)^d$ with $\mathbf{n}\cdot\mathbf{u}_h^R - g \le 0$ on Γ_C from the solution of (2.7) will be explained further below.

Starting from the first term in (3.5) and inserting $\mathcal{A}\boldsymbol{\sigma} = \boldsymbol{\varepsilon}(\mathbf{u})$ lead to

$$\begin{aligned} \|\mathcal{A}\boldsymbol{\sigma}_h - \boldsymbol{\varepsilon}(\mathbf{u}_h^R)\|^2 &= \|\mathcal{A}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h) - \boldsymbol{\varepsilon}(\mathbf{u}-\mathbf{u}_h^R)\|^2 \\ &= \|\mathcal{A}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\|^2 + \|\boldsymbol{\varepsilon}(\mathbf{u}-\mathbf{u}_h^R)\|^2 - 2(\mathcal{A}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h), \boldsymbol{\varepsilon}(\mathbf{u}-\mathbf{u}_h^R)) \\ &= \|\mathcal{A}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\|^2 + \|\boldsymbol{\varepsilon}(\mathbf{u}-\mathbf{u}_h^R)\|^2 - \frac{1}{\mu}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h, \boldsymbol{\varepsilon}(\mathbf{u}-\mathbf{u}_h^R)) \\ &\qquad + \frac{\lambda}{\mu(d\lambda+2\mu)}(\mathrm{tr}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h), \mathrm{div}(\mathbf{u}-\mathbf{u}_h^R))\,, \end{aligned} \tag{3.6}$$

where we have used the specific form of $\mathcal{A}$ from (2.1). For the first mixed term in (3.6), integration by parts leads to

$$\begin{aligned} (\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \boldsymbol{\varepsilon}(\mathbf{u}-\mathbf{u}_h^R)) &= (\boldsymbol{\sigma}-\boldsymbol{\sigma}_h, \nabla(\mathbf{u}-\mathbf{u}_h^R)) - (\boldsymbol{\sigma}-\boldsymbol{\sigma}_h, \mathbf{as}\,\nabla(\mathbf{u}-\mathbf{u}_h^R)) \\ &= \langle(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\cdot\mathbf{n}, \mathbf{u}-\mathbf{u}_h^R\rangle_{\Gamma_C} + (\mathbf{as}\,\boldsymbol{\sigma}_h, \nabla(\mathbf{u}-\mathbf{u}_h^R))\,. \end{aligned} \tag{3.7}$$

We split the boundary term in (3.7) further into its normal and tangential parts, respectively, and obtain

$$\begin{aligned} \langle(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\cdot\mathbf{n}, \mathbf{u}-\mathbf{u}_h^R\rangle_{\Gamma_C} &= \langle \mathbf{n}\cdot((\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\cdot\mathbf{n}), \mathbf{n}\cdot(\mathbf{u}-\mathbf{u}_h^R)\rangle_{\Gamma_C} \\ &\quad + \langle \mathbf{n}\times((\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\cdot\mathbf{n}), \mathbf{n}\times(\mathbf{u}-\mathbf{u}_h^R)\rangle_{\Gamma_C}\,. \end{aligned} \tag{3.8}$$

The first term on the right-hand side in (3.8) can be bounded as

$$\begin{aligned} &\langle \mathbf{n}\cdot((\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\cdot\mathbf{n}), \mathbf{n}\cdot(\mathbf{u}-\mathbf{u}_h^R)\rangle_{\Gamma_C} \\ &\quad = \langle \mathbf{n}\cdot(\boldsymbol{\sigma}\cdot\mathbf{n}), \mathbf{n}\cdot\mathbf{u} - g\rangle_{\Gamma_C} - \langle \mathbf{n}\cdot(\boldsymbol{\sigma}\cdot\mathbf{n}), \mathbf{n}\cdot\mathbf{u}_h^R - g\rangle_{\Gamma_C} \\ &\qquad - \langle \mathbf{n}\cdot(\boldsymbol{\sigma}_h\cdot\mathbf{n}), \mathbf{n}\cdot\mathbf{u} - g\rangle_{\Gamma_C} + \langle \mathbf{n}\cdot(\boldsymbol{\sigma}_h\cdot\mathbf{n}), \mathbf{n}\cdot\mathbf{u}_h^R - g\rangle_{\Gamma_C} \\ &\quad \le \langle \mathbf{n}\cdot(\boldsymbol{\sigma}_h\cdot\mathbf{n}), \mathbf{n}\cdot\mathbf{u}_h^R - g\rangle_{\Gamma_C} \end{aligned} \tag{3.9}$$

due to the first equation in (3.4) and the negative signs of $\mathbf{n}\cdot(\boldsymbol{\sigma}\cdot\mathbf{n})$, $\mathbf{n}\cdot(\boldsymbol{\sigma}_h\cdot\mathbf{n})$, $\mathbf{n}\cdot\mathbf{u}-g$, and $\mathbf{n}\cdot\mathbf{u}_h^R-g$, respectively. For the second term on the right-hand side in (3.8), we obtain

$$\begin{aligned}
&\langle\mathbf{n}\times((\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\cdot\mathbf{n}),\mathbf{n}\times(\mathbf{u}-\mathbf{u}_h^R)\rangle_{\Gamma_C}\\
&\quad=\langle\mathbf{n}\times(\boldsymbol{\sigma}\cdot\mathbf{n}),\mathbf{n}\times\mathbf{u}\rangle_{\Gamma_C}-\langle\mathbf{n}\times(\boldsymbol{\sigma}\cdot\mathbf{n}),\mathbf{n}\times\mathbf{u}_h^R\rangle_{\Gamma_C}\\
&\qquad-\langle\mathbf{n}\times(\boldsymbol{\sigma}_h\cdot\mathbf{n}),\mathbf{n}\times\mathbf{u}\rangle_{\Gamma_C}+\langle\mathbf{n}\times(\boldsymbol{\sigma}_h\cdot\mathbf{n}),\mathbf{n}\times\mathbf{u}_h^R\rangle_{\Gamma_C}\\
&\quad\le\nu_F\langle\mathbf{n}\cdot(\boldsymbol{\sigma}\cdot\mathbf{n}),|\mathbf{n}\times\mathbf{u}|\rangle_{\Gamma_C}-\nu_F\langle\mathbf{n}\cdot(\boldsymbol{\sigma}\cdot\mathbf{n}),|\mathbf{n}\times\mathbf{u}_h^R|\rangle_{\Gamma_C}\\
&\qquad-\nu_F\langle\mathbf{n}\cdot(\boldsymbol{\sigma}_h\cdot\mathbf{n}),|\mathbf{n}\times\mathbf{u}|\rangle_{\Gamma_C}+\langle\mathbf{n}\times(\boldsymbol{\sigma}_h\cdot\mathbf{n}),\mathbf{n}\times\mathbf{u}_h^R\rangle_{\Gamma_C}\,,
\end{aligned}\tag{3.10}$$

where the second equation in (3.4) and the inequality constraint in the definition of $\boldsymbol{\Sigma}(\boldsymbol{\sigma})$ and $\boldsymbol{\Sigma}_h(\boldsymbol{\sigma}_h)$ were used. This implies

$$\begin{aligned}
&\langle\mathbf{n}\times((\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\cdot\mathbf{n}),\mathbf{n}\times(\mathbf{u}-\mathbf{u}_h^R)\rangle_{\Gamma_C}\\
&\quad\le\nu_F\langle\mathbf{n}\cdot((\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\cdot\mathbf{n}),|\mathbf{n}\times\mathbf{u}|-|\mathbf{n}\times\mathbf{u}_h^R|\rangle_{\Gamma_C}\\
&\qquad+\langle\mathbf{n}\times(\boldsymbol{\sigma}_h\cdot\mathbf{n}),\mathbf{n}\times\mathbf{u}_h^R\rangle_{\Gamma_C}-\nu_F\langle\mathbf{n}\cdot(\boldsymbol{\sigma}_h\cdot\mathbf{n}),|\mathbf{n}\times\mathbf{u}_h^R|\rangle_{\Gamma_C}\,.
\end{aligned}\tag{3.11}$$

Finally, the second mixed term in (3.6) may be treated as

$$\begin{aligned}
(\mathrm{tr}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),\mathrm{div}(\mathbf{u}-\mathbf{u}_h^R))&=(\mathrm{tr}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),\frac{1}{d\lambda+2\mu}\mathrm{tr}\,\boldsymbol{\sigma}-\mathrm{div}\,\mathbf{u}_h^R))\\
&=\frac{1}{d\lambda+2\mu}\|\mathrm{tr}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\|^2+(\mathrm{tr}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),\frac{1}{d\lambda+2\mu}\mathrm{tr}\,\boldsymbol{\sigma}_h-\mathrm{div}\,\mathbf{u}_h^R)\\
&\ge(\mathrm{tr}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),\frac{1}{d\lambda+2\mu}\mathrm{tr}\,\boldsymbol{\sigma}_h-\mathrm{div}\,\mathbf{u}_h^R)\,.
\end{aligned}\tag{3.12}$$

Combining (3.7), (3.8), (3.9), and (3.11) and inserting this together with (3.12) into (3.6) gives

$$\begin{aligned}
&\mu\|\mathcal{A}\boldsymbol{\sigma}_h-\boldsymbol{\varepsilon}(\mathbf{u}_h^R)\|^2+\langle\mathbf{n}\cdot\mathbf{u}_h^R-g,\mathbf{n}\cdot(\boldsymbol{\sigma}_h\cdot\mathbf{n})\rangle_{\Gamma_C}\\
&\qquad+\langle\mathbf{n}\times\mathbf{u}_h^R,\mathbf{n}\times(\boldsymbol{\sigma}_h\cdot\mathbf{n})\rangle_{\Gamma_C}-\nu_F\langle|\mathbf{n}\times\mathbf{u}_h^R|,\mathbf{n}\cdot(\boldsymbol{\sigma}_h\cdot\mathbf{n})\rangle_{\Gamma_C}\\
&\quad\ge\mu\|\mathcal{A}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\|^2+\mu\|\boldsymbol{\varepsilon}(\mathbf{u}-\mathbf{u}_h^R)\|^2-(\mathbf{as}\,\boldsymbol{\sigma}_h,\boldsymbol{\nabla}(\mathbf{u}-\mathbf{u}_h^R))\\
&\qquad-\nu_F\langle\mathbf{n}\cdot((\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\cdot\mathbf{n}),|\mathbf{n}\times\mathbf{u}|-|\mathbf{n}\times\mathbf{u}_h^R|\rangle_{\Gamma_C}\\
&\qquad-\frac{\lambda}{d\lambda+2\mu}(\mathrm{tr}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),\mathrm{div}\,\mathbf{u}_h^R-\frac{1}{d\lambda+2\mu}\mathrm{tr}\,\boldsymbol{\sigma}_h)\,.
\end{aligned}\tag{3.13}$$

Our goal is to derive a computable a posteriori error estimator for $\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|$ from (3.13) for sufficiently small friction parameter ν_F. This generalizes corresponding results for the first-order system least squares functional in the linear elasticity case in [8] and the Signorini problem without friction in [2, 15]. To this end, an "inverse triangle-type inequality"

$$\nu_F \| |\mathbf{n} \times \mathbf{u}| - |\mathbf{n} \times \mathbf{u}_h^R| \|_{1/2,\Gamma_C} \leq \gamma_I(\nu_F) \|\mathbf{n} \times \mathbf{u} - \mathbf{n} \times \mathbf{u}_h^R\|_{1/2,\Gamma_C} \tag{3.14}$$

is required to hold with γ_I being a continuous function which is independent of h and satisfies $\gamma_I(0) = 0$. Such an inequality trivially holds with $\gamma_I(\nu_F) = \nu_F$ if $H^{1/2}(\Gamma_C)$ is replaced by $L^2(\Gamma_C)$ which occurs in the context of the regularized friction law in [9, Chap. 8]. It can, however, not be expected to hold in $H^{1/2}(\Gamma_C)$ without making additional assumptions (as will become clear from the example in Sect. 4). In fact, in the unregularized treatment of Coulomb friction, the validity of (3.14) would imply uniqueness of the solution of (2.3) for sufficiently small ν_F which is not known, in general. In Sect. 4, we will show that (3.14) does in fact hold under additional assumptions which are reasonable for situations like those treated in our test cases in Sect. 5. Our assumptions on the tangential displacement trace on Γ_C are similar to those in the study of uniqueness in [16].

Theorem 3.1 *Assume that (3.14) is satisfied. Then, for sufficiently small friction parameter ν_F,*

$$\begin{aligned} \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\| \leq C \Bigg(& \|\mathcal{A}\boldsymbol{\sigma}_h - \boldsymbol{\varepsilon}(\mathbf{u}_h^R)\|^2 + \| \frac{1}{d\lambda + 2\mu} \operatorname{tr} \boldsymbol{\sigma}_h - \operatorname{div} \mathbf{u}_h^R \|^2 \\ & + \langle \mathbf{n} \cdot \mathbf{u}_h^R - g, \mathbf{n} \cdot (\boldsymbol{\sigma}_h \cdot \mathbf{n}) \rangle_{\Gamma_C} \\ & + \langle \mathbf{n} \times \mathbf{u}_h^R, \mathbf{n} \times (\boldsymbol{\sigma}_h \cdot \mathbf{n}) \rangle_{\Gamma_C} - \nu_F \langle |\mathbf{n} \times \mathbf{u}_h^R|, \mathbf{n} \cdot (\boldsymbol{\sigma}_h \cdot \mathbf{n}) \rangle_{\Gamma_C} \Bigg)^{1/2} \end{aligned} \tag{3.15}$$

holds with a constant C which is independent of λ and h.

Proof In order to derive (3.15) from (3.13), we need to use some additional estimates. Firstly,

$$\|\operatorname{tr}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|^2 \leq C_D \|\mathcal{A}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|^2 \tag{3.16}$$

holds with a constant C_D since $(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h) \cdot \mathbf{n} = \mathbf{0}$ on the boundary segment Γ_N (cf. [7, Sect. 5]). Moreover, a Korn inequality of the type

$$\|\nabla(\mathbf{u} - \mathbf{u}_h^R)\|^2 \leq C_K \|\boldsymbol{\varepsilon}(\mathbf{u} - \mathbf{u}_h^R)\|^2 \tag{3.17}$$

holds with a constant C_K due to the fact that $\mathbf{u} - \mathbf{u}_h^R$ vanishes on the boundary segment Γ_D. And, finally, we can bound the antisymmetric stress in the form

$$\|\mathbf{as}\boldsymbol{\sigma}_h\| = 2\mu \|\mathbf{as}\mathcal{A}\boldsymbol{\sigma}_h\| = 2\mu \|\mathbf{as}(\mathcal{A}\boldsymbol{\sigma}_h - \boldsymbol{\varepsilon}(\mathbf{u}_h^R))\| \leq 2\mu \|\mathcal{A}\boldsymbol{\sigma}_h - \boldsymbol{\varepsilon}(\mathbf{u}_h^R)\|. \tag{3.18}$$

Using these inequalities in combination with (3.14), (3.13) leads to

$$\begin{aligned}
&\mu\|\mathcal{A}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\|^2+\mu\|\boldsymbol{\varepsilon}(\mathbf{u}-\mathbf{u}_h^R)\|^2\\
&\quad\le \mu\|\mathcal{A}\boldsymbol{\sigma}_h-\boldsymbol{\varepsilon}(\mathbf{u}_h^R)\|^2+\langle \mathbf{n}\cdot\mathbf{u}_h^R-g,\mathbf{n}\cdot(\boldsymbol{\sigma}_h\cdot\mathbf{n})\rangle_{\Gamma_C}\\
&\quad+\langle \mathbf{n}\times\mathbf{u}_h^R,\mathbf{n}\times(\boldsymbol{\sigma}_h\cdot\mathbf{n})\rangle_{\Gamma_C}-\nu_F\langle|\mathbf{n}\times\mathbf{u}_h^R|,\mathbf{n}\cdot(\boldsymbol{\sigma}_h\cdot\mathbf{n})\rangle_{\Gamma_C}\\
&\quad+\frac{2\mu^2}{\delta_1}\|\mathcal{A}\boldsymbol{\sigma}_h-\boldsymbol{\varepsilon}(\mathbf{u}_h^R)\|^2+\frac{\delta_1}{2}C_K\|\boldsymbol{\varepsilon}(\mathbf{u}-\mathbf{u}_h^R)\|^2\\
&\quad+\frac{\delta_2}{2}\|\mathbf{n}\cdot(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\cdot\mathbf{n}\|^2_{-1/2,\Gamma_C}+\frac{\gamma_I(\nu_F)^2}{2\delta_2}\|\mathbf{n}\times(\mathbf{u}-\mathbf{u}_h^R)\|^2_{1/2,\Gamma_C}\\
&\quad+\frac{\delta_3}{2}\left(\frac{\lambda}{d\lambda+2\mu}\right)^2\|\mathrm{tr}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\|^2+\frac{1}{2\delta_3}\|\mathrm{div}\,\mathbf{u}_h^R-\frac{1}{d\lambda+2\mu}\mathrm{tr}\,\boldsymbol{\sigma}_h\|^2
\end{aligned}\tag{3.19}$$

with positive numbers δ_1,δ_2, and δ_3 to be chosen appropriately. With the trace inequalities

$$\begin{aligned}
\|\mathbf{n}\cdot((\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\cdot\mathbf{n})\|^2_{-1/2,\Gamma_C}&\le\|(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\cdot\mathbf{n}\|^2_{-1/2,\Gamma_C}\le C_T\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|^2\\
\|\mathbf{n}\times(\mathbf{u}-\mathbf{u}_h^R)\|_{1/2,\Gamma_C}&\le\|\mathbf{u}-\mathbf{u}_h^R\|_{1/2,\Gamma_C}\le C_U\|\boldsymbol{\varepsilon}(\mathbf{u}-\mathbf{u}_h^R)\|^2
\end{aligned}$$

and the observation that

$$\begin{aligned}
\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|^2&=\|2\mu\mathcal{A}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)+\frac{\lambda}{d\lambda+2\mu}\mathrm{tr}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\mathbf{I}\|^2\\
&\le 8\mu^2\|\mathcal{A}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\|^2+2d\left(\frac{\lambda}{d\lambda+2\mu}\right)^2\|\mathrm{tr}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\|^2\\
&\le\left(8\mu^2+2d\left(\frac{\lambda}{d\lambda+2\mu}\right)^2C_D\right)\|\mathcal{A}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\|^2
\end{aligned}$$

holds, choosing δ_1,δ_2, and δ_3 sufficiently small finishes the proof. □

For the displacement reconstruction, we basically adopt the approach in [18] which itself relies on the procedure proposed in [17] as follows. From (3.1) we obtain $\nabla\mathbf{u}=\mathcal{A}\boldsymbol{\sigma}+\boldsymbol{\theta}$ which means that $\mathbf{z}_h=\mathcal{A}\boldsymbol{\sigma}_h+\boldsymbol{\theta}_h$ (piecewise polynomial) is an approximate gradient and may be used as a starting point for the construction of $\mathbf{u}_h^R$. The displacement reconstruction is done in the following steps:

(i) For each $T\in\mathcal{T}_h$, determine $\mathbf{u}_h^\circ|_T\in P_k(T)^d$, polynomial of degree k, such that

$$\begin{aligned}
(\nabla\mathbf{u}_h^\circ,\nabla\mathbf{v}_h)_{0,T}&=(\mathbf{z}_h,\nabla\mathbf{v}_h)_{0,T}\ \text{for all}\ \mathbf{v}_h\in P_k(T)^d\,,\\
(\mathbf{u}_h^\circ,\mathbf{e})_{0,T}&=(\mathbf{u}_h,\mathbf{e})_{L^2(T)}\ \text{for all}\ \mathbf{e}\in P_0(T)^d\,.
\end{aligned}\tag{3.20}$$

(ii) A conforming reconstruction $\mathbf{u}_h^R$ is constructed by averaging,

$$\mathbf{u}_h^R(\mathbf{x}) = \frac{1}{\#\{T : \mathbf{x} \in T\}} \sum_{T:\mathbf{x}\in T} \mathbf{u}_h^\circ\big|_T(\mathbf{x}) \,, \tag{3.21}$$

and enforcing the boundary conditions on $\mathbf{u}_h^R = \mathbf{u}^D$ on Γ_D. If we also enforce the complementarity conditions strongly by setting

$$\begin{aligned} \mathbf{n}\cdot\mathbf{u}_h^R - g = 0 \text{ on all edges } E \subset \Gamma_C \text{ with } \mathbf{n}\cdot(\boldsymbol{\sigma}_h\cdot\mathbf{n}) \neq 0 \,, \\ \mathbf{n}\times\mathbf{u}_h^R = 0 \text{ on all edges } E \subset \Gamma_C \text{ with } |\mathbf{n}\times(\boldsymbol{\sigma}_h\cdot\mathbf{n})| \neq -\nu_F \mathbf{n}\cdot(\boldsymbol{\sigma}_h\cdot\mathbf{n}) \,, \end{aligned} \tag{3.22}$$

then the only remaining terms on the right-hand side in (3.15) are

$$\begin{aligned} &\|\mathcal{A}\boldsymbol{\sigma}_h - \boldsymbol{\varepsilon}(\mathbf{u}_h^R)\|^2 + \|\frac{1}{d\lambda+2\mu}\operatorname{tr}\boldsymbol{\sigma}_h - \operatorname{div}\mathbf{u}_h^R\|^2 \\ &= \sum_{T\in\mathcal{T}_h} (\|\mathcal{A}\boldsymbol{\sigma}_h - \boldsymbol{\varepsilon}(\mathbf{u}_h^R)\|_{0,T}^2 + \|\frac{1}{d\lambda+2\mu}\operatorname{tr}\boldsymbol{\sigma}_h - \operatorname{div}\mathbf{u}_h^R\|_{0,T}^2) =: \sum_{T\in\mathcal{T}_h} \eta_T^2 \,, \end{aligned} \tag{3.23}$$

which we will use as error estimator.

4 A Norm Equivalence in $H^{1/2}(\Gamma)$

Our purpose in this section is to show that (3.14) is indeed fulfilled under certain conditions on $\mathbf{n}\times\mathbf{u}_h^R$ which can be verified from our numerical experiments and under certain assumptions on $\mathbf{n}\times\mathbf{u}$ for the exact solution. We restrict ourselves to the two-dimensional case and assume that Γ_C constitutes a smooth boundary segment.

Theorem 4.1 *Assume that Γ_C contains a segment Γ_C° where $\mathbf{n}\times\mathbf{u}_h^R \equiv 0$ and that the tangential derivative satisfies*

$$m \le \partial_{\mathbf{t}}(\mathbf{n}\times\mathbf{u}_h^R) \le M \text{ (or } m \le -\partial_{\mathbf{t}}(\mathbf{n}\times\mathbf{u}_h^R) \le M\text{) uniformly on } \Gamma_C\backslash\Gamma_C^\circ \,, \tag{4.1}$$

with positive constants m and M and $|\Gamma_C\backslash\Gamma_C^\circ| \ge \gamma > 0$ uniformly in h. Moreover, assume that $\mathbf{n}\times\mathbf{u} \in H^1(\Gamma_C)$ holds and that the subsets

$$\Gamma_C^+ = \{\mathbf{x}\in\Gamma_C : \mathbf{n}\times\mathbf{u} > 0\} \,,\ \Gamma_C^- = \{\mathbf{x}\in\Gamma_C : \mathbf{n}\times\mathbf{u} < 0\} \tag{4.2}$$

are separated by a segment of length at least ℓ, where $\mathbf{n} \times \mathbf{u} \equiv 0$ holds. Then,

$$\| |\mathbf{n} \times \mathbf{u}| - |\mathbf{n} \times \mathbf{u}_h^R| \|_{1/2,\Gamma_C} \leq \frac{C_I}{\ell^{1/2}} \|\mathbf{n} \times \mathbf{u} - \mathbf{n} \times \mathbf{u}_h^R\|_{1/2,\Gamma_C} \tag{4.3}$$

holds with a constant C_I that depends only on the ratio M/m.

Note that the assumption on $\mathbf{n} \times \mathbf{u}_h^R$ is apparently fulfilled in our numerical results presented in Sect. 5. The proof will be based on real interpolation between $L^2(\Gamma_C)$ and $H^1(\Gamma_C)$ and the crucial part consists in showing the analogous estimate to (4.3) in $H^1(\Gamma_C)$ which is the content of the following lemma.

Lemma 4.2 *Let $\mathbf{n} \times \mathbf{u}_h^R$ satisfy the assumptions from Theorem 4.1. Then, there exists a constant $C_I \geq 0$, depending only on the ratio M/m in (4.1), such that*

$$\| |\psi| - |\mathbf{n} \times \mathbf{u}_h^R| \|_{1,\Gamma_C} \leq \frac{C_I}{\ell} \|\psi - \mathbf{n} \times \mathbf{u}_h^R\|_{1,\Gamma_C} \tag{4.4}$$

holds for all $\psi \in H^1(\Gamma_C)$ with the property that the subsets

$$\Gamma_C^+ = \{\mathbf{x} \in \Gamma_C : \psi > 0\}\,, \ \Gamma_C^- = \{\mathbf{x} \in \Gamma_C : \psi < 0\}$$

are separated by a segment of length at least ℓ, where $\psi \equiv 0$ holds.

Proof Without loss of generality, we may restrict ourselves to a subset of Γ_C, parametrized by $\mathbf{x}(s)$, $s \in I = [0, 1]$ such that the function $\phi = (\mathbf{n} \times \mathbf{u}_h^R) \circ \mathbf{x} \in H^1(I)$ vanishes on $[0, \xi]$ with $0 \leq \xi \leq 1$ and $m \leq \phi'(s) \leq M$ for $\xi < s < 1$. Define $I_- = \{s \in I : \psi(s) \leq 0\}$ and note that for all $s \in I \backslash I_-$ we have $|\psi| - |\phi| = \psi - \phi$ and therefore

$$\| |\psi| - |\phi| \|_{1,I\backslash I_-} = \|\psi - \phi\|_{1,I\backslash I_-}\,. \tag{4.5}$$

I_- is the (finite) union of subintervals $I_-^{(i)} = [\xi_-^{(i)}, \xi_+^{(i)}]$, $i = 1, 2, 3, \ldots$, where $\psi \leq 0$ holds. We are left with showing

$$\| |\psi| - |\phi| \|_{1,I_-^{(i)}} \leq C_I \|\psi - \phi\|_{1,I_-^{(i)}} \text{ for } i = 1, 2, 3, \ldots\,. \tag{4.6}$$

For each i, we have $I_-^{(i)} = I_l^{(i)} \cup I_r^{(i)}$ such that the following holds:

$$\phi \equiv 0 \text{ on } I_l^{(i)}\,,\ m \leq \phi' \leq M \text{ on } I_r^{(i)} \text{ and } \psi \leq 0 \text{ on } I_-^{(i)}\,. \tag{4.7}$$

The left part $I_l^{(i)}$ of the interval may be empty. The length of the right part $I_r^{(i)}$, however, can be chosen to be either zero or at least ℓ by our assumption on ψ and ϕ if we select $\xi_+^{(i)}$ appropriately. This implies that

$$\frac{\||\psi|-|\phi|\|^2_{1,I_-^{(i)}}}{\|\psi-\phi\|^2_{1,I_-^{(i)}}}=\frac{\|\psi+\phi\|^2_{1,I_-^{(i)}}}{\|\psi-\phi\|^2_{1,I_-^{(i)}}}=\frac{\|\psi+\phi\|^2_{0,I_-^{(i)}}+\|\psi'+\phi'\|^2_{0,I_-^{(i)}}}{\|\psi-\phi\|^2_{0,I_-^{(i)}}+\|\psi'-\phi'\|^2_{0,I_-^{(i)}}}\,, \tag{4.8}$$

which we need to bound from above by a constant. The first terms in the numerator and the denominator of (4.8) are related by

$$\|\psi+\phi\|^2_{0,I_-^{(i)}}\le\|\psi-\phi\|^2_{0,I_-^{(i)}}\,,\quad\text{since } |\psi+\phi|\le|\psi|+|\phi|=-\psi+\phi=|\psi-\phi|$$

holds pointwise on $I_-^{(i)}$. The second term in the numerator may be bounded as

$$\begin{aligned}\|\psi'+\phi'\|^2_{0,I_-^{(i)}}&=\|\psi'-\phi'+2\phi'\|^2_{0,I_-^{(i)}}\le 2\|\psi'-\phi'\|^2_{0,I_-^{(i)}}+8\|\phi'\|^2_{0,I_-^{(i)}}\\&\le 2\|\psi'-\phi'\|^2_{0,I_-^{(i)}}+8M^2\|1\|^2_{0,I_r^{(i)}}\le 2\|\psi'-\phi'\|^2_{0,I_-^{(i)}}+24\frac{M^2}{m^2\ell^2}\|\phi\|^2_{0,I_r^{(i)}}\\&\le 2\|\psi'-\phi'\|^2_{0,I_-^{(i)}}+24\frac{M^2}{m^2\ell^2}\|\psi-\phi\|^2_{0,I_-^{(i)}}\,,\end{aligned}$$

where we used the fact that ϕ lies above the linearly increasing function with slope m on $I_r^{(i)}$. Inserting the last two estimates into (4.8) leads to

$$\frac{\||\psi|-|\phi|\|^2_{1,I_-^{(i)}}}{\|\psi-\phi\|^2_{1,I_-^{(i)}}}\le\max\left\{1+24\frac{M^2}{m^2\ell^2},2\right\}=1+24\frac{M^2}{m^2\ell^2}\,, \tag{4.9}$$

which completes the proof. □

Before turning to the Proof of Theorem 4.1, let us look at an example that (4.4) does not hold without the additional assumption on $\mathbf{n}\times\mathbf{u}$.

Example Let Γ_C be the interval $[0,1]$ and consider $\mathbf{n}\times\mathbf{u}_h^R$ to be the piecewise linear function which vanishes on $[0,1/2]$ and is monotonically increasing from 0 to 1/2 on $[1/2,1]$. With ψ being the function defined by

$$\psi(s)=\begin{cases}-2\delta s & ,\ s\in[0,1/2]\,,\\ s-1/2-\delta & ,\ s\in[1/2,1]\end{cases}$$

we get

$$\begin{aligned}&\|\psi'-(\mathbf{n}\times\mathbf{u}_h^R)'\|^2_{0,\Gamma_C}=2\delta^2\,,\ \|\psi-(\mathbf{n}\times\mathbf{u}_h^R)\|^2_{0,\Gamma_C}=\frac{2}{3}\delta^2\,,\\&\||\psi|'-|\mathbf{n}\times\mathbf{u}_h^R|'\|^2_{0,\Gamma_C}=2\delta^2+4\delta\,,\ \||\psi|-|\mathbf{n}\times\mathbf{u}_h^R|\|^2_{0,\Gamma_C}=\frac{2}{3}\delta^2-\frac{2}{3}\delta^3\,.\end{aligned}$$

Therefore, the ratio

$$\frac{\| |\psi| - |\mathbf{n} \times \mathbf{u}_h^R| \|_{1,\Gamma_C}^2}{\|\psi - \mathbf{n} \times \mathbf{u}_h^R\|_{1,\Gamma_C}^2} = \frac{3}{2\delta} + 1 - \frac{\delta}{4}$$

becomes arbitrarily large as $\delta \to 0$.

