Optimal Control of Static Contact in Finite Strain Elasticity

Von der Universität Bayreuth
zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigte Abhandlung

von

Matthias Stöcklein

aus Bamberg

1. Gutachter: Prof. Dr. Anton Schiela
2. Gutachter: Prof. Dr. Daniel Wachsmuth

Tag der Einreichung: 03.08.2020
Tag des Kolloquiums: 15.12.2020

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor Prof. Dr. Anton Schiela for his continuous support and valuable guidance throughout my research. His expertise, unfailing motivation, and joyful nature made the completion of this thesis possible.
Furthermore, I would like to thank Prof. Dr. Daniel Wachsmuth, who agreed to evaluate this thesis as a second reviewer. I would also like to thank the other members of the examination board Prof. Dr. Lars Grüne and Prof. Dr. Thomas Kriecherbauer.
I thank all members of the Chair of Applied Mathematics and further colleagues: Sigrid Kinder, Dr. Michael Baumann, Dr. Philipp Braun, Matthias Höger, Rüdiger Kempf, Dr. Simon Pirkelmann, Bastian Pötzl, Dr. Marleen Stieler, Arthur Fleig, and Lisa Krügel.
Special thanks go to Dr. Georg Müller, Dr. Robert Baier, Tobias Sproll, and Manuel Schaller, whose assistance with the software development part of my thesis proved invaluable. Without their continuous help, this thesis would not have been possible.
I also have to mention Dr. Julian Alberto Ortiz Lopez, with whom I had the privilege of sharing our office. He was always keeping up the spirit and sometimes believed more in my work than I did. Thank you for the great time we had.
I also have to thank Philipp Dull, Christian Brandt, and Sven Rupp, who accompanied me throughout my studies, and who are now teaching the next generation of mathematicians.
Besides my math studies, many people have been with me over the past ten years in Bayreuth, especially, Tanja Höfer, Tobias Haupt, Tobias Michlik, Oliver Gunzelmann, Andre Kreyer, Markus Hennemann, and Peter Neundorfer. Thank you all for making my time in Bayreuth so enjoyable.
To all my friends, your care helped me to overcome setbacks and to stay focused on my study. I greatly value your friendship and I deeply appreciate your belief in me.

Finally, I must express my gratitude to my family and, especially to my parents. You provided me with unfailing support and continuous encouragement during my years of study. This accomplishment would not have been possible without you.

Zusammenfassung

Optimale Steuerung von nichtlinear elastischen Kontaktproblemen führt zu einem nichtkonvexen, beschränkten Bilevel-Optimierungsproblem. Die Lösungen des untergeordneten Problems müssen nicht eindeutig sein und die Bedingungen erster Ordnung gelten nur für sehr eingeschränkte Settings. Zudem implizieren die Kontaktbeschränkungen eine Nicht-Glattheit, was zu einem höchst anspruchsvollen Problem mit wenig Struktur führt. Ziel dieser Arbeit ist es, die vorhandenen Ergebnisse zur Optimalsteuerung für nichtlineare Elastizität auf den Fall mit Kontaktbeschränkungen zu erweitern und spezialisierte und effiziente Lösungsalgorithmen zu entwickeln.

Zunächst werden die Kontaktbeschränkungen mithilfe einer *normal compliance*-Regularisierung relaxiert. Für das regularisierte elastische Kontaktproblem wird die Konvergenz der Lösungen gezeigt und es werden entsprechende Konvergenzraten ermittelt, die auch in der Optimalsteuerung Anwendung finden. Zusätzlich ergibt sich daraus auch ein regularisiertes Optimalsteuerungsproblem. Die Existenz von Lösungen wird sowohl für das regularisierte als auch für das ursprüngliche Problem nachgewiesen. Im Gegensatz zu der vorherigen Analyse ist der Nachweis der Konvergenz der Lösungen hier weitaus schwieriger und zwei mögliche Ansätze werden vorgestellt, um diesen zu erbringen. Unter strikten Annahmen können die strukturellen Probleme überwunden werden und die Konvergenz von Lösungen kann gezeigt werden. Diese Annahmen sind jedoch bei Anwendungen schwer zu verifizieren. Daher wird eine modifizierte Regularisierung eingeführt, um ähnliche Ergebnisse ohne derartige Einschränkungen zu erreichen.

Das numerische Lösen von Optimalsteuerungsproblemen mit nichtlinearer Elastizität erfordert robuste nichtlineare Löser. Daher ist es erforderlich die Energieminimierung im untergeordneten Problem durch die formale Bedingung erster Ordnung zu ersetzen, um ein bewährtes affin-kovariantes *composite step*-Verfahren anzuwenden. Des Weiteren wird zum Lösen der resultierenden linearen Systeme ein neuer iterativer Löser vorgestellt, der auf einem projizierten CG-Verfahren basiert. Dieser Algorithmus berücksichtigt mögliche Ungenauigkeiten und Nicht-Konvexitäten und hat die gleichen Konvergenzeigenschaften wie ein allgemeines Gradientenverfahren. Die Kombination mit einem Pfad-Verfolgungs-Verfahren ermöglicht es, Lösungen des ursprünglichen Optimalsteuerungsproblems mit Kontaktbeschränkungen zu approximieren. Außerdem wird eine neue nichtlineare Update-Strategie für nichtlinear-elastische Probleme vorgestellt.

Abstract

Optimal control of nonlinear elasticity with contact constraints yields a non-convex constrained bilevel optimization problem. For the lower level problem, solutions do not have to be unique, and corresponding first order conditions only hold for very restrictive settings. Further, the contact constraints add non-smoothness, resulting in a highly challenging problem with a severe lack of structure. The main goal of this thesis is to extend existing results in optimal control of nonlinear elasticity to the contact constrained case and to develop specialized and efficient solution algorithms.

First, the contact constraints are relaxed by deploying a variant of the normal compliance method. For the regularized elastic contact problem, the convergence of solutions is shown, and corresponding rates are established, which also contribute to the analysis of optimal control. Additionally, this also yields a regularized optimal control problem. The existence of solutions is proven for the original optimal control problem and the regularized one. In contrast to before, verifying convergence of solutions is a delicate matter, and two approaches are presented to achieve this. Under strong assumptions, the lack of structure can be overcome, and convergence is shown. However, these assumptions are difficult to verify in applications. Therefore, a modified regularization is introduced to establish similar results without these restrictions.

Solving optimal control problems of nonlinear elasticity requires robust nonlinear solvers. Here, the energy minimizing property in the lower level problem is replaced by its formal first order condition to apply a proven affine covariant composite step method. Further, to solve the arising linear systems, a new iterative solver based on a projected CG method is introduced. This algorithm takes into account the possible inexactness and non-convexity and has the same convergence properties as a general gradient method. Inserting these approaches into a path-following algorithm facilitates the approximation of solutions to the original contact constrained optimal control problem. Also, a new nonlinear update strategy for nonlinear elastic problems is presented and tested.

Contents

Chapter 1

Introduction

A common problem setting in mechanics is considering a body under load and computing the resulting deformation. However, sometimes the inverse direction is of interest. For a given desired deformation y_d, an external force is required that causes a deformation approximating y_d w.r.t. a suitable objective functional. Those kinds of problems have received increased attention recently. Possible applications include implant design [64], deriving biological models [39, 40], and shape optimization [80, 87]. In those settings, we obtain a bilevel problem of the form

$$\begin{aligned} &\min_{(y,u)\in Y\times U} J(y,u) \\ &s.t.\ y \in \operatorname*{argmin}_{v\in\mathcal{A}} I(v,u), \end{aligned}$$

where

$$Y \times U := W^{1,p}(\Omega) \times L^2(\Gamma_N)$$

and $\mathcal{A}$ is a suitable admissible set for deformations. The lower level problem ensures that each optimal state has to be a minimizer of a hyperelastic energy functional I. This minimizing property describes the deformation of a body subject to an external applied force, denoted by u. Here, we choose a tracking-type functional for J, measuring the distance to a desired deformation. Such optimal control problems pose an interesting and challenging problem since hyperelasticity by itself is already rich in complexity. In particular, solutions do not have to be unique since the corresponding energy functional is usually non-convex. Additionally, any deformation y has to satisfy the orientation-preserving condition

$$\det y > 0$$

at each point in the domain to guarantee at least local invertibility. As a result, this requirement rules out the derivation of first order conditions of I, except for very restrictive settings. Alternative conditions were derived in [7]. Unfortunately, their application in numerical simulations seems to be out of reach so far. Still, the techniques elaborated there can be utilized to study regularization approaches for contact problems.

A thorough theoretical analysis of optimal control of hyperelasticity was first conducted in [64, 66]. Notably, the existence of optimal solutions was shown. The aim of this work is to extend those results to optimal control of contact problems. Contact problems in nonlinear elasticity are already highly challenging and the reader is referred to [18, 56] for an extensive overview of this topic. In order to embed contact problems into an optimal control setting and conduct meaningful numerical experiments, a suitable regularization method is required. A proven approach is the normal compliance method [69, 76], which is applied in this work. This results in a regularized optimal control problem:

$$\begin{aligned} &\min_{(y,u)\in Y\times U} J(y,u) \\ &s.t.\ y \in \operatorname*{argmin}_{v\in\mathcal{A}} I_\gamma(v,u), \end{aligned} \tag{1.1}$$

where $\gamma > 0$ denotes the normal compliance parameter and I_γ the regularized total energy functional. The goal is to verify that solutions of this problem approach solutions of the original one as $\gamma \to \infty$. Common regularization methods in optimal control are the Lavrentiev regularization, cf. [48, 49, 70, 71, 86], and the Moreau-Yosida regularization, cf. [45, 46, 47, 52, 72]. However, due to the bilevel structure and the non-uniqueness of solutions for the lower level problem, established techniques to show convergence do not apply here, forcing us to seek alternative paths. Two approaches are presented here. The first one relies on strong structural assumptions while the second one modifies the normal compliance method to ensure convergence.
To solve (1.1) numerically for large parameters γ, we deploy basic path-following scheme. Path-following methods are widely applied to solve parameter-dependent problems, see, e.g., [23, 45, 46, 47, 93]. As inner solver, the affine covariant composite step method that was developed in [64, 67] is chosen. After applying a suitable discretization, the linear systems in the composite step method are represented by saddle point matrices of the form:

$$H := \begin{pmatrix} M & C^T \\ C & 0 \end{pmatrix}.$$

Projected preconditioned conjugate gradient (PPCG) algorithms provide a proven approach to solve these systems, cf. [33, 64]. For the setting chosen here, these methods require solving certain subsystems exactly, limiting the size of the problems that can be considered. Thus, we construct a specially tailored iterative solver, based on a PPCG method, to overcome the limitations. This algorithm was originally developed by Anton Schiela and Alexander Siegl in cooperation with the author, cf. [96]. As a further requirement, the newly developed method also needs to work for subsystems that are not positive definite, which is the case for nonlinear elastic problems.

Outline

The goal of this thesis is to elaborate a mathematical theory for optimal control of nonlinear elastic contact problems. Further, solution algorithms are developed, based on a

suitable regularization scheme. The validity of these approaches is shown in theory as well as in numerical simulations. The remainder of this chapter presents an outline of this thesis.

Chapter 2 - Contact Problems in Nonlinear Elasticity. This chapter introduces the general setting and the most important theoretical results regarding nonlinear elastic contact problems. Moreover, we conduct a thorough analysis of nonlinear elasticity to obtain a detailed understanding of the problem structure. Many results derived here will be required when considering optimal control problems in Chapter 4.

Chapter 3 - Regularization of the Contact Constraints. To overcome the non-smoothness due to the contact conditions, we will apply the normal compliance method, yielding a regularized contact problem. Additionally, corresponding convergence results and rates are derived. First order conditions in elasticity are briefly addressed as well. The regularization approach presented here is necessary to make the optimal control problems, which are discussed in the next chapter, numerically treatable.

Chapter 4 - Optimal Control of Nonlinear Elastic Contact Problems. Based on the previous results, we conduct a detailed theoretical analysis of optimal control of nonlinear elastic contact problems. Analogously to Chapter 3, the contact conditions are relaxed, yielding a regularized optimal control problem. For this regularization approach, corresponding convergence results are established. At this, we consider two strategies. First, under strong structural assumptions, a convergence result can be shown. Alternatively, we introduce a modified regularization, which yields similar results without too restrictive requirements. This chapter concludes with a brief study of KKT conditions.

Chapter 5 - Numerical Algorithms. This chapter is dedicated to an algorithmic examination. First, a cubic regularization approach is presented to solve regularized contact problems. Also, a new nonlinear update strategy is worked out, aiming to increase performance. This strategy was developed by Julián Ortiz. For optimal control of nonlinear elastic problems, we present an affine covariant composite step method. To address the regularized optimal control problem, a simple path-following method is introduced.

Chapter 6 - A Corrected Inexact Projected Preconditioned Conjugate Gradient Method. To solve the large scale linear systems in the composite step method, we introduce a corrected inexact projected preconditioned conjugate gradient (CIPPCG) algorithm. This algorithm describes an iterative solver that has the same convergence properties as a standard gradient method. Further, it also applies to non-convex problems such as optimal control of nonlinear elasticity.

Chapter 7 - Numerical Examples. For the numerical tests, all problems are discretized via a finite element method. In the first part, we deploy the cubic regularization

approach from Chapter 5 to verify the convergence rates derived in Chapter 3. Additionally, the newly developed nonlinear update strategy is tested.
Thereafter, we couple the composite step algorithm with the CIPPCG method and test the performance of the combined approach for optimal control in nonlinear elasticity. Also, we discuss the limits of this method and possible extensions for future research.
To conclude this chapter, path-following is applied to solve regularized optimal control problems and to approximate solutions to the original problem. At this, the composite step method functions as the inner solver, where the CIPPCG method is utilized to solve the arising linear systems.

Chapter 8 - Conclusion and Outlook. This chapter contains a summary of the presented work and an outlook for future research.

Chapter 2

Contact Problems in Nonlinear Elasticity

Models derived from linear elasticity offer an accessible approach to describe the deformations of materials. Therefore, they apply to a wide range of real-world applications. However, those models are restricted to problems with small deformations. For problems involving large deformations, linear elasticity no longer reflects the physical reality and has to be replaced by more sophisticated approaches. In those cases, nonlinear elasticity can be applied to obtain an accurate representation of real-world problems, see, e.g., [24, 104]. Incorporating nonlinear elasticity yields another layer of complexity in form of nonlinear and non-convex energy functions. Additionally, in the setting of contact problems, the required constraints add non-smoothness to an already challenging problem. The aim of this chapter is to derive a complete mathematical description of contact problems in nonlinear elasticity. The established results serve as the starting point for the analysis of optimal control problems in Chapter 4.

This chapter is structured as follows. In Section 2.1, we introduce the settings and notation that are necessary to analyze general problems in nonlinear elasticity. In Section 2.2, we discuss the equilibrium equations of a deformed body which lay the theoretical foundation to describe material behavior in continuum mechanics.

Section 2.3 addresses different properties of materials which are essential to model real-world materials accurately. At this, we restrict our analysis to hyperelastic materials. In the context of hyperelasticity, the respective equilibrium equations can be transformed into an energy minimization problem. The explicit formulation of the corresponding energy functionals is also discussed in detail. Deriving such formulations is necessary to conduct numerical simulations. Further, the required material properties have considerable consequences for the theoretical analysis. First and foremost, the loss of convexity significantly impedes the theoretical examination of hyperelastic problems. Particularly, the uniqueness of solutions can no longer be expected. As a result, the mathematical implications and the relation to real-world problems have to be addressed.

Section 2.4 extends the setting to incorporate contact constraints. Finally, in Section 2.5, we examine the existence of solutions to hyperelastic contact problems by utilizing

the concept of polyconvex functions, which was established in [6].
In our analysis, we mainly rely on the results from [15, 18]. Additional overviews and introductions to nonlinear elasticity can be found in, e.g., [24, 68, 77, 100]. An examination that is more focused on numerics was conducted in [17]. For the analysis of contact problems particularly, see, e.g., [54, 56, 61, 81, 110, 110, 111].
Parts of this chapter have been published in [95].

2.1 Deformations of three-dimensional bodies

First, we introduce the theoretical setting and notation for general nonlinear elastic problems. Throughout this work, $\Omega \subset \mathbb{R}^3$ denotes a bounded Lipschitz domain in the sense of [74, pp. 4-6]. The closure $\overline{\Omega}$ represents a three-dimensional undeformed body in an equilibrium state and is usually referred to as the reference configuration. The corresponding boundary Γ consists of two disjoint relatively open subsets Γ_D and Γ_N with

$$\Gamma = \overline{\Gamma_D \cup \Gamma_N}$$

such that each segment has a non-zero boundary measure. The first set Γ_D denotes the part of the boundary where Dirichlet boundary conditions are enforced while external pressure loads, modeled as Neumann boundary conditions, are applied on the segment Γ_N. Further, the function

$$y : \overline{\Omega} \to \mathbb{R}^3$$

denotes the deformation of a body to its new deformed configuration $\overline{\Omega}_\mathrm{d}$. Accordingly, we define the deformed boundary

$$\Gamma_\mathrm{d} = \overline{\Gamma_{\mathrm{d},D} \cup \Gamma_{\mathrm{d},N}}$$

as the image of Γ under y. The setting is illustrated in Figure 2.1. In order to be physically meaningful, a deformation y is required to be sufficiently smooth, orientation-preserving, and injective on Ω. We only require injectivity in the interior since self-contact on the boundary must be allowed.

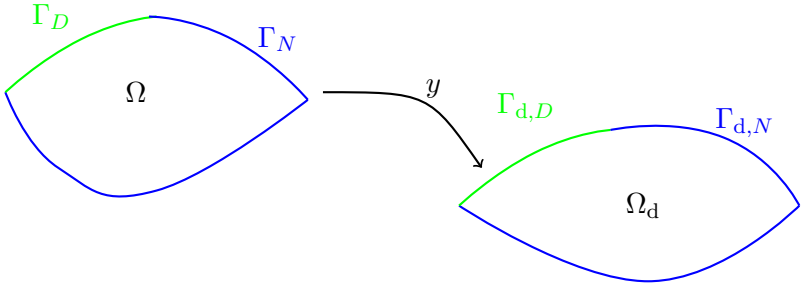

Figure 2.1: Deformation of a body.