Proof of Theorem 4.1

(i) As already indicated, the proof uses the interpretation of $H^{1/2}(\Gamma_C)$ as interpolation space $[L^2(\Gamma_C), H^1(\Gamma_C)]_{1/2,2}$ (cf. [1, Theorem 7.23]). The norm in $H^{1/2}(\Gamma_C)$ is therefore given by

$$\|\chi\|_{1/2,\Gamma_C} = \left(\int_0^\infty t^{-2} K(t, \chi)^2 \, dt \right)^{1/2} \tag{4.10}$$

with

$$K(t, \chi) = \inf_{\vartheta \in H^1(\Gamma_C)} \left(\|\chi - \vartheta\|_{0,\Gamma_C} + t \|\vartheta\|_{1,\Gamma_C} \right) . \tag{4.11}$$

We can therefore prove (4.3) by showing that

$$K(t, |\mathbf{n} \times \mathbf{u}| - |\mathbf{n} \times \mathbf{u}_h^R|) \le C_I \, K(t, \mathbf{n} \times \mathbf{u} - \mathbf{n} \times \mathbf{u}_h^R)$$

holds for all $t \in [0, \infty)$. Equivalently, we will use

$$\hat{K}(t, \chi) = \inf_{\vartheta \in H^1(\Gamma_C)} \left(\|\chi - \vartheta\|_{0,\Gamma_C}^2 + t^2 \|\vartheta\|_{1,\Gamma_C}^2 \right)^{1/2} \tag{4.12}$$

and show that

$$\hat{K}(t, |\mathbf{n} \times \mathbf{u}| - |\mathbf{n} \times \mathbf{u}_h^R|) \le C_I \, \hat{K}(t, \mathbf{n} \times \mathbf{u} - \mathbf{n} \times \mathbf{u}_h^R) \tag{4.13}$$

is satisfied for all $t \in [0, \infty)$.

(ii) For $\chi \in H^1(\Gamma_C)$, the infimum in (4.12) is attained, for $t > 0$, by the solution $\theta_\chi(t) \in H^1(\Gamma_C)$ of

$$\langle \theta_\chi(t), \rho \rangle_{0,\Gamma_C} + t^2 \langle \theta_\chi(t), \rho \rangle_{1,\Gamma_C} = \langle \chi, \rho \rangle_{0,\Gamma_C} \text{ for all } \rho \in H^1(\Gamma_C) \tag{4.14}$$

which leads to

$$\hat{K}(t, \chi)^2 = \|\chi - \theta_\chi(t)\|_{0,\Gamma_C}^2 + t^2 \|\theta_\chi(t)\|_{1,\Gamma_C}^2 =: Q_\chi(t) . \tag{4.15}$$

We have $\theta_\chi(0) = \chi$ and $\theta_\chi \to 0$ for $t \to \infty$ which leads to $Q_\chi(0) = 0$ and $Q_\chi(t) \to \|\chi\|^2_{0,\Gamma_C}$ for $t \to \infty$. Moreover, from (4.12) and (4.15), we deduce that $Q_\chi(t)$ is monotonically increasing in $[0, \infty)$. Differentiating (4.14) with respect to t gives

$$\langle \theta'_\chi(t), \rho\rangle_{0,\Gamma_C} + 2t\langle \theta_\chi(t), \rho\rangle_{1,\Gamma_C} + t^2\langle \theta'_\chi(t), \rho\rangle_{1,\Gamma_C} = 0 \text{ for all } \rho \in H^1(\Gamma_C) \tag{4.16}$$

and, in particular, $\theta'_\chi(0) = 0$. Differentiating (4.15) with respect to t implies

$$Q'_\chi(t) = 2\langle \chi - \theta_\chi(t), \theta'_\chi(t)\rangle_{0,\Gamma_C} + 2t\|\theta_\chi(t)\|^2_{1,\Gamma_C} + 2t^2\langle \theta_\chi(t), \theta'_\chi(t)\rangle_{1,\Gamma_C} \tag{4.17}$$

leading to

$$Q'_\chi(0) = 2\langle \chi - \theta_\chi(0), \theta'_\chi(0)\rangle_{0,\Gamma_C} = 0\,. \tag{4.18}$$

In order to access the behavior of $Q_\chi(t)$ near 0, the second derivative $Q''_\chi(0)$ is therefore needed. Differentiating (4.17) once more, we obtain

$$\begin{aligned} Q''_\chi(t) = {} & 2\langle \chi - \theta_\chi(t), \theta''_\chi(t)\rangle_{0,\Gamma_C} - 2\|\theta'_\chi(t)\|^2_{0,\Gamma_C} + 2\|\theta_\chi(t)\|^2_{1,\Gamma_C} \\ & + 8t\langle \theta_\chi(t), \theta'_\chi(t)\rangle_{1,\Gamma_C} + 2t^2\|\theta'_\chi(t)\|^2_{1,\Gamma_C} + 2t^2\langle \theta_\chi(t), \theta''_\chi(t)\rangle_{1,\Gamma_C} \end{aligned} \tag{4.19}$$

and, in particular,

$$Q''_\chi(0) = 2\langle \chi - \theta_\chi(0), \theta''_\chi(0)\rangle_{0,\Gamma_C} - 2\|\theta'_\chi(0)\|^2_{0,\Gamma_C} + 2\|\theta_\chi(0)\|^2_{1,\Gamma_C} = 2\|\chi\|^2_{1,\Gamma_C}\,. \tag{4.20}$$

The pointwise inverse triangle inequality gives us

$$\begin{aligned} \lim_{t\to\infty} Q_{|\mathbf{n}\times\mathbf{u}|-|\mathbf{n}\times\mathbf{u}_h^R|}(t) &= \||\mathbf{n}\times\mathbf{u}| - |\mathbf{n}\times\mathbf{u}_h^R|\|^2_{0,\Gamma_C} \\ &\le \|\mathbf{n}\times\mathbf{u} - \mathbf{n}\times\mathbf{u}_h^R\|^2_{0,\Gamma_C} = \lim_{t\to\infty} Q_{\mathbf{n}\times\mathbf{u}-\mathbf{n}\times\mathbf{u}_h^R}(t) \end{aligned} \tag{4.21}$$

and from Lemma 4.2, we deduce

$$\begin{aligned} Q''_{|\mathbf{n}\times\mathbf{u}|-|\mathbf{n}\times\mathbf{u}_h^R|}(0) &= 2\||\mathbf{n}\times\mathbf{u}| - |\mathbf{n}\times\mathbf{u}_h^R|\|^2_{1,\Gamma_C} \\ &\le 2\frac{C_I^2}{\ell^2}\|\mathbf{n}\times\mathbf{u} - \mathbf{n}\times\mathbf{u}_h^R\|^2_{1,\Gamma_C} = \frac{C_I^2}{\ell^2} Q''_{\mathbf{n}\times\mathbf{u}-\mathbf{n}\times\mathbf{u}_h^R}(0)\,. \end{aligned} \tag{4.22}$$

If we define $\tilde{Q}_\chi(s) = Q_\chi(s(\ell/C_I))$, then $\tilde{Q}''_\chi(0) = (\ell/C_I)^2 Q''_\chi(0)$ and therefore

$$\begin{aligned} \tilde{Q}''_{|\mathbf{n}\times\mathbf{u}|-|\mathbf{n}\times\mathbf{u}_h^R|}(0) &\le Q''_{\mathbf{n}\times\mathbf{u}-\mathbf{n}\times\mathbf{u}_h^R}(0)\,, \\ \lim_{s\to\infty} \tilde{Q}_{|\mathbf{n}\times\mathbf{u}|-|\mathbf{n}\times\mathbf{u}_h^R|}(s) &\le \lim_{t\to\infty} Q_{\mathbf{n}\times\mathbf{u}-\mathbf{n}\times\mathbf{u}_h^R}(t)\,. \end{aligned} \tag{4.23}$$

This implies the existence of a constant C (independent of ℓ) such that

$$\tilde{Q}_{|\mathbf{n}\times\mathbf{u}|-|\mathbf{n}\times\mathbf{u}_h^R|}(t) \le C\, Q_{\mathbf{n}\times\mathbf{u}-\mathbf{n}\times\mathbf{u}_h^R}(t) \text{ for all } t \in [0,\infty) \tag{4.24}$$

holds leading to

$$\begin{aligned} \||\mathbf{n}\times\mathbf{u}| - |\mathbf{n}\times\mathbf{u}_h^R|\|^2_{1/2,\Gamma_C} &= \int_0^\infty t^{-2} Q_{|\mathbf{n}\times\mathbf{u}|-|\mathbf{n}\times\mathbf{u}_h^R|}(t)\,dt \\ &= \int_0^\infty \left(\frac{\ell}{C_I}s\right)^{-2} Q_{|\mathbf{n}\times\mathbf{u}|-|\mathbf{n}\times\mathbf{u}_h^R|}\left(\frac{\ell}{C_I}s\right)\frac{\ell}{C_I}ds \\ &= \frac{C_I}{\ell}\int_0^\infty s^{-2}\tilde{Q}_{|\mathbf{n}\times\mathbf{u}|-|\mathbf{n}\times\mathbf{u}_h^R|}(s)\,ds \\ &\le \frac{C_I C}{\ell}\int_0^\infty s^{-2} Q_{\mathbf{n}\times\mathbf{u}-\mathbf{n}\times\mathbf{u}_h^R}(s)\,ds = \frac{C_I C}{\ell}\|\mathbf{n}\times\mathbf{u}-\mathbf{n}\times\mathbf{u}_h^R\|^2_{1/2,\Gamma_C} \end{aligned} \tag{4.25}$$

which completes the proof. □

The assumptions on $\mathbf{n}\times\mathbf{u}_h^R$ can be checked numerically, and they do indeed hold in our numerical examples in Sect. 5. The restriction to situations for which the solution satisfies $\mathbf{n}\times\mathbf{u} \in H^1(\Gamma_C)$ is certainly an unpleasant limitation. However, this is justified, at least for the test examples in Sect. 5, by the behavior of $\mathbf{n}\times\mathbf{u}_h^R$ in our numerical experiments. The assumption on $\mathbf{n}\times\mathbf{u}$ vanishing on a segment of length ℓ between sign changes is physically reasonable and corresponds to the sign change of the tangential traction force due to friction. Such an assumption also occurs in the uniqueness study for contact with Coulomb friction in [16] and is used for finite element error analysis in [12]. In order to deduce the assumption $\gamma_I(0) = 0$ from the simple upper bound

$$\gamma_I(v_F) \le C_I \frac{v_F}{\ell^{1/2}} \tag{4.26}$$

obtained from (3.14) and Theorem 4.1, we would need to have $\ell \approx v_F^{2-\varepsilon}$ as $v_F \to 0$ with $\varepsilon > 0$. The actual dependence of ℓ on v_F is expected to depend on the geometry of the domain and of the obstacle considered. From the numerical evidence, it seems that this dependence is not quite reached in our test example. We may, however, still

achieve (3.14) with $\gamma_I(0)$ by the introduction of an approximation $\tilde{\mathbf{u}} \in H^{1/2}(\Gamma_C)$ such that $\mathbf{n} \times \tilde{\mathbf{u}}$ does indeed vanish on an interval of length $\ell \approx \nu_F^{2-\varepsilon}$ and $\|\mathbf{n} \times \tilde{\mathbf{u}} - \mathbf{n} \times \mathbf{u}\|_{1/2,\Gamma_C} \leq C\nu_F \|\mathbf{n} \times \mathbf{u}\|_{1/2,\Gamma_C}$ holds. From the symmetry in this special situation, we may actually deduce that (3.14) is fulfilled with $\gamma_I(\nu_F) = \nu_F$ if the sticky zone around the lowest part of the half-disk is adequately resolved. We will perform adaptive finite element computations with the error estimator from Sect. 3 and obtain good results even for rather large friction coefficients ν_F as we shall see in the next section.

We also want to remark that the three-dimensional situation with Γ_C being a surface area appears to be considerably more complicated. The derivation of a three-dimensional analogue to Theorem 4.1 is the object of our current investigation.

5 Numerical Experiments

In this section, we present some results achieved by the numerical implementation of the discussed finite element method combined with the reconstruction based error estimator and the following adaptive mesh refinement strategy. A Dörfler marking strategy is applied, which consists of finding the smallest set of triangles $\tilde{\mathcal{T}}_h \subset \mathcal{T}_h$ such that

$$\sum_{T \in \tilde{\mathcal{T}}_h} \eta_T^2 \geq \theta^2 \sum_{T \in \mathcal{T}_h} \eta_T^2 \tag{5.1}$$

holds for a chosen parameter θ. All triangles in this set are then refined as well as those adjacent triangles necessary to avoid hanging nodes.

Example 1 (Hertzian Contact - Half-Disk on Line) In this first example, we consider the domain Ω of the lower half-disk with center at the origin and radius $R = 0.5$. The shear modulus μ is scaled to 1 and the coefficient of friction ν_F equals 0.4. Both the compressible case with Lamé parameter $\lambda \approx 1.27$ and the incompressible case ($\lambda = \infty$) will be treated. The body is constrained by a rigid foundation represented by the horizontal line at $y = -0.5$, and the potential contact boundary is $\Gamma_C := \left\{ R(\cos(\varphi), \sin(\varphi)) : \varphi \in (\frac{4}{3}\pi, \frac{5}{3}\pi) \right\}$. Displacement on $\Gamma_D := (-R, R) \times \{0\}$ is prescribed by $\mathbf{u}^D = (0, -0.01)$ while the volume forces $\mathbf{f}$ on Ω are set to zero.

Table 1 shows the results of the compressible case on a sequence of adaptively refined triangulations obtained with Dörfler parameter $\theta = 0.8$. Table 2 represents the incompressible case. The number of active constraints for the conditions on the normal and tangential boundary stress are denoted by A_n and A_t, respectively. The number for the tangential constraint increases faster in the compressible case due to the smaller sticky zone. Since the displacement was reconstructed in a way such that the boundary terms of the error estimator vanished, we did not list them in the table.

Table 1 Results for Example 1 (compressible case)

l	$\dim\boldsymbol{\Sigma}_h$	$\dim\mathbf{Z}_h$	$\dim\mathbf{X}_h$	$\|\mathcal{A}\boldsymbol{\sigma}_h - \boldsymbol{\varepsilon}(\mathbf{u}_h^R)\|^2$	A_n	A_t
0	3624	2166	205	7.499e − 07	16	18
1	3748	2238	212	3.080e − 07	16	20
2	4248	2532	239	1.217e − 07	22	30
3	5016	2982	281	4.873e − 08	30	46
4	7052	4194	390	1.259e − 08	38	60
5	11, 440	6804	619	4.688e − 09	50	86
6	20, 576	12, 252	1092	1.518e − 09	80	136
7	37, 256	22, 206	1960	5.641e − 10	104	203
8	65, 064	38, 814	3389	2.082e − 10	136	318
9	112, 692	67, 230	5851	8.306e − 11	172	527
10	185, 508	110, 628	9615	3.426e − 11	216	905

Table 2 Results for Example 1 (incompressible case)

$\dim\boldsymbol{\Sigma}_h$	$\dim\mathbf{Z}_h$	$\dim\mathbf{X}_h$	$\|\mathcal{A}\boldsymbol{\sigma}_h - \boldsymbol{\varepsilon}(\mathbf{u}_h^R)\|^2$	$\|\mathrm{div}\,\mathbf{u}_h^R\|^2$	A_n	A_t
3624	2166	205	1.248e − 06	5.817e − 07	16	16
3932	2346	222	3.325e − 07	1.799e − 07	20	22
4632	2760	263	1.057e − 07	5.936e − 08	24	24
6540	3894	364	4.216e − 08	2.398e − 08	32	34
8912	5298	491	2.058e − 08	1.274e − 08	38	40
13, 972	8304	756	9.898e − 09	5.913e − 09	52	56
22, 644	13, 452	1218	4.676e − 09	2.888e − 09	62	70
37, 444	22, 248	1997	2.108e − 09	1.356e − 09	82	98
61, 156	36, 324	3250	9.428e − 10	5.825e − 10	102	132
98, 640	58, 530	5237	4.322e − 10	2.563e − 10	116	162
162, 852	96, 534	8656	2.063e − 10	1.265e − 10	160	233

If one chooses not to impose boundary conditions on $\mathbf{n} \times \mathbf{u}_h^R$, the frictional terms in the error estimator will not vanish but the resulting refinement and convergence are comparable.

Figure 1 depicts the initial and deformed configurations of the compressible case after 5 steps of adaptive refinement. The deformed configuration is obtained using the reconstructed displacement $\mathbf{u}_h^R$. Besides the contact zone refinement concentrates at the corner singularities. In the contact zone, the refinement concentrates at the transition points from stick to slip and from contact to separation as can be seen in Fig. 2.

The surface forces in the highly resolved contact zone (10 adaptive refinements) as well as the corresponding reconstructed displacements are shown in Fig. 3. In the incompressible case, the interval where contact occurs is only slightly larger than in the compressible case while the sticky zone is much larger. Also the magnitude of the contact pressure is increased by approximately 45%. Another interesting feature

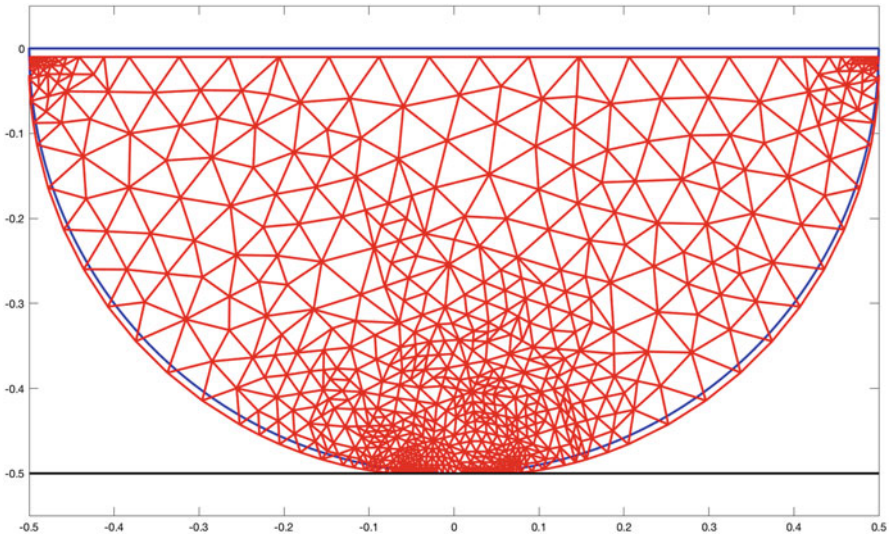

Fig. 1 Example 1: Reference and deformed configuration (compressible case)

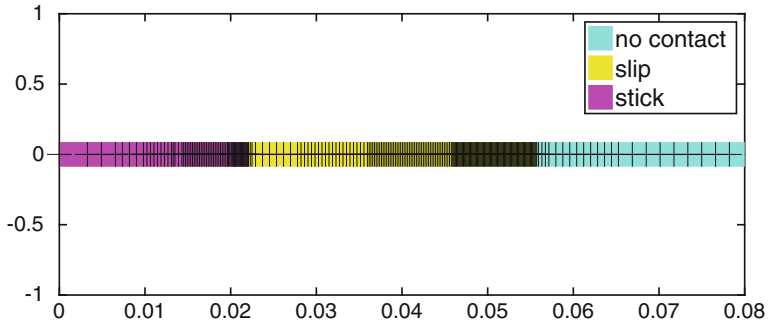

Fig. 2 Example 1: Right half of contact zone after 10 refinement steps (compressible case)

which can be observed in the incompressible case is that the shear stress changes its sign twice in the sticky part. We suspect this to be due to the material's resistance to volume change, causing it to want to move away from the central point of contact at $(0, -0.5)$ but being constrained by friction. Further away from the central point of contact, the body tends to slide inward and the sign of the constraining frictional shear stress changes. It then increases until sliding occurs. This behavior can best be captured with an adequately high resolution in the contact zone which is provided by the adaptive refinement strategy.

A comparison of the reduction of η for uniform and adaptive refinement is given in Fig. 4. Even though the convergence behavior for adaptive refinement is clearly better than for uniform refinement, the optimal convergence behavior achievable

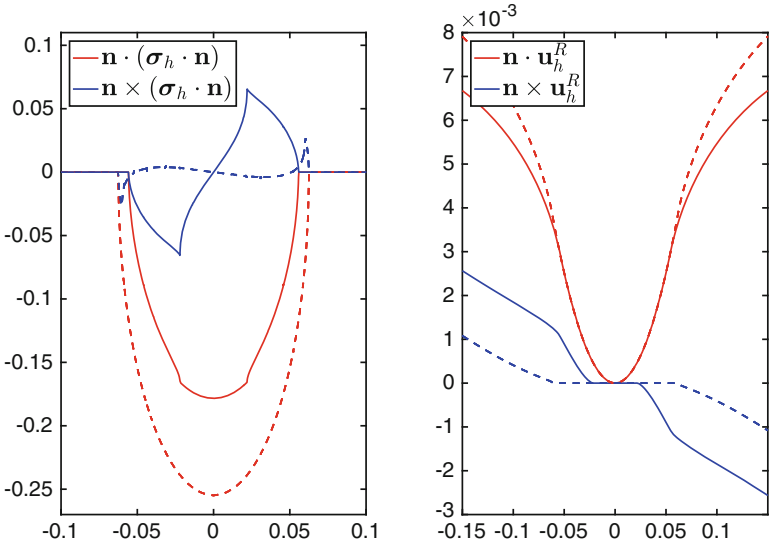

Fig. 3 Example 1: Stress and displacement in contact zone (dashed lines for incompressible case)

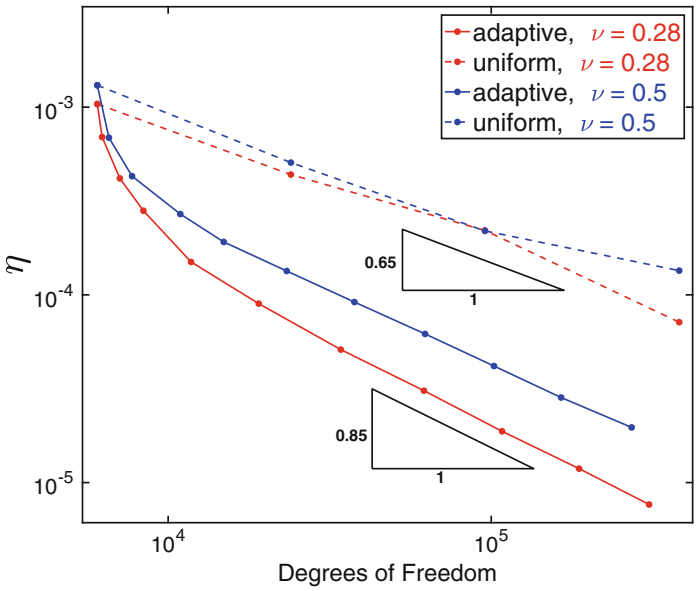

Fig. 4 Example 1: Adaptive vs. uniform refinement

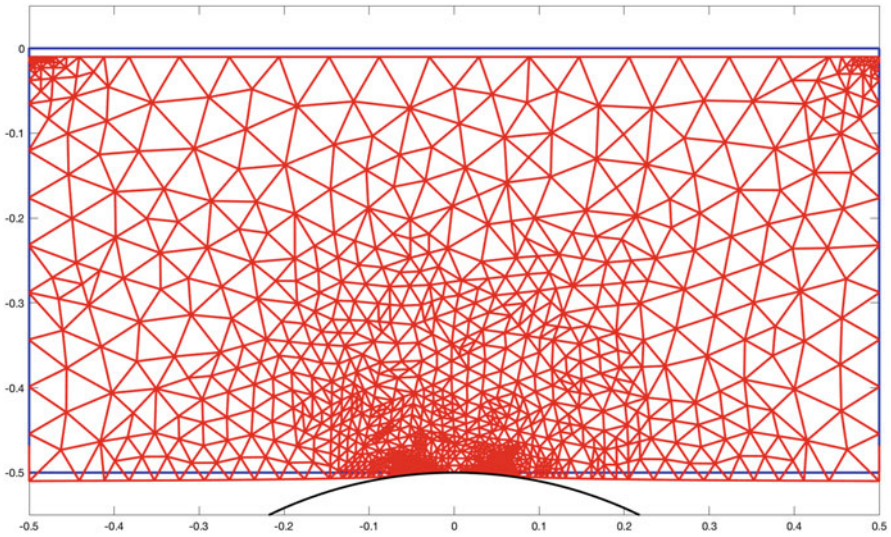

Fig. 5 Example 2: Reference and deformed configurations (incompressible case)

$\eta \sim N_h^{-1}$ (for N_h being the number of degrees of freedom) is almost, but not quite, reached. As the second example illustrates this is probably due to the curved boundary being resolved only by a piecewise linear curve. For optimal convergence rates, it would be necessary to use a piecewise quadratic approximation of the boundary and parametric Raviart–Thomas elements for the stress approximation (cf. [3]).

Example 2 (Hertzian Contact - Rectangle on Semicircle) Our next example is basically the situation from the first one flipped upside down. The body of interest is now a rectangle with length $2R$ and width R while the rigid foundation takes the shape of a semicircle with radius R. Material, friction, and Dörfler parameters as well as Dirichlet data are the same as in Example 1. Initial and deformed configurations of the incompressible case after 5 refinement steps are depicted in Fig. 5. Looking at the stresses and displacements in the contact zone we observe that, while the behavior in the compressible case is comparable to the results in Example 1, the incompressible case provides a notable qualitative difference: The sign of the shear stress changes only once but is completely reversed compared to the compressible case as can be seen in Fig. 6. Figure 7 presents again a comparison of the reduction of η for uniform and adaptive refinement, where now the optimal rate $\eta \sim N_h^{-1}$ is achieved, illustrating the efficiency of our error estimator. The detailed numerical results are summarized in Tables 3 and 4.

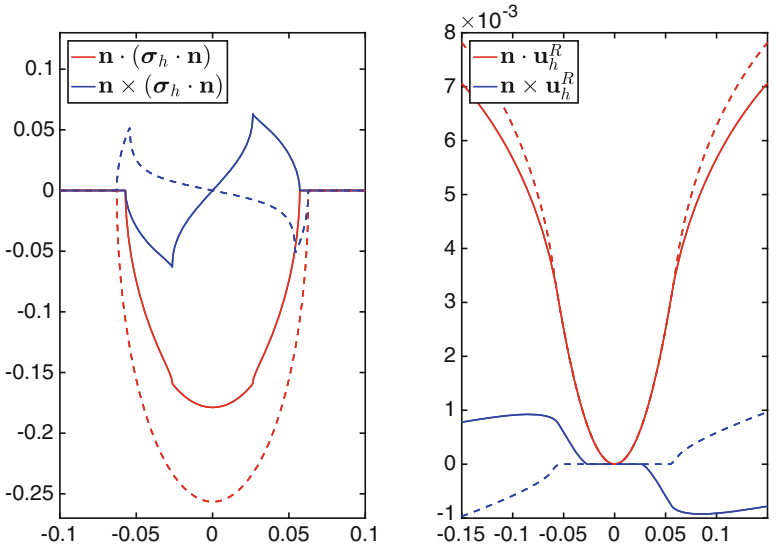

Fig. 6 Example 2: Stress and displacement in contact zone (dashed lines for incompressible case)

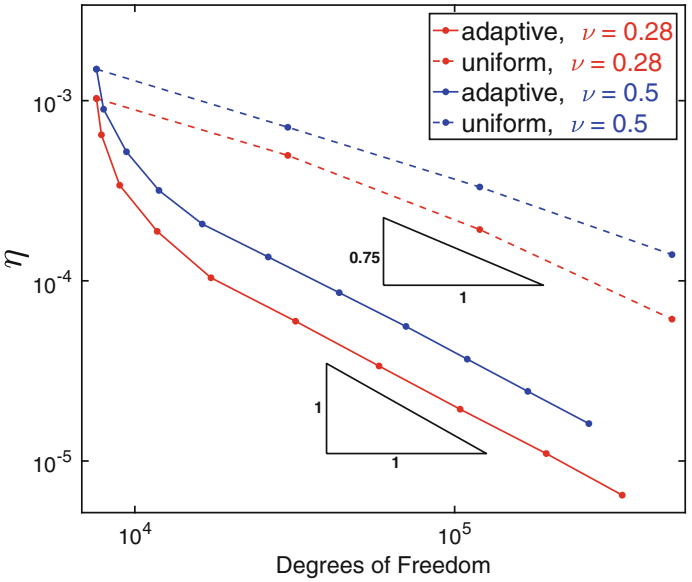

Fig. 7 Example 2: Adaptive vs. uniform refinement

Table 3 Results for Example 2 (compressible case)

l	$\dim\boldsymbol{\Sigma}_h$	$\dim\mathbf{Z}_h$	$\dim\mathbf{X}_h$	$\|\mathcal{A}\boldsymbol{\sigma}_h - \boldsymbol{\varepsilon}(\mathbf{u}_h^R)\|^2$	A_n	A_t
0	4552	2712	253	7.939e − 07	30	32
1	4716	2808	262	2.500e − 07	32	36
2	5376	3198	297	6.619e − 08	38	44
3	7040	4182	387	2.090e − 08	48	64
4	10, 388	6180	560	6.300e − 09	56	76
5	19, 124	11, 382	1012	2.096e − 09	96	126
6	34, 964	20, 862	1830	6.715e − 10	118	158
7	62, 760	37, 488	3245	2.261e − 10	179	246
8	116, 832	69, 876	5996	7.291e − 11	232	316
9	202, 036	120, 942	10, 302	2.517e − 11	292	408

Table 4 Results for Example 2 (incompressible case)

$\dim\boldsymbol{\Sigma}_h$	$\dim\mathbf{Z}_h$	$\dim\mathbf{X}_h$	$\|\mathcal{A}\boldsymbol{\sigma}_h - \boldsymbol{\varepsilon}(\mathbf{u}_h^R)\|^2$	$\|\mathrm{div}\,\mathbf{u}_h^R\|^2$	A_n	A_t
4552	2712	253	1.317e − 06	5.700e − 07	30	30
4788	2850	266	4.766e − 07	2.256e − 07	32	32
5644	3354	313	1.688e − 07	1.088e − 07	38	42
7132	4236	393	5.969e − 08	4.310e − 08	46	50
9752	5796	531	2.538e − 08	1.905e − 08	54	64
15, 700	9336	842	1.103e − 08	8.442e − 09	86	97
26, 228	15, 642	1389	4.556e − 09	3.569e − 09	98	112
42, 464	25, 350	2220	1.896e − 09	1.461e − 09	130	152
66, 088	39, 492	3425	8.340e − 10	6.464e − 10	162	190
102, 312	61, 200	5270	3.647e − 10	2.847e − 10	188	220
158, 872	95, 082	8133	1.578e − 10	1.221e − 10	248	294

Acknowledgments We thank the anonymous reviewer for helpful suggestions. In particular, we are grateful for pointing out a gap in an earlier version of the Proof of Theorem 4.1.

References

1. R. A. Adams and J. F. Fournier. *Sobolev Spaces*. Academic Press, New York, 2nd edition, 2003.
2. F. S. Attia, Z. Cai, and G. Starke. First-order system least squares for the Signorini contact problem in linear elasticity. *SIAM J. Numer. Anal.*, 47:3027–3043, 2009.
3. F. Bertrand and G. Starke. Parametric Raviart-Thomas elements for mixed methods on domains with curved surfaces. *SIAM J. Numer. Anal.*, 54:3648–3667, 2016.
4. D. Boffi, F. Brezzi, and M. Fortin. Reduced symmetry elements in linear elasticity. *Commun. Pure Appl. Anal.*, 8:95–121, 2009.
5. V. Bostan and W. Han. A posteriori error analysis for finite element solutions of a frictional contact problem. *Comput. Methods Appl. Mech. Engrg.*, 195:1252–1274, 2006.
6. M. Bürg and A. Schröder. A posteriori error control of *hp*-finite elements for variational inequalities of the first and second kind. *Comp. Maths. with Appls.*, 70:2783–2802, 2015.

7. Z. Cai and G. Starke. First-order system least squares for the stress-displacement formulation: Linear elasticity. *SIAM J. Numer. Anal.*, 41:715–730, 2003.
8. Z. Cai and G. Starke. Least squares methods for linear elasticity. *SIAM J. Numer. Anal.*, 42:826–842, 2004.
9. A. Capatina. *Variational Inequalities and Frictional Contact Problems*. Springer, Cham, 2014.
10. P. Hild. An example of nonuniqueness for the continuous static unilateral contact model with Coulomb friction. *C. R. Acad. Sci. Paris, Ser. I*, 337:685–688, 2003.
11. P. Hild and V. Lleras. Residual error estimators for Coulomb friction. *SIAM J. Numer. Anal.*, 47:3550–3583, 2009.
12. P. Hild and Y. Renard. An error estimate for the Signorini problem with Coulomb friction approximated by finite elements. *SIAM J. Numer. Anal.*, 45:2012–2031, 2007.
13. I. Hlaváček, J. Haslinger, J. Nečas, and J. Lovíšek. *Solution of Variational Inequalities in Mechanics*. Springer, New York, 1988.
14. N. Kikuchi and J. T. Oden. *Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods*. SIAM, Philadelphia, 1988.
15. R. Krause, B. Müller, and G. Starke. An adaptive least-squares mixed finite element method for the Signorini problem. *Numer. Methods Partial Differential Equations*, 33:276–289, 2017.
16. Y. Renard. A uniqueness criterion for the Signorini problem with Coulomb friction. *SIAM J. Math. Anal.*, 38:452–467, 2006.
17. R. Stenberg. Postprocessing schemes for some mixed finite elements. *Math. Model. Numer. Anal.*, 25:151–167, 1991.
18. M. Vohralík. Unified primal formulation-based a priori and a posteriori error analysis of mixed finite element methods. *Math. Comp.*, 79:2001–2032, 2010.

An Inexact Bundle Method and Subgradient Computations for Optimal Control of Deterministic and Stochastic Obstacle Problems

Lukas Hertlein, Anne-Therese Rauls, Michael Ulbrich, and Stefan Ulbrich

Abstract The aim of this work is to develop an inexact bundle method for nonsmooth nonconvex minimization in Hilbert spaces and to investigate its application to optimal control problems with deterministic or stochastic obstacle problems as constraints. A central requirement is that (approximate) subgradients can be obtained at given points. The second part of the paper thus studies in detail how subgradients can be obtained for optimal control problems governed by (stochastic) obstacle problems.

Keywords Optimal control · Nonsmooth optimization · Generalized derivatives · Bundle method · Obstacle problem · Stochastic obstacle problem · Variational inequalities

Mathematics Subject Classification (2020) Primary 26A24, 49J40, 49J52, 65K05; Secondary 49K20, 90C56

1 Introduction

We consider optimal control problems governed by obstacle problems of the form

$$\text{Find } y \in K_\psi, \quad \langle Ly - F(u), z - y\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \geq 0 \quad \text{for all } z \in K_\psi \tag{VI}$$

L. Hertlein · M. Ulbrich (✉)
Technische Universität München, München, Germany
e-mail: hertlein@ma.tum.de; mulbrich@ma.tum.de

A.-T. Rauls · S. Ulbrich
Technische Universität Darmstadt, Darmstadt, Germany
e-mail: rauls@mathematik.tu-darmstadt.de; ulbrich@mathematik.tu-darmstadt.de

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_19

as well as by stochastic versions of (VI), see ($\mathbf{VI_s}$) below. Here, $\Omega \subseteq \mathbb{R}^d$ is a bounded open domain and K_ψ denotes the closed convex set $K_\psi := \{z \in H_0^1(\Omega) : z \geq \psi \text{ q.e. on } \Omega\}$, where ψ is a given quasi upper-semicontinuous obstacle such that $K_\psi \neq \emptyset$. Additional regularity assumptions on ψ are stated if necessary. The abbreviation "q.e." stands for "quasi-everywhere" and will be used frequently in this paper. It describes that the respective property holds everywhere except on a subset of Ω which has capacity zero. For the notion of capacity and the corresponding definitions, we refer the reader to, e.g., [1, 2, 8, 21]. Furthermore, $L \in \mathcal{L}(H_0^1(\Omega), H^{-1}(\Omega))$ is a coercive and T-monotone operator. The operator $F : U \to H^{-1}(\Omega)$ is assumed to be Lipschitz continuous on bounded sets, continuously differentiable, and monotone, defined on a partially ordered Banach space U. The precise assumptions on the space U are given in Sect. 5, prototypes include $U = L^2(\Omega)$, $U = H^{-1}(\Omega)$, or $U = \mathbb{R}^n$. Due to the operator F and the assumptions on U, the variational inequality (VI) represents a general class of obstacle problems. It is well known that for each $u \in U$ the variational inequality (VI) has a unique solution. We denote the solution operator by $S_F : U \to H_0^1(\Omega)$. If F is the identity map on $H^{-1}(\Omega)$, we omit the subscript and write $S : H^{-1}(\Omega) \to H_0^1(\Omega)$.

We consider the following optimal control problem governed by (VI):

$$\min_{u \in U_{\text{ad}}} J(S_\iota(u)) + \tfrac{\alpha}{2}\|u\|_U^2, \tag{P}$$

where $U_{\text{ad}} \subset U$ is a closed convex subset of the Hilbert space U, $J : H_0^1(\Omega) \to \mathbb{R}$ is the objective function, and $\iota \in \mathcal{L}(U, H^{-1}(\Omega))$ is a compact and injective operator. Here, S_ι denotes the solution operator of (VI) when choosing $F = \iota$.

The nonconvexity and nondifferentiability of the solution operator S_ι require the application of nonsmooth optimization methods. An alternative is to view the VI as a constraint, which results in a mathematical program with equilibrium constraints (MPEC). Most methods for MPECs use regularization or smoothing; we refer to [39] for a survey of numerical methods for the optimal control of elliptic variational inequalities.

Here, we propose to use a variant of the bundle method developed in [22] which is tailored for this use. This method is posed in an appropriate function space setting and can handle inexact function values, inexact subgradients, and inexact solutions of the bundle subproblem. We extend the method of [22], allowing for quite general sets of approximate subgradients. Furthermore, we provide a global convergence result for general locally Lipschitz functions, provided there exists a subsequence of iterations in which the new model is sufficiently much improved over the old model (cf. Theorem 2.6). To ensure this, one usually requires approximate convexity [22, 31] or semismoothness [28]. Our generalization is motivated by the fact that, in general, there can exist points where these properties do not hold. Already in finite dimensions, there is not much literature on bundle methods for nonconvex optimization with inexact function values and subgradients [19, 27, 31]. Our work in this paper and in [22] is inspired by Noll [31] and seems to be the only inexact bundle method for infinite dimensional nonconvex problems.

The bundle method requires an approximate subgradient of the reduced objective function in each iteration. Therefore, we derive a formula for an element of a generalized differential for the solution operator of the obstacle problem in each point in U, from which a Clarke subgradient for the reduced objective function can be extracted. The generalized derivative that we construct for the solution operator S_F of (VI) in an arbitrary $u \in U$ is the operator $\Sigma_F(u;\cdot) \in \mathcal{L}(U, H_0^1(\Omega))$, where $\Sigma_F(u;h) = \eta$ solves

$$\text{Find } \eta \in H_0^1(I(u)), \quad \langle L\eta - F'(u;h), z\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} = 0 \quad \text{for all } z \in H_0^1(I(u)). \tag{1.1}$$

Here, $I(u) := \{\omega \in \Omega : S_F(u)(\omega) > \psi(\omega)\}$ is the inactive set, which is a quasi-open subset of Ω and $H_0^1(I(u)) = \{v \in H_0^1(\Omega) : v = 0 \text{ q.e. outside } I(u)\}$ is a closed subspace of $H_0^1(\Omega)$. The variational equation (1.1) is also a characterization of the directional derivative in points where S_F is Gâteaux differentiable. In such points, (1.1) is obtained from the variational inequality for the directional derivative of S_F established by Mignot [29]. Since the generalized differentials we consider for S_F contain limits of Gâteaux derivatives w.r.t. certain topologies in points $(u_n)_{n\in\mathbb{N}}$ converging to u, we derive the generalized derivative by pursuing a convergence analysis for such problems considering appropriate sequences $(u_n)_{n\in\mathbb{N}}$. We are not aware of any work that establishes generalized derivatives for the solution operator of (VI) in infinite dimensions apart from our presentation in this paper and in [35]. For the case that F is the identity mapping on $H^{-1}(\Omega)$, a characterization of the entire generalized differential is possible. We review the results of [36] for this problem.

We also are interested in the optimal control of the stochastic obstacle problem. Let $(\Xi, \mathcal{A}, P)$ be a probability space and denote by $\mathbf{Y} := L^2(\Xi, H_0^1(\Omega))$ the Bochner space of square integrable functions with values in $H_0^1(\Omega)$ (cf. [24, Def. 1.2.15]). The stochastic obstacle problem ($\mathbf{VI_s}$) is given by the variational inequality

$$\text{Find } \mathbf{y} \in \mathbf{K}_{\boldsymbol{\psi}}, \quad \langle \mathbf{L}\mathbf{y} - \mathbf{b}, \mathbf{z} - \mathbf{y}\rangle_{\mathbf{Y}^*,\mathbf{Y}} \geq 0 \quad \text{for all } \mathbf{z} \in \mathbf{K}_{\boldsymbol{\psi}}, \tag{$\mathbf{VI_s}$}$$

where $\mathbf{L} \in \mathcal{L}(\mathbf{Y}, \mathbf{Y}^*)$, $\mathbf{b} \in \mathbf{Y}^*$, $\boldsymbol{\psi} \in \bar{\mathbf{Y}} := L^2(\Xi, H^1(\Omega))$ and

$$\mathbf{K}_{\boldsymbol{\psi}} := \{\mathbf{y} \in \mathbf{Y} : \mathbf{y}(\xi) \in K_{\psi_\xi} \text{ for } P\text{-almost all } (P\text{-a.a.}) \ \xi \in \Xi\}. \tag{1.2}$$

For the rest of this paper, bold notation refers to the variables in the stochastic setting, whereas non-bold variables refer to the deterministic setting. Under suitable assumptions on the data, cf. Sect. 7, the Lions–Stampacchia theorem [26, Thm. 2.1] implies that the stochastic obstacle problem admits a unique solution and the solution operator $\mathbf{S} : \mathbf{Y}^* \to \mathbf{Y}$ is Lipschitz continuous (cf. Theorem 7.3). For P-a.e. $\xi \in \Xi$, this defines the operators $S_\xi : Z = H^{-1}(\Omega) \to H_0^1(\Omega)$ via $S_\xi(z) := \mathbf{S}(\hat{\boldsymbol{\iota}}z)(\xi)$, where $(\hat{\boldsymbol{\iota}}z)(\xi) := z$, $z \in H^{-1}(\Omega)$. We study the following class of optimal control problems for the stochastic obstacle problem:

$$\min_{u \in U_{\text{ad}}} \mathbb{E}\left[J_\xi(S_\xi(\iota u))\right] + \frac{\alpha}{2}\|u\|_U^2, \tag{$\mathbf{P}_\text{s}$}$$

where $J_\xi : H_0^1(\Omega) \to \mathbb{R}$ is the parametric objective function such that $\xi \mapsto J_\xi(S_\xi(z))$ is integrable for all $z \in H^{-1}(\Omega)$, $\mathbb{E}\left[J_\xi(S_\xi(\cdot))\right]$ is locally Lipschitz and $\mathbb{E}$ denotes the expectation with respect to ξ. The goal is to find a Clarke-stationary point, i.e. a point $\bar{u} \in U$ which satisfies

$$0 \in \partial_C(\mathbb{E}\left[J_\xi(S_\xi(\iota(\cdot)))\right])(\bar{u}) + \alpha\bar{u} + N_{U_{\text{ad}}}(\bar{u}), \tag{1.3}$$

where ∂_C denotes Clarke's subdifferential. The chain rule [11, Thm. 2.3.10] implies that

$$\partial_C(\mathbb{E}\left[J_\xi(S_\xi(\iota(\cdot)))\right])(u) \subset \iota^* \partial_C(\mathbb{E}\left[J_\xi(S_\xi(\cdot))\right])(\iota u) \quad \text{for } P\text{-a.e. } \xi \in \Xi \text{ and all } u \in U.$$

If $J_\xi \circ S_\xi$ or $-J_\xi \circ S_\xi$ is regular at ιu in the sense of Clarke (cf. [11, Def. 2.3.4]), then equality holds at this point. Under suitable assumptions, [11, Thm. 2.7.2] implies

$$\partial_C(\mathbb{E}\left[J_\xi(S_\xi(\cdot))\right])(z) \subset \mathbb{E}\left[\partial_C(J_\xi(S_\xi(\cdot)))(z)\right] \qquad \text{for all } z \in Z$$

with equality if $J_\xi \circ S_\xi$ or $-J_\xi \circ S_\xi$ is regular at z for each $\xi \in \Xi$. Here, the set $\mathbb{E}\left[\partial_C(J_\xi(S_\xi(\cdot)))(z)\right] \subset Z^*$ is defined as

$$\{\mathbb{E}\,[g(\xi)] : g \in L^1(\Xi, Z^*) \text{ is a measurable selection of } \partial_C(J_\xi(S_\xi(\cdot)))(z)\}. \tag{1.4}$$

This formula allows to reuse the subgradients (1.1) of the deterministic problem. However, the reduced objective function might not be regular at all admissible points. In this case, the available calculus rules for the Clarke subdifferential, which often take the form of inclusions, make it difficult to calculate the subdifferential $\partial_C(\mathbb{E}\left[J_\xi(S_\xi(\iota(\cdot)))\right])(\bar{u})$. Thus, we search for weak stationary points (cf. [43]), i.e. points $\bar{u} \in U$ which fulfill

$$0 \in \iota^* \mathbb{E}\left[\partial_C(J_\xi(S_\xi(\cdot)))(\iota\bar{u})\right] + \alpha\bar{u} + N_{U_{\text{ad}}}(\bar{u}). \tag{1.5}$$

However, under additional assumptions on the regularity of the data, in Sect. 7.4, we give a formula for exact subgradients $g \in \partial_C(\mathbb{E}\left[J_\xi(S_\xi(\iota(\cdot)))\right])(\bar{u})$.

The rest of the paper is organized as follows: In Sect. 2 we present a variant of the bundle method of [22] to solve both problems (P) and ($\mathbf{P}_\text{s}$). In Sect. 3, we introduce sets of generalized derivatives that will be used in this article for operators between infinite dimensional spaces. Section 4 deals with the obstacle problem (VI) and its properties, in particular, properties concerning monotonicity and differentiability. We derive a formula for a generalized derivative for the solution operator of the obstacle problem in Sect. 5. In Sect. 6, characterizations of the entire generalized differentials are established for an easier instance of the obstacle problem. In Sect. 7

we discuss the stochastic obstacle problem and derive both formulas for exact subgradients $g \in \partial_C(\mathbb{E}\left[J_\xi(S_\xi(\iota(\cdot)))\right])(\bar{u})$ as well as conditions under which the weak subgradients $g \in \iota^*\mathbb{E}\left[\partial_C(J_\xi(S_\xi(\cdot)))(\iota\bar{u})\right]$ can be used in the bundle method.

2 Inexact Bundle Method

Since the optimal control problems (P) and ($\mathbf{P}_s$) for the deterministic and stochastic obstacle problem are nonsmooth, nonconvex optimization problems in Hilbert spaces, we employ a tailored bundle method to solve them. We adopt the approach of [22], which itself draws from ideas in [31], to the given setting. In particular, we allow for more general choices of approximate subgradients, and we outline a convergence theory for functions which are not approximately convex. Our problem setting is as follows:

$$\min_{u\in U} f(u) + w(u) \quad \text{s.t.} \quad u \in \mathcal{F},$$

where $f(u)$ corresponds to the cost term involving the state and $w(u)$ to the regularization term. The feasible set $\mathcal{F} \subset U$ is nonempty, closed, convex, and U is a Hilbert space. The function $f : \mathcal{F}_U \to \mathbb{R}$, $\mathcal{F}_U \supset \mathcal{F}$ convex and open in U, has the form $f = p \circ \iota$. Here $\iota \in \mathcal{L}(U, Z)$ is a compact and injective operator into the Hilbert space Z and $p : \mathcal{F}_Z \to \mathbb{R}$ is Lipschitz on bounded sets with $\mathcal{F}_Z \subset Z$ convex and open, $\iota(\mathcal{F}_U) \subset \mathcal{F}_Z$. Further, let $w : \mathcal{F}_U \to \mathbb{R}$ be continuously differentiable, Lipschitz on bounded sets, and μ-strongly convex, $\mu > 0$, i.e., for all $u \in U$ with $w(u) < \infty$ there holds

$$w(u + s) - w(u) \geq \langle w'(u), s\rangle_{U^*,U} + \frac{\mu}{2}\|s\|_U^2 \quad \text{for all } s \in U.$$

Note that this implies that w is also weakly sequentially lower semicontinuous.

This setting is applicable to a quite comprehensive class of optimal control problems, in particular, it includes both optimal control problems (P) and ($\mathbf{P}_s$) by setting $w := \frac{\alpha}{2}\|\cdot\|_U$, $\mathcal{F} := U_{\text{ad}}$, $Z := H^{-1}(\Omega)$, $p := J(S(\cdot))$, $f := J(S_\iota(\cdot))$, or for the stochastic problem $p := \mathbb{E}\left[J_\xi(S_\xi(\cdot))\right]$, $f := \mathbb{E}\left[J_\xi(S_\xi(\iota(\cdot)))\right]$.

To find stationary points, bundle methods use subgradient information to build a local model of the nonsmooth part p around the current iterate ιu. Usually, a subgradient g at a point $\iota u \in Z$ is an element of a subdifferential $G(\iota u) \subset Z^*$ such as Clarke's subdifferential $G(\iota u) = \partial_C p(\iota u)$ or the convex subdifferential $G(\iota u) = \partial p(\iota u)$ if p is convex. However, in certain situations it might not be possible to calculate such an element. Therefore, we pose minimal requirements that a multifunction $G : Z \rightrightarrows Z^*$ has to fulfill for being a suitable approximate subdifferential (cf. Assumption 2.1). The multifunction G then also appears in the stationarity condition of our convergence result.

Bundle methods often require a certain regularity of the objective function f beyond Lipschitz continuity. For example, in [28] the objective function is assumed to be semismooth while in [22, 31] approximate convexity [13] is required (being related to the lower C^1 property). These assumptions are needed to ensure that *all possible subgradients* in a neighborhood of the serious iterate improve the quality of the local model. However, for convergence of the algorithm, it is sufficient that the *computed subgradients* improve the local model. We thus introduce a measure for the quality of the local model, cf. (2.13). Depending on this measure, we prove convergence to approximate stationary points in Theorem 2.6. This concept also justifies why the bundle method often returns good results, even when applied to problems which do not satisfy regularity beyond Lipschitz continuity. To increase the flexibility of the algorithm, we also allow for subgradients to be drawn at arbitrary points in a neighborhood of the trial iterate. This implies that there always exists a model with sufficiently high quality to guarantee convergence to stationary points (cf. Remark 2.7).

The general procedure of the bundle method is as follows: In outer iteration j and inner iteration k, a finite set $\mathcal{M}_{j,k}$ of affine linear functions, called cutting planes, is selected. The convex function $\phi_{j,k} := \max\{m(\cdot) : m \in \mathcal{M}_{j,k}\}$ is chosen as the local model of f at the serious iterate u_j. The bundle method subproblem is given by

$$\min_{y \in \mathcal{F}} \phi_{j,k}(y) + w(y) + \tfrac{1}{2}\langle (Q_j + \tau_{j,k} E)\iota(y - u_j), \iota(y - u_j)\rangle_{Z^*,Z}.$$

Here, $\tau_{j,k} > 0$ is the proximity parameter, $Q_j \in \mathcal{L}(Z, Z^*)$ may represent curvature information of p at u_j, and $E \in \mathcal{L}(Z, Z^*)$ denotes the Riesz map. $\tau_{j,k}$ and Q_j are chosen such that the third term in the cost function of the bundle subproblem is strictly convex w.r.t. y, cf. Sect. 2.3. The unique minimizer of the subproblem $y_{j,k}$ is called *inner iterate*. Often it is difficult or impossible to calculate an exact solution of the bundle method subproblem. Therefore, we introduce the *trial iterate* $\tilde{y}_{j,k}$ as an approximation of $y_{j,k}$. If this trial iterate $\tilde{y}_{j,k}$ fulfills a certain decrease condition, it is accepted as the new serious iterate u_{j+1} and the inner loop is terminated. Otherwise, a new cutting plane is selected which enriches the old model. If the new model is not sufficiently improved, the proximity parameter is increased to gather more cutting plane information close to the serious iterate u_j. Then the next inner iteration is started.

When the j index is clear from the context, we often drop this index and refer to the quantities introduced above by u, ϕ_k, $\mathcal{M}_k$, Q, τ_k, y_k, and $\tilde{y}_k$, respectively.

2.1 Trial Iterates, Function Values, and Subgradients

Typically, bundle methods use function values and subgradients (or approximations thereof) to build a model of the objective function. In this paper, we work with a general concept of approximate function values and subgradients. Given a point

$y_k \in \mathcal{F}$, we need to find a point $\tilde{y}_k$, called *trial iterate*, in a neighborhood of y_k at which we can compute a function value approximation and an approximate subgradient. The trial iterate $\tilde{y}_k$ has to satisfy

$$\tilde{y}_k \in \bar{B}_U(y_k, R) \cap \mathcal{F} \quad \text{and} \quad \|\iota(\tilde{y}_k - y_k)\|_Z \leq \min\{M\|\iota(y_k - u)\|_Z, a_k\}, \tag{2.1}$$

where $R, M \geq 0$ are fixed constants and $(a_k)_{k\in\mathbb{N}} \subset \mathbb{R}$ is a forcing sequence such that $a_k \to 0$ as $k \to \infty$. Furthermore, $\tilde{y}_k$ needs to achieve at least a fraction $0 < \theta < 1$ of the model reduction provided by y_k:

$$\Phi_k(u) - \Phi_k(\tilde{y}_k) \geq \theta\,(\Phi_k(u) - \Phi_k(y_k))\,. \tag{2.2}$$

A *function value approximation* $f_{\tilde{y}_k} \in \mathbb{R}$ of $f(\tilde{y}_k)$ is assumed to fulfill

$$|f_{\tilde{y}_k} - f(\tilde{y}_k)| \leq \Delta, \quad \text{where } \Delta > 0 \text{ is a constant.} \tag{2.3}$$

Similar to the trial iterate we introduce the point $v_k \in \bar{B}_U(y_k, \hat{R}) \cap \mathcal{F}$, $\hat{R} \geq 0$, called *subgradient base point*, at which approximate subgradients are drawn. The enlargement of the set of points at which subgradients can be obtained might be helpful to find new subgradients which improve the local model. However, although it is possible to draw subgradients at points v_k, which can be far away from the trial iterate $\tilde{y}_k$, these subgradients might not be useful. See also Remark 2.7 for a discussion on this topic. Denote by $\mathcal{V} \subset U$ the set of all subgradient base points. We define an *approximate subgradient* of the function p at the point $v \in \hat{\mathcal{V}} := \operatorname{cl}\iota(\mathcal{V})$ as an element $\tilde{g} \in G(v)$, where G fulfills:

Assumption 2.1 The multifunction $G : \hat{\mathcal{V}} \rightrightarrows Z^*$ has the following properties:

1. For all $v \in \hat{\mathcal{V}}$, the image $G(v)$ is nonempty and convex.
2. For all bounded sets $B \subset Z$, the set $G(B \cap \hat{\mathcal{V}}) := \cup_{v\in B\cap\hat{\mathcal{V}}} G(v)$ is bounded in Z^*.
3. G has a weakly closed graph, i.e., for all sequences $(v_n)_{n\in\mathbb{N}} \subset \hat{\mathcal{V}}$ and $(g_n)_{n\in\mathbb{N}} \subset Z^*$ such that $v_n \to \bar{v}$ in Z, $g_n \rightharpoonup g$ in Z^* and $g_n \in G(v_n)$ $\forall n \in \mathbb{N}$, it holds $g \in G(\bar{v})$.

These are exactly the requirements on the subgradients needed to prove convergence of the bundle algorithm. In Sect. 7.2, we show that (1.4) fulfills this assumption.

Remark 2.2 For a function $p : Z \to \mathbb{R}$ that is Lipschitz on bounded sets, Clarke's differential $\partial_C p : Z \rightrightarrows Z^*$ satisfies Assumption 2.1. This follows from [11, Prop. 2.1.2 and Prop. 2.1.5]. In [22], $G = \partial_C p + C$ is used, where $C \subset Z^*$ is a closed convex set with $0 \in C$.

2.2 The Cutting Plane Model

For $u, \tilde{y}, v \in U$ and $\tilde{g} \in Z^*$ define the *downshift* $s_{\tilde{y},v,\tilde{g},u} \in \mathbb{R}$, the *tangent* $t_{\tilde{y},\tilde{g},u} : U \to \mathbb{R}$, and the *downshifted tangent* $m_{\tilde{y},v,\tilde{g}}(\cdot, u) : U \to \mathbb{R}$ by

$$\begin{aligned} s_{\tilde{y},v,\tilde{g},u} &:= [f_{\tilde{y}} + \langle \tilde{g}, \iota(u - \tilde{y})\rangle_{Z^*,Z} - f_u]_+ + c\|\iota(v - u)\|_Z^2, \\ t_{\tilde{y},\tilde{g},u}(\cdot) &:= f_{\tilde{y}} + \langle \tilde{g}, \iota(\cdot - \tilde{y})\rangle_{Z^*,Z}, \quad m_{\tilde{y},v,\tilde{g}}(\cdot, u) := t_{\tilde{y},\tilde{g},u}(\cdot) - s_{\tilde{y},v,\tilde{g},u}. \end{aligned} \tag{2.4}$$

Here, the downshift parameter $c > 0$ is fixed. At the serious iterate u we compute the *exactness subgradient* $\tilde{g}_0 \in G(\iota(u))$ and define the *exactness plane* $m_0(\cdot, u) : U \to \mathbb{R}$ by

$$m_0(\cdot, u) := m_{u,u,\tilde{g}_0}(\cdot, u) = f_u + \langle \tilde{g}_0, \iota(\cdot - u)\rangle_{Z^*,Z}.$$

Let $\mathcal{B}_k$ denote the set of all *bundle information of previous iterations* (including all information in previous outer iterations), i.e. all triples of the form $(\tilde{y}_k, v_k, \tilde{g}_k)$ where $\tilde{y}_k$, v_k, and $\tilde{g}_k$ are the trial iterate, the base point, and the subgradient of iteration k. Let $\mathcal{D}_k$ denote the set of *previous downshifted tangents*:

$$\mathcal{D}_k := \{m_{\tilde{y},v,\tilde{g}}(\cdot, u) : (\tilde{y}, v, \tilde{g}) \in \mathcal{B}_k\}. \tag{2.5}$$

We choose a finite subset $\mathcal{M}_k$ of $\operatorname{co} \mathcal{D}_k$ (co = convex hull) to build the *cutting plane model* $\phi_k : U \to \mathbb{R}$ by

$$\phi_k(y) := \max\{m(y) : m \in \mathcal{M}_k\}.$$

Choosing $\mathcal{M}_{k+1} = \{m_\nu(\cdot, u), \nu = 0, \ldots, k\}$ yields the full model

$$\phi_{k+1}^{\text{full}} := \max\{m_\nu(\cdot, u),\ \nu = 0, \ldots, k\}.$$

However, large k might lead to an expensive cutting plane model. Therefore, we allow $\mathcal{M}_k \subset \operatorname{co} \mathcal{D}_k$ to be chosen according to Assumption 2.3 below.

2.3 Proximity Control

If there is curvature information of $p : Z \to \mathbb{R}$ around $\iota(u)$ available, we want to incorporate this into the model. Fix the constants $0 < \underline{q} < \bar{q}$ and denote by $E \in \mathcal{L}(Z, Z^*)$, $Ev = (v, \cdot)_Z$, the Riesz map. We assume that $Q \in \mathcal{L}(Z, Z^*)$ and $q \in (\underline{q}, \bar{q})$ are chosen such that

$$\langle (Q + qE)v, v\rangle_{Z^*,Z} \geq \underline{q}\|v\|_Z^2 \quad \text{for all } v \in Z \qquad \text{and} \qquad \|Q\|_{\mathcal{L}(Z,Z^*)} \leq \bar{q}, \tag{2.6}$$

and that Q is symmetric, i.e., $\langle Qx, y\rangle_{Z^*,Z} = \langle Qy, x\rangle_{Z^*,Z}$ for all $x, y \in Z$. For any *proximity parameter* $\tau \geq q$, the positive definite symmetric bilinear form $\langle (Q + \tau E)\cdot, \cdot\rangle_{Z^*,Z}$ defines a norm on Z via $\|\cdot\|_{Q+\tau E}^2 := \langle (Q + \tau E)\cdot, \cdot\rangle_{Z^*,Z}$.

2.4 The Subproblem of the Bundle Method

The *subproblem of the bundle method* is given by

$$\min_{y\in\mathcal{F}} \Psi_k(y) := \phi_k(y) + w(y) + \tfrac{1}{2}\|\iota(y - u)\|_{Q+\tau_k E}^2. \tag{2.7}$$

Since Ψ_k is strongly convex on $\mathcal{F}$, this problem has a unique minimum $y_k \in \mathcal{F}$ which is called *inner iterate*. Denote the indicator function of $\mathcal{F}$ by $\delta_{\mathcal{F}} : U \to \mathbb{R} \cup \{\infty\}$. We define the *local model* $\Phi_k : U \to \mathbb{R} \cup \{\infty\}$ via $\Phi_k := \phi_k + w + \delta_{\mathcal{F}}$. The sum rule of convex analysis [5, Cor. 16.50] can be applied and yields $\partial\Phi_k = \partial\phi_k + w' + N_{\mathcal{F}}$, where $N_{\mathcal{F}} = \partial\delta_{\mathcal{F}}$ is the normal cone of $\mathcal{F}$ and ∂ denotes the convex subdifferential. The fact that y_k minimizes the subproblem of the bundle method can equivalently be expressed as

$$\begin{aligned} 0 &\in \partial(\Phi_k + \tfrac{1}{2}\|\iota(\cdot - u)\|_{Q+\tau_k E}^2)(y_k) \\ &= \partial\phi_k(y_k) + w'(y_k) + N_{\mathcal{F}}(y_k) + \iota^*(Q + \tau_k E)\iota(y_k - u). \end{aligned}$$

Therefore there exist elements $g_k^* \in \partial\phi_k(y_k)$ and $n_k \in N_{\mathcal{F}}(y_k)$ such that

$$e_k := \iota^*(Q + \tau_k E)\iota(u - y_k) = g_k^* + w'(y_k) + n_k \in \partial\Phi_k(y_k). \tag{2.8}$$

For $m \in \mathcal{M}_k$ denote by $g_m := m'(0) \in U^*$ the derivative of the affine linear function $m : U \to \mathbb{R}$. As the set $\mathcal{M}_k$ is finite, by Clarke [11, Prop. 2.3.12] it holds for all $y \in U$ that

$$\partial\phi_k(y) = \operatorname{co}\{g_m : m \in \mathcal{M}_k,\ m(y) = \phi_k(y)\}. \tag{2.9}$$

Since $g_k^* \in \partial\phi_k(y_k)$, there exist numbers $\lambda_m \geq 0$ with $\sum_{m\in\mathcal{M}_k} \lambda_m = 1$ and $g_k^* = \sum_{m\in\mathcal{M}_k} \lambda_m g_m$. We define the *aggregate cutting plane* $m_k^* \in \operatorname{co}\mathcal{D}_k$ by

$$m_k^*(\cdot, u) := \sum_{m\in\mathcal{M}_k} \lambda_m m(\cdot). \tag{2.10}$$

For the convergence analysis we only require the following properties of the models ϕ_k:

Assumption 2.3 For each $k \geq 0$ the set $\mathcal{M}_{k+1} \subset \operatorname{co} \mathcal{D}_{k+1}$ is chosen such that

$$\text{a) } m_0(\cdot, u) \leq \phi_{k+1}(\cdot), \quad \text{b) } m_k^*(\cdot, u) \leq \phi_{k+1}(\cdot), \tag{2.11}$$

where we set $m_0^*(\cdot, u) := m_0(\cdot, u)$.