At each point $x \in \Omega$, we define the deformation gradient:

$$\nabla y(x) := \begin{pmatrix} \frac{\partial y_1}{\partial x_1}(x) & \frac{\partial y_1}{\partial x_2}(x) & \frac{\partial y_1}{\partial x_3}(x) \\ \frac{\partial y_2}{\partial x_1}(x) & \frac{\partial y_2}{\partial x_2}(x) & \frac{\partial y_2}{\partial x_3}(x) \\ \frac{\partial y_3}{\partial x_1}(x) & \frac{\partial y_3}{\partial x_2}(x) & \frac{\partial y_3}{\partial x_3}(x) \end{pmatrix}.$$

As a result, the orientation-preserving condition

$$\det \nabla y(x) > 0 \quad \text{for all } x \in \Omega$$

ensures at least local injectivity. For incompressible materials, the stronger condition

$$\det \nabla y(x) = 1 \quad \text{for all } x \in \Omega$$

is necessary to accurately model their behavior. Materials that are not required to satisfy this condition are called compressible materials. Over the course of the following analysis, we will study only compressible materials. In the theory of elasticity, it is often convenient to use the displacement notation. The displacement function

$$\phi : \overline{\Omega} \to \mathbb{R}^3 \tag{2.1}$$

is defined by

$$\phi(x) := y(x) - \mathrm{id}(x) \quad \text{for all } x \in \overline{\Omega},$$

where $\mathrm{id} : \overline{\Omega} \to \overline{\Omega}$ denotes the identity mapping. Again, we define the respective gradient by

$$\nabla \phi(x) := \begin{pmatrix} \frac{\partial \phi_1}{\partial x_1}(x) & \frac{\partial \phi_1}{\partial x_2}(x) & \frac{\partial \phi_1}{\partial x_3}(x) \\ \frac{\partial \phi_2}{\partial x_1}(x) & \frac{\partial \phi_2}{\partial x_2}(x) & \frac{\partial \phi_2}{\partial x_3}(x) \\ \frac{\partial \phi_3}{\partial x_1}(x) & \frac{\partial \phi_3}{\partial x_2}(x) & \frac{\partial \phi_3}{\partial x_3}(x) \end{pmatrix},$$

which satisfies the relation

$$\nabla y(x) = \mathrm{Id} + \nabla \phi(x).$$

Here, Id denotes the identity matrix in $\mathbb{R}^{3\times 3}$. To denote the partial derivative of a function f w.r.t. a direction v, we use the notation ∂_v. If there is no risk of ambiguity, we just write f_v. In the case that f only depends on one argument, the derivative is denoted by f'. The space of all $m \times n$ matrices is denoted by $\mathbb{M}^{m\times n}$. If not stated otherwise, the Frobenius norm

$$\|M\| := \sqrt{\mathrm{tr}(M^T M)} \quad \text{for } M \in \mathbb{M}^{m\times n}$$

is chosen as the matrix norm. For the space of quadratic matrices $\mathbb{M}^{n\times n}$, we use the abbreviation $\mathbb{M}^n$. The subspace of matrices with positive determinant is denoted by

$$\mathbb{M}^n_+ := \{M \in \mathbb{M}^n \mid \det M > 0\}.$$

Additionally, $\mathbb{S}^n$ denotes the space of symmetric matrices with the corresponding subspace of positive definite matrices $\mathbb{S}^n_>$. Further, the space of orthogonal matrices is denoted by $\mathbb{O}^n$ and the subspace of all rotations by

$$\mathbb{O}^n_+ := \{M \in \mathbb{O}^n \mid \det M = 1\}.$$

For the set of normed vectors, we write

$$S_1 := \{v \in \mathbb{R}^3 \mid \|v\| = 1\}.$$

The standard matrix scalar product between two matrices A and B is defined via

$$\langle A, B\rangle := \operatorname{tr}(A^T B).$$

For this, we use the short notation $A \cdot B$. As usual, the standard matrix multiplication is written as AB. Next, we introduce the cofactor matrix.

Definition 2.1 (Cofactor matrix)**.** Let $M \in \mathbb{M}^n$. For each pair of indices (i, j), we denote by $M'_{ij} \in \mathbb{M}^{n-1}$ the matrix that results from deleting the ith row and the jth column of M. Then, the cofactor matrix is defined via

$$(\operatorname{Cof} M)_{ij} := (-1)^{i+j} \det M'_{ij}.$$

In the case of M being invertible, the cofactor matrix satisfies

$$\operatorname{Cof} M = (\det M) M^{-T}.$$

Under certain assumptions, the stored energy function of a hyperelastic material depends on the cofactor matrix of the deformation gradient. This issue is discussed in detail in Section 2.3. Next, we define the principal invariants of a matrix.

Definition 2.2 (Principal invariants)**.** Consider a matrix $M \in \mathbb{M}^3$. Further, let λ_1, λ_2, and λ_3 be the corresponding eigenvalues. Then, the three principal invariants $\mathcal{I}_1(M)$, $\mathcal{I}_2(M)$, and $\mathcal{I}_3(M)$ of the matrix M are defined as follows:

$$\begin{aligned}
\mathcal{I}_1(M) &:= \operatorname{tr} M = \lambda_1 + \lambda_2 + \lambda_3,\\
\mathcal{I}_2(M) &:= \operatorname{tr} \operatorname{Cof} M = \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_1,\\
\mathcal{I}_3(M) &:= \det M = \lambda_1\lambda_2\lambda_3.
\end{aligned}$$

If the relation is clear from the context, we use the short notation $\mathcal{I}_1$, $\mathcal{I}_2$, and $\mathcal{I}_3$. The triple of the principal invariants is denoted by $\mathcal{I}(M)$, or just $\mathcal{I}$, respectively.

Further, we introduce the **right Cauchy-Green tensor**

$$C(y) := \nabla y^T \nabla y,$$

which can be interpreted as a measure of strain. This interpretation is reflected in the fact that for translations and rotations of the reference configuration around the origin,

the right Cauchy-Green strain tensor simplifies to the identity matrix Id. Deformations satisfying these conditions are referred to as rigid deformations. If it is clear from the context, we use the short notation C. Related to that, we define the **Green-St Venant strain tensor**

$$E(\phi) := \frac{1}{2}(\nabla\phi^T + \nabla\phi + \nabla\phi^T\nabla\phi).$$

The Green-St Venant strain tensor measures the deviation between a given deformation and a rigid motion. This becomes apparent through the equivalent definition

$$E(\phi) := \frac{1}{2}(C(\mathrm{id}+\phi) - \mathrm{Id}).$$

Note that the tensor E is invariant under rotations and translations. When considering only small deformations, it can be sufficient to reduce the Green-St Venant strain tensor E to its linearization

$$\varepsilon(\phi) := \frac{1}{2}(\nabla\phi^T + \nabla\phi).$$

Besides being restricted to problems exhibiting small deformations, this linearized tensor is also no longer invariant under rotations. Nevertheless, there exists a wide field of applications utilizing the linearized strain tensor, see, e.g., [34, 41, 89, 90].
Next, we consider how to describe deformations of bodies as solutions to mathematical problems. At this, deriving equilibrium equations serves as the starting point.

2.2 Equilibrium equations

The static equilibrium of a body, subjected to external forces, is described by the stress principle of Euler and Cauchy, which forms the foundation of continuum mechanics, cf. [15, Axiom 2.2-1].

Axiom 2.3. *Let $\overline{\Omega}_\mathrm{d}$ denote the deformed configuration of a body, where the respective applied forces are represented by densities $f_\mathrm{d} : \Omega_\mathrm{d} \to \mathbb{R}^3$ and $u_\mathrm{d} : \Gamma_{\mathrm{d},N} \to \mathbb{R}^3$. Then, there exists a vector field*

$$t_\mathrm{d} : \overline{\Omega}_\mathrm{d} \times S_1 \to \mathbb{R}^3$$

such that:

1. *For any subdomain $A_\mathrm{d} \subseteq \overline{\Omega}_\mathrm{d}$, and at any point $x_\mathrm{d} \in \Gamma_{\mathrm{d},N} \cap \partial A_\mathrm{d}$ where the unit outer normal vector n_d to $\Gamma_{\mathrm{d},N} \cap \partial A_\mathrm{d}$ exists, the equation*

$$t_\mathrm{d}(x_\mathrm{d}, n_\mathrm{d}) = u_\mathrm{d}(x_\mathrm{d})$$

   holds.

2. ***Axiom of force balance:*** *For any subdomain $A_\mathrm{d} \subseteq \overline{\Omega}_\mathrm{d}$,*

$$\int_{A_\mathrm{d}} f_\mathrm{d}(x_\mathrm{d})\, dx_\mathrm{d} = -\int_{\partial A_\mathrm{d}} t_\mathrm{d}(x_\mathrm{d}, n_\mathrm{d})\, ds_\mathrm{d}.$$

3. ***Axiom of moment balance:*** *For any subdomain* $A_\mathrm{d} \subseteq \Omega_\mathrm{d}$,

$$\int_{A_\mathrm{d}} x_\mathrm{d} \times f_\mathrm{d}(x_\mathrm{d})\, dx_\mathrm{d} = -\int_{\partial A_\mathrm{d}} x_\mathrm{d} \times t_\mathrm{d}(x_\mathrm{d}, n_\mathrm{d})\, ds_\mathrm{d}.$$

Here, $\times$ *denotes the cross product.*

From here, we obtain one of the most significant results in continuum mechanics.

Theorem 2.4 (Cauchy's theorem)**.** *Let* $f_\mathrm{d} : \overline{\Omega}_\mathrm{d} \to \mathbb{R}^3$ *and* $u_\mathrm{d} : \Gamma_{\mathrm{d},N} \to \mathbb{R}^3$ *be force densities, where* f_d *is continuous. Further, the Cauchy stress vector field*

$$t_\mathrm{d} : \overline{\Omega}_\mathrm{d} \times S_1 \to \mathbb{R}^3$$

is assumed to be continuously differentiable w.r.t. the variable $x_\mathrm{d} \in \overline{\Omega}_\mathrm{d}$ *for each* $n \in S_1$*. Additionally, it is assumed to be continuous w.r.t. the variable* $n \in S_1$ *for each* $x_\mathrm{d} \in \overline{\Omega}_\mathrm{d}$*. Then, the axioms of force and moment balance imply the existence of a continuously differentiable tensor field*

$$T_\mathrm{d} : \overline{\Omega}_\mathrm{d} \to \mathbb{M}^3$$

such that the Cauchy stress vector satisfies

$$t_\mathrm{d}(x_\mathrm{d}, n) = T_\mathrm{d}(x_\mathrm{d})n \qquad \text{for all } x_\mathrm{d} \in \overline{\Omega}_\mathrm{d} \text{ and all } n \in S_1.$$

In addition, the following equations hold:

$$-\operatorname{div}_\mathrm{d} T_\mathrm{d}(x_\mathrm{d}) = f_\mathrm{d}(x_\mathrm{d}) \qquad \text{for all } x_\mathrm{d} \in \Omega_\mathrm{d}, \tag{2.2}$$

$$T_\mathrm{d}(x_\mathrm{d}) = T_\mathrm{d}(x_\mathrm{d})^T \qquad \text{for all } x_\mathrm{d} \in \Omega_\mathrm{d}, \tag{2.3}$$

$$T_\mathrm{d}(x_\mathrm{d})n_\mathrm{d} = u_\mathrm{d}(x_\mathrm{d}) \qquad \text{for all } x_\mathrm{d} \in \Gamma_{\mathrm{d},N}. \tag{2.4}$$

Here, n_d *denotes the unit outer normal vector along the deformed boundary segment* $\Gamma_{\mathrm{d},N}$*. The tensor* $T_\mathrm{d}(x_\mathrm{d})$ *is called the Cauchy stress tensor at the point* $x_\mathrm{d} \in \Omega_\mathrm{d}$*.*

Proof. See [15, Proof of Theorem 2.3-1]. □

The main consequence of this theorem is the coupling of the external applied forces f_d and u_d with the tensor T_d by partial differential equations and boundary conditions. Additionally, we obtain the divergence structure of the resulting equations which allows variational formulations. However, these equations are formulated in dependence of the unknown deformed configuration Ω_d. In order to transfer these equations into the reference configuration Ω, the Piola transform is applied.

2.2.1 Piola transform

Definition 2.5 (Piola transform)**.** Consider a mapping $\widetilde{T}_\mathrm{d} : \overline{\Omega}_\mathrm{d} \to \mathbb{M}^3$. Then, the Piola transform $T : \overline{\Omega} \to \mathbb{M}^3$ of $\widetilde{T}_\mathrm{d}$ at a point $x \in \overline{\Omega}$ is defined by

$$T(x) := (\det \nabla y(x))\widetilde{T}_\mathrm{d}(x_\mathrm{d})\nabla y(x)^{-T}, \quad x_\mathrm{d} = y(x).$$

In the case that $\widetilde{T}_{\mathrm{d}}$ is the Cauchy stress tensor, its Piola transform is called the first Piola-Kirchhoff stress tensor. Note that this tensor is not symmetric in general. However, it can sometimes be convenient to work with symmetric tensors in order to simplify the constitutive equations. This can be achieved by utilizing the second Piola-Kirchoff stress tensor

$$\Sigma(x) := \nabla y(x)^{-1} T(x) = (\det \nabla y(x)) \nabla y(x)^{-1} T_{\mathrm{d}}(x_{\mathrm{d}}) \nabla y(x)^{-T}, \quad x_{\mathrm{d}} = y(x).$$

Nevertheless, we restrict our examination mostly to the first Piola-Kirchhoff stress tensor due to its relevance in hyperelasticity. Before we can apply the Piola transform to the equilibrium equations (2.2)-(2.4), we have to consider the retraction of the applied forces.

2.2.2 Applied forces

Our goal is to associate the applied forces f_{d} and u_{d} with forces, denoted by f and u, that act on the reference configuration. Moreover, this association has to be consistent with the Piola transform applied to Equations (2.2)-(2.4). With that in mind, we define $f : \Omega \to \mathbb{R}^3$ and $u : \Gamma_N \to \mathbb{R}^3$ as follows:

$$f(x) := (\det \nabla y(x)) f_{\mathrm{d}}(x_{\mathrm{d}}), \quad x \in \Omega,\, x_{\mathrm{d}} = y(x)$$

and

$$u(x) := \det \nabla y(x) |\nabla y(x)^{-T} n(x)| u_{\mathrm{d}}(x_{\mathrm{d}}), \quad x \in \Gamma_N,\, x_{\mathrm{d}} = y(x),$$

where n denotes the unit outer normal vector field of $\overline{\Omega}$. This definition yields the equalities

$$f(x)\, dx = f_{\mathrm{d}}(x_{\mathrm{d}})\, dx_{\mathrm{d}} \quad \text{and} \quad u(x)\, ds = u_{\mathrm{d}}(x_{\mathrm{d}})\, ds_{\mathrm{d}}.$$

Here, we use the notation from [15, Chapter 2] and the formulas for the deformed volume elements

$$dx_{\mathrm{d}} = \det \nabla y(x)\, dx \quad \text{and} \quad ds_{\mathrm{d}} = \det \nabla y(x) |\nabla y(x)^{-T} n(x)|\, ds.$$

With these definitions at hand, we can transfer the equilibrium equations back to the reference configuration $\overline{\Omega}$.

Theorem 2.6. *Consider the setting of the boundary value problem* (2.2)-(2.4)*, whereby y denotes the respective deformation of the body. Further, assume that the applied body forces $f : \Omega \to \mathbb{R}^3$ and $u : \Gamma_N \to \mathbb{R}^3$ satisfy $f\, dx = f_{\mathrm{d}}\, dx_{\mathrm{d}}$ and $u\, ds = u_{\mathrm{d}}\, ds_{\mathrm{d}}$. Then, the first Piola-Kirchoff stress tensor, defined by*

$$T(x) := (\det \nabla y(x)) T_{\mathrm{d}}(x_{\mathrm{d}}) \nabla y(x)^{-T},$$

satisfies the following equations:

$$-\operatorname{div} T(x) = f(x) \qquad \text{for all } x \in \Omega, \tag{2.5}$$

$$\nabla y(x) T(x)^T = T(x) \nabla y(x)^T \qquad \text{for all } x \in \Omega, \tag{2.6}$$

$$T(x) n(x) = u(x) \qquad \text{for all } x \in \Gamma_N. \tag{2.7}$$

Here, n denotes the unit outer normal vector field of Γ_N.

Proof. For the proof, see [15, Proof of Theorem 2.6-1]. □

For the analysis of hyperelastic problems in Section 2.3, we have to analyze the structure of applied forces in detail.

Definition 2.7 (Dead load)**.** An applied body force $f_{\mathrm{d}} : \Omega_{\mathrm{d}} \to \mathbb{R}^3$ is called a dead load if the associated function $f : \Omega \to \mathbb{R}^3$, acting on the reference configuration, is independent of the corresponding deformation y.

This definition applies to boundary forces analogously.

Remark 2.8. *Note here that modeling applied forces as dead loads is a mathematical simplification which only holds for a limited number of practical problems. Considering volume forces, the most common example is the gravity field. An example for boundary forces is the simple choice*

$$u_{\mathrm{d}}(x_{\mathrm{d}}) = 0 \quad \textit{for all } x_{\mathrm{d}} \in \Gamma_{\mathrm{d},N}.$$

Consequently, the associated boundary force on the reference configuration has the trivial form

$$u(x) = 0 \quad \textit{for all } x \in \Gamma_N.$$

However, besides this simple setting, the problem structure usually does not allow to model boundary forces as dead loads.

Another important class are conservative forces.

Definition 2.9 (Conservative forces)**.** Let $f_{\mathrm{d}} : \Omega_{\mathrm{d}} \to \mathbb{R}^3$ and $u_{\mathrm{d}} : \Gamma_{\mathrm{d},N} \to \mathbb{R}^3$ be applied body forces acting in the deformed configuration. Assume there exist mappings $\hat{f} : \Omega \times \mathbb{R}^3 \to \mathbb{R}^3$ and $\hat{g} : \Gamma_N \times \mathbb{M}^3_+ \to \mathbb{R}^3$ such that the associated forces in the reference configuration are of the form

$$f(x) = \hat{f}(x, y(x)) \quad \text{for all } x \in \Omega$$

and

$$u(x) = \hat{u}(x, \nabla y(x)) \quad \text{for all } x \in \Gamma_N.$$

Additionally, define the set of sufficiently smooth test functions by

$$V := \{v : \overline{\Omega} \to \mathbb{R}^3 \mid \ v(x) = 0 \quad \text{for all } x \in \Gamma_D\}.$$

Then, f_{d} and u_{d} are called conservative if there exist functions $\hat{F} : \Omega \times \mathbb{R}^3 \to \mathbb{R}$ and $\hat{U} : \Gamma_N \times \mathbb{R}^3 \times \mathbb{M}^3_+ \to \mathbb{R}$ with corresponding integrals

$$E_{\hat{F}}(y) := \int_\Omega \hat{F}(x, y(x))\ dx$$

and

$$E_{\hat{U}}(y) := \int_{\Gamma_N} \hat{U}(x, y(x), \nabla y(x))\ ds$$

such that the Gâteaux derivatives of $E_{\hat{F}}$ and $E_{\hat{U}}$ satisfy

$$E'_{\hat{F}}(y)v = \int_\Omega \hat{f}(x, y(x))v(x)\,dx$$

and

$$E'_{\hat{U}}(y)v = \int_{\Gamma_N} \hat{u}(x, \nabla y(x))v(x)\,ds$$

for all $v \in V$.

Obviously, dead loads are conservative forces while conservative forces provide a more general class of functions. The question of the required structure of applied forces reemerges when we study hyperelastic problems in Section 2.3.
So far, we have succeeded in retracting the problem back to the reference configuration $\overline{\Omega}$. Still, the system (2.5)-(2.7) describes three equations with nine variables. These are the three components of the deformation and the six components of the first Piola-Kirchhoff stress tensor, considering the symmetry of the Cauchy stress tensor. Consequently, the corresponding system is underdetermined. This discrepancy corresponds to the physical interpretation that Equations (2.5)-(2.7) are entirely independent of the material. In order to close this gap, we introduce additional assumptions on the class of admissible materials in order to obtain a well-posed problem.

2.3 Material properties

In this section, we study the necessary physical and mathematical requirements to derive realistic models that describe materials and the corresponding deformations. Thereby, we focus on elastic materials and, in particular, hyperelastic ones. For hyperelastic materials, the system of equilibrium equations can be transformed into an energy minimization problem. From there, we discuss the non-convexity of the corresponding energy functional, and, consequently, the non-uniqueness of solutions, both in theory and in real-world applications.
To overcome the lack of convexity, polyconvex functions are introduced, which provide the theoretical foundation for proving the existence of solutions to hyperelastic problems in Section 2.5. Finally, an explicit model for the elastic energy is elaborated. This model is applied in the numerical tests conducted in Chapter 7.