In Algorithm 1, the inexact bundle method is presented.

Algorithm 1: Inexact bundle method

Parameters : $0 < \gamma < \tilde{\gamma} < 1$, $0 < \theta < 1$, $\Delta > 0$, $R > 0$, $M > 0$, $0 < \underline{q} < \bar{q} \leq T$. Forcing sequence $(a_k)_{k \in \mathbb{N}}$, gradient approximation multifunction $G : \mathcal{F} \rightrightarrows Z^*$ fulfilling Assumption 2.1.

Initialization: Choose an initial iterate $u_1 \in \mathcal{F}$ and function value approximation $f_{u_j} \in \bar{B}(f(u_j), \Delta)$.

for $j = 1, \ldots$ **do**

  Compute $\tilde{g}_0 \in G(\iota(u_j))$ and set $J_{u_j} = f_{u_j} + w(u_j)$. Choose a symmetric operator $Q_j \in \mathcal{L}(Z, Z^*)$ and $q_j \in (\underline{q}, \bar{q})$ satisfying (2.6). Choose $\tau_1 \in [q_j, T]$. Set $m_0(\cdot, u_j) = f_{u_j} + \langle \tilde{g}_{j,0}, \iota(\cdot - u_j) \rangle_{Z^*, Z}$ and $\Phi_1 = m_0(\cdot, u_j) + w$.

  for $k = 1, \ldots$ **do**

    Trial iterate generation. Define the inner iterate y_k by

$$y_k := \arg\min_{y \in \mathcal{F}} \phi_k(y) + w(y) + \tfrac{1}{2} \|\iota(y - u_j)\|^2_{Q_j + \tau_k E}.$$

    Find a trial iterate $\tilde{y}_k \in \bar{B}_U(y_k, R) \cap \mathcal{F}$ which fulfills (2.1) and (2.2). Compute $f_{\tilde{y}_k} \in \bar{B}(f(\tilde{y}_k), \Delta)$ and set $J_{\tilde{y}_k} = f_{\tilde{y}_k} + w(\tilde{y}_k)$.

    Stop if $\tilde{y}_k = u_j$.

    Acceptance test. Set

$$\rho_k = \frac{J_{u_j} - J_{\tilde{y}_k}}{J_{u_j} - \Phi_k(\tilde{y}_k)}.$$

    if $\rho_k \geq \gamma$ **then**

      Set $u_{j+1} = \tilde{y}_k$, $f_{u_{j+1}} = f_{\tilde{y}_k}$ and quit the inner loop.

    end

    Update local model. Enrich the set of bundle information $\mathcal{B}_{k+1}$ by computing a function value approximation $f_{\tilde{y}_k}$ at the trial iterate and a subgradient $\tilde{g}_k \in G(v_k)$ at the base point v_k. Possibly add more bundle information to $\mathcal{B}_{k+1}$. Possibly delete or aggregate old cutting planes such that the new cutting planes $\mathcal{M}_{k+1}$ fulfill (2.11). Set $\Phi_{k+1} = \max\{m : m \in \mathcal{M}_{k+1}\} + w$.

    Update proximity parameter. Set $\tilde{\rho}_k = \frac{J_{u_j} - \Phi_{k+1}(\tilde{y}_k)}{J_{u_j} - \Phi_k(\tilde{y}_k)}$ and update

$$\tau_{k+1} = \begin{cases} 2\tau_k & \text{if } \tilde{\rho}_k \geq \tilde{\gamma} \\ \tau_k & \text{if } \tilde{\rho}_k < \tilde{\gamma} \end{cases}.$$

  end

end

2.5 Global Convergence Result

Definition 2.4 A point $\bar{u} \in U$ is called ϵ-stationary, $\epsilon \geq 0$, if $0 \in w'(\bar{u}) + N_{\mathcal{F}}(\bar{u}) + \iota^*(G(\iota(\bar{u})) + \bar{B}_{Z^*}(0, \epsilon))$. A point which is 0-stationary is called stationary.

Remark 2.5 If $G = \partial_C p$ is the Clarke subdifferential and p or $-p$ is regular at $\iota\bar{u}$, the chain rule [11, Thm. 2.3.10] implies that $\partial_C f(\bar{u}) = \iota^* \partial_C p(\iota\bar{u})$ and stationarity is equivalent to $0 \in \partial_C f(\bar{u}) + w'(\bar{u}) + N_{\mathcal{F}}(\bar{u})$.

For each "bad" iteration (j, k), i.e. (we return to double indexing)

$$\rho_{j,k} < \gamma \quad \text{and} \quad \tilde{\rho}_{j,k} \geq \tilde{\gamma}, \tag{2.12}$$

define the accuracy measure

$$\epsilon_{j,k} := \frac{J_{\tilde{y}_{j,k}} - \Phi_{j,k+1}(\tilde{y}_{j,k})}{\|\iota(\tilde{y}_{j,k} - u_j)\|_Z}. \tag{2.13}$$

Theorem 2.6 (Convergence of the Bundle Method) *Let the initial point $u_1 \in \mathcal{F}$ be such that $\mathcal{F}_1 := \{x \in \mathcal{F} : J(x) \leq J(u_1) + 2\Delta\}$ is bounded in U and define $\epsilon_{j,k}$ as in (2.13).*

1. *If Algorithm 1 produces only finitely many serious iterates and the sequence of proximity parameters $(\tau_k)_k$ is bounded, then the last serious iterate u is stationary.*
2. *If Algorithm 1 produces only finitely many serious iterates and $(\tau_k)_k$ is unbounded, then there exists a subsequence of iterations $((j, k_i))_{i \in \mathbb{N}}$ of the type (2.12) such that $\tau_{j,k_i} \to \infty$, $\tilde{y}_{j,k_i} \rightharpoonup u$ and u is ϵ-stationary with $\epsilon = (M+1)/(\theta(\tilde{\gamma} - \gamma)) \liminf_i \epsilon_{j,k_i}$.*
3. *If Algorithm 1 generates infinitely many serious iterates, $(u_{j_i})_{i \in \mathbb{N}}$ is a subsequence converging weakly to $\bar{u}$, and $\liminf_i \|e_{j_i,k(i)}\|_{U^*} = 0$, cf. (2.8), where $k(i)$ is the last inner iteration in outer iteration j_i (i.e., $u_{j_i+1} = \tilde{y}_{j_i,k(i)}$), then $\bar{u}$ is stationary.*
4. *If Algorithm 1 generates infinitely many serious iterates, $(u_{j_i})_{i \in \mathbb{N}}$ is a subsequence converging weakly to $\bar{u}$, and $\liminf_i \|e_{j_i,k(i)}\|_{U^*} > 0$ with $k(i)$ as in part 3, then for all i sufficiently large there exists a largest k_i such that (j_i, k_i) is of type (2.12) and $\bar{u}$ is ϵ-stationary, where $\epsilon = (M+1)/(\theta(\tilde{\gamma} - \gamma)) \liminf_i \epsilon_{j_i,k_i}$.*

Proof Due to space limitations, it is not possible to give a proof here. The result can be shown by adapting and extending our convergence theory in [22]. □

Remark 2.7

1. As in [22, Rem. 5.7], Theorem 2.6 still holds true if the function w is set to zero ($w \equiv 0$) and the feasible set $\mathcal{F}$ is bounded in U.
2. In the setting of [22], which is a special case of the situation considered here, we recover the statement of [22, Thm. 5.6]. There, for parts 2 and 4, ϵ-stationarity

with $\epsilon \leq (M+1)/(\theta(\tilde{\gamma}-\gamma))(\epsilon_1+\epsilon_2)$ is obtained under the following assumptions: The function value approximation condition

$$f_{\tilde{y}_k} - f(\tilde{y}_k) \leq f_u - f(u) + \epsilon_1 \|\iota(\tilde{y}_k - u)\|_Z + \Xi \|\iota(\tilde{y}_k - u)\|_Z^2$$

holds for $\epsilon_1 \geq 0$, $\Xi \geq 0$ with $\epsilon_1 + \Xi > 0$; $p : Z \to \mathbb{R}$ is approximately convex [13, 22] at $\iota(u)$ (part 2) or $\iota(\bar{u})$ (part 4); $G = \partial_C p + \bar{B}_{Z^*}(0, \epsilon_2)$, where $\epsilon_2 \geq 0$; $\|\iota(v_k - u)\|_Z \leq \hat{M}\|\iota(y_k - u)\|_Z$ with $\hat{M} \geq 0$; and $\Phi_k(\cdot) \geq m_k(\cdot, u)$.

3. Theorem 2.6 gives an indicator on how to refine the model. If in line 11 of Algorithm 1 the term $\epsilon_{j,k}$ is not sufficiently small, refine the function value approximation and the approximate subgradients or calculate more cutting planes to enrich the new model.
4. If exact function values (i.e. $f_u = f(u)\ \forall u \in \mathcal{F}$) and exact subgradients (i.e. $G(u) = \partial_C p(u)\ \forall u \in \iota(\mathcal{F})$) are used, there always exists a new model Φ_{k_i+1} such that the limit point $\hat{u}$ in Theorem 2.6 (i.e. $\hat{u} = u$ or $\hat{u} = \bar{u}$) is stationary. In fact, at iteration (j_i, k_i) there then exists $v_{j_i,k_i} \in [u_{j_i}, \tilde{y}_{j_i,k_i}]$ and $\tilde{g}_{j_i,k_i} \in G(\iota(v_{j_i,k_i}))$ with $p(\iota(\tilde{y}_{j_i,k_i})) - p(\iota(u_j)) \leq \langle \tilde{g}_{j_i,k_i}, \iota(\tilde{y}_{j_i,k_i} - u_{j_i})\rangle_{Z^*,Z}$ (Lebourg's mean value theorem [11, Thm. 2.3.7] shows that even "=" can be achieved). Assume we can find v_{j_i,k_i} and $\tilde{g}_{j_i,k_i}$ with this property and that the cutting plane $m^i := m_{\tilde{y}_{j_i,k_i}, v_{j_i,k_i}, \tilde{g}_{j_i,k_i}}(\cdot, u_{j_i})$ is included in the new model, i.e., $\Phi_{j_i,k_i+1} \geq m^i$, then we find

$$\begin{aligned} J_{\tilde{y}_{j_i,k_i}} - \Phi_{j_i,k_i+1}(\tilde{y}_{j_i,k_i}) &\leq f_{\tilde{y}_{j_i,k_i}} - m^i(\tilde{y}_{j_i,k_i}) \\ &= [f(\tilde{y}_{j_i,k_i}) + \langle \tilde{g}_{j_i,k_i}, \iota(u_{j_i} - \tilde{y}_{j_i,k_i})\rangle_{Z^*,Z} - f(u_{j_i})]_+ + c\|\iota(v_{j_i,k_i} - u_{j_i})\|_Z^2 \\ &\leq c\|\iota(\tilde{y}_{j_i,k_i} - u_{j_i})\|_Z^2. \end{aligned}$$

In the case that $\tilde{y}_{j_i,k_i} \rightharpoonup \hat{u}$, this shows that $\liminf \epsilon_{j_i,k_i} \leq \liminf c\|\iota(\tilde{y}_{j_i,k_i} - u_{j_i})\|_Z = 0$. According to Theorem 2.6, this means that $0 \in \iota^* \partial_C p(\iota\hat{u}) + w'(\hat{u}) + N_{\mathcal{F}}(\hat{u})$.

3 Generalized Derivatives

In this section, we will define the sets of generalized derivatives that we will consider for the solution operator of the obstacle problem which is defined between infinite dimensional spaces. A generalization of the so-called subdifferentials for functions mapping to $\mathbb{R}$ is necessary. Due to the choice of weak and strong topologies in infinite dimensional spaces, we obtain four different generalized differentials as generalizations of the Bouligand subdifferential consisting of different combinations of topologies on the involved spaces and there is no unique generalization of the concepts in finite dimension, see also [10, 36]. For the finite dimensional case, see, e.g., [32, Def. 2.12], [14, Def. 4.6.2].

Definition 3.1 Let the operator $T\colon X \to Y$ be a locally Lipschitz continuous operator between a separable Banach space X and a Hilbert space Y. Denote the subset of X on which T is Gâteaux differentiable by D_T and let T' be the respective Gâteaux derivative. We define the following sets of (Bouligand) generalized derivatives of T in u

$$\partial_B^{ij} T(u) := \{\Sigma \in \mathcal{L}(X, Y) : T'(u_n) \to \Sigma \text{ in the sense of } \mathrm{OT}(i) \\ \text{for some } (u_n)_{n\in\mathbb{N}} \subset D_T \text{ with } u_n \to u \text{ in the sense of } \mathrm{T}(j)\}.$$

Here, $i, j \in \{s, w\}$ and $T(s), T(w)$ means convergence in the strong, respective weak, sense in X, while $OT(s)$ means convergence in the strong operator topology and $OT(w)$ convergence in the weak operator topology and, in addition, $T(u_n) \rightharpoonup T(u)$ in Y.

Recall that convergence of operators $(T_n)_{n\in\mathbb{N}} \subseteq \mathcal{L}(X, Y)$ in the strong operator topology means pointwise convergence of $(T_n)_{n\in\mathbb{N}}$ to T in Y, convergence in the weak operator topology means pointwise weak convergence of $(T_n)_{n\in\mathbb{N}}$ to T in Y.

Remark 3.2

1. Assume that T fulfills the assumptions of Definition 3.1 and let, in addition, Y be separable. Then the following relations between the differentials hold for all $u \in X$, see also [36, Prop. 2.11],

$$\partial_B^{ss} T(u) \subseteq \partial_B^{sw} T(u) \subseteq \partial_B^{ww} T(u) \quad \text{and} \quad \partial_B^{ss} T(u) \subseteq \partial_B^{ws} T(u) \subseteq \partial_B^{ww} T(u).$$

2. The set $\partial_B^{sw} T(u)$, and thus also $\partial_B^{ww} T(u)$, is nonempty. See also [35, Rem. 1.1].
3. Let $S\colon X \to Y$ be a solution operator of a partial differential equation or of a variational inequality, which is Lipschitz continuous on bounded sets. Let $J : Y \times X \to \mathbb{R}$ be a continuously differentiable objective function and denote by $\hat{J}(u) = J(S(u), u)$ the corresponding reduced objective function. Then

$$\left\{\Sigma^* J_y\,(S(u), u) + J_u\,(S(u), u) : \Sigma \in \partial_B^{sw} S(u)\right\} \subset \partial_B^{sw} \hat{J}(u) \subset \partial_C \hat{J}(u).$$

4. It directly follows that if T is Gâteaux differentiable in u with Gâteaux derivative $T'(u)$, then $T'(u)$ belongs to all generalized differentials defined in Definition 3.1.

4 Properties of the Obstacle Problem

In this section, we deal with the variational inequality

$$\text{Find } y \in K_\psi, \quad \langle Ly - F(u), z - y\rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \geq 0 \quad \text{for all } z \in K_\psi. \qquad \text{(VI)}$$

The variational inequality (VI) is a basic but, due to the operator F, quite general form of an obstacle problem. We assume that $\Omega \subseteq \mathbb{R}^d$ is an open and bounded domain and that $L \in \mathcal{L}(H_0^1(\Omega), H^{-1}(\Omega))$ is a coercive and T-monotone operator, i.e., the inequality $\langle Ly, y\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \geq C_L \|y\|^2_{H_0^1(\Omega)}$ holds for some positive constant $C_L > 0$, as well as $\langle L(y-z), (y-z)_+\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} > 0$ for all $y, z \in H_0^1(\Omega)$ with $(y-z)_+ \neq 0$, see [37].

Furthermore, $F\colon U \to H^{-1}(\Omega)$ is a continuously differentiable and monotone operator on a separable partially ordered Banach space U which is Lipschitz continuous on bounded subsets of U. We will specify the precise assumptions on U in Assumption 5.5, but let us note that the class of Banach spaces U we consider includes the important examples $H^{-1}(\Omega)$, $L^2(\Omega)$, or $\mathbb{R}^n$. The closed convex set K_ψ is of the form $K_\psi := \left\{z \in H_0^1(\Omega) : z \geq \psi\right\}$ and the quasi upper-semicontinuous obstacle ψ is chosen such that K_ψ is nonempty. The inequality "$z \geq \psi$" is to be understood pointwise quasi-everywhere (q.e.) in Ω (see e.g. [1, 2, 8, 21]).

It is well known that under these assumptions the obstacle problem (VI) has a unique solution and that the solution operator $S_F\colon U \to H_0^1(\Omega)$ that assigns the solution of the variational inequality to a given $u \in U$ is Lipschitz continuous on bounded sets, see, e.g., [4, 15, 26]. If F is the identity mapping on $H^{-1}(\Omega)$, we just write $S\colon H^{-1}(\Omega) \to H_0^1(\Omega)$ for the corresponding solution operator. Note that $S_F = S \circ F$.

The following lemma establishes monotonicity properties of the solution operator S_F. This result can be found in [15, Prob. 3, p. 30] and [37, Thm. 5.1].

Lemma 4.1 *The solution operator $S_F\colon U \to H_0^1(\Omega)$ of the obstacle problem (VI) is increasing: If u_1, u_2 are elements of U such that $u_1 \geq u_2$, then the inequality $S_F(u_1) \geq S_F(u_2)$ holds a.e. and q.e. in Ω.*

In the following sections, we often write $\langle \cdot, \cdot \rangle$ for the dual pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$, omitting the subscript specifying the spaces.

4.1 Differentiability of the Solution Operator

A classical result by Mignot [29] states that the solution operator $S\colon H^{-1}(\Omega) \to H_0^1(\Omega)$ is directionally differentiable, and the directional derivative is given by a variational inequality. Based on the directional differentiability of S in the sense of Hadamard, see, e.g., [38] for this notion of directional differentiability and its relation to other notions, we can apply a chain rule and obtain the directional derivative $S_F'(u; h)$ in $u \in U$ and in direction $h \in U$ for the composite mapping $S_F = S \circ F$

$$\text{Find } \eta \in \mathcal{K}_{K_\psi}(F(u)), \quad \langle L\eta - F'(u; h), z - \eta\rangle \geq 0 \quad \text{for all } z \in \mathcal{K}_{K_\psi}(F(u)). \tag{4.1}$$

Here, $\mathcal{K}_{K_\psi}(F(u)) := \mathcal{T}_{K_\psi}(S_F(u)) \cap \mu^\perp$ is called the critical cone and $\mathcal{T}_{K_\psi}(S_F(u))$ denotes the tangent cone of K_ψ at $S_F(u) \in K_\psi$, the set $\mu^\perp = \{z \in H_0^1(\Omega) : \langle \mu, z\rangle = 0\}$ is the annihilator with respect to the functional $\mu = LS_F(u) - F(u) \in H^{-1}(\Omega)$. With the help of capacity theory, one can find the following characterization of the critical cone:

$$\begin{aligned}\mathcal{K}_{K_\psi}(F(u)) &= \left\{z \in H_0^1(\Omega) : z \geq 0 \text{ q.e. on } A(u), \langle \mu, z\rangle = 0\right\} \\ &= \left\{z \in H_0^1(\Omega) : z \geq 0 \text{ q.e. on } A(u), z = 0 \text{ q.e. on } A_s(u)\right\}.\end{aligned} \tag{4.2}$$

Here, $A(u) := \{\omega \in \Omega : S_F(u)(\omega) = \psi(\omega)\}$ denotes the active set, the set where $S_F(u)$ touches the obstacle ψ, and $A_s(u)$ denotes the strictly active set. The second characterization in (4.2) gives an implicit representation of the strictly active set, while it can also be defined explicitly as the fine support of the regular Borel measure associated with $\mu = LS_F(u) - F(u) \in H^{-1}(\Omega)^+$. For details we refer to [29], [8, Sect. 6.4], [42, App. A]. Both sets, the active set as well as the strictly active set, are quasi-closed subsets of Ω that are defined up to a set of zero capacity.

We now specify the behavior of S_F' in points where S_F is Gâteaux differentiable. Therefore, we cite the following lemma from [35, Lem. 3.3].

Lemma 4.2 *Suppose that S_F is Gâteaux differentiable in $u \in U$ and let $h \in U$ be arbitrary. Then the directional derivative $S_F'(u; h)$ is determined by the solution to the problem*

$$\text{Find } \eta \in H_0^1(D(u)), \quad \langle L\eta - F'(u; h), z\rangle = 0 \quad \text{for all } z \in H_0^1(D(u)). \tag{4.3}$$

Here, any quasi-open set $D(u)$ satisfying $\Omega \setminus A(u) \subseteq D(u) \subseteq \Omega \setminus A_s(u)$ up to a set of zero capacity is admissible in (4.3) and provides the same solution η.

Remark 4.3

1. The sets $H_0^1(D(u))$ are Sobolev spaces on quasi-open domains. For a thorough introduction to such spaces, we refer to [25]. The space $H_0^1(O)$ for a quasi-open set $O \subset \Omega$ can be defined as $H_0^1(O) := \{z \in H_0^1(\Omega) : z = 0 \text{ q.e. outside } O\}$.
2. $H_0^1(\Omega \setminus A(u))$ is the largest linear subset and $H_0^1(\Omega \setminus A_s(u))$ is the linear hull of the critical cone $\mathcal{K}_{K_\psi}(F(u))$. This describes the relation between (4.3) and (4.1).
3. Observe that whenever $A(u) = A_s(u)$ holds up to a set of zero capacity, i.e., when the strict complementarity condition is fulfilled in u, then up to disagreement on a set of capacity zero, there is only one set $D(u) = \Omega \setminus A(u) = \Omega \setminus A_s(u)$ admissible in Lemma 4.2. Nevertheless, due to the generality of the operator F in the variational inequality (VI), there might be points where S_F is Gâteaux differentiable and where the strict complementarity condition does not hold. This cannot happen for S.

The analysis we will carry out relies on the characterization of the Gâteaux derivative of S_F given as the solution of (4.3) with the choice $D(u) = \Omega \setminus A(u)$.

In the following, we will write $I(u) := \Omega \setminus A(u)$ for the inactive set. We will find a similar description as in (4.3) also for a generalized derivative of S_F in points where S_F is not Gâteaux differentiable. In such points, different choices of $D(u)$ in (4.3) generally yield different solutions and solution operators. Therefore, we have to distinguish sequences $(u_n)_{n\in\mathbb{N}} \subseteq U$ with different properties and carefully analyze the resulting behavior and the stability of the sets $H_0^1(D(u))$ and the corresponding solution operators of (4.3). The dependency of the solutions of variational inequalities, such as (4.3), on the set of test functions, such as $H_0^1(D(u))$, will be clarified in the first part of the next section.

5 An Element of the Bouligand Generalized Differential

In this section, we will construct an element of $\partial_B^{ss} S_F(u)$. In Lemma 4.2, we have seen that the Gâteaux derivatives of the solution operator S_F in differentiability points evaluated in a direction h solve a variational equation. Since the generalized differentials from Definition 3.1 contain limits of Gâteaux derivatives, we need to study the convergence of solutions (4.3). The tool will be the following definition, see [30, 37].

Definition 5.1 (Mosco Convergence) Let X be a Banach space and denote by $(C_n)_{n\in\mathbb{N}}$ a sequence of nonempty, closed, convex subsets of X. We say that $(C_n)_{n\in\mathbb{N}}$ converges to a closed convex set C in the sense of Mosco if and only if the following conditions hold:

1. For all $c \in C$ there is $(c_n)_{n\in\mathbb{N}}$ with $c_n \in C_n$ for all $n \in \mathbb{N}$ as well as $c_n \to c$ in X.
2. For any sequence $(c_{n_k})_{k\in\mathbb{N}}$ satisfying $c_{n_k} \in C_{n_k}$ for a subsequence $(n_k)_{k\in\mathbb{N}}$ of $(n)_{n\in\mathbb{N}}$ as well as $c_{n_k} \rightharpoonup c$ in X, it follows $c \in C$.

Based on Definition 5.1, the following result can be obtained, see [37, Thm. 4.1].

Proposition 5.2 *Let $L \in \mathcal{L}\left(H_0^1(\Omega), H^{-1}(\Omega)\right)$ be coercive and let $(C_n)_{n\in\mathbb{N}}, C$ be closed convex subsets of $H_0^1(\Omega)$. Assume that $C_n \to C$ in the sense of Mosco and $h_n \to h$ in $H^{-1}(\Omega)$, then the unique solutions of*

$$\text{Find } \eta_n \in C_n, \quad \langle L\eta_n - h_n, z - \eta_n\rangle \geq 0 \quad \text{for all } z \in C_n$$

converge to the solution η of the limit problem with C_n, h_n replaced by C, h, respectively.

5.1 The Set-valued Map $u \mapsto H_0^1(I(u))$

In this subsection, we analyze the set-valued map $u \mapsto H_0^1(I(u))$ and establish a Mosco convergence result for the spaces $H_0^1(I(u_n))$. We show the convergence of $\left(S_F'(u_n; h)\right)_{n\in\mathbb{N}} \subset H_0^1(\Omega)$ for Gâteaux differentiability points u_n of S_F converging from below toward u and identify the limit. This will give us an element of the Bouligand generalized differential $\partial_B^{ss} S_F(u)$. At this point, let us recall the variational equation

$$\text{Find } \eta \in H_0^1(I(u)), \quad \langle L\eta - F'(u; h), z\rangle = 0 \quad \text{for all } z \in H_0^1(I(u)). \tag{5.1}$$

for the Gâteaux derivatives that we have developed in Lemma 4.2.

The crucial point to be considered when examining the convergence of solution operators of (5.1), which is by Proposition 5.2 linked to the Mosco convergence of the sets $H_0^1(I(u_n))$, is to avoid sudden jumps in the inactive sets. As Fig. 1 shows, these jumps can occur suddenly in the limit active set $I(u)$ and Mosco convergence of $H_0^1(I(u_n))$ to $H_0^1(I(u))$ cannot be expected. Figure 1 also shows the influence of monotonicity of the sequence $(u_n)_{n\in\mathbb{N}}$ on the active and strictly active sets. More precisely, different solutions of the obstacle problem are depicted in Fig. 1. The associated values of u_i are chosen constant and equal to zero, respectively >0 and <0. We can also see the corresponding active sets $A(u_i)$ and the strictly active sets $A_s(u_i)$ underneath. In $u = 0$ with the respective solution in red, since the isolated point in $A(0)$ belongs to the set $A(0)$, but is not contained in $A_s(0)$, the strict complementarity condition does not hold, i.e., $A(0) \neq A_s(0)$. Note that a single point has capacity strictly positive in the one-dimensional case. Therefore,

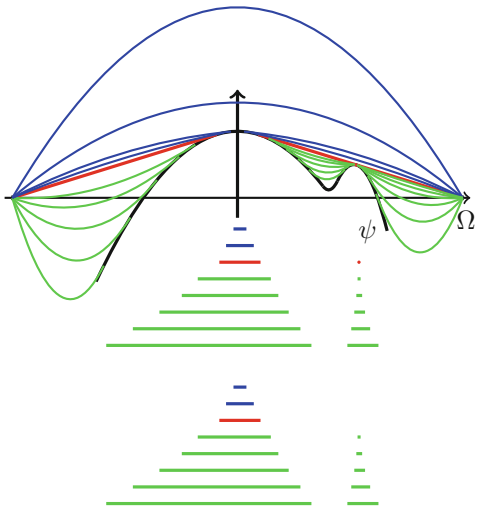

Fig. 1 Top: An instance of the obstacle problem for a piecewise quadratic obstacle ψ. The solution $S(0)$ is plotted in red, while solutions for $S(u)$ with different parameters for $u \leq 0$ are plotted in green and for $u \geq 0$ in blue. Middle: The corresponding active sets $A(u)$ for the different values of u. Bottom: The corresponding strictly active sets $A_s(u)$ for the different values of u

$u = 0$ is a point where the respective solution operator is potentially non-Gâteaux differentiable.

Example Let us consider the sets $(H_0^1(I(u_n)))_{n\in\mathbb{N}}$. We argue that the Mosco limit will, in general, not be $H_0^1(I(u))$ for a decreasing sequence $(u_n)_{n\in\mathbb{N}}$ with $u_n \to u$. In the situation of Fig. 1, choose an element $v \in H^1(\mathbb{R}^d)$ with $\{v > 0\} = \Omega \setminus A_s(0)$ up to a set of zero capacity, see [41, Prop. 2.3.14] or [18, Lem. 3.6], and define $v_n := v$ for all $n \in \mathbb{N}$. Then, it holds $v_n \in H_0^1(I(u_n))$ for all $n \in \mathbb{N}$ as well as $v_n \to v$. Nevertheless, v is not an element of $H_0^1(I(0))$. Therefore, the Mosco limit of the sequence $(H_0^1(I(u_n)))_{n\in\mathbb{N}}$ is not $H_0^1(I(0))$ (but rather $H_0^1(\Omega \setminus A_s(0))$).

This idea to consider increasing sequences $(u_n)_{n\in\mathbb{N}}$ converging to u in order to obtain Mosco convergence of the sets $H_0^1(I(u_n))$ to $H_0^1(I(u))$ is formalized in the following theorem, which is taken from [35, Thm. 5.2].

Theorem 5.3 *Let $(u_n)_{n\in\mathbb{N}} \subset U$ be an increasing sequence such that $u_n \uparrow u$. Then, the sequence $(H_0^1(I(u_n)))_{n\in\mathbb{N}}$ converges to $H_0^1(I(u))$ in the sense of Mosco. If, furthermore, S_F is Gâteaux differentiable in u_n for all $n \in \mathbb{N}$, then $\left(S_F'(u_n;\cdot)\right)_{n\in\mathbb{N}}$ converges in the strong operator topology to $\Sigma_F(u;\cdot)$, where, for a given $h \in U$, the element $\Sigma_F(u;h)$ is given by the unique solution of (5.1).*

5.2 Existence of Points of Gâteaux Differentiability in the Positive Cone

Next, we argue that an increasing sequence $(u_n)_{n\in\mathbb{N}}$ converging to an arbitrary $u \in U$ in which S_F is Gâteaux differentiable always exists. With this result, we can infer that $\Sigma_F(u;\cdot) \in \mathcal{L}\left(U, H_0^1(\Omega)\right)$ is in $\partial_B^{ss} S_F(u)$. The argument is based on the following theorem.

Theorem 5.4 *Every map from a separable Banach space to a Hilbert space which is Lipschitz continuous on bounded sets is Gâteaux differentiable on a dense subset of its domain.*

A proof can be found in, e.g., [6, Thm. 6.42]. We also refer the reader to [29, Thm. 1.2], where the same result is shown for the case that only Hilbert spaces appear.