2.3.1 Elastic materials

Elasticity can be used to model a wide range of materials such as steel, rubber, aluminum, and biological soft tissue. Therefore, it is frequently applied to describe problems related to real-world applications, cf. [24, 34, 41, 89, 90, 104].
First, we introduce the mathematical definition of elasticity and discuss how it affects the results established so far. We call a material elastic if at each point $x_\mathrm{d} = y(x)$ in the deformed domain $\overline{\Omega}_\mathrm{d}$, the Cauchy stress tensor $T_\mathrm{d}(x_\mathrm{d})$ is entirely determined by the

deformation gradient $\nabla y(x)$ at the respective point $x \in \overline{\Omega}$. This yields the following definition.

Definition 2.10 (Elastic material)**.** A material is called elastic if there exists a mapping

$$\widehat{T} : \overline{\Omega} \times \mathbb{M}^3_+ \to \mathbb{S}^3$$

such that for each point $x_\mathrm{d} := y(x)$ in the deformed domain $\overline{\Omega}_\mathrm{d}$, the equation

$$T_\mathrm{d}(x_\mathrm{d}) = \widehat{T}(x, \nabla y(x))$$

is satisfied.

The mapping $\widehat{T}$ is called the response function for the Cauchy stress. As mentioned above, at a fixed point $x_\mathrm{d} \in \overline{\Omega}_\mathrm{d}$, the Cauchy stress tensor T_d does not depend on the function values of the deformation $y(x)$ but only on its gradient. This condition is consistent with our prior analysis since such a dependency would imply that rigid translations can affect the Cauchy stress tensor.
One possible extension to this definition is the inclusion of the dependency on the deformation gradient evaluated at all other points x in $\overline{\Omega}$. This approach gives rise to the theory of nonlocal elasticity, see [28] or [15, Chapter 3]. Nevertheless, classic elasticity has proved to be a suitable tool in describing real-world problems, and thus, it will be the only definition considered in this work. For a detailed discussion of elastic materials, the reader is referred to [15, 77, 100] and the references therein.
Sometimes, only materials are considered whose response is the same at each point. Such materials are called homogeneous. Mathematically, this property can be defined as follows.

Definition 2.11 (Homogeneous material)**.** A material in a reference configuration $\overline{\Omega}$ is called homogeneous if its response function is independent of the particular point $x \in \overline{\Omega}$. Thus, for each $x_\mathrm{d} := y(x) \in \overline{\Omega}_\mathrm{d}$, the response function for the Cauchy stress satisfies

$$T_\mathrm{d}(x_\mathrm{d}) = \widehat{T}(\nabla y(x)).$$

Note that this definition only applies for the reference configuration. If the deformed state is chosen as the reference configuration, this property is not necessarily satisfied anymore.
Again, we want to study elastic problems defined on the reference configuration $\overline{\Omega}$. Thus, we have to examine how the elasticity property affects the first and second Piola-Kirchhoff stress tensor. The respective implications are addressed in the following theorem.

Theorem 2.12. *The elasticity property stated in Definition 2.10 implies the existence of two mappings*

$$\widetilde{T} : \overline{\Omega} \times \mathbb{M}^3_+ \to \mathbb{M}^3 \quad \textit{and} \quad \widetilde{\Sigma} : \overline{\Omega} \times \mathbb{M}^3_+ \to \mathbb{S}^3$$

such that

$$\widetilde{T}(x, M) = (\det M)\widehat{T}(x, M)M^{-T}$$

and

$$\widetilde{\Sigma}(x, M) = (\det M)M^{-1}\widehat{T}(x, M)M^{-T}$$

for all $x \in \overline{\Omega}$ *and* $M \in \mathbb{M}^3_+$. *Additionally, the first and second Piola-Kirchhoff stress tensor satisfy the equations*

$$T(x) = \widetilde{T}(x, \nabla y(x))$$

and

$$\Sigma(x) = \widetilde{\Sigma}(x, \nabla y(x))$$

for all $x \in \overline{\Omega}$.

Proof. See [15, Chapter 3]. □

The mapping $\widetilde{T}$ is called the response function for the first Piola-Kirchoff stress. Accordingly, $\widetilde{\Sigma}$ is referred to as the response function for the second Piola-Kirchhoff stress . Due to its relevance in hyperelasticity, we mainly focus on the first one. The response function $\widetilde{T}$ can be incorporated into equilibrium equations defined on the reference configuration $\overline{\Omega}$.

Lemma 2.13. *Consider the setting of Theorem 2.6 where* y *denotes the deformation of the body. In addition, Dirichlet boundary conditions on* Γ_D *with the corresponding function* $y_D : \Gamma_D \to \mathbb{R}^3$ *are required. Then, there exist functions* $\tilde{f} : \Omega \times \mathbb{R}^3 \to \mathbb{R}^3$ *and* $\tilde{u} : \Gamma_N \times \mathbb{M}^3_+ \to \mathbb{R}^3$ *such that*

$$-\operatorname{div} \widetilde{T}(x, \nabla y(x)) = \tilde{f}(x, y(x)) \qquad \text{for all } x \in \Omega, \tag{2.8}$$

$$\widetilde{T}(x, \nabla y(x))n(x) = \tilde{u}(x, \nabla y(x)) \qquad \text{for all } x \in \Gamma_N, \tag{2.9}$$

$$y(x) = y_D(x) \qquad \text{for all } x \in \Gamma_D. \tag{2.10}$$

Proof. See [15, Chapter 4]. □

Analogously, we can derive equilibrium equations for the response function $\widetilde{\Sigma}$ for the second Piola-Kirchoff stress.

Lemma 2.14. *Consider the setting of Lemma 2.13. Then, there exist functions* $\tilde{f} : \Omega \times \mathbb{R}^3 \to \mathbb{R}^3$ *and* $\tilde{u} : \Gamma_N \times \mathbb{M}^3_+ \to \mathbb{R}^3$ *such that*

$$-\operatorname{div} \nabla y(x)\widetilde{\Sigma}(x, \nabla y(x)) = \tilde{f}(x, y(x)) \qquad \text{for all } x \in \Omega, \tag{2.11}$$

$$\nabla y(x)\widetilde{\Sigma}(x, \nabla y(x))n(x) = \tilde{u}(x, \nabla y(x)) \qquad \text{for all } x \in \Gamma_N, \tag{2.12}$$

$$y(x) = y_D(x) \qquad \text{for all } x \in \Gamma_D. \tag{2.13}$$

Proof. See [15, Chapter 4]. □

Equations (2.11)-(2.13) are of particular interest since they can be utilized to show the existence of solutions, at least under strong structural assumptions. An analysis of this issue is considered in Section 2.5.

2.3.2 Hyperelastic materials

An intuitive interpretation of hyperelasticity is assuming the existence of an inner energy of the body. At this, deformations are the natural consequences of minimizing such an energy when the body is under stress. Consequently, the question arises whether the equilibrium equations can be transformed into a minimization problem.
As the subsequent analysis will show, such a transformation is possible if the response function $\widetilde{T}$ can be written as the derivative of an energy function. This motivates the following definition.

Definition 2.15 (Hyperelastic material)**.** An elastic material is called hyperelastic if there exists a function

$$\hat{W} : \overline{\Omega} \times \mathbb{M}^3_+ \to \mathbb{R}$$

that is differentiable w.r.t. the variable $M \in \mathbb{M}^3_+$ for each $x \in \overline{\Omega}$. Further, it satisfies

$$\widetilde{T}(x, M) = \frac{\partial \hat{W}}{\partial M}(x, M) \quad \text{for all } x \in \overline{\Omega} \text{ and } M \in \mathbb{M}^3_+.$$

The function $\hat{W}$ is commonly referred to as the stored energy function.

Although this definition of hyperelastic materials seems to be motivated purely by mathematical arguments, it is equivalent to the more physical interpretation that the work in closed processes should be positive. This is a widely accepted property of real-world materials. For a more detailed discussion of this issue, the reader is referred to [100] and the references therein.
Given a hyperelastic material and the corresponding conservative applied forces f_d and u_d, then the respective energy functional I reads as follows

$$I(y) := \int_\Omega \hat{W}(x, \nabla y(x))\ dx - \int_\Omega \hat{F}(x, y(x))\ dx - \int_{\Gamma_N} \hat{U}(x, y(x), \nabla y(x))\ ds,$$

where $\hat{F}$ and $\hat{U}$ are defined as in Definition 2.9. In literature, I is usually called the total energy functional. The first term of I is called the strain energy of the material, and we write

$$I_{\mathrm{strain}}(y) := \int_\Omega \hat{W}(x, \nabla y(x))\ dx.$$

It can be shown that minimizers of the energy I also satisfy the equilibrium equations (2.8)-(2.10).

Theorem 2.16. *Consider the setting of Theorem 2.6 for a hyperelastic material subjected to applied conservative forces $f_\mathrm{d} : \Omega \times \mathbb{R}^3 \to \mathbb{R}^3$ and $u_\mathrm{d} : \Gamma_N \times \mathbb{M}^3_+ \to \mathbb{R}^3$. The associated forces are denoted by $\hat{f} : \Omega \times \mathbb{R}^3 \to \mathbb{R}^3$ and $\hat{g} : \Gamma_N \times \mathbb{M}^3_+ \to \mathbb{R}^3$. Further, let $y_D : \Gamma_D \to \mathbb{R}^3$ denote the function corresponding to the Dirichlet boundary conditions and let I denote the total energy functional. Then, each sufficiently smooth mapping ψ from the set*

$$\Psi := \{v : \overline{\Omega} \to \mathbb{R}^3 \mid\ v(x) = y_D(x) \quad \textit{for all } x \in \Gamma_D\}$$

that satisfies

$$I(\psi) = \inf_{v \in \Psi} I(v)$$

solves the following system:

$$\begin{aligned} -\operatorname{div} \frac{\partial \hat{W}}{\partial M}(x, \nabla\psi(x)) &= \hat{f}(x, \psi(x)) && \text{for all } x \in \Omega, \\ \frac{\partial \hat{W}}{\partial M}(x, \nabla\psi(x))n &= \hat{u}(x, \nabla\psi(x)) && \text{for all } x \in \Gamma_N, \\ \psi(x) &= y_D(x) && \text{for all } x \in \Gamma_D. \end{aligned}$$

Proof. See [15, Proof of Theorem 4.1-2]. □

Describing deformations as energy minimizers allows the derivation of existence results while not relying on too regular settings. This topic is discussed thoroughly in Section 2.5. Although hyperelasticity corresponds to the nature of real-world materials, it is not sufficient to derive an explicit formulation of the respective energy functions. In order to achieve this, further material properties have to be taken into consideration.

2.3.3 Material frame-indifference

In contrast to the assumptions made so far, material frame-indifference is an axiomatic property. It states that the Cauchy stress tensor is independent of the particular orthogonal basis in which it is computed. This axiom also has a more general counterpart in physics where this property is assumed to hold for any observable quantity.
In the case of elasticity, only rotations of the chosen and fixed basis have to be considered. Translations of the origin can be ignored since they do not affect the deformation gradient. Describing frame-indifference in purely mathematical terms yields the following formulation.

Axiom 2.17 (Axiom of material frame-indifference)**.** *Let y be a deformation with its corresponding deformed domain $\overline{\Omega}_{\mathrm{d}}$. Further, consider a rotation $R \in \mathbb{O}^3_+$ and the corresponding new deformation $y_{\mathrm{r}} : \overline{\Omega} \to \mathbb{R}^3$, defined by $y_{\mathrm{r}} := Ry$. The rotated domain is denoted by $\overline{\Omega}_{\mathrm{r}}$ and the respective points by $x_{\mathrm{r}} := y_{\mathrm{r}}(x)$. Then, the respective Cauchy stress vector fields $t_{\mathrm{d}} : \overline{\Omega}_{\mathrm{d}} \times S_1 \to \mathbb{R}^3$ and $t_{\mathrm{r}} : \overline{\Omega}_{\mathrm{r}} \times S_1 \to \mathbb{R}^3$ satisfy*

$$t_{\mathrm{r}}(x_{\mathrm{r}}, Rn) = Rt_{\mathrm{d}}(x_{\mathrm{d}}, n) \quad \text{for all } x \in \overline{\Omega} \text{ and } n \in S_1.$$

Frame-indifference naturally adds further requirements for the stored energy function $\hat{W}$, which are discussed in the next theorem.

Theorem 2.18. *The stored energy function $\hat{W}$ of a hyperelastic material satisfies the axiom of frame-indifference if and only if for each point $x \in \overline{\Omega}$,*

$$\hat{W}(x, RM) = \hat{W}(x, M) \quad \text{for all } M \in \mathbb{M}^3_+ \text{ and } R \in \mathbb{O}^3_+,$$

or equivalently, there exists a function

$$\overline{W} : \overline{\Omega} \times \mathbb{S}^3_> \to \mathbb{R}$$

such that

$$\hat{W}(x, M) = \overline{W}(x, M^T M) \quad \textit{for all } M \in \mathbb{M}^3_+.$$

Proof. See [15, Proof of Theorem 4.2-1]. □

The second condition is of particular interest since it states that the stored energy function can be expressed as a function of the right Cauchy-Green strain tensor C. As the analysis in Subsection 2.3.6 will show, frame-indifference already excludes convex functions as candidates for the stored energy function $\hat{W}$, see also [15, Theorem 4.8-1].

2.3.4 Isotropic materials

In physical terms, isotropy means that at each point, the response of a given material does not depend on the direction. In mathematical terms, this property can be described as follows.

Definition 2.19 (Isotropic material)**.** Let $\widehat{T}$ be the response function for the Cauchy stress. An elastic material is isotropic at a point x in $\overline{\Omega}$ if

$$\widehat{T}(x, MR) = \widehat{T}(x, M) \quad \text{for all } M \in \mathbb{M}^3_+ \text{ and } R \in \mathbb{O}^3_+.$$

An elastic material occupying a reference configuration $\overline{\Omega}$ is isotropic if it is isotropic at each point x in $\overline{\Omega}$.

This definition implies that the Cauchy stress tensor remains unchanged when the reference configuration $\overline{\Omega}$ is rotated around the point x. Note that isotropy in the reference configuration is not necessarily carried over to the deformed configuration. Isotropy also yields a more specific characterization of the response function for the Cauchy stress.

Theorem 2.20. *The response function $\widehat{T}$ for the Cauchy stress tensor is isotropic at a point $x \in \overline{\Omega}$ if and only if there exists a mapping $\overline{T}(x, \cdot) : \mathbb{S}^3_> \to \mathbb{S}^3$ such that*

$$\widehat{T}(x, M) = \overline{T}(x, MM^T) \quad \textit{for all } M \in \mathbb{M}^3_+.$$

Proof. See [15, Proof of Theorem 3.4-1]. □

Analogously to the axiom of frame-indifference, isotropy leads to additional conditions for the stored energy function $\hat{W}$. In the context of hyperelasticity, we obtain the following characterization.

Theorem 2.21. *The stored energy function $\hat{W}$ of a hyperelastic material is called isotropic at $x \in \overline{\Omega}$ if and only if*

$$\hat{W}(x, M) = \hat{W}(x, MR) \quad \textit{for all } M \in \mathbb{M}^3_+ \textit{ and } R \in \mathbb{O}^3_+.$$

Proof. The proof can be found in [15, Proof of Theorem 4.3-1]. □

Although isotropy is not necessary to show the existence of solutions to hyperelastic problems, it is a key property to derive an explicit model for stored energy function $\hat{W}$. The first step to go from purely theoretical properties to an explicit representation of the corresponding tensors is achieved in the Rivlin-Ericksen representation theorem.

Theorem 2.22 (Rivlin-Ericksen representation theorem)**.** *Consider a mapping $\check{T} : \mathbb{M}^3_+ \to \mathbb{S}^3$. Then, the conditions*

$$\check{T}(RM) = R\check{T}(M)R^T \quad \text{and} \quad \check{T}(MR) = \check{T}(M) \quad \text{for all } M \in \mathbb{M}^3_+ \text{ and } R \in \mathbb{O}^3_+$$

are equivalent to

$$\check{T}(M) = \check{T}(MM^T) \quad \text{for all } M \in \mathbb{M}^3_+,$$

where the mapping $\check{T} : \mathbb{S}^3_> \to \mathbb{S}^3$ is of the form

$$\check{T}(A) = \alpha_0(\mathcal{I}(A))\,\mathrm{Id} + \alpha_1(\mathcal{I}(A))A + \alpha_2(\mathcal{I}(A))A^2 \quad \text{for all } A \in \mathbb{S}^3_>.$$

Here, α_0, α_1, and α_2 are real-valued functions of the three principal invariants of the matrix A as defined in Definition 2.2.

Proof. See [15, Proof of Theorem 3.6-1]. □

The incorporation of the Rivlin-Ericksen representation theorem into our setting allows a first explicit description of the response function $\widetilde{\Sigma}$.

Theorem 2.23. *Consider an elastic, isotropic, and frame-indifferent material. Then,*

$$T_{\mathrm{d}}(x_{\mathrm{d}}) = \widehat{T}(x, \nabla y(x)) = \check{T}(x, \nabla y(x)\nabla y(x)^T)$$

with $\check{T}(x, \cdot) : \mathbb{S}^3_> \to \mathbb{S}^3$ defined by

$$\check{T}(x, A) := \alpha_0(x, \mathcal{I}(A))\,\mathrm{Id} + \alpha_1(x, \mathcal{I}(A))A + \alpha_2(x, \mathcal{I}(A))A^2 \quad \text{for all } A \in \mathbb{S}^3_>,$$

where $\alpha_0(x, \cdot)$, $\alpha_1(x, \cdot)$, and $\alpha_2(x, \cdot)$ are real-valued functions of the three principal invariants of the matrix A. Further, the second Piola-Kirchhoff stress tensor is of the form

$$\Sigma(x) = \tilde{\Sigma}(x, \nabla y(x)) = \check{\Sigma}(x, \nabla y(x)^T \nabla y(x))$$

such that the response function $\check{\Sigma}(x, \cdot) : \mathbb{S}^3_> \to \mathbb{S}^3$ is of the form

$$\check{\Sigma}(x, B) = \beta_0(x, \mathcal{I}(B))\,\mathrm{Id} + \beta_1(x, \mathcal{I}(B))B + \beta_2(x, \mathcal{I}(B))B^2 \quad \text{for all } B \in \mathbb{S}^3_>.$$

The corresponding functions $\beta_0(x, \cdot)$, $\beta_1(x, \cdot)$, and $\beta_2(x, \cdot)$ are real-valued, whereby their arguments are the three principal invariants of the matrix B. In reverse, if at least one of the response functions $\check{T}$ or $\check{\Sigma}$ is of the stated form, then the axiom of material frame-indifference holds and the material is isotropic at the point x.

Proof. For the proof, see [15, Proof of Theorem 3.6-2]. □

Theorem 2.23 already provides a rough structure of the explicit representation of the response functions.

2.3.5 Material behavior for large strains

A physically intuitive condition is to require that extreme strains correspond to large stresses. For hyperelastic materials, this condition translates to the stored energy function $\hat{W}$ approaching infinity if one of the eigenvalues of the matrix $C = M^T M$ approaches zero or infinity. Denoting the respective eigenvalues by λ_1, λ_2, and λ_3, we can restrict the analysis to keeping λ_2 and λ_3 in a compact interval in $]0, \infty[$. Then, we obtain:

$$\begin{aligned}
&\lambda_1 \to 0^+ \Leftrightarrow \det M \to 0^+,\\
&\lambda_1 \to \infty \Leftrightarrow \|M\| \to \infty,\\
&\lambda_1 \to \infty \Leftrightarrow \|\operatorname{Cof} M\| \to \infty,\\
&\lambda_1 \to \infty \Leftrightarrow \det M \to \infty.
\end{aligned}$$

The first condition yields the following implication for the stored energy function $\hat{W}$:

$$\det M \to 0^+ \Rightarrow \hat{W}(x, M) \to \infty,\ M \in \mathbb{M}^3_+. \tag{2.14}$$

This condition reflects the idea that for realistic materials, compressing a given volume to zero requires an infinite amount of energy. Additionally, the last three conditions describe the implication

$$(\|M\| + \|\operatorname{Cof} M\| + \det M) \to \infty \Rightarrow \hat{W}(x, M) \to \infty,\ M \in \mathbb{M}^3_+.$$

A sharper version of this assumption, which is required for the existence theorem in Section 2.5, leads to the following coerciveness condition.

Assumption 2.24. *There exist constants $a, p, s, r > 0$, and $b \in \mathbb{R}$ such that at each point $x \in \overline{\Omega}$, the coerciveness inequality*

$$\hat{W}(x, M) \geq a(\|M\|^p + \|\operatorname{Cof} M\|^s + (\det M)^r) + b \quad \textit{for all } M \in \mathbb{M}^3_+$$

holds.

The coerciveness inequality represents a measure of the material's strength and a necessary growth condition required for the stored energy function. For a detailed analysis of this topic, see [15, Chapter 4].
Assumption 2.24 is naturally embedded into the setting to show existence of solutions to hyperelastic problems since coerciveness is often a necessary requirement when studying minimization problems. On the contrary, Condition (2.14) adds significant restrictions on possible candidates for the stored energy function such as the exclusion of convex functions.

2.3.6 Non-convexity of the stored energy function

We recall that Theorem 2.16 yields an energy minimizing approach for hyperelastic problems. Techniques to show the existence of solutions to minimization problems usually

require the convexity of the considered objective function. In that context, the question arises whether the stored energy $\hat{W}$ can be chosen as a convex function. However, as mentioned above, the priorly introduced physical restrictions already rule out this possibility.

Theorem 2.25. *Consider $x \in \overline{\Omega}$ such that the function*

$$\hat{W}(x, \cdot) : \mathbb{M}^3_+ \to \mathbb{R}$$

is convex. Then:

1. *The convexity of $\hat{W}(x, \cdot)$ implies that Condition (2.14) cannot hold.*
2. *The convexity of $\hat{W}(x, \cdot)$ contradicts the axiom of frame-indifference.*

The proof of this theorem is rather technical and does not yield further insight into the problem structure. Therefore, the reader is referred to [15, Proof of Theorem 4.8-1]. The lack of convexity poses significant difficulties in the theoretical analysis of hyperelastic problems. However, with the application of polyconvex functions in [6], this issue was resolved and a rigorous existence theory has been established. Still, we cannot expect the uniqueness of solutions.

Non-uniqueness of solutions

The possible non-uniqueness of solutions is not just a mathematical artifact due to an unsuitable or incomplete problem description. In fact, it has a real-world interpretation. For this reason, we want to study an intuitive example. Consider a horizontal cantilever fixed at a wall, as depicted in Figure 2.2. On the frontal face, a constant boundary force is applied, which is represented by red arrows. In addition, this boundary force acts orthogonally to the surface, and the weight of the cantilever is neglected. For sufficiently large forces, the following behavior can be expected. First, we obtain an unstable solution with very small displacements as illustrated in Figure 2.3. In numerical simulations, the instability is reflected in a non-positive definite Hessian matrix of the total energy function at the current iterate. Also, the slightest perturbation of the boundary force can lead to significant changes of the solution.
In addition to the unstable solution, we also obtain two stable solutions which are shown in Figure 2.4. There, the boundary forces make the body snap upwards or downwards, respectively. Due to symmetry, the two solutions cannot be distinguished w.r.t. the elastic energy. In literature, this effect is usually referred to as buckling and is observed for a wide range of problems. Buckling and other effects that yield non-uniqueness of solutions of real-world problems were discussed in, e.g., [15, 24]. Regarding non-uniqueness in elasticity in general, there exists a large amount of literature. For a deeper analysis of this issue, the reader is referred to [27, 83, 98, 99, 103]. This example already shows that even in a very simple setting, the uniqueness of solutions is ruled out. Consequently, the possible non-uniqueness has to be taken into account in the theoretical analysis and the numerical simulations.

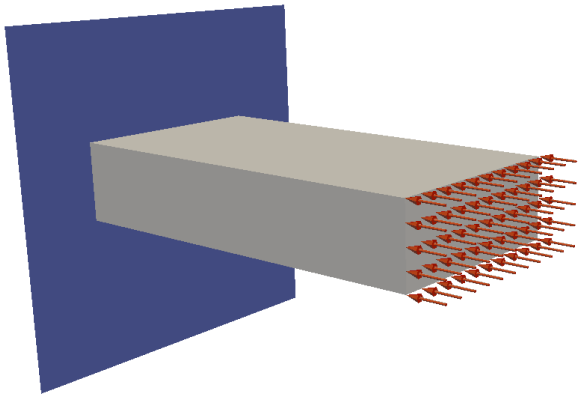

Figure 2.2: Undeformed cantilever.

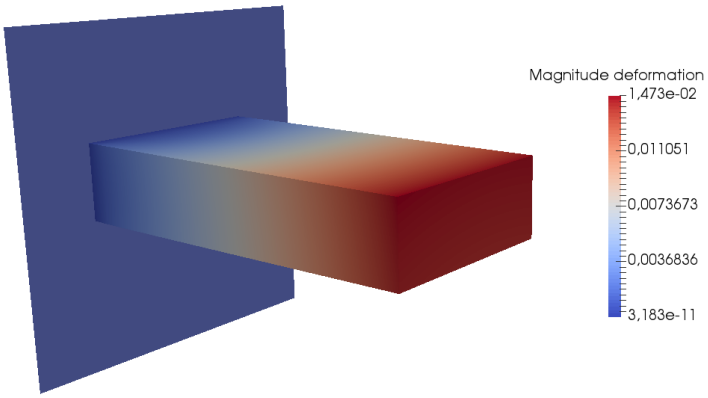

Figure 2.3: Unstable deformation.

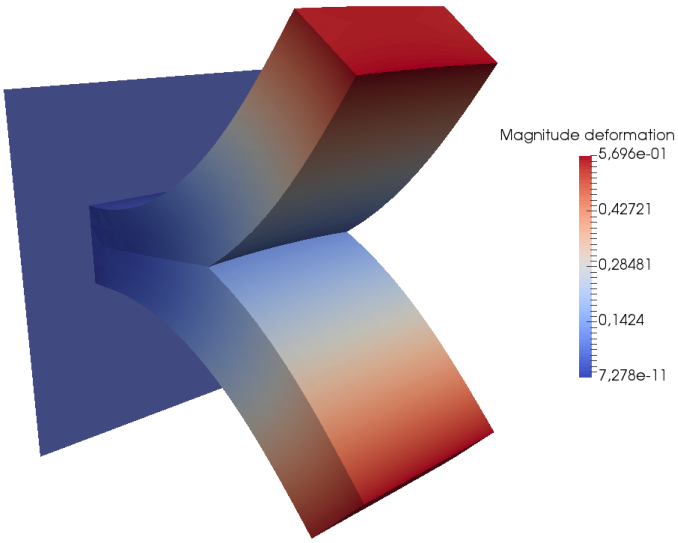

Figure 2.4: Stable deformations.

Computing multiple non-unique solutions analytically seems to be out of reach in the general setting of three-dimensional hyperelasticity so far. Nevertheless, there exist several methods to compute multiple solutions numerically. One approach that seems promising is the deflation technique developed in [29] to find distinct solutions of nonlinear partial differential equations. However, applying this approach to optimal control problems combined with elasticity is beyond the scope of this work and remains a subject of future research.
In summary, we have shown that the convexity of the stored energy function is incompatible with the physical restrictions required to model real materials. Additionally, the uniqueness of energy minimizers has been ruled out for general settings in theory as well as in real-world applications. The lack of uniqueness becomes a major issue for optimal control of hyperelastic problems in Chapter 4. To show at least the existence of solutions to hyperelastic problems, we continue by studying an approach that does not rely on convex energy functions.

2.3.7 Polyconvex functions

The concept of polyconvex functions is one possible way to compensate for the lack of convexity and show the existence of solutions, though not uniqueness. This approach was elaborated in [6]. First, we define polyconvex functions.

Definition 2.26 (Polyconvex function)**.** Consider sets $Z \subset \mathbb{M}^3$ and

$$\mathcal{Z} := \{(M, \operatorname{Cof} M, \det M) \in \mathbb{M}^3 \times \mathbb{M}^3 \times \mathbb{R} \mid M \in Z\}.$$

Then, a function $W : Z \to \mathbb{R}$ is called polyconvex if there exists a convex function

$\mathbb{W} : \mathcal{Z} \to \mathbb{R}$ such that

$$W(M) = \mathbb{W}(M, \operatorname{Cof} M, \det M) \quad \text{for all } M \in Z.$$

In contrast to convexity, polyconvexity is consistent with the physical assumptions made so far, see Subsection 2.3.8. In [6], techniques to show the existence of solutions to hyperelastic problems were introduced. Unlike other approaches, the choice of polyconvex functions allows to prove existence of solutions without relying on strong structural requirements that rule out many problem classes. As a result, choosing polyconvex stored energy functions is the common approach to prove the existence of solutions to hyperelastic problems, and we restrict ourselves here to this technique. The analysis conducted in [6] and an alternative technique are studied in Section 2.5.

2.3.8 Models for the stored energy function

In the previous analysis, we introduced physical assumptions to narrow the choice of the stored energy function $\hat{W}$. Although the previous results are sufficient to show the existence of solutions, we still lack an explicit model of $\hat{W}$.
So far, we have restricted our analysis only to a certain class of materials, i.e., hyperelastic and isotropic ones. Thus, there is still some freedom of choice when it comes to the specific materials used. This additional information is sufficient to derive an explicit model of the stored energy function $\hat{W}$.
In the subsequent examination, we discuss how such a model can be constructed. Thereby, only homogeneous materials are considered. First, we introduce the following definition, cf. [15, Chapter 3].

Definition 2.27 (Natural state)**.** A reference configuration $\overline{\Omega}$ is called a natural state if the response function $\widehat{T}$ of the Cauchy stress tensor satisfies

$$\widehat{T}(x, \mathrm{Id}) = 0 \quad \text{for all } x \in \overline{\Omega}.$$

This definition corresponds to the physical interpretation that the reference configuration is a stress-free state. The existence of such stress-free states is a physically reasonable assumption for solid materials.
For a given material, and under suitable smoothness assumptions, it is possible to derive an explicit description of the material behavior near a natural state. The material dependency enters this description via two positive constants λ and μ.

Theorem 2.28. *Consider an elastic, isotropic, and homogeneous material whose reference configuration is a natural state. Further, assume that for each $x \in \overline{\Omega}$, the functions $\beta_0(x, \cdot)$, $\beta_1(x, \cdot)$, and $\beta_2(x, \cdot)$ defined in Theorem 2.23 are differentiable at the point $(3, 3, 1)$, which corresponds to the three principal invariants of the identity matrix. Then, there exist constants λ and μ such that the response function $\widetilde{\Sigma}$ of the second Piola stress tensor can be expressed as a function of the Green-St Venant strain tensor E and satisfies*

$$\widetilde{\Sigma}(M) = \overline{\Sigma}(E) := \lambda(\operatorname{tr} E)\,\mathrm{Id} + 2\mu E + o(E), \;\; M \in \mathbb{M}^3_+ \tag{2.15}$$

with

$$E := \frac{1}{2}(M^T + M + M^T M).$$

Proof. See, [15, Proof of Theorem 3.8-1]. □

In our setting, the constants λ and μ coincide with the Lamé constants of the respective material. Thus, we do not distinguish between them and the material parameters λ and μ. These parameters are related to the well-known Poisson ratio ν and the Young modulus E_y via the relations

$$\lambda = \frac{E_y \nu}{(1+\nu)(1-2\nu)} \quad \text{and} \quad \mu = \frac{E_y}{2(1+\nu)}.$$

For the inverse relation, we obtain

$$\nu = \frac{\lambda}{2(\lambda+\mu)} \quad \text{and} \quad E_y = \frac{\mu(3\lambda+2\mu)}{\lambda+\mu}.$$

These parameters have a direct physical interpretation. In linear elasticity, and for a given material, the respective Poisson ratio is a measure of lateral contraction while the Young modulus is a measure of stiffness. In practice, all these constants can be determined experimentally. For a detailed discussion, the reader is referred to [15, Chapter 3].

Recall Equation (2.15). The simplest choice that satisfies this equation and is consistent with the assumptions made in Theorem 2.28 is

$$\overline{\Sigma}(E) := \lambda(\operatorname{tr} E)\operatorname{Id} + 2\mu E, \quad \operatorname{Id} + 2E \in \mathbb{S}^3_{>}. \tag{2.16}$$

An elastic material whose response function is of the stated form is called a St Venant-Kirchhoff material. One major drawback of these models is that in the case of hyperelastic materials, the corresponding stored energy function is not polyconvex [15, Theorem 4.10-1]. Consequently, the existence of solutions has not been proven yet in a general setting. In addition, the term $\det \nabla y$ is not a part of the model. Thus, Condition (2.14) cannot be enforced. This lack of structure restricts the usefulness of these models to problems with small strains. Still, St Venant-Kirchhoff materials contain more information than models that arise from linearized elasticity. For this reason, and due to their simplicity, they are applied in a wide range of applications, cf. [75].

Going back to Theorem 2.28, the elaborated results yield a similar characterization of the stored energy function $\hat{W}$.

Theorem 2.29. *Consider an elastic, isotropic, and homogeneous material whose reference configuration is a natural state. In addition, let the regularity assumptions defined in [15, Theorem 4.5-1] be satisfied. Then, the stored energy function $\hat{W}$ can be written as a function of the Green-St Venant strain tensor E, satisfying*

$$\hat{W}(M) = \tilde{W}(E) := \frac{\lambda}{2}(\operatorname{tr} E)^2 + \mu \operatorname{tr} E^2 + o(\|E\|^2), \ M \in \mathbb{M}^3_+,$$

with

$$E := \frac{1}{2}(M^T + M + M^T M).$$

Proof. The proof can be found in [15, Proof of Theorem 4.5-1]. □

With these requirements on $\hat{W}$ at hand, we can finally derive an explicit model that is consistent with all the previous assumptions made.

Theorem 2.30. *Consider two given Lamé constants $\lambda > 0$ and $\mu > 0$. Then, there exist constants $a, b, c, d, \alpha > 0$, $\beta \in \mathbb{R}$, and $e \in \mathbb{R}$ such that the polyconvex stored energy function $\hat{W}$ defined by*

$$\hat{W}(M) := a\|M\|^2 + b\|\operatorname{Cof} M\|^2 + c(\det M)^2 - d\ln(\det M) + e \tag{2.17}$$

satisfies the equation

$$\hat{W}(M) = \tilde{W}(E) = \frac{\lambda}{2}(\operatorname{tr} E)^2 + \mu \operatorname{tr} E^2 + O(\|E\|^3)$$

and the coerciveness inequality

$$\hat{W}(M) \geq \alpha(\|M\|^2 + \|\operatorname{Cof} M\|^2 + (\det M)^2) + \beta,$$

for all $M \in \mathbb{M}^3_+$ and

$$E := \frac{1}{2}(M^T + M + M^T M).$$

Proof. For the proof, see [15, Proof of Theorem 4.10-2]. □

The model defined in Equation (2.17) explicitly depends on the term $\det M$. This property is essential in order to incorporate the compression condition stated in (2.14). Last, we have to address how to compute the parameters a, b, c, and d from given Lamé constants λ and μ. One possible approach to do this has been analyzed in [15, Proof of Theorem 4.10-2]. However, the proof is rather technical and does not yield further insight into the problem. Therefore, we only consider models with given parameters, whereby we skip the explicit derivation. The type of model introduced in Theorem 2.30 describes compressible Mooney-Rivlin materials, cf. [16]. Similarly, compressible neo-Hookean materials are described by the model

$$\hat{W}(M) = a\|M\|^2 + b(\det M)^2 - c\log(\det M),$$

with positive constants a, b, and c. For St Venant-Kirchhoff materials, we obtain

$$\hat{W}(M) = \overline{W}(E) := \frac{\lambda}{2}(\operatorname{tr} E)^2 + \mu \operatorname{tr} E^2.$$

A more detailed analysis on different types of models can be found in, e.g., [21, 77]. In this thesis, the numerical simulations are restricted to compressible Mooney-Rivlin materials.

In summary, all necessary physical and mathematical properties have been established to sufficiently describe the real-world behavior of isotropic and hyperelastic materials. Further, we have derived an explicit representation of the respective energy functionals.

2.4 Contact problems

First, let us introduce the setting. Analogously to before, $\Omega \subset \mathbb{R}^3$ denotes a bounded Lipschitz domain (in the sense of [74, pp. 4-6]) representing the three-dimensional body. Additionally, the respective material is assumed to be isotropic and hyperelastic. The boundary Γ consists of three disjoint and relatively open subsets such that

$$\Gamma = \overline{\Gamma_D \cup \Gamma_N \cup \Gamma_C},$$

whereby each segment has a non-zero boundary measure. Here, Γ_D denotes the segment where Dirichlet boundary conditions are enforced and Γ_N denotes the Neumann boundary where the external pressure load acts. For the remainder of this work, we make the simplifying assumption that all applied forces are dead loads.
Further, we extend this framework to contact problems where deformations of the body are restricted by an obstacle. The corresponding non-penetration conditions are imposed on the contact boundary Γ_C. The deformation function

$$y : \overline{\Omega} \to \mathbb{R}^3$$

is an element of the vector valued Sobolev space $W^{1,p}(\Omega, \mathbb{R}^3)$ with $p \in [2, \infty)$. For the sake of brevity, we just write $W^{1,p}(\Omega)$ and omit the image space for all vector valued spaces. Also, we will suppress the explicit notation of trace operators if it is not required for the analysis. Further, we do not distinguish between sequences and their elements when it is clear from the context. In Sobolev spaces, the orientation-preserving condition can only be satisfied in the weaker form

$$\det \nabla y(x) > 0 \quad \text{for a.e. } x \in \Omega. \tag{2.18}$$

If y is not continuously differentiable, this condition does not even guarantee local invertibility. In [19], the authors studied deformations that are almost everywhere injective. This result was achieved by introducing the additional requirement

$$\int_\Omega \det \nabla y(x)\, dx \leq \operatorname{vol} y(\Omega).$$

However, it is unclear how to transfer this condition to a numerical approach and, thus, we do not consider it here. For further studies on local and global invertibility, the reader is referred to [8, 30, 31, 106].
Regarding boundary conditions, the deformation y is required to be the identity mapping on Γ_D. On the segment Γ_N, we consider an applied force $u \in L^q(\Gamma_N, \mathbb{R}^3)$ with $q \in (1, \infty)$. The boundary force also acts as control in the optimization setting discussed in Chapter 4. With that in mind, we introduce the notation $Y := W^{1,p}(\Omega)$ and $U := L^q(\Gamma_N)$ to denote the state space and the control space, respectively. Although volume forces are not considered here, they could be included in our framework.