In order to ensure that Theorem 5.3 yields an element of $\partial_B^{ss} S_F(u)$, we make the following assumptions on the size of the positive cone in U.

Assumption 5.5 We assume that V is a partially ordered space such that the positive cone $\mathcal{P} := \{v \in V : v \geq 0\}$ has nonempty interior. Let V be embedded into the space U. The embedding $\iota\colon V \to U$ is assumed to be continuous, dense, and compatible with the order structures of V and U, i.e., if $v_1, v_2 \in V$ with $v_1 \leq v_2$ then $\iota(v_1) \leq \iota(v_2)$ in U.

Note that Assumption 5.5 is satisfied for, e.g., $U = L^2(\Omega)$, $U = H^{-1}(\Omega)$ and $U = \mathbb{R}^n$. Now, we can show the following proposition, which is taken from [35, Prop. 5.5]:

Proposition 5.6 *Let u be an arbitrary element of U and assume that Assumption 5.5 is satisfied for U. Then there exists a sequence $(u_n)_{n\in\mathbb{N}}$ such that the solution operator S_F is Gâteaux differentiable in each u_n and $u_n \uparrow u$.*

5.3 Characterization of a Generalized Derivative

The preceding results imply the following characterization of a generalized derivative in $\partial_B^{ss} S_F(u) \subseteq \partial_B^{sw} S_F(u)$ for arbitrary elements $u \in U$, see [35, Thm. 5.6].

Theorem 5.7 *Let Assumption 5.5 be fulfilled for U and let $u \in U$ be arbitrary. Then the operator $\Sigma_F(u;\cdot) \in \mathcal{L}(U, H_0^1(\Omega))$, where $\Sigma_F(u;h)$ is given by the unique solution to the variational equation (5.1), is in the Bouligand generalized differential $\partial_B^{ss} S_F(u)$ of S_F in u.*

Let us now consider an optimal control problem where the obstacle problem describes the constraint set, such as $\min_u \hat{J}(u) = J(S_F(u), u)$. Here, $J : H_0^1(\Omega) \times U \to \mathbb{R}$ is a continuously differentiable objective function. An element of Clarke's generalized gradient $\partial_C \hat{J}(u)$ at the point $u \in U$ can be obtained in the following way, see [35, Thm.5.7].

Theorem 5.8 *Let q be the unique solution of the variational equation*

$$\text{Find } q \in H_0^1(I(u)), \quad \left\langle L^* q, v \right\rangle = \left\langle J_y\left(S_F(u), u\right), v \right\rangle \quad \text{for all } v \in H_0^1(I(u)). \tag{5.2}$$

Then, $F'(u)^* q + J_u\left(S_F(u), u\right)$ is in $\partial_C \hat{J}(u)$. Here, J_y and J_u denote the continuous Fréchet derivatives of J with respect to y and u, respectively, $F'(u)^* \in \mathcal{L}\left(H_0^1(\Omega), U^*\right)$ is the (Banachian) adjoint operator of $F'(u) \in \mathcal{L}\left(U, H^{-1}(\Omega)\right)$, and $L^* \in \mathcal{L}\left(H_0^1(\Omega), H^{-1}(\Omega)\right)$ is the (Banachian) adjoint operator of $L \in \mathcal{L}\left(H_0^1(\Omega), H^{-1}(\Omega)\right)$.

6 Characterization of the Entire Generalized Differentials

Now, we reduce the generality of (VI) and consider the obstacle problem

$$\text{Find } y \in K_\psi, \quad \langle Ly - u, z - y\rangle \geq 0 \quad \text{for all } z \in K_\psi \tag{VI_{id}}$$

and the corresponding solution operator $S\colon H^{-1}(\Omega) \to H_0^1(\Omega)$. Here, $L = -\Delta$, i.e., $\langle Ly, v\rangle = \int_\Omega \nabla y \cdot \nabla v \, dx$. The operator F from (VI) is realized by the identity on $H^{-1}(\Omega)$.

The generalized derivative formula obtained in Theorem 5.7 applies in this special case and we have already computed an element of $\partial_B^{ss} S(u)$ for this setting. Nevertheless, we can make use of the simplified structure of ($\mathrm{VI}_{\mathrm{id}}$) and obtain, by abstract arguments, more than just one generalized derivative. Indeed, it is possible to characterize $\partial_B^{ss} S(u)$, $\partial_B^{sw} S(u)$, and $\partial_B^{ws} S(u)$. This section is based on [36].

The preimage space of S is the whole space $H^{-1}(\Omega)$, while the operator F entering the more general variational inequality (VI) often realizes only a smaller subset of $H^{-1}(\Omega)$, since the range of F is, in general, smaller than $H^{-1}(\Omega)$. For arbitrary $u \in H^{-1}(\Omega)$, given a quasi-open set $D(u)$ with $I(u) \subseteq D(u) \subseteq \Omega \setminus A_s(u)$, the availability of all elements in $H^{-1}(\Omega)$ allows to construct a sequence $u_n \to u$ in $H^{-1}(\Omega)$ such that $S'(u_n)$ is the solution operator to

$$\text{Find } \eta \in H_0^1(D(u)), \quad \langle L\eta - h, z\rangle = 0 \quad \text{for all } z \in H_0^1(D(u)),$$

i.e., the sequence $(S'(u_n))_{n\in\mathbb{N}}$ is constant and converges. This is a strategy entering the proof of Theorem 6.1 in [36]. It indicates why we are able to characterize generalized differentials for solution operators of ($\mathrm{VI}_{\mathrm{id}}$), while the situation is much more complicated for the general variational inequality (VI). Already in finite dimensions, the authors of [20] impose a local surjectivity assumption on the analog of the operator F in finite dimension, in order to characterize a generalized differential. We obtain the following characterization of $\partial_B^{ss} S(u)$ and $\partial_B^{ws} S(u)$. For the proofs, see [36].

Theorem 6.1 *Let $u \in H^{-1}(\Omega)$ be arbitrary. The Bouligand generalized differentials $\partial_B^{ss} S(u)$ and $\partial_B^{ws} S(u)$ contain all solution operators of (4.3) for any quasi-open set $D(u)$ with $I(u) \subseteq D(u) \subseteq \Omega \setminus A_s(u)$ and any element of $\partial_B^{ss} S(u)$ and $\partial_B^{ws} S(u)$ is of this form.*

Remark 6.2

1. The characterization of Theorem 6.1 applies independent from differentiability. If and only if there is no gap between $A(u)$ and $A_s(u)$ in the sense of capacity, the operator S is Gâteaux differentiable in u, and the differentials $\partial_B^{ss} S(u)$ and $\partial_B^{ws} S(u)$ contain only the Gâteaux derivative.
2. The result in Theorem 6.1 supports the conjecture that by focusing on the sets $\Omega \setminus A_s(u_n)$ instead of $I(u_n)$, carrying out the appropriate analysis and using the approach from Sect. 5, one would indeed obtain a further generalized derivative, also for the variational inequality (VI) invoking the monotone operator F.

6.1 Capacitary Measures and the Differentials Involving the Weak Operator Topologies

As already mentioned in Remark 3.2, the generalized differentials using the weak operator topology are supersets of the differentials characterized in Theorem 6.1. In this subsection, we will see that they are in fact larger and get to know the objects they contain in addition.

Definition 6.3 We denote by $\mathcal{M}_0(\Omega)$ the set of all regular Borel measures μ on Ω with the property that $\mu(B) = 0$ holds for every Borel set $B \subseteq \Omega$ with $\operatorname{cap}(B) = 0$. Here, regularity of μ means that $\mu(B) = \inf\{\mu(O) : O \text{ quasi-open}, B \subseteq O\}$ holds for every Borel set $B \subseteq \Omega$. The set $\mathcal{M}_0(\Omega)$ is called the set of capacitary measures on Ω.

The convergence of solution operators of

$$\text{Find } \eta \in H_0^1(O), \quad \langle L\eta - h, z\rangle = 0 \quad \text{for all } z \in H_0^1(O) \tag{6.1}$$

for quasi-open sets $O \subseteq \Omega$ in the weak operator topology is metrizable, see [12, Prop. 4.9], but the resulting metric space is not a complete space. Recall that the Gâteaux derivative operators of S in points of differentiability are exactly of this form with $I(u) \subseteq O \subseteq \Omega \setminus A_s(u)$, see Lemma 4.2. For $\mu \in \mathcal{M}_0(\Omega)$, denote by $X_\mu(\Omega)$ the space $H_0^1(\Omega) \cap L_\mu^2(\Omega)$, where $L_\mu^2(\Omega)$ is the Lebesgue space of square integrable functions on Ω w.r.t. the measure μ. As shown in, e.g., [12], the completion contains exactly the solution operators of

$$\text{Find } \eta \in X_\mu(\Omega), \quad \int_\Omega \nabla\eta \cdot \nabla z \, dx + \int_\Omega \eta\, z \, d\mu = \langle h, z\rangle \quad \text{for all } z \in X_\mu(\Omega) \tag{6.2}$$

for all capacitary measures μ on Ω. Thus, it is not surprising that a subset of these solution operators of (6.2) enter the set $\partial_B^{sw} S(u)$. For the details, see [36].

Theorem 6.4 *Under some regularity assumptions on the obstacle and on $S(u)$, the Bouligand differential $\partial_B^{sw} S(u)$ in $u \in H^{-1}(\Omega)$ contains exactly all solution operators of (6.2) for any capacitary measure μ fulfilling $\mu(I(u)) = 0$ and $\mu = +\infty$ on $A_s(u)$. Here, $\mu = +\infty$ on $A_s(u)$ means that $v = 0$ q.e. on $A_s(u)$ holds for all $v \in H_0^1(\Omega) \cap L_\mu^2(\Omega)$.*

Remark 6.5

1. For a quasi-open set $O \subset \Omega$, we can define for each Borel set $B \subset \Omega$

$$\infty_{\Omega\setminus O}(B) := \begin{cases} 0, & \text{if } \operatorname{cap}(B \setminus O) = 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

With this definition, $\infty_{\Omega\setminus O}$ is a capacitary measure and L_O is the solution operator of (6.1) if and only if L_O is the solution operator of (6.2) with $\mu = \infty_{\Omega\setminus O}$. The condition $I(u) \subseteq O \subseteq A_s(u)$ can be expressed as $\infty_{\Omega\setminus O}(I(u)) = 0$ and $\infty_{\Omega\setminus O} = +\infty$ on $A_s(u)$. In this sense, $\partial_B^{ss} S(u) = \partial_B^{ws} S(u) \subseteq \partial_B^{sw} S(u)$.

2. The notion of convergence for the solution operators of (6.2) based on the weak operator topology is also called γ-convergence of the respective measures. The study of this convergence and the so-called relaxed Dirichlet problems, such as (6.2), is interesting also in shape optimization, see [9] or [2].
3. Mosco convergence of sets $H_0^1(O_n)$ to $H_0^1(O)$ for quasi-open sets $O_n, O \subseteq \Omega$ is equivalent to the γ-convergence of the measures $\infty_{\Omega\setminus O_n}$ to $\infty_{\Omega\setminus O}$. This gives the link to the approach in Sect. 5.

In [36], an example is given which illustrates that the generalized differential $\partial_B^{ww} S(u)$ is very large, even when S is Gâteaux differentiable in u and $A(u) = A_s(u)$.

Based on the characterization of the generalized differentials for S, necessary optimality conditions for the optimal control of the obstacle problem with control constraints can be obtained, see [36].

7 The Stochastic Obstacle Problem

The subject of this section is the optimal control problem ($\mathbf{P}_s$) governed by the stochastic obstacle problem ($\mathbf{VI}_s$). We want to find stationary points of the reduced objective function by applying the bundle method, developed in Sect. 2. To do so, we need to calculate a subgradient of the reduced objective function or an approximate subgradient in the sense of Assumption 2.1.

7.1 Problem Setting

Let $(\Xi, \mathcal{A}, P)$ be a measure space and set $Y := H_0^1(\Omega)$. For $\xi \in \Xi$, we consider a variational inequality of type (VI). In particular, let $L_\xi \in \mathcal{L}(Y, Y^*)$ be an operator, $\psi_\xi \in H^1(\Omega)$ an obstacle, $b \in Y^*$, define the set $K_{\psi_\xi} := \{y \in H_0^1(\Omega) : y \geq \psi_\xi\}$ and the parametric obstacle problem

$$\text{Find } y \in K_{\psi_\xi}, \quad \langle L_\xi y - b, z - y\rangle_{Y^*,Y} \geq 0 \quad \text{for all } z \in K_{\psi_\xi}. \tag{VI$_\xi$}$$

We want to relate the solutions to (VI$_\xi$) and ($\mathbf{VI}_s$), see [16, 17] for related results. Using standard techniques, one can show that the projection onto the set $\mathbf{K}_{\boldsymbol{\psi}}$, defined in (1.2), agrees pointwise P-a.e. with the projection onto K_{ψ_ξ}:

Lemma 7.1 *If $\boldsymbol{\psi} \in \bar{\mathbf{Y}}$ such that $K_{\psi_\xi} \neq \emptyset$ for P-a.a. $\xi \in \Xi$ then $\mathbf{K}_{\boldsymbol{\psi}}$ is a nonempty closed convex subset of $\mathbf{Y}$ and $P_{\mathbf{K}_{\boldsymbol{\psi}}}(\mathbf{y})(\xi) = P_{K_{\psi_\xi}}(\mathbf{y}(\xi))$ for P-a.a. $\xi \in \Xi$ and for all $\mathbf{y} \in \bar{\mathbf{Y}}$.*

Using this result, we can show that the solution operator of ($\mathbf{VI_s}$) agrees pointwise P-a.e. with the solution operator of (VI_ξ). We need the following definition:

Definition 7.2 A family of operators $(L_\xi)_{\xi\in\Xi} \subset \mathcal{L}(Y, Y^*)$ is called uniformly coercive, if there exists a parameter $C_L > 0$ such that $\langle L_\xi x, x\rangle_{Y^*,Y} \geq C_L \|x\|_Y^2$ for P-a.a. $\xi \in \Xi$.

Theorem 7.3 *Assume that $\xi \mapsto L_\xi y$ is measurable for every $y \in Y$, $\xi \mapsto \|L_\xi\|_{\mathcal{L}(Y,Y^*)}$ is in $L^\infty(\Xi)$, and $\boldsymbol{\psi} : \xi \mapsto \psi_\xi$ is in $\bar{\mathbf{Y}}$. Then, for every $\mathbf{y} \in \mathbf{Y}$, the map $\mathbf{L_y} : \xi \mapsto L_\xi(\mathbf{y}(\xi))$ is in $\mathbf{Y}^*$ and $\mathbf{L} : \mathbf{y} \mapsto \mathbf{L_y}$ is in $\mathcal{L}(\mathbf{Y}, \mathbf{Y}^*)$. Moreover, suppose that $(L_\xi)_{\xi\in\Xi}$ is uniformly coercive and that $K_{\psi_\xi} \neq \emptyset$ for P-a.a. $\xi \in \Xi$. Then, for P-a.a. $\xi \in \Xi$ and all $b \in Y^*$, (VI_ξ) has a unique solution $y_{\xi,b}$ and the solution operator $S_\xi : Y^* \to Y$, $S_\xi(b) := y_{\xi,b}$, is Lipschitz with modulus $1/C_L$. Furthermore, for all $\mathbf{b} \in \mathbf{Y}^*$, ($\mathbf{VI_s}$) has a unique solution $\mathbf{y_b}$, the solution operator $\mathbf{S} : \mathbf{Y}^* \to \mathbf{Y}$, $\mathbf{S}(\mathbf{b}) := \mathbf{y_b}$ is Lipschitz with modulus $1/C_L$, and $(\mathbf{S}(\mathbf{b}))(\xi) = S_\xi(\mathbf{b}(\xi))$ for P-a.a. $\xi \in \Xi$.*

Proof Under the given integrability assumptions, one can show that $\mathbf{L} \in \mathcal{L}(\mathbf{Y}, \mathbf{Y}^*)$ is well defined and coercive with constant C_L. The Lions–Stampacchia theorem, cf. [26, Thm. 2.1] implies that both problems are uniquely solvable. Since $\mathbf{y} \in \mathbf{Y}$, defined by $\mathbf{y}(\xi) := S_\xi(\mathbf{b}(\xi)) \in K_{\psi_\xi}$, fulfills ($\mathbf{VI_s}$) and the solution of ($\mathbf{VI_s}$) is unique, we deduce $\mathbf{y} = \mathbf{S}(\mathbf{b})$. □

7.2 Approximate Subgradients of the Stochastic Reduced Objective Function

In this section we show that the weak subgradients (1.4) can be used in the bundle method since they fulfill Assumption 2.1. We work in the following setting:

Assumption 7.4 Let $\mathcal{F}_Z$ be an open subset of a separable reflexive Banach space Z. Suppose that for all $\xi \in \Xi$ the functions $p_\xi : \mathcal{F}_Z \to \mathbb{R}$ satisfy the following conditions:

1. For all $z \in \mathcal{F}_Z$, the map $\xi \mapsto p_\xi(z)$ is measurable.
2. There exists a $z \in \mathcal{F}_Z$ such that $\int_\Xi |p_\xi(z)|\, dP(\xi) < \infty$.
3. For all bounded sets $B \subset Z$ there exists a function $L_B \in L^1(\Xi)$ such that

$$|p_\xi(z_1) - p_\xi(z_2)| \leq L_B(\xi)\|z_1 - z_2\|_Z \quad \text{for all } z_1, z_2 \in B \cap \mathcal{F}_Z \text{ and for } P\text{-a.a. } \xi \in \Xi.$$

Let $\bar{\mathcal{F}}$ be a closed subset of $\mathcal{F}_Z$ and consider the map $G : \bar{\mathcal{F}} \rightrightarrows Z^*$ defined by

$$G(z) := \left\{ \int_\Xi g(\xi)\, dP(\xi) : g \in L^1(\Xi, Z^*),\, g(\xi) \in \partial_C p_\xi(z)\ P\text{-a.e.} \right\}. \tag{7.1}$$

Theorem 7.5 *Under Assumption 7.4, the multifunction $G : \bar{\mathcal{F}} \rightrightarrows Z^*$, defined in (7.1), fulfills Assumption 2.1 and it holds $\partial_C p(z) \subset G(z)$ for all $z \in \bar{\mathcal{F}}$, where $p : \bar{\mathcal{F}} \to \mathbb{R}$ is defined by $p(z) := \int_\Xi p_\xi(z)\, dP(\xi)$.*

***Proof* $\partial_C p(z) \subset G(z)$.** Let $z \in \bar{\mathcal{F}}$ be arbitrary. By Clarke [11, Thm. 2.7.2], p is well defined, locally Lipschitz and for every $g \in \partial_C p(z)$ there is a corresponding mapping $\xi \mapsto g_\xi$ from Ξ to Z^* with $g_\xi \in \partial_C p_\xi(z)$ P-a.e. and such that for every $v \in Z$, the function $\xi \mapsto \langle g_\xi, v\rangle_{Z^*,Z}$ belongs to $L^1(\Xi)$ and one has $\langle g, v\rangle_{Z^*,Z} = \int_\Xi \langle g_\xi, v\rangle_{Z^*,Z}\, dP(\xi)$. Consequently, by Hytönen et al. [24, Cor. 1.1.2], the map $\xi \mapsto g_\xi$ is measurable. Denote by $L_B \in L^1(\Xi)$ the function according to property 3 of Assumption 7.4 for $B := \bar{B}_X(z, 1)$. From $\int_\Xi \|g_\xi\|_{X^*}\, dP(\xi) \le \int_\Xi L_B(\xi)\, dP(\xi) < \infty$ we deduce that $\xi \mapsto g_\xi$ is in $L^1(\Xi, Z^*)$ which shows $\partial_C p(z) \subset G(z)$.

1. For arbitrary $z \in \bar{\mathcal{F}}$ it holds $\emptyset \neq \partial_C p(z) \subset G(z)$. Therefore $G(z)$ is nonempty. Since $\partial_C p_\xi(z)$ is convex P-a.e. , the set $G(z)$ is convex.
2. Let $B \subset Z$ be a bounded set and denote

$$\hat{G} := \{\hat{g} \in L^1(\Xi, Z^*) : z \in B \cap \bar{\mathcal{F}},\, \hat{g}(\xi) \in \partial_C p_\xi(z)\ P\text{-a.e.}\}. \tag{7.2}$$

   Choose a neighborhood $\hat{B} \subset Z$ of $B \cap \bar{\mathcal{F}}$ and denote by $L_{\hat{B}} \in L^1(\Xi)$ the function which fulfills property 3 of Assumption 7.4. By Clarke [11, Prop. 2.1.2], there holds $\partial_C p_\xi(z) \subset \bar{B}_{Z^*}(0, L_{\hat{B}}(\xi))$ for all $z \in B \cap \bar{\mathcal{F}}$. Consequently, $\hat{G}$ is bounded in $L^1(\Xi, Z^*)$ by the constant $\int_\Xi L_{\hat{B}}(\xi)\, dP(\xi) < \infty$ and we find for arbitrary $g \in G(B \cap \bar{\mathcal{F}})$ that there exists a $\hat{g} \in \hat{G}$ such that $g = \int_\Xi \hat{g}(\xi)\, dP(\xi)$ and it holds

$$\|g\|_{Z^*} = \left\| \int_\Xi \hat{g}(\xi)\, dP(\xi) \right\|_{Z^*} \le \int_\Xi \|\hat{g}(\xi)\|_{Z^*}\, dP(\xi) \le \int_\Xi L_{\hat{B}}(\xi)\, dP(\xi).$$

3. We verify the assumptions of [34, Thm. 4.2]. Since $\bar{\mathcal{F}}$ is a closed subset of a complete metric space, $(\bar{\mathcal{F}}, \|\cdot\|_Z)$ is a complete metric space. By Clarke [11, Prop. 2.1.2] the map $(\xi, z) \mapsto \partial_C p_\xi(z)$ is nonempty, closed, and convex valued. Using [11, Lem. 2.7.2], [3, Thm. 8.2.11 and Thm. 8.2.9] one sees that the multifunction $\xi \mapsto \partial_C p_\xi(z)$ is measurable for all $z \in \mathcal{F}_Z$, i.e. for every open set $\mathcal{O}$ the inverse image $\{\xi \in \Xi : \partial_C p_\xi(z) \cap \mathcal{O} \neq \emptyset\}$ is measurable. By Clarke [11, Prop. 2.1.5], for all $\xi \in \Xi$, $\partial_C p_\xi$ has a weakly closed graph. Now let $B \subset Z$ be a compact set and denote by $L_{\hat{B}} \in L^1(\Xi)$ a function which fulfills property 3 of Assumption 7.4 for a neighborhood $\hat{B}$ of B. Define $G_B : \Xi \rightrightarrows Z^*$ to be the multifunction $G_B(\xi) := \text{w-cl co} \cup_{z \in B} \partial_C p_\xi(z)$. First note that, since B is

bounded, [11, Prop. 2.1.2] implies that $\cup_{z\in B}\partial_C p_\xi(z)$ is bounded by $L_{\hat{B}}(\xi)$ P-almost everywhere, i.e. G_B is integrably bounded. Also, for fixed $\xi \in \Xi$, the set $G_B(\xi)$ is bounded. Consequently, by Alaoglu's theorem, $G_B(\xi)$ is weakly compact, and obviously nonempty and convex. As $\partial_C p_\xi(z) \subset G_B(\xi)$ P-a.e. and for all $z \in B$, [34, Thm. 4.2] yields that $z \mapsto \tilde{G}(z)$ is weakly upper semicontinuous, i.e. for every weakly closed set $C \subset Y$ the set $G^-(C) := \{x \in \mathcal{F} : G(x) \cap C \neq \emptyset\}$ is closed in $\mathcal{F}$. By Papageorgiou [33, Cor. 3.1], the multifunction $\tilde{G}$ is weakly closed valued. Therefore, [23, Thm. 2.5] implies that $G : \tilde{\mathcal{F}} \rightrightarrows Z^*$ has a weakly closed graph. □

Example For all $\xi \in \Xi$, let $J_\xi : Y \to \mathbb{R}$ be given via $J_\xi(\cdot) := \frac{1}{2}\|O_\xi(\cdot) - y_\xi^d\|_H^2$, where $O_\xi \in \mathcal{L}(Y, H)$ is the stochastic observation operator and $y_\xi^d \in H$ is the stochastic desired state. Under the assumptions of Theorem 7.3 and if additionally $(\Xi, \mathcal{A}, P)$ is a probability space, $\xi \mapsto O_\xi y$ is measurable for all $y \in Y$, $\xi \mapsto \|O_\xi\|_{\mathcal{L}(Y,H)}$ is in $L^\infty(\Xi)$ and $\xi \mapsto y_\xi^d$ is in $L^2(\Xi, H)$, then the functions $p_\xi := J_\xi(S_\xi(\cdot))$ satisfy Assumption 7.4. Consequently, $G(z) := \mathbb{E}[\partial_C(J_\xi(S_\xi(\cdot)))(z)]$ fulfills Assumption 2.1 and can be used as a subdifferential for the bundle method.

7.3 Computation of Exact Subgradients

Although the approach in the previous section is very versatile, it uses an approximate subdifferential G that is possibly larger than the Clarke differential of the cost function. The resulting (ϵ-)stationarity, cf. Theorem 2.6, then corresponds to weak (ϵ-)stationarity (cf. (1.5) for weak stationarity). If possible, it would be favorable to search for Clarke-stationary points (1.3). To do so, the bundle method requires elements of the Clarke subdifferential $\partial_C(\mathbb{E}\left[J_\xi(S_\xi(\iota(\cdot)))\right])(u)$ (or approximations thereof). In this section we derive a formula to compute such subgradients under additional assumptions on the regularity of the problem data. We do not make use of chain rules for Clarke's subdifferential since they require a certain regularity, see [11].

Assume that Ξ is a separable Banach space, and let $(\Xi, \mathcal{A}, P)$ be a finite measure space. We assume that there is a nondegenerate Gaussian measure $\mathbb{P}$ on Ξ, such that P is absolutely continuous w.r.t. $\mathbb{P}$. For the notion of nondegenerate Gaussian measures we refer to [6] and Definition 7.7. For $\xi \in \Xi$, let $L_\xi \in \mathcal{L}(H_0^1(\Omega), H^{-1}(\Omega))$ be a T-monotone operator. We assume that the family of operators $(L_\xi)_{\xi\in\Xi}$ is uniformly coercive in the sense of Definition 7.2. Furthermore, let $\Xi \ni \xi \mapsto L_\xi \in \mathcal{L}(H_0^1(\Omega), H^{-1}(\Omega))$ be Lipschitz continuous on bounded sets. Then the maps $\Xi \ni \xi \mapsto L_\xi y \in H^{-1}(\Omega)$ are measurable for all $y \in H_0^1(\Omega)$. We also assume that $\xi \mapsto \|L_\xi\|_{\mathcal{L}(H_0^1(\Omega),H^{-1}(\Omega))}$ is in $L^\infty(\Xi)$. Let $F\colon \Xi\times U \to H^{-1}(\Omega)$ be an operator that is Lipschitz continuous on bounded sets and satisfies $F(\cdot, u) \in L^2(\Xi, H^{-1}(\Omega))$ for all $u \in U$. As before, assume that $F(\xi, \cdot)$ is monotone and continuously differentiable for almost all $\xi \in \Xi$. We keep the assumptions on U

previously considered in Sect. 5, see Assumption 5.5. We consider the following subclass of parametric obstacle problems (VI_ξ)

$$\text{Find } y \in K_\psi, \quad \langle L_\xi y - F(\xi, u), z - y\rangle \geq 0 \quad \text{for all } z \in K_\psi. \tag{$\mathrm{VI}_{\xi'}$}$$

Here, the obstacle $\psi \in H^1(\Omega)$ is chosen such that $K_\psi \neq \emptyset$. In contrast to (VI_ξ), the obstacle does not depend on the parameter ξ. We denote the solution operator of the family in ($\mathrm{VI}_{\xi'}$) by $S_{F,\xi} \colon U \to H_0^1(\Omega)$. Note that $S_{F,\xi}$ is defined only for almost all $\xi \in \Xi$ and that Theorem 7.3 implies that $\xi \mapsto S_{F,\xi}(u)$ is in $L^2(\Xi, H_0^1(\Omega))$ for all $u \in U$. Let $(J_\xi \colon H_0^1(\Omega) \times U \to \mathbb{R})_{\xi\in\Xi}$ be a family of parametrized objective functions, such that almost all J_ξ are continuously differentiable and such that $(J_\xi)_{\xi\in\Xi}$ is uniformly Lipschitz continuous on bounded sets for almost all $\xi \in \Xi$. Furthermore, let $\xi \mapsto J_\xi(S_{F,\xi}(u), u)$ be integrable for all $u \in U$. For example, consider the parameter dependent objective functions $J_\xi(y, u) = \frac{1}{2}\|y - y_\xi^d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_U$, for a family $(y_\xi^d)_{\xi\in\Xi} \subseteq H_0^1(\Omega)$ such that $\xi \mapsto y_\xi^d$ is in $L^2(\Xi, H_0^1(\Omega))$. Now, we are interested in the optimal control of the stochastic obstacle problem of the form

$$\min_{u\in U_{ad}} \hat{J}(u) = \int_\Xi J_\xi(S_{F,\xi}(u), u)\, dP(\xi). \tag{$\mathbf{P}'$}$$

The set $U_{\mathrm{ad}} \subset U$ is a closed convex subset of U.

We verify the following Lipschitz continuity of $S_{F,\xi}$.

Lemma 7.6 *Under the above assumptions, the mapping $T \colon \Xi \times U \ni (\xi, u) \mapsto S_{F,\xi}(u) \in H_0^1(\Omega)$ is Lipschitz continuous on bounded subsets of $\Xi \times U$.*

Proof Let $B_\Xi \times B_U$ be a bounded subset. By assumption $\xi \mapsto L_\xi$ is Lipschitz continuous on B_Ξ and F is Lipschitz continuous on $B_\Xi \times B_U$ with Lipschitz constants c_1 and c_2, respectively. Moreover, L_ξ is coercive with a common coercivity constant C_L for almost all ξ and by continuity for all ξ.

For $i = 1, 2$, let $(\xi_i, u_i) \in B_\Xi \times B_U$ and denote $F_i := F(\xi_i, u_i)$, $L_i := L_{\xi_i}$ and $y_i := S_{F,\xi_i}(u_i)$. Now, we estimate

$$\begin{aligned}
C_L\|y_1 - y_2\|^2_{H_0^1(\Omega)} &\leq \langle L_1 y_1 - L_2 y_2 + (L_2 - L_1)y_2, y_1 - y_2\rangle \\
&\leq \langle F_1 - F_2, y_1 - y_2\rangle + \langle (L_2 - L_1)y_2, y_1 - y_2\rangle \\
&\leq (\|F_1 - F_2\|_{H^{-1}(\Omega)} + \|(L_2 - L_1)y_2\|_{H^{-1}(\Omega)})\|y_1 - y_2\|_{H_0^1(\Omega)} \\
&\leq \Big(c_1\|u_1 - u_2\|_U + (c_1 + c_2\|y_2\|_{H_0^1(\Omega)})\|\xi_1 - \xi_2\|_\Xi\Big)\, \|y_1 - y_2\|_{H_0^1(\Omega)}.
\end{aligned}$$

We obtain

$$\|y_1 - y_2\|_{H_0^1(\Omega)} \leq \frac{1}{C_L}\left(c_1\|u_1 - u_2\|_U + (c_1 + c_2\|y_2\|_{H_0^1(\Omega)})\|\xi_1 - \xi_2\|_\Xi\right). \tag{7.3}$$

Since this inequality holds for fixed $y_2 = S_{F,\xi_2}(u_2)$ and arbitrary $y_1 = S_{F,\xi_1}(u_1)$ with $(\xi_1, u_1) \in B_\Xi \times B_u$, this shows that $\|y\| \leq c$ holds for some constant $c > 0$ and for all $y \in S_{F,B_\Xi}(B_U)$. Inequality (7.3) also shows the desired Lipschitz continuity on $B_\Xi \times B_U$. □

7.4 Construction of a Subgradient for $\hat{J}$

Similar to the argument in Sect. 5.2, using the Lipschitz continuity of T established in Lemma 7.6, we can argue that T is Gâteaux differentiable on a large set. For our analysis, we need the notion of Gaussian measures on separable Banach spaces, see also [6, 7].

Definition 7.7

1. Let X be a separable Banach space. A Borel probability measure $\mathbb{P}$ on X is called a Gaussian measure if for every $x^* \in X^*$ the pushforward measure $\mathbb{P}_{x^*}$ of $\mathbb{P}$ with respect to x^*, i.e., $\mathbb{P}_{x^*}(A) = \mathbb{P}((x^*)^{-1}(A))$ for any measurable set $A \subset \mathbb{R}$, has a Gaussian distribution.
2. The Gaussian measure $\mathbb{P}$ is called nondegenerate, if $\mathbb{P}_{x^*}$ is nondegenerate for every $X^* \ni x^* \neq 0$, i.e., if it is not a Dirac measure.

Lemma 7.8 *Let U fulfill the conditions of Assumption 5.5 and let $\mathbb{V}$ be an arbitrary nondegenerate Gaussian measure on V. As in Lemma 7.6, consider the operator T with $T(\xi, u) = S_{F,\xi}(u)$. For an arbitrary $u \in U$ let $\bar{T}\colon \Xi \times V \to H_0^1(\Omega)$ be defined by $\bar{T}(\xi, v) = T(\xi, v + u)$. Then $\bar{T}$ is Gâteaux differentiable except on a $P \otimes \mathbb{V}$-null set in $\Xi \times V$.*

Proof By the properties of V, the operator $\bar{T}$ is Lipschitz continuous on bounded subsets of $\Xi \times V$. Benyamini and Lindenstrauss [6, Theorem 6.42] and the equivalence of notions of negligible sets developed in [6, Chap. 6.3] imply that $\bar{T}$ is Gâteaux differentiable on all points of its domain $\Xi \times V$ except on a Gauss null set, i.e., all nondegenerate Gaussian measures on $\Xi \times V$ vanish on this set. Note that the results from [6] easily carry over to operators which are Lipschitz continuous only on bounded subsets of their domain.

Since $\mathbb{P}, \mathbb{V}$ are nondegenerate Gaussian measures on Ξ, respectively V, the measure $\mathbb{P} \otimes \mathbb{V}$ is a Gaussian measure on $\Xi \times V$, see [7, Cor. 2.2.6]. It is also nondegenerate. To see this, let $(\xi^*, v^*) \in (\Xi \times V)^*$ be an arbitrary element of the dual space. Denote the density of $\mathbb{P}_{\xi^*}$, respectively $\mathbb{V}_{v^*}$, w.r.t. the Lebesgue measure

by ρ_{ξ^*}, respectively ρ_{v^*}. Then, for all measurable sets A it holds

$$
\begin{aligned}
(\mathbb{P}\otimes\mathbb{V})_{(\xi^*,v^*)}(A) &= \int_{\Xi} \mathbb{V}(\{v\in V : v^*(v)\in(A-\xi^*(\xi))\})\,d\mathbb{P}(\xi)\\
&= \int_{\Xi} \mathbb{V}_{v^*}(A-\langle\xi^*,\xi\rangle)\,d\mathbb{P}(\xi) = \int_{\mathbb{R}} \mathbb{V}_{v^*}(A-t)\,d\mathbb{P}_{\xi^*}(t)\\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} \chi_{A-t}(s)\,\rho_{v^*}(s)\,ds\,\rho_{\xi^*}(t)\,dt = \int_{\mathbb{R}}\int_{\mathbb{R}} \chi_A(r)\rho_{v^*}(r-t)\,dr\,\rho_{\xi^*}(t)\,dt\\
&= \int_{\mathbb{R}} \chi_A(r)\int_{\mathbb{R}} \rho_{v^*}(r-t)\rho_{\xi^*}(t)\,dt\,dr = \int_{\mathbb{R}} \chi_A(r)(\rho_{v^*}*\rho_{\xi^*})(r)\,dr.
\end{aligned}
$$

Since the convolution of two normal distributions is again a normal distribution, the conclusion follows. Thus, the set of points where $\bar{T}$ is not Gâteaux differentiable is a $\mathbb{P}\otimes\mathbb{V}$-null set. Since P is absolutely continuous w.r.t. $\mathbb{P}$, the lemma is proved. □

Remark 7.9 By Benyamini and Lindenstrauss [6, Prop. 6.18, 6.20] there is a nondegenerate Gaussian measure on V.

Lemma 7.10 *Let U fulfill the conditions of Assumption 5.5 and let $u \in U$ be arbitrary. Then there is an increasing sequence $(u_n)_{n\in\mathbb{N}} \subseteq U$ converging to u where $T(\xi,\cdot) = S_{F,\xi}$ is differentiable for P-almost all $\xi \in \Xi$.*

Proof Let $\mathbb{V}$ be an arbitrary nondegenerate Gaussian measure on V and let N be the set of points in $\Xi \times V$ where $\bar{T}$, defined as in Lemma 7.8, is not differentiable. Then, Lemma 7.8 implies

$$0 = (P\otimes\mathbb{V})(N).$$

We want to proceed as in the proof of Proposition 5.6, see also [35, Prop. 5.1] for the proof, and construct a sequence $(u_n)_{n\in\mathbb{N}} \subseteq V$, such that each u_n is taken from a specified set with interior points, to ensure the monotonicity of the sequence. Thus, we have to ensure that sets with interior points contain common points of Gâteaux differentiability of the family $(\bar{T}(\xi,\cdot))_{\xi\in\Xi}$ for P-almost all $\xi \in \Xi$.

Therefore, let $n \in \mathbb{N}$ and let O_n be a set in V with interior points. In [40], the support of Gaussian measures is discussed. Nondegenerate Gaussian measures on separable spaces have full support, i.e., any nondegenerate Gaussian measure has a positive measure on any measurable set with interior points. This implies $\mathbb{V}(O_n) > 0$. For each $v \in V$ define $N_v := \{\xi \in \Xi : (\xi, v) \in N\}$ and consider $V_0 := \{v \in V : P(N_v) = 0\}$. Then, we have

$$0 = (P\otimes\mathbb{V})(N) = \int_V P(N_v)\,d\mathbb{V}(v),$$

i.e., $\mathbb{V}(V \setminus V_0) = 0$. Let us now consider the set $\tilde{O}_n := O_n \cap V_0 = \{v \in O_n : P(N_v) = 0\}$. Then, it holds $\mathbb{V}(O_n) = \mathbb{V}(\tilde{O}_n) > 0$. Choose an arbitrary $v_n \in$

$\tilde{O}_n$. By definition, $\bar{T}$ is Gâteaux differentiable in (ξ, v_n) for P-almost all $\xi \in \Xi$. In particular, $\bar{T}(\xi, \cdot)$ is Gâteaux differentiable in v_n for P-almost all $\xi \in \Xi$. We can thus find a sequence $(v_n)_{n\in\mathbb{N}} \subseteq V$, where $\bar{T}(\xi, \cdot)$ is Gâteaux differentiable for almost all $\xi \in \Xi$.

We can again argue as in the proof of Proposition 5.6 to conclude that also $T(\xi, \cdot) = S_{F,\xi}$ defined on the whole space U is Gâteaux differentiable in $v_n + u$ for P-almost all $\xi \in \Xi$.

□

Theorem 7.11 *Let U fulfill the conditions of Assumption 5.5 and let $u \in U$ be arbitrary. Then, under the assumptions on the data specified in Sect. 7.3, a subgradient for $\hat{J}$ as defined in ($\mathbf{P}'$) is given by $\int_\Xi \Sigma_\xi(u)\, dP(\xi)$, where $\Sigma_\xi(u)$ is the subgradient of the parameter dependent reduced objective function $\hat{J}_\xi$ in u constructed in Theorem 5.8.*

Proof We consider the reduced objective function

$$\hat{J}(u) = \int_\Xi \hat{J}_\xi(u)\, dP(\xi) = \int_\Xi J_\xi(S_{F,\xi}(u), u)\, dP(\xi) = \int_\Xi J_\xi(T(\xi, u), u)\, dP(\xi).$$

Since $\hat{J}_\xi$ is Lipschitz continuous on bounded sets with a common Lipschitz constant for almost all $\xi \in \Xi$, we can exchange integration and differentiation and obtain the differentiability of $\hat{J}$ in each u_n and it holds $\hat{J}_u(u_n) = \int_\Xi (\hat{J}_\xi)_u(u_n)\, dP(\xi)$. Since $(\hat{J}_\xi)_u(u_n)$ is bounded by the common Lipschitz constant of $\hat{J}_\xi$ on a bounded set containing $(u_n)_{n\in\mathbb{N}}$, we can again exchange limits and obtain

$$\lim_{n\to\infty} \hat{J}_u(u_n) = \int_\Xi \lim_{n\to\infty} (\hat{J}_\xi)_u(u_n)\, dP(\xi) = \int_\Xi \Sigma_\xi(u)\, dP(\xi),$$

where for each $\xi \in \Xi$, the integrand $\Sigma_\xi(u)$ is the subgradient of $\hat{J}_\xi$ given in Theorem 5.8. □

Acknowledgments The authors would like to acknowledge the support by the DFG through the Priority Programme SPP 1962 within Project 23.

References

1. D. Adams, L. Hedberg, *Function spaces and potential theory.* Springer, 1996.
2. H. Attouch, G. Buttazzo, G. Michaille, *Variational analysis in Sobolev and BV spaces.* SIAM, 2014.
3. J.-P. Aubin, H. Frankowska, *Set-valued analysis.* Birkhäuser, 2009.
4. V. Barbu, *Optimal control of variational inequalities.* Pitman, 1984.
5. H.H. Bauschke, P.L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces.* Springer, 2011.
6. Y. Benyamini, J. Lindenstrauss, *Geometric nonlinear functional analysis.* AMS, 2000.

7. V. Bogachev, *Gaussian measures.* AMS, 1998.
8. J.F. Bonnans, A. Shapiro, *Perturbation analysis of optimization problems.* Springer, 2000.
9. D. Bucur, G. Buttazzo, *Variational methods in shape optimization problems.* Birkhäuser, 2005.
10. C. Christof, C. Meyer, S. Walther, C. Clason, *Optimal control of a non-smooth semilinear elliptic equation.* Math. Contr. Rel. Fields **8** (2018), 247–276.
11. F.H. Clarke, *Optimization and nonsmooth analysis.* SIAM, 1998.
12. G. Dal Maso, U. Mosco, *Wiener's criterion and* Γ*-convergence.* Appl. Math. Optim. **15** (1987), 15–63.
13. A. Daniilidis, P. Georgiev, *Approximate convexity and submonotonicity.* J. Math. Anal. Appl. **291** (2004), 292–301.
14. F. Facchinei, J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems.* Springer, 2007.
15. A. Friedman, *Variational principles and free-boundary problems.* Courier Corporation, 2010.
16. J. Gwinner, *A note on random variational inequalities and simple random unilateral boundary value problems: well-posedness and stability results.* In: Advances in convex analysis and global optimization, N. Hadjisavvas, P.M. Pardalos, eds., Springer (2001), 531–543. Springer US, Boston, MA, 2001.
17. J. Gwinner, F. Raciti, *On a class of random variational inequalities on random sets.* Numer. Funct. Anal. Optim. **27** (2006), 619–636.
18. F. Harder, G. Wachsmuth, *Comparison of optimality systems for the optimal control of the obstacle problem.* GAMM-Mitt. **40** (2018), 312–338.
19. W. Hare, C. Sagastizábal and M. Solodov, *A proximal bundle method for nonsmooth nonconvex functions with inexact information.* Comput. Optim. Appl. **63** (2016), 1–28.
20. J. Haslinger, T. Roubíček, *Optimal control of variational inequalities. Approximation theory and numerical realization.* Appl. Math. Optim. **14** (1986), 187–201.
21. J. Heinonen, T. Kilpeläinen, O. Martio, *Nonlinear potential theory of degenerate elliptic equations.* Oxford University Press, 1993.
22. L. Hertlein, M. Ulbrich, *An inexact bundle algorithm for nonconvex nonsmooth minimization in Hilbert space.* SIAM J. Control Optim., **57** (2019), 3137–3165.
23. S.H. Hou, *On property* (Q) *and other semicontinuity properties of multifunctions.* Pacific J. Math. **103** (1982), 39–56.
24. T. Hytönen, J. van Neerven, M. Veraar, L. Weis, *Analysis in Banach spaces, Vol. I: Martingales and Littlewood-Paley theory.* Springer, 2016.
25. T. Kilpelainen, J. Malý, *Supersolutions to degenerate elliptic equations on quasi open sets.* Comm. Partial Differential Equations **17** (1992), 371–405.
26. D. Kinderlehrer, G. Stampaccia, *An introduction to variational inequalities and their applications.* SIAM, 2000.
27. J. Lv, L.-P. Pang, F.-Y. Meng, *A proximal bundle method for constrained nonsmooth nonconvex optimization with inexact information.* J. Glob. Optim. **70** (2018), 517–549.
28. R. Mifflin, *An algorithm for constrained optimization with semismooth functions.* Math. Oper. Res. **2** (1977), 191–207.
29. F. Mignot, *Contrôle dans les inéquations variationelles elliptiques.* J. Funct. Anal. **22** (1976), 130–185.
30. U. Mosco *Convergence of convex sets and of solutions of variational inequalities.* Adv. Math. **3** (1969), 510–585.
31. D. Noll, *Bundle method for non-convex minimization with inexact subgradients and function values.* In: Computational and analytical mathematics, D. Bailey, H. Bauschke, et al., eds., Springer (2013), 555–592.
32. J. Outrata, M. Kočvara, J. Zowe, *Nonsmooth approach to optimization problems with equilibrium constraints. Theory, applications and numerical results.* Springer, 2013.
33. N.S. Papageorgiou, *On the theory of Banach space valued multifunctions. I. Integration and conditional expectation.* J. Multivariate Anal. **17** (1985), 185–206.
34. N.S. Papageorgiou, *On Fatou's lemma and parametric integrals for set-valued functions.* Proc. Indian Acad. Sci. Math. Sci., **103** (1993), 181–195.

35. A.-T. Rauls, S. Ulbrich, *Computation of a Bouligand generalized derivative for the solution operator of the obstacle problem.* SIAM J. Control Optim., **57** (2019), 3223–3248.
36. A.-T. Rauls, G. Wachsmuth *Generalized derivatives for the solution operator of the obstacle problem.* Set-Valued Var. Anal. (2018), 1–27.
37. J.-F. Rodrigues, *Obstacle problems in mathematical physics.* Elsevier, 1987.
38. A. Shapiro, *On concepts of directional differentiability.* J. Optim. Theory Appl. **66** (1990), 477–487.
39. T.M. Surowiec, *Numerical optimization methods for the optimal control of elliptic variational inequalities.*, In: Frontiers in PDE-constrained optimization, H. Antil, D.P. Kouri, et al., eds., Springer (2018), 123–170.
40. N. N. Vakhania, *The topological support of Gaussian measure in Banach space.* Nagoya Math. J. **57** (1975), 59–63.
41. B. Velichkov, *Existence and regularity results for some shape optimization problems.* Springer, 2015.
42. G. Wachsmuth, *Strong stationarity for optimal control of the obstacle problem with control constraints.* SIAM J. Optim. **24** (2014), 1914–1932.
43. H. Xu, D. Zhang, *Smooth sample average approximation of stationary points in nonsmooth stochastic optimization and applications.* Math. Program. **119** (2009), 371–401.

Maxwell Variational Inequalities in Type-II Superconductivity

Malte Winckler and Irwin Yousept

Abstract This report is concerned with the mathematical and numerical analysis of Bean's critical state model for type-II superconductivity in combination with the Maxwell equations. We review three different approaches to prove the well-posedness of the resulting nonsmooth problem. At first, we discuss a direct proof and an equivalent representation by means of hyperbolic variational inequalities. Thereafter, we present a well-posedness result for a general class of hyperbolic Maxwell variational inequalities and show the well-posedness of the Bean-Maxwell system the other way around. Although this result allows a flexible choice of the nonlinearity, it is not suitable for the case where the nonlinearity depends explicitly on the time-variable. Therefore, we conclude the analysis by discussing a fully discrete approximation of the underlying variational inequality with temperature effects. Moreover, this method is the foundation for the numerical algorithm since it includes a strong convergence result and a priori error estimates. We close this report by presenting 3D numerical experiments for type-II superconductors.

Mathematics Subject Classification (2020) 35L85, 35Q60

1 Introduction

The analysis of variational inequalities has a long history. With first mathematical contributions going back to the 1960s, VIs have found a wide range of applications in the modeling of mechanics, electromagnetics, and fluid-dynamics. They are characterized by a nonsmooth nonlinearity that usually comes in form of a convex and lower semicontinuous (l.s.c.) function. Therefore, the analysis is notoriously difficult and often demands the use of regularization techniques in both analysis and numerics. In this review, we concentrate on hyperbolic Maxwell variational

M. Winckler · I. Yousept (✉)
University of Duisburg-Essen, Essen, Germany
e-mail: malte.winckler@uni-due.de; irwin.yousept@uni-due.de

M. Hintermüller et al. (eds.), *Non-Smooth and Complementarity-Based Distributed Parameter Systems*, International Series of Numerical Mathematics 172,
https://doi.org/10.1007/978-3-030-79393-7_20

inequalities of the second kind and its application to the physical phenomenon of type-II (HTS) superconductivity. In general, superconductors exhibit two underlying effects when they are cooled down below a certain critical temperature:

1. The electrical resistance of the material drops to zero, i.e., an electrical current travels through the material without energy dissipation
2. The material does not allow any penetration by weak magnetic fields, i.e., they are completely expelled from the superconductor.

The latter is widely known as the *Meissner–Ochsenfeld effect*. Superconductors are classified into two different types. A characteristic property of type-I superconductors is the sharp transition between the normal and the superconducting state. As soon as the critical temperature or a critical magnetic field strength is exceeded, the Meissner–Ochsenfeld effect is no longer observable and the superconducting state breaks down. In contrast, type-II superconductors possess a mixed state. That is, above a first magnetic field strength H_{c1}, the field lines start penetrating the material partially. The superconducting state is only destroyed if the field strength exceeds a second critical value H_{c2}. Usually, the critical field strength in type-II is significantly higher than in type-I. Moreover, in the mixed state, a superconductor has almost zero electrical resistance. For these reasons, type-II superconductors are vital for many modern technological applications such as magnetic resonance imaging (MRI), magnetic confinement fusion technologies, magnetic levitation trains (MAGLEV), and many more.

In the 1960s, Bean [4, 5] developed a critical state model that postulates a nonsmooth constitutive relation between the electric field $\boldsymbol{E}$ and the current density $\boldsymbol{J}$ as follows:

(A1) The current density strength $|\boldsymbol{J}|$ cannot exceed some critical value $j_c \in \mathbb{R}^+$;
(A2) the electric field $\boldsymbol{E}$ vanishes if the current density strength $|\boldsymbol{J}|$ is strictly less than j_c;
(A3) the electric field $\boldsymbol{E}$ is parallel to the current density $\boldsymbol{J}$.

Under the assumption that the temperature of the superconductor is constantly below the critical one, the Maxwell equations in combination with Bean's model describe the evolution of the electromagnetic waves in the medium Ω:

$$\begin{cases} \epsilon \partial_t \boldsymbol{E} - \mathbf{curl}\, \boldsymbol{H} + \boldsymbol{J} = \boldsymbol{f} & \text{in } \Omega \times (0,T), \\ \mu \partial_t \boldsymbol{H} + \mathbf{curl}\, \boldsymbol{E} = 0 & \text{in } \Omega \times (0,T), \\ \boldsymbol{E} \times \boldsymbol{n} = 0 & \text{on } \partial\Omega \times (0,T), \\ (\boldsymbol{E}, \boldsymbol{H})(\cdot, t) = (\boldsymbol{E}_0, \boldsymbol{H}_0) & \text{in } \Omega. \end{cases} \tag{1.1a}$$

along with

$$\begin{cases} \boldsymbol{J}(x,t)\cdot\boldsymbol{E}(x,t) = j_c|\boldsymbol{E}(x,t)| & \text{a.e. in } \Omega_{sc}\times(0,T), \\ |\boldsymbol{J}(x,t)| \le j_c & \text{a.e. in } \Omega_{sc}\times(0,T), \\ \boldsymbol{J}(x,t) = 0 & \text{a.e. in } \Omega\backslash\Omega_{sc}\times(0,T). \end{cases} \tag{1.1b}$$

In our context, $\Omega \subset \mathbb{R}^3$ is an open set, $j_c \in \mathbb{R}^+$ denotes the critical current density, and $\Omega_{sc} \subset \Omega$ stands for the domain of the superconductor (cf. (A1)).

For a comprehensive review on Bean's critical state law along with other mathematical models, we refer the reader to [6]. The first to analyze the system (1.1a)–(1.1b) was Prigozhin [22] under the so-called eddy current approximation. If the displacement current $\epsilon\partial_t \boldsymbol{E}$ is significantly smaller than $-\mathbf{curl}\,\boldsymbol{H} + \boldsymbol{J}$, then (1.1a) can be approximated by neglecting $\epsilon\partial_t \boldsymbol{E}$. This results in a parabolic variational inequality of the first kind for the magnetic field. For other contributions concerned with the analysis and numerics for the eddy current approximation, we refer the reader to [2, 8, 9].

In this report we will review the existing literature concerned with the analysis and the numerics of (1.1a)–(1.1b) (including the displacement current) and the progress that was made during our project within the DFG priority program 1962. After introducing the notation and the necessary function spaces in Sect. 2, we will analyze three different approaches to prove the well-posedness of (1.1a)–(1.1b). The first one relies on the semigroup theory for Maxwell's equations and proves the existence of a unique solution to (1.1a)–(1.1b). Thereafter, we will discuss an extension of this result for general hyperbolic mixed variational inequalities of the second kind. Last but not least, an approach for a nonlinearity that depends explicitly on the time is presented. As this method uses a fully discrete approximation of the underlying variational inequality, it is also the foundation for the numerical algorithm. We conclude by presenting 3D numerical experiments.

2 Preliminaries

For a given Hilbert space V, we denote a standard norm by $\|\cdot\|_V$ and a standard scalar product by $(\cdot,\cdot)_V$. A bold typeface is used to indicate a three-dimensional vector function or a Hilbert space of three-dimensional vector functions. We introduce

$$\boldsymbol{H}(\mathbf{curl}) := \left\{\boldsymbol{q} \in \boldsymbol{L}^2(\Omega) \;\middle|\; \mathbf{curl}\,\boldsymbol{q} \in \boldsymbol{L}^2(\Omega)\right\},$$

where the operator **curl** is understood in the sense of distributions. As usual, $\boldsymbol{\mathcal{C}}_0^\infty(\Omega)$ stands for the space of all infinitely differentiable three-dimensional vector functions with compact support contained in Ω. We denote the closure of $\boldsymbol{\mathcal{C}}_0^\infty(\Omega)$ with respect to the $\boldsymbol{H}(\mathbf{curl})$-topology by $\boldsymbol{H}_0(\mathbf{curl})$. It is well-known that the Hilbert space

$\boldsymbol{H}_0(\mathbf{curl})$ admits the following characterization (see e.g. [28, Appendix A]):

$$\boldsymbol{H}_0(\mathbf{curl}) = \big\{\boldsymbol{q} \in \boldsymbol{H}(\mathbf{curl}) \mid (\boldsymbol{q}, \mathbf{curl}\, \boldsymbol{v})_{\boldsymbol{L}^2(\Omega)} = (\mathbf{curl}\, \boldsymbol{q}, \boldsymbol{v})_{\boldsymbol{L}^2(\Omega)}\ \forall \boldsymbol{v} \in \boldsymbol{H}(\mathbf{curl})\big\}. \tag{2.1}$$

The material parameters ϵ and μ stand for the electric permittivity and the magnetic permeability, respectively. They are assumed to be of class $L^\infty(\Omega)^{3\times 3}$, symmetric and uniformly positive-definite in the sense that there exist constants $\underline{\epsilon}, \underline{\mu} > 0$ such that

$$\xi^T \epsilon(x)\xi \geq \underline{\epsilon}|\xi|^2 \quad \text{and} \quad \xi^T \mu(x)\xi \geq \underline{\mu}|\xi|^2 \quad \text{for a.e. } x \in \Omega \text{ and all } \xi \in \mathbb{R}^3. \tag{2.2}$$

We note that, physically speaking, (A1)–(A3) is not suitable for matrix-valued material parameters. Due to (A3) and Ohm's law

$$\boldsymbol{E} = \rho \boldsymbol{J},$$

the resistivity ρ has to be scalar-valued. This cannot be guaranteed if we consider the case where ϵ and μ are matrix-valued. However, the mathematical analysis is not affected by this issue. Therefore, we retain the more general assumption.

Given a symmetric and uniformly positive-definite matrix-valued function $\alpha \in L^\infty(\Omega)^{3\times 3}$, let $\boldsymbol{L}^2_\alpha(\Omega)$ denote the weighted $\boldsymbol{L}^2(\Omega)$-space endowed with the weighted scalar product $(\alpha\cdot, \cdot)_{\boldsymbol{L}^2(\Omega)}$. Based on this notation, let us introduce the pivot Hilbert space used in our analysis:

$$\mathbf{X} := \boldsymbol{L}^2_\epsilon(\Omega) \times \boldsymbol{L}^2_\mu(\Omega),$$

equipped with the scalar product

$$((\boldsymbol{e}, \boldsymbol{h}), (\boldsymbol{v}, \boldsymbol{w}))_{\mathbf{X}} = (\epsilon \boldsymbol{e}, \boldsymbol{v})_{\boldsymbol{L}^2(\Omega)} + (\mu \boldsymbol{h}, \boldsymbol{w})_{\boldsymbol{L}^2(\Omega)} \quad \forall (\boldsymbol{e}, \boldsymbol{h}), (\boldsymbol{v}, \boldsymbol{w}) \in \mathbf{X}. \tag{2.3}$$

We close this section by introducing the (unbounded) Maxwell operator

$$\mathcal{A} : D(\mathcal{A}) \subset \mathbf{X} \to \mathbf{X}, \quad \mathcal{A} := -\begin{pmatrix} \epsilon & 0 \\ 0 & \mu \end{pmatrix}^{-1} \begin{pmatrix} 0 & -\mathbf{curl} \\ \mathbf{curl} & 0 \end{pmatrix}, \tag{2.4}$$

with

$$D(\mathcal{A}) := \boldsymbol{H}_0(\mathbf{curl}) \times \boldsymbol{H}(\mathbf{curl}).$$

The choice of the domain $D(\mathcal{A})$ is motivated by the perfectly conducting electric boundary condition, which specifies that the tangential component of the electric

field vanishes on the boundary. Obviously, $\mathcal{A} : D(\mathcal{A}) \subset \mathbf{X} \to \mathbf{X}$ is a densely defined and closed operator. More importantly, it is skew-adjoint, i.e., $D(\mathcal{A}) = D(\mathcal{A}^*)$ and $\mathcal{A} = -\mathcal{A}^*$. Therefore, thanks to Stone's theorem [21, Theorem 10.8], $\mathcal{A}$ generates a strongly continuous group $\{\mathbb{T}_t\}_{t\geq 0}$ of unitary operators in X.

3 Analysis

In this section, we will review different approaches to prove the well-posedness of (1.1a)–(1.1b). In the following, let $\Omega \subset \mathbb{R}^3$ be an open set and $T \in \mathbb{R}^+$.

3.1 Direct Approach

The first study of the well-posedness of (1.1a)–(1.1b) goes back to Jochmann [16, 17]. His proof mainly relies on the maximal monotone structure related to (1.1b). We summarize his results in the following theorem (see [16, Theorem 1] and [17, Lemma 4.3]).

Theorem 3.1 *Let* $\boldsymbol{f} \in W^{1,\infty}((0,T), \boldsymbol{L}^2(\Omega))$ *and* $(\boldsymbol{E}_0, \boldsymbol{H}_0) \in D(\mathcal{A})$. *Then, there exists a unique* $(\boldsymbol{E}, \boldsymbol{H}) \in L^\infty((0,T), D(\mathcal{A})) \cap W^{1,\infty}((0,T), \mathbf{X})$ *and a unique* $\boldsymbol{J} \in L^\infty((0,T), \boldsymbol{L}^\infty(\Omega))$ *that solve* (1.1a)–(1.1b).

He also generalized his result [17] to the case of $j_c = j_c\big(x, \boldsymbol{H}(x,t)\big)$ by a regularization technique and local compactness results for the magnetic field. These results were taken up by Yousept [26] who proved equivalence of (1.1a)–(1.1b) and a mixed hyperbolic variational inequality of the second kind of the form:

$$\begin{cases} \displaystyle\int_\Omega \epsilon \frac{d}{dt}\boldsymbol{E}(t) \cdot (\boldsymbol{v} - \boldsymbol{E}(t)) + \mu \frac{d}{dt}\boldsymbol{H}(t) \cdot (\boldsymbol{w} - \boldsymbol{H}(t))\, dx \\ \displaystyle + \int_\Omega \mathbf{curl}\, \boldsymbol{E}(t) \cdot \boldsymbol{w} - \boldsymbol{H}(t) \cdot \mathbf{curl}\, \boldsymbol{v}\, dx + j(\boldsymbol{v}) - j(\boldsymbol{E}(t)) \\ \displaystyle \geq \int_\Omega \boldsymbol{f}(t) \cdot (\boldsymbol{v} - \boldsymbol{E}(t))\, dx, \\ \text{for a.e. } t \in (0,T) \text{ and all } (\boldsymbol{v}, \boldsymbol{w}) \in \boldsymbol{H}_0(\mathbf{curl}) \times \boldsymbol{L}^2(\Omega), \\ (\boldsymbol{E}, \boldsymbol{H})(0) = (\boldsymbol{E}_0, \boldsymbol{H}_0) \end{cases} \tag{VI$_\mathrm{B}$}$$

where $j : \boldsymbol{L}^2_\epsilon(\Omega) \to \mathbb{R}$ is defined by

$$j(\boldsymbol{v}) := \int_{\Omega_{sc}} j_c |\boldsymbol{v}|\, dx. \tag{3.1}$$

Therefore, the existence of a solution to (VI_B) follows immediately from the well-posedness of (1.1a)–(1.1b). The proof is given in [26, Theorem 3.1]. In the case of $j_c = j_c(x, \boldsymbol{H}(x,t))$ this results in a quasi-variational inequality. Moreover, the optimal control of (1.1a)–(1.1b) was the subject of [25] (cf. also [24]).