2.4.1 Contact constraints

Physically speaking, we want to add an obstacle, denoted by Ψ, which the body cannot penetrate. Therefore, non-penetration constraints on the contact boundary Γ_C are introduced. For simplicity, only constraints of the form

$$y_3 \geq 0 \quad \text{a.e. on } \Gamma_C \tag{2.19}$$

are considered, meaning that the third component y_3 of y should be non-negative. This restriction describes the plane spanned by the first two canonical basis vectors acting as an obstacle. By incorporating these constraints, we can model static contact problems without friction. Note that the set

$$\mathcal{C} = \{v \in Y \mid v_3 \geq 0 \quad \text{a.e. on } \Gamma_C\}$$

is weakly closed in Y. Our setting is illustrated in Figure 2.5.

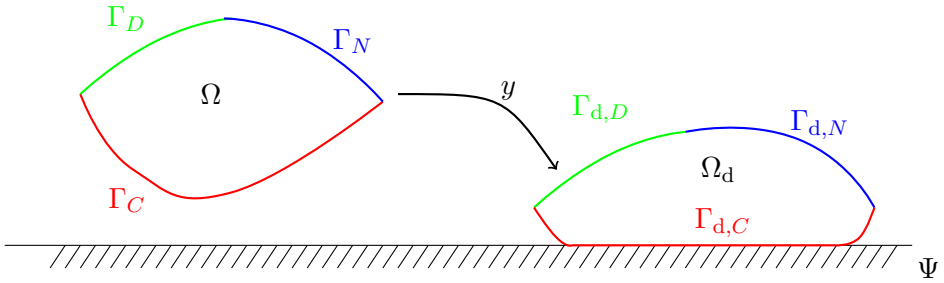

Figure 2.5: Contact problem.

Contact constraints add significant difficulties to an already highly nonlinear and non-convex problem. Among those is the non-smoothness due to the sudden change of behavior in the contact region. Additionally, the area of contact is unknown before the computation, which impedes enforcing contact constraints for more complicated geometries. The type of constraints studied here significantly simplifies the theoretical and numerical examinations. Therefore, a brief survey of alternative contact conditions is given.

Bilateral contact

Let n denote the outer normal vector field of the boundary segment Γ_C. One way to define contact constraints is to require that the normal displacements vanish:

$$\phi_n = \phi \cdot n = 0 \quad \text{a.e. on } \Gamma_C.$$

This description corresponds to a setting where the body is glued to the obstacle, see [43, Chapter 5]. Bilateral contact conditions were studied in [63, 79]. For general cases, this kind of constraint is far too restrictive. By contrast, the Signorini contact condition allows for modeling a wider range of problems.

Signorini contact condition

In this setting, deformations are restricted a priori according to the distance between the obstacle Ψ and the body Ω. Let $g : \Gamma_C \to \mathbb{R}$ denote the corresponding function which measures the distance at each point $x \in \Gamma_C$ along the unit outer normal vector n. Then, enforcing a non-penetration condition along the normal n corresponds to requiring that

$$\phi_n \leq g \quad \text{a.e. on } \Gamma_C. \tag{2.20}$$

The Signorini contact condition was first analyzed in [97] and has been extensively studied ever since, see, e.g., [20, 53, 56]. Apparently, Condition (2.20) is a strong simplification since it is far from obvious that contact will always occur along the outer normal n. The validity of such conditions depends on the geometry and the initial setting. Thus, the question arises whether the Signorini condition sufficiently describes the underlying contact problem. At least for small displacements, and if the boundary segment Γ_C and the obstacle Ψ are very close and essentially parallel, cf. [56, Chapter 2], Condition (2.20) is a suitable approximation. A more detailed discussion of this topic can be found in [25, 73]. For settings involving large deformations, and in particular for multi-body problems, closest point projections are widely applied to model contact conditions, see, e.g., [59, 60, 109, 110].
Additional complications arise in the case of multi-body contact problems. These problems require specialized discretization schemes that apply to non-matching grids such as the mortar method, see, e.g., [12, 50, 81, 91, 108, 110, 111]. A further option to model contact conditions is the normal compliance method, which is studied in Chapter 3.
In summary, the study of more sophisticated contact constraints and geometries is still the subject of current research and remains beyond the scope of this work. A more detailed discussion about modeling contact constraints, though not for nonlinear elasticity, can be found in [73, 91]. Applying contact constraints in nonlinear elasticity has been discussed in [81, 110, 111]. For an overview of contact problems in general, the reader is referred to [43, 56].

2.4.2 Contact problems in hyperelasticity

Recalling Theorem 2.16, we know that under certain smoothness assumptions minimizers of the total energy functional I describe deformations of a body. In contrast to before, the total energy functional

$$I : Y \times U \to \mathbb{R} \cup \{\infty\}$$

is a function of the deformation y and of the applied force u. The corresponding boundary force u_{d} acting on the deformed segment $\Gamma_{\mathrm{d},N}$ is a dead load by assumption. As a result, u does not depend on the deformation y. Consequently, we write

$$I(y,u) := \int_\Omega \hat{W}(x, \nabla y(x))\, dx - \int_{\Gamma_N} y(x)u(x)\, ds. \tag{2.21}$$

This notation is later required to analyze optimal control problems combined with elasticity. We also define the splitting

$$I(y,u) = I_{\text{strain}}(y) - I_{\text{out}}(y,u)$$

with

$$I_{\text{strain}}(y) := \int_\Omega \hat{W}(x, \nabla y(x))\,dx \quad \text{and} \quad I_{\text{out}}(y,u) := \int_{\Gamma_N} y(x)u(x)\,ds. \tag{2.22}$$

Further restrictions on the function spaces and the stored energy function $\hat{W}$ are required since we have to ensure that the resulting energy minimization problem is well-posed and that our setting corresponds to the physical model elaborated in the previous sections. The necessary properties are summarized in the following assumption.

Assumption 2.31. *Let $\hat{W} : \Omega \times \mathbb{M}^3_+ \to \mathbb{R}$ be the stored energy function of a hyperelastic material and let $(p,r,s,q,q') \in]1,\infty[^5$ be fixed with $p \geq 2$, $s \geq \frac{p}{p-1}$, and $r > 1$. Assume that the following properties hold.*

1. *Polyconvexity: For almost all $x \in \Omega$, there exists a convex function $\mathbb{W}(x,\cdot,\cdot,\cdot) : \mathbb{M}^3 \times \mathbb{M}^3 \times]0,\infty[\to \mathbb{R}$ such that*

   $$\hat{W}(x,M) = \mathbb{W}(x, M, \operatorname{Cof} M, \det M) \quad \text{for all } M \in \mathbb{M}^3_+.$$

   Further, the function

   $$\mathbb{W}(\cdot, M, \operatorname{Cof} M, \det M) : \Omega \to \mathbb{R}$$

   is measurable for all $(M, \operatorname{Cof} M, \det M) \in \mathbb{M}^3 \times \mathbb{M}^3 \times]0,\infty[$.

2. *For almost all $x \in \Omega$, the implication*

   $$\det M \to 0^+ \Rightarrow \hat{W}(x,M) \to \infty$$

   holds.

3. *The sets of admissible deformations are defined by*

   $$\begin{aligned}\mathcal{A}_c := \{&y \in W^{1,p}(\Omega),\ \operatorname{Cof} \nabla y \in L^s(\Omega),\ \det \nabla y \in L^r(\Omega),\\ &y = \mathrm{id} \ \text{a.e. on } \Gamma_D,\ \det \nabla y > 0 \ \text{a.e. in } \Omega,\ y_3 \geq 0 \ \text{a.e. on } \Gamma_C\}\end{aligned}$$

   and

   $$\begin{aligned}\mathcal{A} := \{&y \in W^{1,p}(\Omega),\ \operatorname{Cof} \nabla y \in L^s(\Omega),\ \det \nabla y \in L^r(\Omega),\\ &y = \mathrm{id} \ \text{a.e. on } \Gamma_D,\ \det \nabla y > 0 \ \text{a.e. in } \Omega\}.\end{aligned}$$

4. *Coerciveness: There exist constants $a \in \mathbb{R}$ and $b > 0$ such that*

   $$\hat{W}(x,M) \geq a + b(\|M\|^p + \|\operatorname{Cof} M\|^s + |\det M|^r) \quad \text{for all } M \in \mathbb{M}^3_+.$$

5. *The exponents* $q, q' \in]1, \infty[$ *satisfy* $q^{-1} + q'^{-1} = 1$,

$$q' < \frac{2p}{3-p} \text{ for } p < 3, \text{ and } q' < \infty \text{ for } p \geq 3.$$

6. *For the zero boundary force* $u_z \in U$, *the identity mapping* $\mathrm{id} : \overline{\Omega} \to \overline{\Omega}$ *satisfies* $\mathrm{id} \in \operatorname{argmin}_{v \in \mathcal{A}_c} I(v, u_z)$ *with* $I(\mathrm{id}, u_z) < \infty$.

The first assumption states the polyconvexity of the stored energy function $\hat{W}$. This property is necessary to compensate for the non-convexity of $\hat{W}$ and show the existence of solutions to hyperelastic problems.
Assumptions 2.31(2) and 2.31(4) describe the physical behavior of the material for large strains. In this context, Assumption 2.31(2) corresponds to the condition that compressing a given volume to zero requires an infinite amount of work. Assumption 2.31(4) is a sharper version of the coerciveness property stated in Assumption 2.24 since the exponents p, s, and r satisfy stronger restrictions here.
Further, from Assumption 2.31(3), we obtain the admissible set of deformations $\mathcal{A}_c$ for the obstacle constrained case. Later, we introduce a regularization approach for the contact constraints which allows us to operate on the relaxed admissible set $\mathcal{A}$.
The definition of the admissible set $\mathcal{A}$ and the first assumption ensure that the integral $I_{\text{strain}}(y)$ is well-defined for all $y \in \mathcal{A}$, cf. [15, Proof of Theorem 7.7-1].
The coerciveness of the strain energy I_{strain} is implied by Assumption 2.31(4), see, e.g., [15, Proof of Theorem 7.3-2].
Assumption 2.31(5) is a technical requirement to apply the Hölder inequality and ensure the existence of compact trace operators.
Finally, Assumption 2.31(6) guarantees that $\mathcal{A}$ and $\mathcal{A}_c$ are not empty. Also, for some fixed boundary force $u \in U$, we obtain $\inf_{v \in \mathcal{A}_c} I(v, u) < \infty$ and $\inf_{v \in \mathcal{A}} I(v, u) < \infty$. Note here that the non-emptiness of the admissible sets essentially depends on the boundary conditions imposed on Γ_D. Since admissible deformations coincide with the identity mapping on Γ_D, the non-emptiness of the admissible sets $\mathcal{A}$ and $\mathcal{A}_c$ follows directly. However, this implication no longer holds for general boundary conditions imposed on Γ_D. In those cases, the non-emptiness of the admissible sets is assumed a priori. This is basically an assumption on the function y_D, which describes the boundary conditions on Γ_D.
For a detailed discussion of these requirements, the reader is referred to [15, Chapter 7] and [18]. We assume that these assumptions are satisfied throughout this work. With Assumption 2.31 at hand, modeling contact problems corresponds to a well-posed minimization problem.

Definition 2.32 (Contact problem)**.** Let $u \in U$ be a fixed boundary force applied to a hyperelastic body. Further, the deformation of the body is restricted by Condition (2.19). Then, the resulting deformation y satisfies

$$y \in \operatorname*{argmin}_{v \in \mathcal{A}_c} I(v, u). \tag{2.23}$$

With Theorem 2.16 in mind, the question arises whether the minimization problem (2.23) also corresponds to solving some kind of equilibrium equations. However, in the case of contact problems, the relation can only be derived formally.

Theorem 2.33. *Let y be a sufficiently smooth solution to the energy minimization problem*

$$y \in \operatorname*{argmin}_{v \in \mathcal{A}_c} I(v, u).$$

Then, y formally solves the following boundary value problem:

$$\begin{aligned}
-\operatorname{div} \widetilde{T}(\nabla y(x)) &= 0 && \text{for all } x \in \Omega,\\
\widetilde{T}(\nabla y(x))n(x) &= u(x) && \text{for all } x \in \Gamma_N,\\
y(x) &= \mathrm{id}(x) && \text{for all } x \in \Gamma_D,\\
y_3(x) &\geq 0 && \text{for all } x \in \Gamma_C,\\
\widetilde{T}(\nabla y(x))n(x) &= 0 && \text{if } x \in \Gamma_C \text{ and } y_3(x) > 0,\\
\widetilde{T}(\nabla y(x))n(x) &= \lambda(x) n_{\mathrm{d}}(y(x)) && \text{with } \lambda(x) \leq 0 \text{ if } x \in \Gamma_C \text{ and } y_3(x) = 0.
\end{aligned}$$

Proof. For the proof, see [15, Proof of Theorem 5.3-1]. □

2.5 Existence theory for nonlinear elastic problems

So far, we have established that deformations of a body can be modeled as energy minimizers of the total energy function

$$I(y) = \int_\Omega \hat{W}(x, \nabla y(x))\, dx - \int_{\Gamma_N} y(x) u(x)\, ds$$

over a suitable admissible set. Proving the existence of minimizers is a delicate matter. As previously mentioned, the restrictions imposed on the stored energy function $\hat{W}$ rule out convexity. Thus, techniques that require the weak lower semi-continuity of the functional no longer apply. As a further consequence, uniqueness of minimizers cannot be expected.
In this section, we derive an existence result for hyperelastic contact problems in the setting of polyconvexity based on the results in [6, 15, 18]. Over the course of this examination, we aim for slightly more general results which are required in the optimal control framework. In view of the optimal control problem studied in Chapter 4, the setting is restricted to obstacle problems where applied forces only act on the boundary. Besides polyconvexity, there exists another approach to show existence of solutions by examining the equilibrium equations directly.

2.5.1 Existence results by differential calculus

Under strong structural requirements, the implicit function theorem can be applied in order to derive the existence of solutions to the equilibrium equations. This idea was first

considered in [101, 102]. However, this approach only applies in very restricted settings. For example, the analysis conducted in [15, Section 6] is restricted to pure displacement problems of the form

$$\begin{aligned} -\operatorname{div}(\mathrm{Id}+\nabla\phi)\widetilde{\Sigma}(E(\phi)) &= f \text{ in } \Omega, \\ \phi &= 0 \text{ on } \Gamma. \end{aligned}$$

This system describes the equilibrium equation derived in Lemma 2.14 for the displacement function ϕ. Then, the existence of solutions can be shown for elastic, isotropic, and homogeneous materials whose reference configuration is a natural state.

Theorem 2.34. *Consider a domain Ω with C^2-boundary and the spaces*

$$\mathbb{E} := \left\{ \tfrac{1}{2}(C - \mathrm{Id}) \in \mathbb{S}^3 \mid C \in \mathbb{S}^3_> \right\}$$

and

$$V^p := \left\{ v \in W^{2,p}(\Omega, \mathbb{R}^3) \mid v = 0 \text{ on } \Gamma \right\}.$$

Further, let $\widetilde{\Sigma} \in C^2(\mathbb{E}, \mathbb{S}^3)$ be a mapping satisfying

$$\widetilde{\Sigma}(E) = \lambda(\operatorname{tr} E)\,\mathrm{Id} + 2\mu E + O(\|E\|^2) \quad \text{with } \lambda > 0 \text{ and } \mu > 0.$$

Then, for $p > 3$, there exists a neighborhood F of the origin in $L^p(\Omega, \mathbb{R}^3)$ and a neighborhood V of the origin in V^p such that for each $f \in F$, the equation

$$-\operatorname{div}((\mathrm{Id}+\nabla v)\widetilde{\Sigma}(E(v))) = f \text{ in } \Omega$$

has exactly one solution in V.

Proof. The proof can found in [15, Proof of Theorem 6.7-1]. □

Although this approach yields smooth solutions, it only holds for small deformations and does not apply in the case of mixed boundary conditions. This rules out a wide range of problems, including contact problems. Therefore, choosing a hyperelastic and polyconvex setting is our method of choice. For a detailed discussion of the alternative approach presented above, the reader is referred to [15, Chapter 6].

2.5.2 Existence theory for polyconvex functions

Here, we only give a brief summary of the proof conducted in [15, 18]. At this, we focus on results that are also necessary for the analysis of optimal control problems. Techniques for showing existence of solutions to hyperelastic problems in the setting of polyconvexity were introduced in [6]. The extension to contact problems was first considered in [18]. There, the authors studied a more general setting compared to the one considered in this work. As discussed in Section 2.3, we cannot expect uniqueness of solutions due to the non-convexity of the stored energy function $\hat{W}$ both in the mathematical setting and in real-world applications. Before we show the existence of solutions to hyperelastic problems, a rough sketch of the proof is presented:

Consider an infimizing sequence $y_n \subset \mathcal{A}_c$ of the total energy function I. From there, the first critical point is proving the existence of a weakly converging subsequence y_k with its weak limit $\overline{y}$ being an element of $\mathcal{A}_c$ again. Further, we have to verify the weak lower semi-continuity of I w.r.t. the sequence y_k. Then, combining these two results yields the estimate

$$I(\overline{y}) \leq \liminf_{k\to\infty} I(y_k)$$

and the existence of at least one global minimizer $\overline{y} \in \mathcal{A}_c$. Since the set $\mathcal{A}_c$ is non-convex, we cannot deduce its weak closedness. Analogously, due to the non-convexity of I, we cannot expect weak lower semi-continuity of I. Nevertheless, it can be shown that both properties hold for infimizing sequences. First, the following result is required.

Theorem 2.35. *Let y_n be a sequence in Y. Then, the following implication holds:*

$$\left.\begin{aligned} y_n &\rightharpoonup \hat{y} \quad \text{in } Y, \\ \operatorname{Cof} \nabla y_n &\rightharpoonup N \quad \text{in } L^s(\Omega), \\ \det \nabla y_n &\rightharpoonup d \quad \text{in } L^r(\Omega). \end{aligned}\right\} \Rightarrow \begin{cases} N = \operatorname{Cof} \nabla \hat{y}, \\ d \ = \det \nabla \hat{y}. \end{cases}$$

Proof. For the proof, see [6, Lemma 6.1 and Theorem 6.2]. □

From there, we can derive the weak lower semi-continuity of I for particular sequences. Additionally, the outer energy $I_{\text{out}}(y,u)$ is even weakly continuous.

Lemma 2.36. *The outer energy I_{out} is weakly continuous. Further, the total energy functional I is weakly lower semi-continuous w.r.t. sequences that leave the strain energy I_{strain} bounded.*

Proof. We start by showing the weak continuity of the outer energy function $I_{\text{out}}(y,u)$. At this, we extend the analysis of [15, Proof of Theorem 7.1-5] to the case where the boundary force is no longer fixed. Consider a weakly converging sequence (y_n, u_n) in $Y \times U$ whose limit $(\overline{y}, \overline{u})$ is again an element of $Y \times U$. Assumption 2.31(5) and [74, Theorem 6.2] imply the existence of a continuous and compact trace operator

$$\tau : Y \to L^{q'}(\Gamma).$$

Thus, $\tau(y_n) \to \tau(\overline{y})$. Additionally, Hölder's inequality and the continuity of the trace operator yield the estimate

$$|I_{\text{out}}(y,u)| \leq \int_{\Gamma_N} |\tau(y)u| \, ds \leq \|\tau(y)\|_{L^{q'}(\Gamma_N)} \|u\|_U \leq C\|y\|_Y \|u\|_U,$$

for some constant $C > 0$. Since the outer energy I_{out} is bilinear, the derived boundedness implies that $I_{\text{out}}(y,u)$ is continuous. We rewrite

$$I_{\text{out}}(y_n,u_n) - I_{\text{out}}(\overline{y},\overline{u}) = I_{\text{out}}(y_n - \overline{y}, u_n) + I_{\text{out}}(\overline{y}, u_n) - I_{\text{out}}(\overline{y},\overline{u}).$$

Combining the boundedness of u_n, the continuity of I_{out} and the existence of a compact trace operator τ yields that the term $I_{\text{out}}(y_n - \overline{y}, u_n)$ converges to zero as n approaches

infinity. Due to the definition of weak convergence, the second term $I_{\text{out}}(\overline{y}, u_n) - I_{\text{out}}(\overline{y}, \overline{u})$ converges to zero as well. This concludes the first part of the proof. For the sake of brevity, the reader is referred to [15, Proof of Theorem 7.7-1] for the proof of the second statement. □

The next lemma is a slightly modified version of the results in [15, Proof of Theorem 7.3-2]. Here, we obtain a lower bound and a modified coerciveness property for the total energy functional I.

Lemma 2.37. *Let $u_n \subset U$ be a bounded sequence. Then, there exist uniform constants $a > 0$ and $b \in \mathbb{R}$ such that the total energy functional satisfies the estimate*

$$I(v, u_n) \geq a\|v\|_Y^p + b \quad \textit{for all } v \in \mathcal{A} \textit{ and all } n \in \mathbb{N}.$$

Proof. Assumption 2.31(4) implies that there exist constants $c > 0$ and $d \in \mathbb{R}$ such that the strain energy satisfies

$$I_{\text{strain}}(v) \geq c\|v\|_Y^p + d \quad \text{for all } v \in \mathcal{A},$$

see also [15, Proof of Theorem 7.3-2]. Following the argumentation in the proof of Lemma 2.36, the existence of a trace operator $\tau : Y \to L^{q'}(\Gamma)$ and Hölder's inequality yield the estimate

$$|I_{\text{out}}(v, u_n)| \leq \int_{\Gamma_N} |\tau(v)u_n| \, ds \leq \|\tau(v)\|_{L^{q'}(\Gamma_N)} \|u_n\|_U \leq C\|v\|_Y,$$

for some $C > 0$. Since $p > 1$,

$$I(v, u_n) \geq I_{\text{strain}}(v, u_n) - |I_{\text{out}}(v, u_n)| \geq a\|v\|_Y^p + b$$

holds for suitable $a > 0$ and $b \in \mathbb{R}$. This concludes the proof. □

Apparently, this result implies the coerciveness of the total energy functional I w.r.t. to the first argument if the sequence u_n is bounded. In order to show the existence of solutions in hyperelasticity, the previous lemma only needs to hold for fixed boundary forces $u \in U$. Nevertheless, considering this more general result for bounded sequences is required in Chapter 4 for studying optimal control problems.
We continue by proving that the weak limit of sequences in $\mathcal{A}_c$ is again contained in the admissible set $\mathcal{A}_c$ if the corresponding strain energy values are bounded. This is far from obvious since the set $\mathcal{A}_c$ is not weakly closed in the space $W^{1,p}(\Omega)$.

Lemma 2.38. *Let $y_n \rightharpoonup \overline{y}$ be a weakly converging sequence in $\mathcal{A}_c$ that leaves the strain energy $I_{\text{strain}}(y_n)$ bounded. Then, $\overline{y} \in \mathcal{A}_c$.*

Proof. The techniques used here follow along the lines of [15, Proof of Theorem 7.7-1]. First, we know from Assumption 2.31(4) that the stored energy function $\hat{W}$ is coercive w.r.t. the sequence $(y_n, \operatorname{Cof} \nabla y_n, \det \nabla y_n)$. Consequently, the sequence

$(y_n, \operatorname{Cof} \nabla y_n, \det \nabla y_n)$ is bounded due to the boundedness of the strain energy $I_{\text{strain}}(y_n)$. By reflexivity, we obtain a weakly converging subsequence

$$(y_{n_k}, \operatorname{Cof} \nabla y_{n_k}, \det \nabla y_{n_k}) \rightharpoonup (\overline{y}, c, d)\,.$$

Thus, Theorem 2.35 applies and

$$(c, d) = (\operatorname{Cof} \overline{y}, \det \nabla \overline{y})$$

holds. Next, we have to verify that $\overline{y}$ satisfies the Dirichlet boundary conditions. Due to the compactness of the trace operator

$$\tau : W^{1,p}(\Omega) \to L^p(\Gamma)\,,$$

cf. [15, Theorem 6.1-7], we can extract a pointwise converging subsequence of y_{n_k} such that its trace converges ds-almost everywhere on Γ. Consequently, $\overline{y}$ satisfies the Dirichlet boundary conditions on Γ_D. The orientation-preserving condition

$$\det \overline{y} > 0 \quad \text{a.e. in } \Omega$$

results from the boundedness of $I_{\text{strain}}(y_{n_k})$. This can be shown via proof by contradiction. Following the argumentation in [15, Proof of Theorem 7.7-1], it suffices to restrict the analysis to the case $\det \overline{y} \geq 0$ a.e. in Ω. However, if this is assumed, we can find a subsequence, denoted again by y_{n_k}, such that

$$I_{\text{strain}}(y_{n_k}) \to \infty,$$

which contradicts the boundedness assumption of $I_{\text{strain}}(y_{n_k})$. Finally, it has to be shown that

$$\overline{y}_3 \geq 0 \quad \text{a.e. on } \Gamma_C.$$

This simply results from the weak closedness of the set $\mathcal{C}$. From there, we conclude $\overline{y} \in \mathcal{A}_c$. □

The same argumentation applies for $\mathcal{A}$. With Lemma 2.38 at hand, we can derive the subsequent existence result, which was first established in [18, Theorem 4.2] for a more general setting.

Theorem 2.39. *Let $u \in U$ be some fixed boundary force. Then, the total energy functional $I(\,\cdot\,, u)$ has at least one minimizer in $\mathcal{A}_c$.*

Proof. Let $y_n \subset \mathcal{A}_c$ be an infimizing sequence. Assumption 2.31(6) and Lemma 2.37 yield the boundedness of $I(y_n, u)$ and y_n. Due to reflexivity of Y, there exists a weakly converging subsequence, also denoted by y_n. The weak limit is denoted by $\overline{y}$. Since $I(y_n, u)$ and y_n are bounded, we obtain the boundedness of the strain energy $I_{\text{strain}}(y_n)$ as well. As previously established, Lemma 2.38 and the weak closedness of the set $\mathcal{C}$ yield $\overline{y} \in \mathcal{A}_c$. Accordingly, Lemma 2.36 yields that the total energy functional I is weakly lower semi-continuous w.r.t. the sequence y_n. Then, the fact that $\overline{y}$ is a minimizer results from

$$\inf_{y \in \mathcal{A}_c} I(y, u) \leq I(\overline{y}, u) \leq \liminf_{n \to \infty} I(y_n, u) = \inf_{y \in \mathcal{A}_c} I(y, u).$$

This concludes the proof. □

2.6 Summary

In summary, we studied how to describe deformations of nonlinear elastic bodies with obstacle constraints mathematically. In our analysis, we started from Cauchy's theorem and incorporated various material properties such as hyperelasticity to obtain a well-posed but non-convex energy minimization problem of the form

$$y \in \operatorname*{argmin}_{v \in \mathcal{A}_c} I(v, u).$$

Also, under some restrictions, an explicit formulation of the corresponding energy functions has been established. By considering polyconvex stored energy functions, it is possible to prove the existence of energy minimizers within a reasonable framework. However, these minimizers are not necessarily unique. This non-uniqueness corresponds to the physical behavior of real-world hyperelastic materials, as discussed in Section 2.3.

Chapter 3

Regularization of the Contact Constraints

Describing a nonlinear elastic body in contact corresponds to a non-convex constrained minimization problem. For those kinds of problems, primal–dual active set strategies are quite popular, as discussed in, e.g., [51]. Recently, truncated non-smooth Newton multigrid (TNNMG) methods, cf. [35, 36], and filter-trust-region methods, cf. [110], are applied more and more frequently. A further approach was pursued in [64]. There, the obstacle condition was replaced by a pressure-type boundary condition on the deformed domain. However, this approach is only valid formally and has further analytical drawbacks. For a detailed analysis, the reader is referred to [15, Chapter 5] and [64]. The analytical difficulties due to non-smoothness are increased even further when we embed nonlinear elastic contact problems into an optimal control setting. For those reasons, we opt for a regularization approach, in particular, the normal compliance method, cf. [69, 76]. Thereby, we remove the non-smoothness and facilitate the application of efficient solution algorithms in the optimal control setting.
This chapter is structured as follows. The normal compliance method is introduced in Section 3.1. Additionally, we show that this approach is consistent w.r.t. the original contact problem (2.23). Section 3.2 addresses equilibrium conditions for regularized contact problems. Finally, Section 3.3 is dedicated to derive asymptotic convergence rates for the normal compliance method applied to contact problems.
Parts of this chapter have been published in [95].

3.1 Normal compliance method

The basic idea of the normal compliance method is to allow violations of the constraints and penalize these violations according to their extent. In this context, we introduce a penalty functional $P : Y \to \mathbb{R}_0^+$ of the form

$$P(v) := \frac{1}{k} \int_{\Gamma_C} [v]_+^k \, ds, \quad k \in \mathbb{N},\ k > 1, \quad v \in Y, \tag{3.1}$$

where

$$[v]_+ := \max(-v_3(\,\cdot\,), 0) \text{ on } \Gamma_C.$$

Additionally, we multiply the penalty functional with a positive parameter γ and add it to the total energy functional I. The resulting penalized total energy functional reads as follows:

$$I_\gamma(y, u) := I(y, u) + \gamma P(y).$$

Further, assume that for $p < 3$, the inequality $k \leq \frac{2p}{3-p}$ holds throughout the subsequent analysis. Under this assumption, there exists a trace operator $\tau : Y \to L^k(\Gamma)$, see [15, Theorem 6.1-7]. Consequently, the penalty function is well-defined, convex, and weakly lower semi-continuous. The normal compliance approach was studied in [69, 76]. Since then, different variants have been applied in, e.g., [4, 5, 26, 43, 56, 57, 58, 85]. For an overview of penalty methods in general, see [56, 109].
With the regularized total energy functional at hand, we can drop the contact constraint and obtain the relaxed admissible set $\mathcal{A}$ as defined in Assumption 2.31(3). For a fixed penalty parameter $\gamma > 0$, this leads to a relaxed minimization problem.

Definition 3.1 (Regularized contact problem)**.** Consider a fixed boundary force $u \in U$ and a positive penalty parameter γ. Then, a deformation y is called a solution to the regularized contact problem w.r.t. γ if y solves the minimization problem

$$y \in \operatorname*{argmin}_{v \in \mathcal{A}} I_\gamma(v, u). \tag{3.2}$$

The motivation behind this approach is that for increasing parameter γ, the resulting solutions yield increasingly better approximations of solutions of the original contact problem (2.23). Before we can prove this statement, we have to examine whether the regularized problem (3.2) admits at least one optimal solution.

Theorem 3.2. *Let $\gamma > 0$ be a fixed penalty parameter and $u \in U$ be some fixed boundary force. Then, the regularized total energy functional $I_\gamma(\,\cdot\,, u)$ has at least one minimizer in $\mathcal{A}$.*

Proof. Since the penalty function P is weakly lower semi-continuous, the argumentation from the proof of Theorem 2.39 applies here as well. □

In order to conclude that the normal compliance method can be applied to nonlinear elastic contact problems, it has to be verified that solutions of the regularized problem (3.2) approach solutions of the original one (2.23) as $\gamma \to \infty$. At this, we use well-established techniques to analyze regularization approaches which can be found in, e.g., [56]. Before that, we need to establish that in hyperelasticity, bounded boundary forces leave the resulting regularized total energy bounded.

Lemma 3.3. *Let $\gamma_n \to \infty$ be a positive sequence of penalty parameters and let u_n be a bounded sequence in U. Additionally, let y_n be a sequence of corresponding energy minimizers satisfying*

$$y_n \in \operatorname*{argmin}_{v \in \mathcal{A}} I_{\gamma_n}(v, u_n).$$

Then, $I_{\gamma_n}(y_n, u_n)$, $I_{\mathrm{strain}}(y_n)$, and $I(y_n, u_n)$ are bounded. Further, y_n is bounded in Y.

Proof. The boundedness from below results from Lemma 2.37. For the boundedness from above, Theorem 2.39 implies the existence of a state $\tilde{y} \in \mathcal{A}_c$ with $I_{\mathrm{strain}}(\tilde{y}) < \infty$. Consequently, there exists a constant $C > 0$ such that

$$I_{\gamma_n}(y_n, u_n) \leq I_{\gamma_n}(\tilde{y}, u_n) = I(\tilde{y}, u_n) \leq I_{\mathrm{strain}}(\tilde{y}) + |I_{\mathrm{out}}(\tilde{y}, u_n)| < C.$$

The last estimate follows from Hölder's inequality as applied in the proof of Lemma 2.37. The boundedness from above of $I(y_n, u_n)$ simply follows from the previous estimate and the property $\gamma_n P(y_n) > 0$. Again, Lemma 2.37, which also holds for

$$I_{\gamma_n}(y_n, u_n) \geq I(y_n, u_n),$$

implies the boundedness of y_n. From there, we can derive that $I_{\mathrm{out}}(y_n, u_n)$ is bounded due to Hölder's inequality. Finally, this yields the boundedness of $I_{\mathrm{strain}}(y_n)$, which concludes the proof. □

A similar result was shown in [66]. With this at hand, we can derive a continuity result for our regularization approach.

Lemma 3.4. *Let $\gamma_n \to \infty$ be a monotonically increasing sequence of penalty parameters. Further, consider a weakly converging sequence $(y_n, u_n) \rightharpoonup (\overline{y}, \overline{u})$ such that*

$$y_n \in \operatorname*{argmin}_{v \in \mathcal{A}} I_{\gamma_n}(v, u_n).$$

Then, $(\overline{y}, \overline{u}) \in \mathcal{A}_c \times U$ and

$$\overline{y} \in \operatorname*{argmin}_{v \in \mathcal{A}_c} I(v, \overline{u}).$$

Additionally,

$$\lim_{n \to \infty} I_{\gamma_n}(y_n, u_n) = I(\overline{y}, \overline{u}).$$

Proof. The weak convergence of (y_n, u_n) implies its boundedness in $Y \times U$. Thus, we also obtain the boundedness of the outer energy $I_{\mathrm{out}}(y_n, u_n)$ analogously to the proof of Lemma 2.37. Consequently, $I_{\mathrm{strain}}(y_n)$ is bounded as well. Hence, Lemma 2.38 applies and $\overline{y} \in \mathcal{A}$. Next, the relation

$$P(y_n) = \frac{I_{\gamma_n}(y_n, u_n) - I(y_n, u_n)}{\gamma_n}$$

yields $\lim_{n \to \infty} P(y_n) = 0$ since the numerator is bounded as previously established. By combining this result with the weak lower semi-continuity of P, we obtain

$$0 \leq P(\overline{y}) \leq \liminf_{n \to \infty} P(y_n) = 0.$$

This yields $\overline{y} \in \mathcal{A}_c$. Finally, we show that $\overline{y}$ is again a solution to the original contact problem (2.23). From Theorem 2.39, we derive the existence of a state $\tilde{y} \in \mathcal{A}_c$ with

$$\tilde{y} \in \operatorname*{argmin}_{v \in \mathcal{A}_c} I(v, \overline{u}).$$

Furthermore, the weak lower semi-continuity of the total energy I w.r.t. (y_n, u_n) follows from the boundedness of $I_{\text{strain}}(y_n)$ and Lemma 2.36. Consequently, we obtain

$$\begin{aligned}\limsup_{n\to\infty} I_{\gamma_n}(y_n, u_n) &\leq \limsup_{n\to\infty} I_{\gamma_n}(\overline{y}, u_n) = \limsup_{n\to\infty} I(\overline{y}, u_n) = \lim_{n\to\infty} I(\overline{y}, u_n) \\ &= I(\overline{y}, \overline{u}) \leq \liminf_{n\to\infty} I(y_n, u_n) \leq \liminf_{n\to\infty} I_{\gamma_n}(y_n, u_n).\end{aligned}$$

Hence,

$$\lim_{n\to\infty} I_{\gamma_n}(y_n, u_n) = I(\overline{y}, \overline{u}).$$

A similar argumentation was applied in [66, Proof of Lemma 3.3]. From the above results, we derive

$$I(\tilde{y}, \overline{u}) \leq I(\overline{y}, \overline{u}) = \lim_{n\to\infty} I_{\gamma_n}(y_n, u_n) \leq \lim_{n\to\infty} I_{\gamma_n}(\tilde{y}, u_n) = I(\tilde{y}, \overline{u}),$$

which shows that $\overline{y}$ is a minimizer of the total energy functional $I(\,\cdot\,, \overline{u})$ in $\mathcal{A}_c$.

□

Finally, we prove that limit points of regularized solutions exist and satisfy the original contact problem.

Proposition 3.5. *Let $u \in U$ and $\gamma_n \to \infty$ be a monotonically increasing sequence of penalty parameters. Further, y_n denotes a corresponding sequence of minimizers satisfying*

$$y_n \in \operatorname*{argmin}_{v \in \mathcal{A}} I_{\gamma_n}(v, u).$$

Then, y_n has a weakly converging subsequence. The limit point $\overline{y}$ of any such sequence satisfies

$$\overline{y} \in \operatorname*{argmin}_{v \in \mathcal{A}_c} I(v, u).$$

In particular, the energy values also converge: $I_{\gamma_n}(y_n, u) \to I(\overline{y}, u)$.

Proof. The boundedness of y_n follows from Lemma 3.3, enabling us to extract a weakly converging subsequence. Application of Lemma 3.4 to each of these subsequences yields the desired result. The convergence of the energy values results from $I_{\gamma_n}(y_n, u)$ being a monotonically increasing and bounded sequence. □

We conclude that the normal compliance method describes a suitable approach to approximate contact problems in hyperelasticity. Sometimes, it is of interest whether an inverse result of Proposition 3.5 holds.

Remark 3.6. *Let $u \in U$ be a fixed boundary force and $\overline{y}$ a corresponding energy minimizer which satisfies*

$$\overline{y} \in \operatorname*{argmin}_{v \in \mathcal{A}_c} I(v, u).$$

Further, let $\gamma_n \to \infty$ be a monotonically increasing sequence of penalty parameters. Then, it remains unclear whether there exists a sequence y_n with

$$y_n \in \operatorname*{argmin}_{v \in \mathcal{A}} I_{\gamma_n}(v, u)$$

and $y_n \rightharpoonup \overline{y}$.

The main obstacle here is the non-uniqueness of energy minimizers. However, in cases where the uniqueness of solutions can be verified, we obtain an inverse version of Proposition 3.5.