3.2 General Hyperbolic Maxwell VIs of the Second Kind

In the previous section, the nonlinearity was explicitly given by (3.1) or characterized by (1.1b). Now, we present a global well-posedness result for hyperbolic Maxwell variational inequalities of the second kind with a proper, convex, and lower semicontinuous (l.s.c.) function

$$\varphi : \mathbf{X} \to \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}. \tag{3.2}$$

Therefore, the variational inequality under consideration reads as follows:

$$\begin{cases} \displaystyle\int_\Omega \epsilon \frac{d}{dt}\boldsymbol{E}(t) \cdot (\boldsymbol{v} - \boldsymbol{E}(t)) + \mu \frac{d}{dt}\boldsymbol{H}(t) \cdot (\boldsymbol{w} - \boldsymbol{H}(t))\, dx \\ \displaystyle + \int_\Omega \mathbf{curl}\, \boldsymbol{E}(t) \cdot \boldsymbol{w} - \mathbf{curl}\, \boldsymbol{H}(t) \cdot \boldsymbol{v}\, dx + \varphi(\boldsymbol{v}, \boldsymbol{w}) - \varphi((\boldsymbol{E}, \boldsymbol{H})(t)) \\ \displaystyle \geq \int_\Omega \boldsymbol{f}(t) \cdot (\boldsymbol{v} - \boldsymbol{E}(t)) + \boldsymbol{g}(t) \cdot (\boldsymbol{w} - \boldsymbol{H}(t))\, dx, \\ \text{for a.e. } t \in (0,T) \text{ and all } (\boldsymbol{v}, \boldsymbol{w}) \in \mathbf{X}, \\ (\boldsymbol{E}, \boldsymbol{H})(0) = (\boldsymbol{E}_0, \boldsymbol{H}_0) \end{cases} \tag{VI}$$

where $(\boldsymbol{f}, \boldsymbol{g})$ is given data and $(\boldsymbol{E}_0, \boldsymbol{H}_0)$ the initial value. Their necessary assumptions for the well-posedness of (VI) will be specified in Theorem 3.3.

Before we discuss the well-posedness of (VI), we also note that (VI) can be concisely reformulated by means of the Maxwell operator $\mathcal{A}$ and the subdifferential $\partial\varphi : \mathbf{X} \to 2^{\mathbf{X}}$ of φ which is given by

$$\begin{aligned} \partial\varphi(\boldsymbol{v}, \boldsymbol{w}) := \{(\boldsymbol{y}, \boldsymbol{z}) \in \mathbf{X} \,|\, \big((\boldsymbol{y}, \boldsymbol{z}), (\boldsymbol{p}, \boldsymbol{q}) - (\boldsymbol{v}, \boldsymbol{w})\big)_{\mathbf{X}} \\ \leq \varphi(\boldsymbol{p}, \boldsymbol{q}) - \varphi(\boldsymbol{v}, \boldsymbol{w}) \quad \forall (\boldsymbol{p}, \boldsymbol{q}) \in \mathbf{X}\}. \end{aligned} \tag{3.3}$$

Let $(\boldsymbol{E}, \boldsymbol{H}) \in L^\infty((0,T), D(\mathcal{A})) \cap W^{1,\infty}((0,T), \mathbf{X})$ be a solution to (VI). Since $(\boldsymbol{E}, \boldsymbol{H})(t) \in D(\mathcal{A}) = \boldsymbol{H}_0(\mathbf{curl}) \times \boldsymbol{H}(\mathbf{curl})$ holds for a.e. $t \in (0,T)$, (2.1) implies for a.e. $t \in (0,T)$ that

$$\int_\Omega \mathbf{curl}\, \boldsymbol{E}(t) \cdot \boldsymbol{H}(t)\, dx = \int_\Omega \boldsymbol{E}(t) \cdot \mathbf{curl}\, \boldsymbol{H}(t)\, dx \tag{3.4}$$

which in turn implies for every $(\boldsymbol{v}, \boldsymbol{w}) \in \mathbf{X}$ that

$$\begin{aligned}
&\int_\Omega \mathbf{curl}\, \boldsymbol{E}(t) \cdot \boldsymbol{w} - \mathbf{curl}\, \boldsymbol{H}(t) \cdot \boldsymbol{v}\, dx \\
&= \int_\Omega \mathbf{curl}\, \boldsymbol{E}(t) \cdot (\boldsymbol{w} - \boldsymbol{H}(t)) - \mathbf{curl}\, \boldsymbol{H}(t) \cdot (\boldsymbol{v} - \boldsymbol{E}(t))\, dx. \\
&= -\big(\mathcal{A}(\boldsymbol{E}, \boldsymbol{H})(t), (\boldsymbol{v}, \boldsymbol{w}) - (\boldsymbol{E}, \boldsymbol{H})(t)\big)_{\mathbf{X}}. \qquad (3.5)
\end{aligned}$$

Therefore, we may insert (3.5) into (VI) and obtain

$$\begin{aligned}
\left(\left(\frac{d}{dt} - \mathcal{A}\right)(\boldsymbol{E}, \boldsymbol{H})(t), (\boldsymbol{v}, \boldsymbol{w}) - (\boldsymbol{E}, \boldsymbol{H})(t)\right)_{\mathbf{X}} + \varphi(\boldsymbol{v}, \boldsymbol{w}) - \varphi((\boldsymbol{E}, \boldsymbol{H})(t)) \\
\geq \int_\Omega \boldsymbol{f}(t) \cdot (\boldsymbol{v} - \boldsymbol{E}(t)) + \boldsymbol{g}(t) \cdot (\boldsymbol{w} - \boldsymbol{H}(t))\, dx.
\end{aligned}$$

Now, (2.2)–(2.3) and simple algebraic rearrangements yield

$$\begin{aligned}
\left(-\left(\frac{d}{dt} - \mathcal{A}\right)(\boldsymbol{E}, \boldsymbol{H})(t) + (\epsilon^{-1}\boldsymbol{f}, \mu^{-1}\boldsymbol{g})(t), (\boldsymbol{v}, \boldsymbol{w}) - (\boldsymbol{E}, \boldsymbol{H})(t)\right)_{\mathbf{X}} \\
\leq \varphi(\boldsymbol{v}, \boldsymbol{w}) - \varphi((\boldsymbol{E}, \boldsymbol{H})(t)). \qquad (3.6)
\end{aligned}$$

Finally, by the definition of the subdifferential (3.3), we conclude that (VI) is nothing but

$$\begin{cases}
-\left(\frac{d}{dt} - \mathcal{A}\right)(\boldsymbol{E}, \boldsymbol{H})(t) + (\epsilon^{-1}\boldsymbol{f}, \mu^{-1}\boldsymbol{g})(t) \in \partial\varphi((\boldsymbol{E}, \boldsymbol{H})(t)) \text{ a.e. in } (0, T), \\
(\boldsymbol{E}, \boldsymbol{H})(0) = (\boldsymbol{E}_0, \boldsymbol{H}_0).
\end{cases} \qquad (3.7)$$

Next, we discuss the global well-posedness result for (VI) and its assumptions.

Assumption 3.2 *For every $M > 0$, there exists a constant $C(M) > 0$ such that*

$$\|(\boldsymbol{y}, \boldsymbol{z})\|_{\mathbf{X}} \leq C(M) \quad \forall (\boldsymbol{y}, \boldsymbol{z}) \in \partial\varphi(\boldsymbol{v}, \boldsymbol{w}), \qquad (3.8)$$

for all $(\boldsymbol{v}, \boldsymbol{w}) \in \mathbf{X}$ satisfying $\|(\boldsymbol{v}, \boldsymbol{w})\|_{\mathbf{X}} \leq M$.

A key tool in the existence analysis for (3.7) is the Yosida approximation of the subdifferential. For $\lambda > 0$, as our pivot space $\mathbf{X}$ is a Hilbert space, we introduce the resolvent $\boldsymbol{J}_\lambda : \mathbf{X} \to \mathbf{X}$ and the Yosida approximation $\boldsymbol{\Phi}_\lambda : \mathbf{X} \to \mathbf{X}$ by

$$\boldsymbol{J}_\lambda := (\boldsymbol{I} + \lambda\partial\varphi)^{-1} \quad \text{and} \quad \boldsymbol{\Phi}_\lambda := \frac{1}{\lambda}(\boldsymbol{I} - \boldsymbol{J}_\lambda), \qquad (3.9)$$

where $\boldsymbol{I}\colon \mathbf{X} \to \mathbf{X}$ stands for the identity operator. Thanks to the maximal monotonicity of $\partial\varphi$, $\boldsymbol{J}_\lambda$ is well-defined as a non-expansive mapping. Moreover, $\boldsymbol{\Phi}_\lambda$ is maximal monotone and Lipschitz-continuous with the Lipschitz-constant λ^{-1}.

Based on the use of the operator $\boldsymbol{\Phi}_\lambda$, [28] considers the following integral equation: For every $n \in \mathbb{N}$, find $(\boldsymbol{E}_n, \boldsymbol{H}_n) \in \mathcal{C}([0,T], \mathbf{X})$ such that

$$(\boldsymbol{E}_n, \boldsymbol{H}_n)(t) = \mathbb{T}_t(\boldsymbol{E}_0, \boldsymbol{H}_0) + \int_0^t \mathbb{T}_{t-s}\left((\epsilon^{-1}\boldsymbol{f}, \mu^{-1}\boldsymbol{g})(s) - \boldsymbol{\Phi}_{\lambda_n}((\boldsymbol{E}_n, \boldsymbol{H}_n)(s))\right) ds$$

for all $t \in [0,T]$ where $\{\lambda_n\}_{n=1}^\infty$ is a sequence of positive real numbers converging to zero and $\{\mathbb{T}_t\}_{t\geq 0}$ is the strongly continuous group generated by the Maxwell operator $\mathcal{A}$. Its well-posedness follows from the classical contraction principle. Thereafter, the uniform boundedness of the sequences $\{(\boldsymbol{E}_n, \boldsymbol{H}_n)\}_{n=1}^\infty$,$\{\boldsymbol{J}_{\lambda_n}(\boldsymbol{E}_n, \boldsymbol{H}_n)\}_{n=1}^\infty$ and $\{\boldsymbol{\Phi}_{\lambda_n}(\boldsymbol{E}_n, \boldsymbol{H}_n)\}_{n=1}^\infty$ is obtained by the virtue of the energy balance equality result (see [28, Lemma 2.1]) and Assumption 3.2. Thus, we may extract weakly-$*$ converging subsequences whose limits turn out to solve the original variational inequality. Uniqueness follows by energy estimates. The detailed proof can be found in [28, Lemma 3.2 and Theorem 3.3].

Theorem 3.3 *Let $\varphi\colon \mathbf{X} \to \overline{\mathbb{R}}$ be a convex, l.s.c. function that fulfills $\partial\varphi(0,0) \neq \emptyset$ as well as Assumption 3.2. Furthermore, let $(\boldsymbol{f}, \boldsymbol{g}) \in W^{1,\infty}((0,T), \mathbf{X})$ and $(\boldsymbol{E}_0, \boldsymbol{H}_0) \in D(\mathcal{A})$. Then, there exists a unique*

$$(\boldsymbol{E}, \boldsymbol{H}) \in L^\infty((0,T), D(\mathcal{A})) \cap W^{1,\infty}((0,T), \mathbf{X})$$

that solves (VI).

As we will see, Assumption 3.2 is satisfied for (VI_B) and hence, Theorem 3.3 can be applied to (VI_B). Unfortunately, Assumption 3.2 has some limitation. For instance, if we choose φ as the indicator function of a nonempty, closed and convex set $\boldsymbol{K} \subset \mathbf{X}$, i.e.,

$$\varphi(\boldsymbol{v}, \boldsymbol{w}) = \mathcal{I}_{\boldsymbol{K}}(\boldsymbol{v}, \boldsymbol{w}) := \begin{cases} 0 & \text{if } (\boldsymbol{v}, \boldsymbol{w}) \in \boldsymbol{K}, \\ +\infty & \text{if } (\boldsymbol{v}, \boldsymbol{w}) \notin \boldsymbol{K}, \end{cases} \tag{3.10}$$

then φ is proper, l.s.c., and convex. However, its subdifferential is given by

$$\partial\mathcal{I}_{\boldsymbol{K}}(\boldsymbol{v}, \boldsymbol{w}) := \{(\boldsymbol{y}, \boldsymbol{z}) \in \mathbf{X} \mid ((\boldsymbol{y}, \boldsymbol{z}), (\boldsymbol{p}, \boldsymbol{q}) - (\boldsymbol{v}, \boldsymbol{w}))_{\mathbf{X}} \leq 0 \quad \forall (\boldsymbol{p}, \boldsymbol{q}) \in \boldsymbol{K}\}, \tag{3.11}$$

which does not necessarily satisfy Assumption 3.2. Therefore, we cannot apply Theorem 3.3 to (3.10). However, it is possible to drop the local boundedness assumption of $\partial\varphi$ in Theorem 3.3 resulting in a more general existence result. This can be proven by taking the minimal section operator associated with the Nemytskii operator of $\partial\varphi$ acting in the Bochner space $L^2((0,T), \mathbf{X})$ into account. Here, the

initial electromagnetic field has to fulfill the additional assumption $(\boldsymbol{E}_0, \boldsymbol{H}_0) \in D(\mathcal{A}) \cap D(\partial\varphi)$. In addition, the test-functions are chosen to be more regular, leading to a weaker existence result without uniqueness. For a detailed proof we refer the reader to [28, Theorem 3.11]. Moreover, we want to emphasize that, in contrast to the previous approach, φ may also depend on the magnetic field $\boldsymbol{H}$ and not solely on the electric field $\boldsymbol{E}$.

In the remainder of this section, we come back to the variational inequality ($\mathrm{VI_B}$) and the Maxwell-Bean system (1.1a)–(1.1b). Thus, we define $\varphi\colon \mathbf{X} \to \mathbb{R}$ by $\varphi(\boldsymbol{v}, \boldsymbol{w}) := j(\boldsymbol{v})$ for $(\boldsymbol{v}, \boldsymbol{w}) \in \mathbf{X}$. Clearly, φ is proper, convex, and lower semicontinuous. It remains to verify that φ fulfills Assumption 3.2. By definition, for every $(\boldsymbol{v}, \boldsymbol{w}) \in \mathbf{X}$, it holds that

$$\partial\varphi(\boldsymbol{v}, \boldsymbol{w}) = \big\{(\boldsymbol{y}, \boldsymbol{z}) \in \mathbf{X} \,|\, (\boldsymbol{y}, \boldsymbol{p} - \boldsymbol{v})_{\boldsymbol{L}^2_\epsilon(\Omega)} + (\boldsymbol{z}, \boldsymbol{q} - \boldsymbol{w})_{\boldsymbol{L}^2_\mu(\Omega)} + j(\boldsymbol{v}) \\ \leq j(\boldsymbol{p}) \quad \forall (\boldsymbol{p}, \boldsymbol{q}) \in \mathbf{X}\big\}.$$

As φ does not depend on the second variable $\boldsymbol{w} \in \boldsymbol{L}^2_\mu(\Omega)$, it holds that $(\boldsymbol{y}, \boldsymbol{z}) \in \partial\varphi(\boldsymbol{v}, \boldsymbol{w})$ if and only if $\boldsymbol{z} = 0$ and $\boldsymbol{y} \in \partial j(\boldsymbol{v})$. Hence

$$\partial\varphi(\boldsymbol{v}, \boldsymbol{w}) = \partial j(\boldsymbol{v}) \times \{0\} \quad \forall (\boldsymbol{v}, \boldsymbol{w}) \in \mathbf{X}. \tag{3.12}$$

In order to verify Assumption 3.2, let $\boldsymbol{v} \in \boldsymbol{L}^2_\epsilon(\Omega)$ and $\boldsymbol{y} \in \partial j(\boldsymbol{v})$. By definition, $\boldsymbol{y}$ satisfies

$$(\boldsymbol{y}, \boldsymbol{p} - \boldsymbol{v})_{\boldsymbol{L}^2_\epsilon(\Omega)} \leq j(\boldsymbol{p}) - j(\boldsymbol{v}) \quad \forall \boldsymbol{p} \in \boldsymbol{L}^2_\epsilon(\Omega).$$

If we insert $\boldsymbol{p} = \boldsymbol{y} + \boldsymbol{v}$, then this yields

$$\|\boldsymbol{y}\|^2_{\boldsymbol{L}^2_\epsilon(\Omega)} \leq \int_{\Omega_{sc}} j_c(|\boldsymbol{y} + \boldsymbol{v}| - |\boldsymbol{v}|)\, dx \leq \int_{\Omega_{sc}} j_c |\boldsymbol{y}|\, dx \leq j_c |\Omega_{sc}|^{\frac{1}{2}} \|\boldsymbol{y}\|_{\boldsymbol{L}^2(\Omega)}.$$

In conclusion, Assumption 3.2 holds true. Furthermore, it is easy to see that $(0, 0) \in \partial\varphi(0, 0)$. Finally, we take $\boldsymbol{g} \equiv 0$, $\boldsymbol{f} \in W^{1,\infty}((0, T), \boldsymbol{L}^2_\epsilon(\Omega))$ as well as $(\boldsymbol{E}_0, \boldsymbol{H}_0) \in \boldsymbol{H}_0(\mathbf{curl}) \times \boldsymbol{H}(\mathbf{curl})$. Hence, Theorem 3.3 implies the existence of a unique

$$(\boldsymbol{E}, \boldsymbol{H}) \in L^\infty\big((0, T), \boldsymbol{H}_0(\mathbf{curl}) \times \boldsymbol{H}(\mathbf{curl})\big) \cap W^{1,\infty}\big((0, T), \boldsymbol{L}^2_\epsilon(\Omega) \times \boldsymbol{L}^2_\mu(\Omega)\big)$$

that fulfills the variational inequality

$$\begin{cases} \displaystyle\int_\Omega \epsilon \frac{d}{dt}\boldsymbol{E}(t)\cdot(\boldsymbol{v}-\boldsymbol{E}(t)) + \mu\frac{d}{dt}\boldsymbol{H}(t)\cdot(\boldsymbol{w}-\boldsymbol{H}(t))\,dx \\ \displaystyle + \int_\Omega \mathbf{curl}\,\boldsymbol{E}(t)\cdot\boldsymbol{w} - \mathbf{curl}\,\boldsymbol{H}(t)\cdot\boldsymbol{v}\,dx + j(\boldsymbol{v}) - j(\boldsymbol{E}(t)) \\ \displaystyle \geq \int_\Omega \boldsymbol{f}(t)\cdot(\boldsymbol{v}-\boldsymbol{E}(t))\,dx, \\ \text{for a.e. } t\in(0,T) \text{ and all } (\boldsymbol{v},\boldsymbol{w})\in \boldsymbol{L}^2_\epsilon(\Omega)\times\boldsymbol{L}^2_\mu(\Omega), \\ (\boldsymbol{E},\boldsymbol{H})(0) = (\boldsymbol{E}_0,\boldsymbol{H}_0). \end{cases} \tag{3.13}$$

Thanks to (2.1), (3.13) is equivalent to ($\mathrm{VI_B}$). Hence, $(\boldsymbol{E},\boldsymbol{H})$ is the unique solution of ($\mathrm{VI_B}$). Moreover, the representation (3.13) allows us to recover Faraday's law

$$\mu\partial_t\boldsymbol{H} + \mathbf{curl}\,\boldsymbol{E} = 0. \tag{3.14}$$

In fact, by simply testing (3.13) with $\boldsymbol{v} = \boldsymbol{E}(t)$, we obtain

$$\int_\Omega \left(\mu\frac{d}{dt}\boldsymbol{H}(t) + \mathbf{curl}\,\boldsymbol{E}(t)\right)\cdot(\boldsymbol{w}-\boldsymbol{H}(t))\,dx \geq 0 \quad \forall \boldsymbol{w}\in\boldsymbol{L}^2_\mu(\Omega).$$

Thus, (3.14) follows. On the other hand, we may insert (3.14) back into (3.13) and we obtain the following variational inequality with Faraday's law:

$$\begin{cases} \displaystyle\int_\Omega \epsilon\frac{d}{dt}\boldsymbol{E}(t)\cdot(\boldsymbol{v}-\boldsymbol{E}(t)) - \mathbf{curl}\,\boldsymbol{H}(t)\cdot(\boldsymbol{v}-\boldsymbol{E}(t))\,dx + j(\boldsymbol{v}) - j(\boldsymbol{E}(t)) \\ \displaystyle \geq \int_\Omega \boldsymbol{f}(t)\cdot(\boldsymbol{v}-\boldsymbol{E}(t))\,dx \quad \text{for a.e. } t\in(0,T) \text{ and all } \boldsymbol{v}\in\boldsymbol{L}^2_\epsilon(\Omega), \\ \mu\frac{d}{dt}\boldsymbol{H}(t) + \mathbf{curl}\,\boldsymbol{E}(t) = 0 \quad \text{for a.e. } t\in(0,T), \\ (\boldsymbol{E},\boldsymbol{H})(0) = (\boldsymbol{E}_0,\boldsymbol{H}_0) \end{cases} \tag{$\mathrm{VI_F}$}$$

Our final step is to construct a solution to the nonsmooth Maxwell system (1.1a)–(1.1b) from the solution of ($\mathrm{VI_B}$). From (3.7) and (3.12), we know that (3.13) can be equivalently reformulated by

$$\begin{cases} -\left(\frac{d}{dt} - \mathcal{A}\right)(\boldsymbol{E},\boldsymbol{H})(t) + (\epsilon^{-1}\boldsymbol{f}(t),0) \in \partial j(\boldsymbol{E}(t))\times\{0\} \text{ a.e. in } (0,T), \\ (\boldsymbol{E},\boldsymbol{H})(0) = (\boldsymbol{E}_0,\boldsymbol{H}_0). \end{cases} \tag{3.15}$$

Note that in this case (3.15) is also an immediate consequence of (VI_F). We define

$$(\boldsymbol{J}, \tilde{\boldsymbol{J}}) := -\begin{pmatrix} \epsilon & 0 \\ 0 & \mu \end{pmatrix} \left(\left(\frac{d}{dt} - \mathcal{A} \right) (\boldsymbol{E}, \boldsymbol{H}) - (\epsilon^{-1} \boldsymbol{f}, 0) \right) \in L^\infty((0,T), \mathbf{X}). \tag{3.16}$$

Since $(\boldsymbol{E}, \boldsymbol{H})$ solves (3.15) for a.e. $t \in (0,T)$, it follows that

$$(\epsilon^{-1}\boldsymbol{J}, \mu^{-1}\tilde{\boldsymbol{J}})(t) = -\left(\frac{d}{dt} - \mathcal{A} \right) (\boldsymbol{E}, \boldsymbol{H}) + (\epsilon^{-1}\boldsymbol{f}, 0) \in \partial j(\boldsymbol{E}(t)) \times \{0\}. \tag{3.17}$$

Hence, (3.17) yields that $\tilde{\boldsymbol{J}}(t) = 0$ and $\epsilon^{-1}\boldsymbol{J}(t) \in \partial j(\boldsymbol{E}(t))$ for almost every $t \in (0,T)$. Moreover, by (3.16) and the definition of $\mathcal{A}$, we obtain that $(\boldsymbol{E}, \boldsymbol{H}, \boldsymbol{J})$ fulfills

$$\begin{cases} \epsilon \dfrac{d}{dt}\boldsymbol{E} - \mathbf{curl}\, \boldsymbol{H} + \boldsymbol{J} = \boldsymbol{f} & \text{in } \Omega \times [0,T], \\ \mu \dfrac{d}{dt}\boldsymbol{H} + \mathbf{curl}\, \boldsymbol{E} = 0 & \text{in } \Omega \times [0,T]. \end{cases}$$

Together with the boundary condition for $\boldsymbol{E}$, this corresponds to (1.1a). Ultimately, it remains to verify that $\boldsymbol{J}$ satisfies (1.1b). As $\epsilon^{-1}\boldsymbol{J}(t) \in \partial j(\boldsymbol{E}(t))$ for almost every $t \in (0,T)$, it holds by definition that

$$(\boldsymbol{J}(t), \boldsymbol{v} - \boldsymbol{E}(t))_{\boldsymbol{L}^2(\Omega)} \le j(\boldsymbol{v}) - j(\boldsymbol{E}(t)) \quad \forall \boldsymbol{v} \in \boldsymbol{L}^2_\epsilon(\Omega). \tag{3.18}$$

We may test (3.18) with $\boldsymbol{v} = 0$ and $\boldsymbol{v} = 2\boldsymbol{E}(t)$, respectively, to obtain

$$(\boldsymbol{J}(t), \boldsymbol{E}(t))_{\boldsymbol{L}^2(\Omega)} = j(\boldsymbol{E}(t)). \tag{3.19}$$

On the other hand, by adding (3.18) and (3.19) it follows that

$$(\boldsymbol{J}(t), \boldsymbol{v})_{\boldsymbol{L}^2(\Omega)} \le j(\boldsymbol{v}) = \int_{\Omega_{sc}} j_c |\boldsymbol{v}|\, dx \quad \forall \boldsymbol{v} \in \boldsymbol{L}^2_\epsilon(\Omega). \tag{3.20}$$

We see that (3.20) implies immediately that $\boldsymbol{J}(t) = 0$ a.e. in $\Omega \setminus \Omega_{sc}$. Furthermore, if we assume that there exists $\omega \subset \Omega_{sc}$ with $|\omega| \neq 0$ such that $|\boldsymbol{J}(x,t)| > j_c$ holds for almost every $x \in \omega$, then we may insert $\boldsymbol{v} := \chi_\omega \frac{\boldsymbol{J}(t)}{|\boldsymbol{J}(t)|} \in \boldsymbol{L}^2_\epsilon(\Omega)$ in (3.20) and obtain a contradiction right away. This implies

$$|\boldsymbol{J}(x,t)| \le j_c \quad \text{for a.e. } (x,t) \in \Omega_{sc} \times (0,T). \tag{3.21}$$

By means of this, we obtain from (3.19) that

$$0 = \int_{\Omega_{sc}} j_c|\boldsymbol{E}(t)| - \boldsymbol{J}(t)\cdot\boldsymbol{E}(t)\,dx \quad \overset{(3.21)}{\Rightarrow} \quad \boldsymbol{J}(t)\cdot\boldsymbol{E}(t) = j_c|\boldsymbol{E}(t)| \quad \text{a.e. in } \Omega_{sc}.$$

Thus, $(\boldsymbol{E}, \boldsymbol{H}, \boldsymbol{J})$ solves (1.1a)–(1.1b) in the sense of Theorem 3.1.

4 Numerical Analysis

In this section, we consider the case where the nonlinearity φ is also explicitly depending on the time-variable t. This is the subject of [23]. Although Theorem 3.3 already allows a wide class of nonlinearities $\varphi\colon \mathbf{X} \to \mathbb{R}$, it is not applicable for this case. According to [23], the time-dependence is given as follows. Let $\theta\colon \Omega \times [0,T] \to \mathbb{R}$ be the operating temperature in the medium Ω. The fundamental properties of superconductivity demand that an accurate model includes a temperature-dependence in the critical current density j_c. Therefore, [23] specifies the nonlinearity $j\colon [0,T] \times \boldsymbol{L}^1(\Omega) \to \mathbb{R}$ by

$$j(t,\boldsymbol{v}) := \int_\Omega j_c\big(x,\theta(x,t)\big)|\boldsymbol{v}(x)|\,dx, \tag{4.1}$$

with $j_c\colon \Omega \times \mathbb{R} \to \mathbb{R}$. The mathematical assumptions for j_c which are given in Theorem 4.1 (cf. [23, Assumption 2.1]) are also justified by physical measurements [1, 7]. With (4.1) in ($\mathrm{VI_B}$), the variational inequality is given by

$$\begin{cases} \displaystyle\int_\Omega \epsilon\frac{d}{dt}\boldsymbol{E}(t)\cdot(\boldsymbol{v}-\boldsymbol{E}(t)) + \mu^{-1}\frac{d}{dt}\boldsymbol{B}(t)\cdot(\boldsymbol{w}-\boldsymbol{B}(t))\,dx \\ \quad + \displaystyle\int_\Omega \mu^{-1}\,\mathbf{curl}\,\boldsymbol{E}(t)\cdot\boldsymbol{w} - \mu^{-1}\boldsymbol{B}(t)\cdot\mathbf{curl}\,\boldsymbol{v}\,dx \\ \quad + j(t,\boldsymbol{v}) - j(t,\boldsymbol{E}(t)) \geq \displaystyle\int_\Omega \boldsymbol{f}(t)\cdot(\boldsymbol{v}-\boldsymbol{E}(t))\,dx \\ \quad \text{for a.e. } t\in(0,T) \text{ and every } (\boldsymbol{v},\boldsymbol{w}) \in \boldsymbol{H}_0(\mathbf{curl}) \times \boldsymbol{L}^2(\Omega), \\ \quad (\boldsymbol{E}(0),\boldsymbol{B}(0)) = (\boldsymbol{E}_0,\boldsymbol{B}_0). \end{cases} \tag{$\mathrm{VI_T}$}$$

Note that we use the equivalent $\boldsymbol{E}$-$\boldsymbol{B}$-formulation for Maxwell's equations here. In ($\mathrm{VI_T}$) the variables are the electric field strength $\boldsymbol{E}$ and the magnetic induction $\boldsymbol{B}$ which is given by the constitutive relation

$$\boldsymbol{B} = \mu\boldsymbol{H}. \tag{4.3}$$

This formulation is more convenient for the spatial discretization if μ is not a piecewise constant function (cf. [19]) and equivalent to (VI_B) with (4.1).

In this section, we will describe the approach and sketch the main results of [23]. Their method relies on a fully discrete approximation of (VI_T) and, unlike the previous approaches, it is not based on the semigroup theory.