Corollary 3.7. *Let $\gamma_n \to \infty$ be a monotonically increasing sequence of penalty parameters. Assume that for a fixed boundary force u, the contact problem (2.23) admits a unique solution, denoted by $\overline{y}$. Then, there exists a subsequence y_{n_k} of regularized solutions corresponding to γ_{n_k} such that $y_{n_k} \rightharpoonup \overline{y}$ and $I_{\gamma_{n_k}}(y_{n_k}, u) \to I(\overline{y}, u)$.*

Proof. The existence of a sequence y_n of solutions to (2.23) follows from Theorem 3.2. Proposition 3.5 yields a subsequence y_{n_k} that converges weakly to a solution y_* of the contact problem (2.23). Since the solution of (2.23) is unique, we obtain $y_* = \overline{y}$. The convergence $I_{\gamma_{n_k}}(y_{n_k}, u) \to I(\overline{y}, u)$ also follows from Proposition 3.5. □

This result only applies to a limited range of problems since the uniqueness of solutions cannot be expected, not even in very simple settings. The question of whether approximating sequences exist reemerges in Chapter 4. There, we will analyze regularization approaches for optimal control problems.

3.2 Equilibrium conditions of local energy minimizers

In view of numerical algorithms to solve general hyperelastic problems, the characterization of minimizers is a central topic. In particular, it is of interest whether a local minimizer $y_* \in \mathcal{A}$ of the regularized total energy functional I_γ satisfies some kind of first order optimality condition. Note that our regularization term P does not alter the properties of the total energy function I. Since the critical term is the strain energy I_{strain}, and since we want to stay in line with already existing literature on this topic, we restrict our analysis to I. Still, the elaborated results apply to the regularized stored energy function I_γ as well. First, we define local minimizers in the following way.

Definition 3.8 (Local minimizer)**.** A deformation $y \in \mathcal{A}$ is a $W^{1,p}(\Omega)$ local minimizer of $I(\,\cdot\,, u)$ if there exists an $\varepsilon > 0$ such that for all $v \in \mathcal{A}$ with

$$\|v - y\|_{W^{1,p}(\Omega)} \leq \epsilon,$$

the relation $I(v, u) > I(y, u)$ holds.

Given a local minimizer $y_* \in \mathcal{A}$, the aim is to verify first order optimality conditions of the following form:

$$\partial_y I(y_*, u)v = 0 \quad \text{for all } v \in Y.$$

However, this condition does not hold for general problems in hyperelasticity, cf. [7, Problem 5]. The central problem is the compression condition of the stored energy function which states that

$$\det \nabla y \to 0^+ \Rightarrow \hat{W}(x, \nabla y(x)) \to \infty.$$

This condition is necessary to avoid local self-penetration of the body. As a result, the set

$$Y_\infty := \left\{ v \in Y \;\middle|\; \int_\Omega \hat{W}(x, \nabla v(x))\, dx = \infty \right\}$$

is a dense subset of $W^{1,p}(\Omega)$ for $p < \infty$. Consequently, proving Gâteaux differentiability seems to be out of reach, at least for spaces weaker than $W^{1,\infty}(\Omega)$. So far, there exist two main approaches to overcome this problem and derive first order optimality conditions. In the first approach, only energy minimizers in the space $W^{1,\infty}(\Omega)$ are considered.

3.2.1 First order conditions for non-degenerate minimizers

Here, minimizers are assumed to satisfy the following non-degeneracy property.

Definition 3.9 (Non-degenerate deformation)**.** Let $y \in \mathcal{A}$ be a deformation. Then, y is called non-degenerate if there exists an $\epsilon > 0$ such that

$$\det \nabla y(x) \geq \epsilon \quad \text{for a.e. } x \in \Omega$$

is satisfied.

For compressible Mooney-Rivlin models, an analysis of first order optimality conditions based on a non-degeneracy property has been conducted in [66, Theorem 4.6]. There, the authors proved that a non-degenerated energy minimizer $y_* \in W^{1,\infty}(\Omega)$ indeed satisfies the optimality condition

$$\partial_y I(y_*, u)v = 0 \quad \text{for all } v \in W^{1,\infty}(\Omega).$$

A major drawback of this approach is that the regularity and non-degeneracy assumptions cannot be proven a priori. In contrast, sometimes it can be shown that minimizers are not in $W^{1,\infty}(\Omega)$, see, e.g., [9]. Although this seems counter-intuitive, in the case of cavitation, the unboundedness of the stored energy function $\hat{W}$ at some point can be more than compensated by smaller values elsewhere. For a detailed discussion of this topic, the reader is referred to [7, Problem 6].
Additionally, operating in the space $W^{1,\infty}(\Omega)$ has consequences in the optimal control setting. There, the requirement that a minimizer y_* is an element of $W^{1,\infty}(\Omega)$ prevents the derivation of KKT conditions. A thorough analysis of this issue is conducted in Chapter 4. The presented lack of structure renders this approach unsuitable for theoretical discussions of energy minimizers and optimal control problems.

3.2.2 Alternative first order conditions

An alternative method has been discussed in [7]. Imposing a suitable growth condition on the stored energy function $\hat{W}$ yields alternative first order optimality conditions. However, it is not clear how to transfer these results to a numerical implementation. Still, the techniques studied there contribute to the analysis of regularization approaches for elastic contact problems. Therefore, we recapitulate the most important results. In the following analysis, we use the short notation $\hat{W}'$ to denote the matrix-valued derivative of $\hat{W}$. First, we introduce the required growth condition on $\hat{W}$ as defined in [7, Assumption C1].

Assumption 3.10. *Let $\hat{W} : \mathbb{M}^3_+ \to \mathbb{R}$ be the stored energy function of a homogeneous, isotropic, and hyperelastic material. Then, assume that $\hat{W}$ satisfies the following growth condition:*

$$\|\hat{W}'(M)M^T\| \leq K(\hat{W}(M) + 1) \quad \text{for all } M \in \mathbb{M}^3_+, \tag{3.3}$$

where K is a positive constant.

This assumption implies that the stored energy function $\hat{W}$ has polynomial growth, cf. [7, Proposition 2.7]. From there, we derive the following result elaborated in [7, Lemma 2.5].

Lemma 3.11. *Assume the stored energy function $\hat{W}$ satisfies Assumption 3.10 with the corresponding constant $K > 0$. Then, there exists an $\varepsilon > 0$ such that for each $C \in \mathbb{M}^3_+$ with*

$$\|C - \mathrm{Id}\| < \varepsilon,$$

the inequality

$$\|\hat{W}'(CM)M^T\| \leq 3K(\hat{W}(M) + 1) \quad \text{for all } M \in \mathbb{M}^3_+$$

holds.

Proof. The proof follows along the lines of [7, Proof of Lemma 2.5]. First, we show that there exists an $\epsilon > 0$ such that for each matrix $C \in \mathbb{M}^3_+$ with

$$\|C - \mathrm{Id}\| < \varepsilon$$

and for all $M \in \mathbb{M}^3_+$, the estimate

$$\hat{W}(CM) + 1 \leq \frac{3}{2}(\hat{W}(M) + 1) \tag{3.4}$$

holds. We define

$$C(t) := tC + (1 - t)\,\mathrm{Id} \quad \text{for } t \in [0, 1].$$

For sufficiently small $\varepsilon \in (0, \frac{1}{6K})$ with $\|C - \mathrm{Id}\| < \varepsilon$, we obtain $\|C(t)^{-1}\| \leq 2$ for all $t \in [0,1]$ due to $\|\mathrm{Id}\| = \sqrt{3} < 2$. From there, we derive

$$\begin{aligned}\hat{W}(CM) - \hat{W}(M) &= \int_0^1 \frac{d}{dt}\hat{W}(C(t)M)\,dt = \int_0^1 \hat{W}'(C(t)M)\cdot((C-\mathrm{Id})M)\,dt \\ &= \int_0^1 \hat{W}'(C(t)M)(C(t)M)^T\cdot((C-\mathrm{Id})C(t)^{-1})\,dt \\ &\overset{(3.3)}{\leq} K\int_0^1 (\hat{W}(C(t)M)+1)\cdot\|C-\mathrm{Id}\|\cdot\|C(t)^{-1}\|\,dt \\ &\leq 2K\varepsilon\int_0^1(\hat{W}(C(t)M)+1)\,dt.\end{aligned}$$

Defining

$$\zeta(M) := \sup_{C:\|C-\mathrm{Id}\|<\varepsilon} \hat{W}(CM)$$

yields

$$\hat{W}(CM) - \hat{W}(M) \leq \zeta(M) - \hat{W}(M) \leq 2K\varepsilon(\zeta(M)+1) \leq \frac{1}{3}(\zeta(M)+1),$$

and consequently,

$$\zeta(M) \leq \frac{1}{3}(\zeta(M)+1) + \hat{W}(M).$$

Solving for $\zeta(M)$ leads to

$$\hat{W}(CM) \leq \zeta(M) \leq \frac{3}{2}\hat{W}(M) + \frac{1}{2},$$

which shows (3.4). Finally, for $\|C - \mathrm{Id}\| < \varepsilon$, we obtain

$$\begin{aligned}\|W'(CM)M^T\| = \|W'(CM)(CM)^TC^{-T}\| &\overset{(3.3)}{\leq} K(W(CM)+1)\|C^{-T}\| \\ &\overset{(3.4)}{\leq} \tfrac{3}{2}K(W(M)+1)\|C^{-T}\| \leq 3K(W(M)+1).\end{aligned}$$

This concludes the proof. □

Lemma 3.11 essentially describes a sensitivity result for the stored energy function $\hat{W}$ for small perturbations induced by a matrix multiplication in its argument. With the necessary estimates at hand, we can derive alternative first order optimality conditions. The following result was elaborated in [7, Theorem 2.4].

Theorem 3.12. *Let u be a fixed boundary force and let the stored energy function $\hat{W}$ satisfy Assumption 3.10. Further, define the admissible set of deformations by*

$$\mathcal{A}_d := \{y \in W^{1,p}(\Omega) \mid y = \mathrm{id} \ \text{ a.e. on } \Gamma_D,\ \det\nabla y > 0 \ \text{ a.e. in } \Omega\}$$

and the set of test functions by

$$\Phi := \{\varphi \in C^1(\mathbb{R}^3, \mathbb{R}^3) \mid \varphi = 0 \text{ on } \Gamma_D,\ \varphi \text{ and } \nabla\varphi \text{ are uniformly bounded }\}.$$

Then, a local minimizer y_* of the total energy function $I(\,\cdot\,, u)$ in the set $\mathcal{A}_d$ satisfies

$$\int_\Omega (\hat{W}'(\nabla y_*(x))\nabla y_*(x)^T) \cdot \nabla\varphi(y_*(x))\,dx - \int_{\Gamma_N} u(x)\varphi(y_*(x))\,dx = 0 \tag{3.5}$$

for all $\varphi \in \Phi$.

Proof. For $t \in \mathbb{R}$, consider the disturbed deformation

$$y_t(x) := y_*(x) + t\varphi(y_*(x)).$$

The goal is to compute the limit

$$\lim_{t\to 0} \frac{1}{t}(I(y_t) - I(y_*)).$$

What sets this approach apart from other methods is the utilization of small perturbations of the deformed domain. Usually, only perturbations of the undeformed domain are studied. The technique considered here allows us to rely solely on the growth condition defined in Assumption 3.10 to derive optimality conditions. Computing the respective deformation gradient yields

$$\nabla y_t(x) = (\mathrm{Id} + t\nabla\varphi(y_*(x)))\nabla y_*(x) \quad \text{for a.e. } x \in \Omega.$$

First, note that for sufficiently small $|t|$, the disturbed gradient satisfies

$$\det \nabla y_t(x) > 0 \quad \text{for a.e. } x \in \Omega.$$

Consequently, for suitable $t \in \mathbb{R}$, the disturbed deformation y_t is again an element of the admissible set $\mathcal{A}_d$. Additionally,

$$\lim_{t\to 0} \|y_t - y_*\|_Y = 0.$$

Since y_* is a local minimizer by definition, $I(y_t) > I(y_*)$ holds for sufficiently small $|t|$. For the outer energy I_{out}, we directly obtain

$$\lim_{t\to 0} \frac{1}{t}(-I_{\mathrm{out}}(y_t) + I_{\mathrm{out}}(y_*)) = -\int_{\Gamma_N} u(x)\,\varphi(y_*(x))\,dx.$$

Next, we compute

$$\begin{aligned}\frac{1}{t}(I_{\mathrm{strain}}(y_t) - I_{\mathrm{strain}}(y_*)) &= \frac{1}{t}\int_\Omega \int_0^1 \frac{d}{ds}\hat{W}((\mathrm{Id} + st\nabla\varphi(y_*(x)))\nabla y_*(x))\,ds\,dx \\ &= \int_\Omega \int_0^1 \hat{W}'((\mathrm{Id} + st\nabla\varphi(y_*(x)))\nabla y_*(x)) \cdot (\nabla\varphi(y_*(x))\nabla y_*(x))\,ds\,dx.\end{aligned}$$

Lemma 3.11 and the uniform boundedness of $\nabla\varphi$ yield the following estimate for the integrand:

$$\|\hat{W}'((\mathrm{Id}+st\nabla\varphi(y_*(x)))\nabla y_*(x))\cdot(\nabla\varphi(y_*(x))\nabla y_*(x))\| \\ \leq 3K(\hat{W}(\nabla y_*(x))+1)\sup_{v\in\mathbb{R}^3}\|\nabla\varphi(v)\|,$$

where K is a positive constant. Thus, the theorem of dominated convergence applies, cf. [62, Theorem 1.8], and passing to the limit $t \to 0$ leads to the desired result. □

Although we have been able to derive optimality conditions in a reasonable setting, it remains unclear how to transfer these results to a numerical approach. The main obstacle is that the test functions in (3.5) are defined on the entire space $\mathbb{R}^3$. Therefore, it is not obvious how this optimality condition translates to an approximation in a finite dimensional setting, e.g., finite elements. Nevertheless, the elaborated techniques offer valuable insight into the problem structure, and they will prove essential for further analyzing our regularization approach in the following section.

3.3 Asymptotic rates of the normal compliance method

Summarizing the previous results, we have introduced a suitable regularization approach (3.2) to approximate the original contact problem (2.23). Correspondingly, a convergence result for sequences of regularized solutions has been established in Proposition 3.5, at least in the sense of subsequences. In addition to the previous results, we would like to derive convergence rates for both the regularized energy values and the corresponding minimizers. For sequences $\gamma_n \to \infty$, we are thus interested in estimates of the form:

$$\min_{v\in\mathcal{A}_c} I(v,u) - \min_{v\in\mathcal{A}} I_{\gamma_n}(v,u) \leq c_1\gamma_n^{-\rho_1}, \tag{3.6}$$

and

$$\|y_n - \overline{y}\|_{L^p(\Omega)} \leq c_2\gamma_n^{-\rho_2}, \tag{3.7}$$

with positive constants c_1, c_2, ρ_1, and ρ_2. Here, y_n and $\overline{y}$ are solutions of the corresponding regularized contact problem (3.2) and the original one (2.23), respectively. Since Proposition 3.5 only yields the weak convergence $y_n \rightharpoonup \overline{y}$ in $W^{1,p}(\Omega)$, strong convergence can only be expected in $L^p(\Omega)$, assuming the compact embedding $W^{1,p}(\Omega) \hookrightarrow L^p(\Omega)$ holds. Still, deriving a convergence result as defined in (3.7) seems to be out of reach with the current tools and techniques at hand. The total energy functional I does not explicitly depend on $\|y\|_{L^p(\Omega)}$, but on the deformation gradient ∇y. Additionally, the non-uniqueness of energy minimizers further impedes a thorough convergence analysis. However, by using refined arguments and an assumption on the geometric setting, we can derive a priori estimates on the rate of convergence, at least for the energy values. Such an estimate is possible due to the uniqueness of the optimal energy values $I_{\gamma_n}(y_n,u)$, in contrast to the corresponding minimizers y_n, which are not unique. Besides this, an

estimate for the maximum constraint violation of the regularized solutions y_n can be derived.
The elaborated convergence rates not only yield an estimate for the quality of our regularization approach, but they also contribute to the analysis of optimal control problems in Chapter 4. To this end, we will use ideas from [7, 47].

3.3.1 Asymptotic rates of the energy

Consider the case $p > 3$. Then $W^{1,p}(\Omega)$ is continuously embedded into the space $C^\beta(\Omega)$ of Hölder continuous functions for some suitable $\beta \in]0,1[$. Further, the following assumption on the geometry of the boundary conditions is made.

Assumption 3.13. *Assume that there is a constant $K > 0$ such that for each $\varepsilon > 0$, there exists an invertible mapping $\psi_\varepsilon \in W^{1,\infty}(\mathbb{R}^3,\mathbb{R}^3) \cap C^1(\mathbb{R}^3,\mathbb{R}^3)$ satisfying*

$$\|\psi_\varepsilon - \mathrm{id}\,\|_{W^{1,\infty}(\mathbb{R}^3,\mathbb{R}^3)} \leq K\varepsilon, \tag{3.8}$$

$\psi_\varepsilon = \mathrm{id}$ on Γ_D, and

$$\psi_\varepsilon(x)_3 \geq 0 \quad \text{for all } x \in \Gamma_C \text{ with } x_3 \geq -\varepsilon.$$

Denoting

$$\mathcal{A}_\varepsilon := \{y \in \mathcal{A} \mid y_3 \geq -\varepsilon \text{ on } \Gamma_C\},$$

the inclusion $y \in \mathcal{A}_\varepsilon$ implies $\psi_\varepsilon \circ y \in \mathcal{C}$. Next, Assumption 3.10 is incorporated. From there, it is possible to derive an upper bound for the change in energy if we perturb the current deformation by applying the transformation ψ_ε from Assumption 3.13.

Lemma 3.14. *Let $u \in U$ be a fixed boundary force and let Assumptions 3.10 and 3.13 hold. If $\varepsilon > 0$ is sufficiently small, then there exist positive constants c_0 and c_1 such that*

$$I(\psi_\varepsilon \circ y, u) - I(y,u) \leq c_0(I_{\mathrm{strain}}(y) + c_1)\varepsilon,$$

for all $y \in \mathcal{A}_\varepsilon$.

Proof. Consider the following splitting

$$I(\psi_\varepsilon \circ y, u) - I(y,u) = I_{\mathrm{strain}}(\psi_\varepsilon \circ y) - I_{\mathrm{strain}}(y) - I_{\mathrm{out}}(\psi_\varepsilon \circ y, u) + I_{\mathrm{out}}(y,u).$$

First, we derive an estimate for the term $I_{\mathrm{strain}}(\psi_\varepsilon \circ y, u) - I_{\mathrm{strain}}(y,u)$. Let $\varepsilon > 0$ be sufficiently small and let $C, M \in \mathbb{M}^3_+$ such that $\|C - \mathrm{Id}\,\| < \varepsilon$. Further, we define

$$\zeta(M) := \sup_{C:\|C-\mathrm{Id}\,\|<\varepsilon} \hat{W}(CM).$$

Following the techniques in the proof of Lemma 3.11, we obtain

$$\hat{W}(CM) - \hat{W}(M) \leq 2K\varepsilon(\zeta(M)+1) \leq 3K\varepsilon(\hat{W}(M)+1). \tag{3.9}$$

We will utilize these results to derive a similar estimate for the total energy functional I. By definition, ψ_ε is continuously differentiable and $y \in W^{1,p}(\Omega)$. Thus, the chain rule applies to $\psi_\varepsilon \circ y$, and the respective derivative is well-defined in $L^p(\Omega)$, see, e.g., [62, Theorem 6.16]. With the notation for the deformation gradient in mind, we denote the derivative of $\psi_\varepsilon \circ y$ by $\nabla\psi_\varepsilon(y)\nabla y$. Further, due to Condition (3.8) combined with the embedding $W^{1,p}(\Omega) \hookrightarrow C^\beta(\Omega)$, there exists a constant $K_1 > 0$ such that

$$\|\nabla\psi_\varepsilon(y(x)) - \mathrm{Id}\,\| < K_1\varepsilon \quad \text{for all } x \in \Omega.$$

The same reasoning yields

$$\det \nabla\psi_\varepsilon(y(x)) > 0 \quad \text{for all } x \in \Omega.$$

For sufficiently small ε, Estimate (3.9) transfers to the strain energy as follows:

$$\begin{aligned} I_{\mathrm{strain}}(\psi_\varepsilon \circ y) - I_{\mathrm{strain}}(y) &= \int_\Omega \hat{W}(\nabla\psi_\varepsilon(y(x))\nabla y(x)) - \hat{W}(\nabla y(x))\,dx \\ &\leq C\varepsilon \int_\Omega \hat{W}(\nabla y(x)) + 1\,dx, \end{aligned}$$

for suitable $C > 0$. Similarly, for the difference of the outer energies, we obtain

$$-I_{\mathrm{out}}(\psi_\varepsilon(y), u) + I_{\mathrm{out}}(y, u) = \int_{\Gamma_N} (-\psi_\varepsilon(y(x)) + y(x))u(x)\,ds \overset{(3.8)}{\leq} c\,\varepsilon,$$

for suitable $c > 0$. Finally, combining the two estimates yields

$$I(\psi_\varepsilon \circ y, u) - I(y, u) \leq C\varepsilon(I_{\mathrm{strain}}(y) + \mathrm{vol}(\Omega)) + c\,\varepsilon,$$

which concludes the proof. □

This lemma implies the following result.