4.1 Fully Discrete Scheme

Let $\Omega \subset \mathbb{R}^3$ be a polyhedral Lipschitz domain with a connected boundary $\partial\Omega$. We consider a family of quasi-uniform triangulations $\{\mathcal{T}_h\}_{h>0}$, i.e.,

$$\overline{\Omega} = \bigcup_{T \in \mathcal{T}_h} T \quad \forall h > 0.$$

The index $h > 0$ denotes the maximal diameter of the tetrahedra in $\mathcal{T}_h$. We choose $\boldsymbol{V}_h$ to be the first family of Nédélec's curl-conforming edge elements defined by

$$\boldsymbol{V}_h := \{\boldsymbol{v}_h \in \boldsymbol{H}_0(\mathbf{curl}) \ : \ \boldsymbol{v}_h|_T = \mathbf{a}_T + \mathbf{b}_T \times x \text{ with } \mathbf{a}_T, \mathbf{b}_T \in \mathbb{R}^3 \quad \forall T \in \mathcal{T}_h\}$$

and $\boldsymbol{W}_h$ as the finite element space of piecewise constant functions which is denoted by

$$\boldsymbol{W}_h := \{\boldsymbol{w}_h \in \boldsymbol{L}^2(\Omega) \ : \ \boldsymbol{w}_h|_T = \mathbf{a}_T \text{ with } \mathbf{a}_T \in \mathbb{R}^3 \quad \forall T \in \mathcal{T}_h\}.$$

Furthermore, the family of triangulations $\{\mathcal{T}\}_{h>0}$ is chosen such that there exists $\bar{h} > 0$ with

$$\boldsymbol{V}_{\tilde{h}} \subset \boldsymbol{V}_h \quad \text{and} \quad \boldsymbol{W}_{\tilde{h}} \subset \boldsymbol{W}_h \quad \forall\, 0 < h \leq \tilde{h} \leq \bar{h}.$$

For instance, this can be practically achieved by repeating bisection of the tetrahedra. For the time-discretization of (VI_T) we focus on the implicit Euler scheme. To this aim, let us fix $N \in \mathbb{N}$ and define an equidistant partition of $[0, T]$ in the following way:

$$\tau := \frac{T}{N}, \quad 0 = t_0 < t_1 < \cdots < t_N = T \quad \text{with} \quad t_n := n\tau$$

for all $n \in \{0, \ldots, N\}$. Furthermore, we define

$$\boldsymbol{f}^n := \boldsymbol{f}(t_n) \in \boldsymbol{L}^2(\Omega), \quad j^n(\boldsymbol{v}) := \int_\Omega j_c(x, \theta(x, t_n))|\boldsymbol{v}(x)|\, dx \quad \forall n \in \{0, \ldots, N\}.$$

The combination of the implicit Euler in time with the mixed finite element method in space leads to the following fully discrete scheme for ($\mathrm{VI_T}$): Find $\{(\boldsymbol{E}_h^n, \boldsymbol{B}_h^n)\}_{n=1}^N \subset \boldsymbol{V}_h \times \boldsymbol{W}_h$ such that

$$\begin{cases} \displaystyle\int_\Omega \epsilon \delta \boldsymbol{E}_h^n \cdot (\boldsymbol{v}_h - \boldsymbol{E}_h^n) + \mu^{-1} \delta \boldsymbol{B}_h^n \cdot (\boldsymbol{w}_h - \boldsymbol{B}_h^n)\, dx \\ \displaystyle + \int_\Omega \mu^{-1} \,\mathbf{curl}\, \boldsymbol{E}_h^n \cdot \boldsymbol{w}_h - \mu^{-1} \boldsymbol{B}_h^n \cdot \mathbf{curl}\, \boldsymbol{v}_h \, dx \\ \displaystyle + j^n(\boldsymbol{v}_h) - j^n(\boldsymbol{E}_h^n) \geq \int_\Omega \boldsymbol{f}^n \cdot (\boldsymbol{v}_h - \boldsymbol{E}_h^n)\, dx \\ \text{for every } (\boldsymbol{v}_h, \boldsymbol{w}_h) \in \boldsymbol{V}_h \times \boldsymbol{W}_h \text{ and } n \in \{1, \dots, N\} \\ (\boldsymbol{E}_h^0, \boldsymbol{H}_h^0) = (\boldsymbol{E}_{0h}, \boldsymbol{H}_{0h}), \end{cases} \tag{$\mathrm{VI}_{N,h}$}$$

where

$$\delta \boldsymbol{E}_h^n := \frac{\boldsymbol{E}_h^n - \boldsymbol{E}_h^{n-1}}{\tau} \quad \text{and} \quad \delta \boldsymbol{B}_h^n := \frac{\boldsymbol{B}_h^n - \boldsymbol{B}_h^{n-1}}{\tau} \quad \forall n \in \{1, \dots, N\}.$$

Moreover, $(\boldsymbol{E}_{0h}, \boldsymbol{B}_{0h}) \in \boldsymbol{V}_h \times \boldsymbol{W}_h$ denotes the finite element approximation of the initial data $(\boldsymbol{E}_0, \boldsymbol{B}_0)$ and has to satisfy a compatibility system (see [23]) to guarantee convergence. In order to recover the time-dependence in ($\mathrm{VI}_{N,h}$), we introduce the quantities

$$\begin{cases} \boldsymbol{E}_{N,h}(0) := \boldsymbol{E}_{0h} \\ \boldsymbol{E}_{N,h}(t) := \boldsymbol{E}_h^{n-1} + (t - t_{n-1}) \delta \boldsymbol{E}_h^n \end{cases} \text{and} \quad \begin{cases} \overline{\boldsymbol{E}}_{N,h}(0) := \boldsymbol{E}_{0h} \\ \overline{\boldsymbol{E}}_{N,h}(t) := \boldsymbol{E}_h^n \end{cases} \tag{4.3}$$

for $t \in (t_{n-1}, t_n]$ and $n \in \{1, \dots, N\}$. In the same way, we define $\boldsymbol{B}_{N,h}$ and $\overline{\boldsymbol{B}}_{N,h}$. The main results of [23] are summarized in the following Theorem.

Theorem 4.1 *Let $\boldsymbol{f} \in \mathcal{C}^{0,1}([0,T], \boldsymbol{L}^2(\Omega))$ and the temperature distribution $\theta \in \mathcal{C}^{0,1}([0,T], L^2(\Omega)) \cap \mathcal{C}([0,T], L^\infty(\Omega))$. Moreover, $j_c \colon \Omega \times \mathbb{R} \to \mathbb{R}$ is assumed to be Lebesgue-measurable and nonnegative in the first variable as well as locally bounded and locally Lipschitz-continuous in the second variable [23, Assumption 2.1]. Then, there exists*

$$(\boldsymbol{E}, \boldsymbol{B}) \in W^{1,\infty}\big((0,T), \boldsymbol{L}^2_\epsilon(\Omega) \times \boldsymbol{H}_0(\mathrm{div} = 0)\big) \cap L^\infty\big((0,T), \boldsymbol{H}_0(\mathbf{curl}) \times \boldsymbol{H}_0(\mathrm{div} = 0)\big)$$

such that

$$\lim_{h \to 0} \|\boldsymbol{E}_{N,h} - \boldsymbol{E}\|_{\mathcal{C}([0,T], \boldsymbol{L}^2_\epsilon(\Omega))} = \lim_{h \to 0} \|\boldsymbol{B}_{N,h} - \boldsymbol{B}\|_{\mathcal{C}([0,T], \boldsymbol{L}^2(\Omega))} = 0,$$

$$\lim_{h \to 0} \|\overline{\boldsymbol{E}}_{N,h} - \boldsymbol{E}\|_{L^\infty((0,T), \boldsymbol{L}^2_\epsilon(\Omega))} = \lim_{h \to 0} \|\overline{\boldsymbol{B}}_{N,h} - \boldsymbol{B}\|_{L^\infty((0,T), \boldsymbol{L}^2(\Omega))} = 0$$

and $(\boldsymbol{E}, \boldsymbol{B})$ is the unique solution of ($\mathrm{VI_T}$).

The basic idea for this proof makes use of a decoupling ansatz as follows: At first, we insert $\boldsymbol{v}_h := \boldsymbol{E}_h^n$ into ($\mathrm{VI}_{N,h}$) and obtain a discrete version of Faraday's law (cf. (3.14))

$$\mu^{-1}\delta \boldsymbol{B}_h^n = -\mu^{-1}\,\mathbf{curl}\,\boldsymbol{E}_h^n \quad \Rightarrow \quad \boldsymbol{B}_h^n = \boldsymbol{B}_h^{n-1} - \tau\,\mathbf{curl}\,\boldsymbol{E}_h^n. \tag{4.4}$$

Thus, we have an explicit formula for $\boldsymbol{B}_h^n$ provided that $\boldsymbol{E}_h^n$ is already computed. Next, setting $\boldsymbol{w}_h = \boldsymbol{B}_h^n$ in ($\mathrm{VI}_{N,h}$) and employing (4.4) yield the variational inequality

$$\begin{aligned}&\int_\Omega \epsilon\delta \boldsymbol{E}_h^n \cdot (\boldsymbol{v}_h - \boldsymbol{E}_h^n)\,dx + \int_\Omega \tau\mu^{-1}\,\mathbf{curl}\,\boldsymbol{E}_h^n \cdot \mathbf{curl}(\boldsymbol{v}_h - \boldsymbol{E}_h^n)\,dx + j^n(\boldsymbol{v}_h)\\ &- j^n(\boldsymbol{E}_h^n) \geq \int_\Omega \boldsymbol{f}^n \cdot (\boldsymbol{v}_h - \boldsymbol{E}_h^n) + \mu^{-1}\boldsymbol{B}_h^{n-1} \cdot \mathbf{curl}(\boldsymbol{v}_h - \boldsymbol{E}_h^n)\,dx \quad \forall \boldsymbol{v}_h \in \boldsymbol{V}_h\end{aligned} \tag{4.5}$$

which is equivalent to an elliptic **curl–curl** variational inequality of the form

$$a(\boldsymbol{E}_h^n, \boldsymbol{v}_h - \boldsymbol{E}_h^n) + j^n(\boldsymbol{v}_h) - j^n(\boldsymbol{E}_h^n) \geq \langle \tilde{\boldsymbol{f}}^n, \boldsymbol{v}_h - \boldsymbol{E}_h^n\rangle \quad \forall \boldsymbol{v}_h \in \boldsymbol{V}_h \tag{4.6}$$

with the continuous and coercive bilinear form $a\colon \boldsymbol{H}_0(\mathbf{curl}) \times \boldsymbol{H}_0(\mathbf{curl}) \to \mathbb{R}$ defined by

$$a(\boldsymbol{u}, \boldsymbol{v}) := \int_\Omega \tau^{-1}\epsilon\boldsymbol{u} \cdot \boldsymbol{v}\,dx + \int_\Omega \tau\mu^{-1}\,\mathbf{curl}\,\boldsymbol{u} \cdot \mathbf{curl}\,\boldsymbol{v}\,dx \quad \forall \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{H}_0(\mathbf{curl})$$

and the right-hand side $\tilde{\boldsymbol{f}}^n \in \boldsymbol{H}_0(\mathbf{curl})^*$ by

$$\langle \tilde{\boldsymbol{f}}^n, \boldsymbol{v}\rangle := \int_\Omega (\boldsymbol{f}^n + \tau^{-1}\epsilon\boldsymbol{E}_h^{n-1}) \cdot \boldsymbol{v}\,dx + \int_\Omega \mu^{-1}\boldsymbol{B}_h^{n-1} \cdot \mathbf{curl}\,\boldsymbol{v}\,dx \quad \forall \boldsymbol{v} \in \boldsymbol{H}_0(\mathbf{curl}).$$

The well-posedness of (4.6) is covered by a classical result in [18, Theorem 2.2] and therefore, the existence of a unique solution $\{(\boldsymbol{E}_h^n, \boldsymbol{B}_h^n)\}_{n=1}^N \subset \boldsymbol{V}_h \times \boldsymbol{W}_h$ of ($\mathrm{VI}_{N,h}$) follows by inductive reasoning. By means of zero- and first-order stability estimates, a weak-* limit $(\boldsymbol{E}, \boldsymbol{B})$ is generated that in fact solves (VI_{B}).

Thereafter, in order to prove the strong convergence, the solution operator $\boldsymbol{\Psi}_h\colon \boldsymbol{H}_0(\mathbf{curl}) \to \boldsymbol{V}_h$ of the following mixed variational problem is introduced: For every $\boldsymbol{y} \in \boldsymbol{H}_0(\mathbf{curl})$ find $\boldsymbol{y}_h \in \boldsymbol{V}_h$

$$\begin{cases} (\mu^{-1}\,\mathbf{curl}\,\boldsymbol{y}_h, \mathbf{curl}\,\boldsymbol{v}_h)_{\boldsymbol{L}^2(\Omega)} = (\mu^{-1}\,\mathbf{curl}\,\boldsymbol{y}, \mathbf{curl}\,\boldsymbol{v}_h)_{\boldsymbol{L}^2(\Omega)} & \forall \boldsymbol{v}_h \in \boldsymbol{V}_h,\\ (\boldsymbol{y}_h, \nabla\psi_h)_{\boldsymbol{L}^2(\Omega)} = (\boldsymbol{y}, \nabla\psi_h)_{\boldsymbol{L}^2(\Omega)} & \forall \psi_h \in \Theta_h \end{cases} \tag{4.7}$$

where Θ_h denotes the space of continuous piecewise linear elements with vanishing traces

$$\Theta_h := \{\phi_h \in H_0^1(\Omega) \,:\, \phi_h|_T = \mathbf{a}_T \cdot x + b_T \text{ with } \mathbf{a}_T \in \mathbb{R}^3, b_T \in \mathbb{R} \quad \forall T \in \mathcal{T}_h\}.$$

The theory of mixed problems (cf. [20, Theorem 2.45]) in combination with the discrete Poincaré–Friedrichs-type inequality [14, Theorem 4.7] and the discrete LBB condition (cf. [27, pp. 2802–2803]) implies that for every $h > 0$ and $\boldsymbol{y} \in \boldsymbol{H}_0(\mathbf{curl})$, (4.7) admits a unique solution $\boldsymbol{y}_h = \boldsymbol{\Psi}_h \boldsymbol{y} \in \boldsymbol{V}_h$ satisfying

$$\|\boldsymbol{\Psi}_h \boldsymbol{y} - \boldsymbol{y}\|_{\boldsymbol{H}(\mathbf{curl})} \le C \left(\inf_{\boldsymbol{\chi}_h \in \boldsymbol{V}_h} \|\boldsymbol{y} - \boldsymbol{\chi}_h\|_{\boldsymbol{H}(\mathbf{curl})} \right) \quad \forall \boldsymbol{y} \in \boldsymbol{H}_0(\mathbf{curl}).$$

with a constant $C > 0$, independent of h and $\boldsymbol{y}$. By the virtue of the best-approximation property of $\boldsymbol{\Psi}_h$ and the density of the families $\boldsymbol{V}_h \subset \boldsymbol{H}_0(\mathbf{curl})$ as well as $\boldsymbol{W}_h \subset \boldsymbol{L}^2(\Omega)$, energy estimates yield the desired strong convergence of the discrete solutions toward the solution $(\boldsymbol{E}, \boldsymbol{B})$ of ($\mathrm{VI_T}$).

Finally, they prove a priori error estimates that basically rely on the mentioned energy estimates and an error estimate with low field regularity for $\boldsymbol{\Psi}_h$ that originally comes from [10, 11].

5 Computations

This method does not only yield the well-posedness for ($\mathrm{VI_T}$) but it is also the basis for the numerical implementation. Thanks to the formulas (4.4) and (4.6), we have to solve an elliptic **curl–curl** variational inequality of the second kind in each time-step. For the sake of simplicity, we will only describe how we solve the system for a fixed time-step $n_0 \in \{1, \dots, N\}$. Therefore, we assume that $(\boldsymbol{E}_h^{n_0-1}, \boldsymbol{B}_h^{n_0-1})$ are already computed. Now, set $\tilde{\boldsymbol{f}} = \tilde{\boldsymbol{f}}^{n_0}$, $j_c(x) = j_c(x, \theta(x, t_{n_0}))$ and let $\boldsymbol{E}_h := \boldsymbol{E}_h^{n_0} \in \boldsymbol{V}_h$ be the unique solution of (4.6) for $n = n_0$. Furthermore, $\boldsymbol{B}_h = \boldsymbol{B}_h^{n_0} \in \boldsymbol{H}_0(\mathrm{div} = 0)$ is given by (4.4).

A classical result from the theory of VIs yields the existence of a Lagrange-multiplier for (4.6) (cf. [12]). Thus, there exists a $\boldsymbol{\theta}_h \in \boldsymbol{L}^\infty(\Omega)$ such that

$$\begin{cases} a(\boldsymbol{E}_h, \boldsymbol{v}_h) + \displaystyle\int_\Omega \boldsymbol{\theta}_h \cdot \boldsymbol{v}_h \, dx = \langle \tilde{\boldsymbol{f}}, \boldsymbol{v}_h \rangle \quad \forall \boldsymbol{v}_h \in \boldsymbol{V}_h \\ |\boldsymbol{\theta}_h(x)| \le j_c(x), \quad \boldsymbol{\theta}_h(x) \cdot \boldsymbol{E}_h(x) = j_c(x)|\boldsymbol{E}_h(x)| \text{ for a.e. } x \in \Omega. \end{cases} \tag{5.1}$$

Due to the non-differentiability and the VI structure, we have to employ an additional regularization technique to j. Therefore, we introduce the Moreau–

Yosida regularization $\eta_\lambda : \mathbb{R}^3 \to \mathbb{R}$ of $|\cdot|$ by

$$\eta_\lambda(x) = \begin{cases} |x| - \dfrac{\lambda}{2} & \text{, if } |x| \geq \lambda \\ \dfrac{1}{2\lambda}|x|^2 & \text{, else} \end{cases} \tag{5.2}$$

which is continuously differentiable for every $\lambda > 0$. For an arbitrary Hilbert space $\{H, (\cdot,\cdot)_H\}$, the Moreau–Yosida regularization $\varphi_\lambda \colon H \to \mathbb{R}$ for a convex and l.s.c. function $\varphi \colon H \to \mathbb{R}$ is given by

$$\varphi_\lambda(x) := \min_{v \in H} \frac{\|v - x\|_H^2}{2\lambda} + \varphi(v). \tag{5.3}$$

Following Fermat's principle, the unique minimizer x_λ of (5.3) for $x \in H$ satisfies

$$0 \in \frac{1}{\lambda}(x_\lambda - x) + \partial\varphi(x_\lambda) \quad \Rightarrow \quad x_\lambda = (\mathbf{I} + \lambda\partial\varphi)^{-1}(x).$$

Therefore, $x \mapsto x_\lambda$ corresponds to the resolvent $\boldsymbol{J}_\lambda$ defined in (3.9) for $H = \mathbf{X}$. This shows that the Moreau–Yosida regularization is related to the Yosida approximation (3.9) used in the existence analysis (Sect. 3). In fact, it is well-known [3, Proposition 12.30] that the Yosida approximation for the subdifferential of every proper, convex, and l.s.c. function φ is the Gâteaux-derivative of the Moreau–Yosida regularization φ_λ.

Let us now briefly discuss the equivalence of φ_λ and η_λ for our case, i.e., with $H = \mathbb{R}^3$ and $\varphi = |\cdot|$. Thus, fix $x \neq 0$. We may split the minimization problem in (5.3) as follows:

$$\begin{aligned} \varphi_\lambda(x) &\underbrace{=}_{(5.3)} \min_{v \in \mathbb{R}^3} \frac{1}{2\lambda}|v|^2 - \frac{1}{\lambda} v \cdot x + \frac{1}{2\lambda}|x|^2 + |v| \\ &= \min_{r \geq 0} \min_{v \in \mathbb{R}^3, |v| = r} \frac{1}{2\lambda} r^2 - \frac{1}{\lambda} v \cdot x + \frac{1}{2\lambda}|x|^2 + r. \end{aligned} \tag{5.4}$$

For every fixed $r \geq 0$ the Cauchy–Schwarz inequality yields that the inner minimization problem in (5.4) obtains the minimizer

$$v_{\min}(r) = r\frac{x}{|x|}.$$

Hence,

$$\varphi_\lambda(x) = \min_{r \geq 0} \frac{1}{2\lambda}(r - |x|)^2 + r. \tag{5.5}$$

Now, by standard calculus arguments, we may compute the minimizer of (5.5). It is given by

$$r_{\min} = \max(0, |x| - \lambda). \tag{5.6}$$

Finally, inserting $r_{\min}$ into (5.5) implies that $\varphi_\lambda(x) = \eta_\lambda(x)$ for every $x \neq 0$. The case $x = 0$ is trivial.

For $\lambda > 0$ let us now define $j_\lambda \colon \boldsymbol{L}^2_\epsilon(\Omega) \to \mathbb{R}$ by

$$j_\lambda(\boldsymbol{v}) \coloneqq \int_\Omega j_c(x)\eta_\lambda(\boldsymbol{v}(x))\,dx$$

and consider the regularized problem of finding $\boldsymbol{E}_h^\lambda \in \boldsymbol{V}_h$ such that

$$a(\boldsymbol{E}_h^\lambda, \boldsymbol{v}_h - \boldsymbol{E}_h^\lambda) + j_\lambda(\boldsymbol{v}_h) - j_\lambda(\boldsymbol{E}_h^\lambda) \geq \langle \tilde{\boldsymbol{f}}, \boldsymbol{v}_h - \boldsymbol{E}_h^\lambda \rangle \quad \forall \boldsymbol{v}_h \in \boldsymbol{V}_h. \tag{5.7}$$

Thanks to the Gâteaux-differentiability of $j_\lambda \colon \boldsymbol{L}^2_\epsilon(\Omega) \to \mathbb{R}$, the unique solution $\boldsymbol{E}_h^\lambda \in \boldsymbol{V}_h$ to (5.7) is uniquely characterized by

$$\begin{cases} a(\boldsymbol{E}_h^\lambda, \boldsymbol{v}_h) + \displaystyle\int_\Omega \boldsymbol{\theta}_h^\lambda \cdot \boldsymbol{v}_h\,dx = \langle \tilde{\boldsymbol{f}}, \boldsymbol{v}_h \rangle \quad \forall \boldsymbol{v}_h \in \boldsymbol{V}_h \\ \boldsymbol{\theta}_h^\lambda(x) = j_c(x) \dfrac{\gamma \boldsymbol{E}_h^\lambda(x)}{\max\{1, \gamma |\boldsymbol{E}_h^\lambda(x)|\}} \quad \text{for a.e. } x \in \Omega. \end{cases} \tag{5.8}$$

Of course, by employing a regularization technique, we generate an additional error. For our specific approach, this error is straightforwardly computed. In fact, if we subtract (5.8) from (5.1), we obtain for $\boldsymbol{v}_h = \boldsymbol{E}_h - \boldsymbol{E}_h^\lambda$

$$a(\boldsymbol{E}_h - \boldsymbol{E}_h^\lambda, \boldsymbol{E}_h - \boldsymbol{E}_h^\lambda) = \int_\Omega \left(\boldsymbol{\theta}_h - \boldsymbol{\theta}_h^\lambda\right) \cdot \left(\boldsymbol{E}_h^\lambda - \boldsymbol{E}_h\right) dx. \tag{5.9}$$

Thus, with the properties of $\boldsymbol{\theta}_h, \boldsymbol{\theta}_h^\lambda$ we may estimate the right-hand side in (5.9) to obtain

$$\|\boldsymbol{E}_h - \boldsymbol{E}_h^\lambda\|_{\boldsymbol{H}_0(\mathbf{curl})} \leq \frac{C}{\sqrt{\lambda}} \quad \forall \lambda > 0$$

with a constant $C > 0$, independent of h and λ.

Now, the numerical approach is to fix $\lambda \gg 0$ and solve the nonlinear system (5.8) in each time-step. An efficient way to compute the solution to (5.8) is to use the semismooth Newton method (see [15]). It is well-known that the max-function from $\mathbb{R}^n$ to $\mathbb{R}^n$ satisfies the necessary regularity (see [13]) such that we may apply the SSN-method for (5.8) and obtain local superlinear convergence.

6 Numerical Experiments

We finish this report by presenting the accuracy of our algorithm with a physical example from type-II superconductivity. We drop the time-dependence of j_c and consider a constant critical current density $j_c \in \mathbb{R}^+$. We specify the numerical setup as follows: The hold-all domain is the cube $\Omega = (-1, 1)^3$ and $T = 10$. We take a wire given by

$$\Omega_p := \left\{(x, y, z) \in \Omega : \sqrt{x^2 + z^2} \leq 0.1\right\}$$

carrying a current. It can be described by the function $\boldsymbol{f} : \Omega \times [0, T] \to \mathbb{R}^3$ with

$$\boldsymbol{f}(x, y, z, t) := \begin{cases} 1/R\,(0, (t + 0.1), 0) & \text{for } (x, y, z) \in \Omega_p, \\ 0 & \text{for } (x, y, z) \notin \Omega_p. \end{cases}$$

The constant $R > 0$ denotes the electrical resistance of the pipe ($R = 10$). Moreover, for the sake of simplicity, the material parameters are chosen $\epsilon = \mu = 1$. If we do not include a superconductor in this setup, then the wire induces magnetic field which comes in the shape of co-centric circles around the wire (see Fig. 1b). The implementation is done with the finite element framework FEniCS and ParaView is our visualization tool.

Let us now place a type-II superconducting box with side length 0.1 next to the wire (see Fig. 1a). For the time-discretization, we choose $\tau = 1/20$. The uniform triangulation with $h = 1/20$ gets refined around the wire and the superconductor such that we obtain roughly 200,000 cells. The corresponding finite element spaces $\boldsymbol{V}_h$ and $\boldsymbol{W}_h$ have roughly 230,000 and 580,000 degrees of freedom (DoFs), respectively. In Fig. 2 we present 2d slices of the original 3d plots where the superconductor is left of the wire. The time-step is denoted by $n \in \{0, \ldots, N - 1\}$ with $N = 1/\tau = 20$.

As we start with a rather weak current strength $|\boldsymbol{f}| = 0.01$, we can observe the full Meissner–Ochsenfeld effect in the first time-step (see Fig. 2a). In Fig. 2b the magnetic field begins penetrating into the material. That is, the superconductor is in its *mixed state*. With an increasing current strength, the magnetic field strength rises

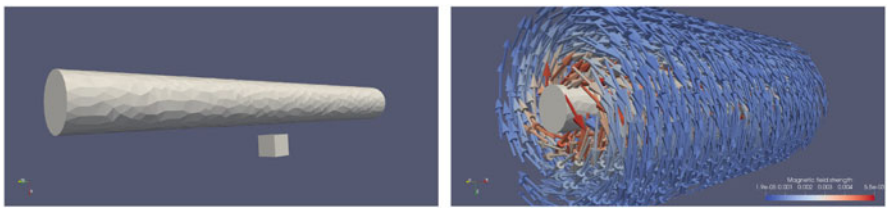

Fig. 1 Left: Wire and superconductor. Right: Magnetic field lines (glyphs) without superconductor

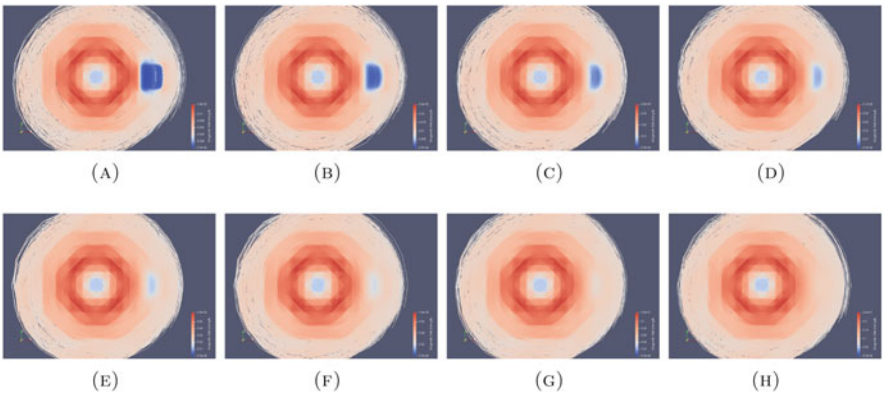

Fig. 2 2d-slices of the magnetic field with a superconductor below its critical temperature. (**a**) $n = 0$. (**b**) $n = 1$. (**c**) $n = 2$. (**d**) $n = 3$. (**e**) $n = 4$. (**f**) $n = 5$. (**g**) $n = 9$. (**h**) $n = 19$

and the superconductor expels less and less field lines (see Fig. 2c–f). For $n = 9$ (Fig. 2g) the superconducting state is almost completely broken down. Finally, in the last iteration (see Fig. 2h) the Meissner–Ochsenfeld-effect is completely broken down. We also note that the superconductor is in its mixed state during most of the time-steps. In fact, the magnetic field strength has to be more than 10 times higher than the initial one in order to destroy the superconducting state.

Acknowledgments This work was supported by the German Research Foundation Priority Program DFG SPP 1962 "Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization", Project YO 159/2-1.

References

1. J. Aponte, H.C. Abache, A. Sa-Neto, and M. Octavio. Temperature dependence of the critical current in high-T_c superconductors. *Phys. Rev. B: Condens. Matter; (United States)*, 39:4, 2 1989.
2. J. W. Barrett and L. Prigozhin. Sandpiles and superconductors: Nonconfoming linear finite element approximations for mixed formulations of quasi-variational inequalities. *IMA J. Numer. Anal.*, 35 (2015), pp. 1–38.
3. H. H. Bauschke and P. L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Springer, 2017
4. C. P. Bean. Magnetization of hard superconductors. *Phys. Rev. Lett.*, 8:250–253, Mar 1962.
5. C. P. Bean, Magnetization of high-field superconductors. *Rev. Modern Phys.*, 36:31–39. 1964.
6. S. Chapman. A hierarchy of models for type-II superconductors. *SIAM Rev.*, 42 (2000), pp.555–598.
7. G. Deutscher and K. A. Müller. Origin of superconductive glassy state and extrinsic critical currents in high-T_c oxides. *Phys. Rev. Lett.; (United States)*, 59:15, 1987.
8. C. M. Elliot and Y. Kashima. A finite-element analysis of critical-state model for type-II superconductivity in 3D. *IMA J. Numer. Anal.*, 27 (2007), pp. 293–331.

9. C. M. Elliot, D. Kay, and V. Styles. A finite element approximation of a variational inequality formulation of Bean's model for superconductivity. *SIAM J. Numer. Anal.*,42 (2004), pp. 1324–1341.
10. A. Ern and J.-L. Guermond. Finite element quasi-interpolation and best approximation. *ESAIM: M2AN*, 51(4):1367–1385, 2017.
11. A. Ern and J.-L. Guermond. Analysis of the edge finite element approximation of the Maxwell equations with low regularity solutions. *Computers & Mathematics with Applications*, 75(3):918 – 932, 2018.
12. R. Glowinski, J. L. Lions, and R. Trémolières. *Numerical Analysis of Variational Inequalities*. Studies in Mathematics and its Applications. Elsevier Science, 1981.
13. M. Hintermüller, K. Ito, K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, 13:3 (2002), pp. 865–888.
14. R. Hiptmair. Finite elements in computational electromagnetism. *Acta Numerica*, 11(1):237–339, 2002.
15. K. Ito and K. Kunisch, Lagrange multiplier approach to variational problems and applications, *Society for Industrial and Applied Mathematics*, 2008.
16. F. Jochmann, On a first-order hyperbolic system including Bean's model for superconductors with displacement currents, *J. Differential Equations*, 246:2151–2191, 2009.
17. F. Jochmann, Well-posedness for Bean's critical state model with displacement current, *J. Math. Anal. Appl.*, 362:505–513, 2010.
18. J. L. Lions and G. Stampacchia. Variational inequalities. *Communications on Pure and Applied Mathematics*, 20(3):493–519, 1967.
19. Ch. G. Makridakis and P. Monk. Time-discrete finite element schemes for Maxwell's equations. *RAIRO Modél. Math. Anal. Numér.*, 29(2):171–197, 1995.
20. P. Monk. Finite Element Methods for Maxwell's Equations. *Numerical Analysis and Scientific Computation*. Clarendon Press, 2003.
21. A. Pazy, Semigroups of linear operators and applications to partial differential equations, *Applied Mathematical Sciences*, vol. 44, Springer-Verlag, New York, 1983.
22. L. Prigozhin. On the Bean critical-state model in superconductivity. *European Journal of Applied Mathematics*, 7(3):237–247, 1996.
23. M. Winckler and I. Yousept, Fully discrete scheme for Bean's critical-state model with temperature effects in superconductivity, *SIAM J. Numer. Anal.*, 57(6):2685–2706, 2019.
24. I. Yousept, Optimal Control of Quasilinear $\boldsymbol{H}(\mathbf{curl})$-Elliptic Partial Differential Equations in Magnetostatic Field Problems, *SIAM J. Control Optim.*, 51(5):3624–3651, 2013.
25. I. Yousept. Optimal control of non-smooth hyperbolic evolution Maxwell equations in type-II superconductivity. *SIAM J. Control Optim.*, 55(4):2305–2332, 2017.
26. I. Yousept, Hyperbolic Maxwell variational inequalities for Bean's critical state model in type-II superconductivity, *SIAM J. Numer. Anal.*, 55(5):2444–2464, 2017.
27. I. Yousept and J. Zou. Edge element method for optimal control of stationary Maxwell system with Gauss law. *SIAM J. Numer. Anal.*, 55(6):2787–2810, 2017.
28. I. Yousept, Hyperbolic Maxwell variational inequalities of the second kind, *ESAIM: COCV*, 26, Paper No. 34, 2020.