Corollary 3.15. *Let $u \in U$ be a fixed boundary force and let Assumptions 3.10 and 3.13 hold. Then, there exists a constant $C > 0$ such that for sufficiently small $\epsilon > 0$, the estimate*

$$0 \leq \min_{v \in \mathcal{A}_c} I(v, u) - \min_{v \in \mathcal{A}_\varepsilon} I(v, u) \leq C\varepsilon$$

holds.

Proof. First, we choose ε as defined in Lemma 3.14. Further, let y_* and y_ε be minimizers of $I(\cdot, u)$ in $\mathcal{A}_c$ and $\mathcal{A}_\varepsilon$, respectively. Note that $\min_{v\in\mathcal{A}_\varepsilon} I(v, u)$ is bounded as ε approaches zero. Consequently, this holds for $I_{\mathrm{strain}}(y_\varepsilon)$ as well. Then, by Lemma 3.14, there exist positive constants c_0, c_1, and c_2 such that

$$0 \leq I(y_*, u) - I(y_\varepsilon, u) \leq I(\psi_\varepsilon(y_\varepsilon), u) - I(y_\varepsilon, u) \leq c_0(I_{\mathrm{strain}}(y_\varepsilon) + c_1)\varepsilon \leq c_2\varepsilon.$$

□

The minimization problem $\min_{v\in\mathcal{A}_\varepsilon} I(v,u)$ can be interpreted as a relaxed contact problem where some violation of the constraints is allowed. In Corollary 3.15, we have proven that the difference between the minimal energy values for the relaxed problem and the original contact problem approaches zero at least linearly with the maximum violation ε of the contact constraints. Thus, by examining the relation between the penalty parameter γ and the maximum violation ε, we can derive a convergence rate w.r.t. γ.

3.3.2 An estimate for the constraint violation

It remains to show that for sufficiently large γ, energy minimizers of I_γ are contained in $\mathcal{A}_\varepsilon$, where $\varepsilon = O(\gamma^{-\rho})$ for a certain rate ρ. For this, we use that the corresponding sequence of minimizers y_γ is bounded in $C^\beta(\Omega)$ for $p > 3$.
Our analysis here is based on the techniques applied in [47, Proposition 2.4]. Considering the maximum constraint violation function $[y]_+$ on Γ_C, it follows that $[y]_+$ enters the regularized contact problem (3.2) via the penalty function P. Given the definition $P(y) := \frac{1}{k}\|[y]_+\|^k_{L^k(\Gamma_C)}$, the question arises whether under suitable regularity assumptions, we can estimate the maximum constraint violation $\|[y]_+\|_{L^\infty(\Gamma_C)}$ against the norm $\|[y]_+\|_{L^k(\Gamma_C)}$. To do so, we first derive an upper bound for the supremum norm in a general setting.
At this point, we require additional assumptions on the boundary segment Γ_C in order to simplify the computations. Thus, the segment Γ_C is assumed to be a flat two-dimensional sub-manifold of $\mathbb{R}^3$. Nonetheless, under suitable assumptions, the following results still hold if Γ_C is curved. However, this is fairly technical and does not yield further insight. Consequently, only the simple case is considered here. Further, the following requirement on the boundary segment Γ_C is added.

Assumption 3.16. *Let Γ_C satisfy a uniform interior cone condition in the following sense: for each point $x \in \Gamma_C$, it is possible to construct a two-dimensional circular sector*

$$S_{R',\theta}(x) \subset \Gamma_C \tag{3.10}$$

with center at x, radius $R' > 0$, and center angle $\theta > 0$. Then, assume that R' and θ can be chosen independently of x.

Here, each circular sector $S_{R',\theta}(x) \subset \Gamma_C$ is interpreted as a two-dimensional sub-manifold in $\mathbb{R}^3$. Due to the flatness assumption on Γ_C, this is a meaningful definition. We assume that Assumption 3.16 holds throughout the subsequent examination. Next, we can derive the following estimate.

Proposition 3.17. *Let $\beta \in]0,1[$ and $s \geq 1$. Further, let $f \in C^\beta(\Gamma_C) \cap L^s(\Gamma_C)$ be a positive function. Additionally, let $\|f\|_{C^\beta(\Gamma_C)} \leq M$ and $\|f\|_{L^s(\Gamma_C)} \leq 1$. Without loss of generality, assume that $0 \in \Gamma_C$ and $f(0) = \|f\|_{L^\infty(\Gamma_C)}$. Due to Assumption 3.16, there exists a circular sector $S_{R',\theta}(0) \subset \Gamma_C$ with $R' \leq 1$. Then, the following estimate holds:*

$$\|f\|_{L^\infty(\Gamma_C)} \leq c(s,\beta,R',\theta,M)\|f\|^{\frac{\beta s}{\beta s+2}}_{L^s(\Gamma_C)},$$

where the positive constant $c(s, \beta, R', \theta, M)$ only depends on the exponents β and s, the angle θ, the radius R', and the upper bound M.

Proof. First, we define

$$R = \left(\frac{f(0)}{\|f\|_{C^\beta(\Gamma_C)}} \right)^{\frac{1}{\beta}} = \left(\frac{\|f\|_{L^\infty(\Gamma_C)}}{\|f\|_{C^\beta(\Gamma_C)}} \right)^{\frac{1}{\beta}}. \tag{3.11}$$

We choose the maximum positive $\alpha \leq 1$ such that for $\tilde{R} := \alpha R$, the inequalities

$$\left(\frac{\tilde{R}}{R}\right)^\beta s \leq 1 \Leftrightarrow \alpha^\beta s \leq 1 \tag{3.12}$$

and $\tilde{R} \leq R'$ hold. As a result,

$$S_{\tilde{R},\theta}(0) \subset S_{R',\theta}(0). \tag{3.13}$$

In addition, recall Bernoulli's inequality [14]:

$$(1+x)^n \geq 1 + nx, \tag{3.14}$$

for real numbers $x \geq -1$ and $n \geq 1$. The Hölder continuity of f yields the estimate

$$f(x) \geq f(0) - \|f\|_{C^\beta(\Gamma_c)} \|x - 0\|^\beta \quad \text{for all } x \in S_{R',\theta}(0). \tag{3.15}$$

Combining the previous results leads to the following estimate:

$$\begin{aligned}
\|f\|^s_{L^s(\Gamma_C)} &= \int_{\Gamma_C} |f(x)|^s \, dx \geq \int_{S_{R',\theta}(0)} |f(x)|^s \, dx \\
&\overset{(3.13)(3.15)}{\geq} \int_{S_{\tilde{R},\theta}(0)} |f(0) - \|f\|_{C^\beta(\Gamma_C)} \|x-0\|^\beta|^s \, dx \\
&\overset{(3.11)}{=} \|f\|^s_{C^\beta(\Gamma_C)} \int_{S_{\tilde{R},\theta}(0)} |R^\beta - \|x-0\|^\beta|^s \, dx \\
&= 2|S_{1,\theta}(0)| \|f\|^s_{C^\beta(\Gamma_C)} \int_0^{\tilde{R}} |R^\beta - r^\beta|^s r \, dr \\
&\overset{(3.12)(3.14)}{\geq} 2|S_{1,\theta}(0)| \|f\|^s_{C^\beta(\Gamma_C)} R^{\beta s} \int_0^{\tilde{R}} \left(1 - s\tfrac{r^\beta}{R^\beta}\right) r \, dr.
\end{aligned}$$

At this point, we have to distinguish between two cases. The first case is $\alpha^\beta s = 1$, which implies $\tilde{R} \leq R'$. In this case, we obtain

$$\begin{aligned}
&2|S_{1,\theta}(0)| \|f\|^s_{C^\beta(\Gamma_C)} R^{\beta s} \int_0^{\tilde{R}} \left(1 - s\tfrac{r^\beta}{R^\beta}\right) r \, dr \\
&= 2|S_{1,\theta}(0)| \|f\|^s_{C^\beta(\Gamma_C)} R^{\beta s} \left[\tfrac{1}{2} r^2 - \tfrac{s r^{\beta+2}}{(\beta+2)R^\beta}\right]_0^{\alpha R} \\
&= 2|S_{1,\theta}(0)| \|f\|^s_{C^\beta(\Gamma_C)} R^{\beta s + 2} \tfrac{\alpha^2}{\beta+2} \left(\tfrac{1}{2}(\beta+2) - s\alpha^\beta\right).
\end{aligned}$$

Due to the condition $\alpha^\beta s = 1$, we know that the constant

$$c_0(s,\beta,R',\theta) := 2|S_{1,\theta}(0)|\tfrac{\alpha^2}{\beta+2}\left(\tfrac{1}{2}(\beta+2) - s\alpha^\beta\right) \overset{(3.12)}{>} 0$$

does not approach zero even if $R \to 0$. Thus, we can insert the definition of R and obtain the estimate

$$\|f\|^s_{L^s(\Gamma_C)} \geq c_0(s,\beta,R',\theta)\|f\|^s_{C^\beta(\Gamma_C)}\left(\frac{\|f\|_{L^\infty(\Gamma_C)}}{\|f\|_{C^\beta(\Gamma_C)}}\right)^{\frac{\beta s+2}{\beta}}.$$

Now, solving for $\|f\|_{L^\infty(\Gamma_C)}$ yields

$$\begin{aligned}\|f\|_{L^\infty(\Gamma_C)} &\leq c_0(s,\beta,R',\theta)^{-\frac{\beta}{\beta s+2}}\|f\|^{\frac{2}{\beta s+2}}_{C^\beta(\Gamma_C)}\|f\|^{\frac{\beta s}{\beta s+2}}_{L^s(\Gamma_C)}\\ &\leq c(s,\beta,R',\theta,M)\|f\|^{\frac{\beta s}{\beta s+2}}_{L^s(\Gamma_C)},\end{aligned}$$

which shows the desired result. For the second case, we have $\alpha^\beta s < 1$, which implies $\tilde{R} = R'$. This leads to the estimate

$$\begin{aligned}&2|S_{1,\theta}(0)|\|f\|^s_{C^\beta(\Gamma_C)}R^{\beta s}\int_0^{\tilde{R}}\left(1 - s\tfrac{r^\beta}{R^\beta}\right)r\,dr\\ &= 2|S_{1,\theta}(0)|\|f\|^s_{C^\beta(\Gamma_C)}R^{\beta s}\left[\tfrac{1}{2}r^2 - \tfrac{sr^{\beta+2}}{(\beta+2)R^\beta}\right]_0^{\tilde{R}}\\ &= 2|S_{1,\theta}(0)|\|f\|^s_{C^\beta(\Gamma_C)}R^{\beta s}\left(\tfrac{1}{2}\tilde{R}^2 - \tfrac{s\tilde{R}^{\beta+2}}{(\beta+2)R^\beta}\right)\\ &\overset{(3.12)}{\geq} 2|S_{1,\theta}(0)|\|f\|^s_{C^\beta(\Gamma_C)}R^{\beta s}\tilde{R}^2\left(\tfrac{1}{2} - \tfrac{1}{(\beta+2)}\right).\end{aligned}$$

By applying $\tilde{R} = R'$, we can define the constant

$$c_0(s,\beta,R',\theta) := 2|S_{1,\theta}(0)|(R')^2\left(\tfrac{1}{2} - \tfrac{1}{(\beta+2)}\right) > 0,$$

which does not depend on R. Analogously to the computations above, we insert the definition of R and obtain

$$\|f\|^s_{L^s(\Gamma_C)} \geq c_0(s,\beta,R',\theta)\|f\|^s_{C^\beta(\Gamma_C)}\left(\frac{\|f\|_{L^\infty(\Gamma_C)}}{\|f\|_{C^\beta(\Gamma_C)}}\right)^{\frac{\beta s}{\beta}}.$$

Solving for $\|f\|_{L^\infty(\Gamma_C)}$ and using $\|f\|_{L^s(\Gamma_C)} \leq 1$ yields

$$\|f\|_{L^\infty(\Gamma_C)} \leq c_0(s,\beta,R',\theta)^{-\frac{1}{s}}\|f\|_{L^s(\Gamma_C)} \leq c(s,\beta,R',\theta)\|f\|^{\frac{\beta s}{\beta s+2}}_{L^s(\Gamma_C)}.$$

Taking both estimates together concludes the proof. □

Inserting $[y]_+$ into Proposition 3.17, we see that the maximum constraint violation can be estimated by the penalty function since $P(y) := \frac{1}{k}\|[y]_+\|^k_{L^k(\Gamma_C)}$. Thus, we successfully shifted the problem of deriving a convergence rate of $\|[y]_+\|_{L^\infty(\Gamma_C)}$ to the penalty function P. Intuitively, it is clear how to proceed. For each sequence $\gamma_n \to \infty$, the corresponding minimal energies $I_{\gamma_n}(y_n, u)$ are bounded due to Lemma 3.3. Consequently, $P(y_n)$ has to approach zero at least at the same rate as γ_n. Proposition 3.17 provides us with the tools to derive an explicit convergence rate.

Corollary 3.18. *Let $\gamma_n \to \infty$ be monotonically increasing and let $y_n \subset Y$ be a bounded sequence. Then, there exists a constant $c > 0$ such that*

$$\|[y_n]_+\|_{L^\infty(\Gamma_C)} \leq cP(y_n)^{\frac{\beta}{k\beta+2}}.$$

Proof. The continuous embedding of $W^{1,p}(\Omega)$ into the space $C^\beta(\Omega)$ yields the boundedness of $[y_n]_+$ in the space $C^\beta(\Gamma_C)$. By definition, $P(y_n) := \frac{1}{k}\|[y_n]_+\|^k_{L^k(\Gamma_C)}$. Thus, Proposition 3.17 applies, and we obtain the stated estimate. □

From this result, we can directly deduce a convergence rate for the regularized total energy.

Theorem 3.19. *Let $u \in U$ be some fixed boundary force. Additionally, let $\gamma_n \to \infty$ be a monotonically increasing sequence of penalty parameters and y_n a sequence of minimizers to the corresponding regularized contact problems (3.2). Further, assume that $W^{1,p}(\Omega)$ is continuously embedded into the space $C^\beta(\Omega)$ and that Assumptions 3.10, 3.13, and 3.16 hold. Then, there exists a constant $c > 0$ such that*

$$\begin{aligned}
\|[y_n]_+\|_{L_\infty(\Gamma_C)} &\leq c\gamma_n^{-\frac{\beta}{(k-1)\beta+2}},\\
\min_{v\in\mathcal{A}_c} I(v,u) - I(y_n,u) &\leq c\gamma_n^{-\frac{\beta}{(k-1)\beta+2}},\\
\min_{v\in\mathcal{A}_c} I(v,u) - \min_{v\in\mathcal{A}} I_{\gamma_n}(v,u) &\leq c\gamma_n^{-\frac{\beta}{(k-1)\beta+2}}.
\end{aligned}$$

Proof. In the proof, we use the generic positive constants c and d which change throughout the estimates. Lemma 3.14 yields the transformation functions ψ_{ε_n} with

$$\varepsilon_n = \|[y_n]_+\|_{L^\infty(\Gamma_C)}.$$

By optimality of y_n and Lemma 3.14, we obtain the estimate

$$\begin{aligned}
\gamma_n P(y_n) = I_{\gamma_n}(y_n,u) - I(y_n,u) &\leq \min_{v\in\mathcal{A}_c} I_{\gamma_n}(v,u) - I(y_n,u) = \min_{v\in\mathcal{A}_c} I(v,u) - I(y_n,u)\\
&\leq I(\psi_{\varepsilon_n}\circ y_n, u) - I(y_n,u) \leq c(I_{\text{strain}}(y_n)+d)\|[y_n]_+\|_{L^\infty(\Gamma_C)},
\end{aligned} \tag{3.16}$$

for sufficiently large γ_n. Taking into account the boundedness of $I_{\text{strain}}(y_n)$ due to Lemma 3.3, we can derive

$$P(y_n) \leq c\gamma_n^{-1}\|[y_n]_+\|_{L^\infty(\Gamma_C)}.$$

Further, due to Corollary 3.18, the estimates

$$\|[y_n]_+\|_{L^\infty(\Gamma_C)} \leq cP(y_n)^{\frac{\beta}{k\beta+2}} \leq c\left(\gamma_n^{-1}\|[y_n]_+\|_{L^\infty(\Gamma_C)}\right)^{\frac{\beta}{k\beta+2}}$$

hold. Finally, solving for $\|[y_n]_+\|_{L^\infty(\Gamma_C)}$ yields

$$\|[y_n]_+\|_{L^\infty(\Gamma_C)} \leq c\gamma_n^{-\frac{\beta}{(k-1)\beta+2}},$$

and combined with (3.16), the desired result. □

For this proof, an anonymous referee contributed valuable remarks that helped to simplify some argumentations. From a theoretical point of view, the convergence of the energy hinges on an a priori bound on the Hölder continuity of the solutions for some $\beta > 0$. In practical computations, β can be quite large, e.g., $\beta \to 1$. A thorough numerical study of this issue will be conducted in Chapter 7. As we will see, the results derived there indicate a faster convergence rate than predicted by the theoretical results established in this chapter.

3.4 Summary

In summary, we have successfully applied the normal compliance approach to nonlinear elastic contact problems to obtain a regularized problem of the form:

$$y \in \operatorname*{argmin}_{v\in\mathcal{A}} I_\gamma(v,u).$$

Relaxing the contact constraints simplifies the numerical treatment significantly. Additionally, we verified that this regularized problem is well-posed and can be utilized to approximate solutions to the original contact problem (2.23). Moreover, under suitable structural assumptions, corresponding convergence rates have been established. For sequences $\gamma_n \to \infty$, we elaborated estimates of the form

$$\min_{v\in\mathcal{A}_c} I(v,u) - \min_{v\in\mathcal{A}} I_{\gamma_n}(v,u) \leq c_1\gamma_n^{-\rho},$$

and

$$\|[y_n]_+\|_{L^\infty(\Gamma_C)} \leq c_2\gamma_n^{-\rho},$$

where y_n denotes the sequence of regularized solutions corresponding to Problem (3.2). However, similar estimates for the norms $\|y_n\|_{W^{1,p}(\Omega)}$ and $\|y_n\|_{L^p(\Omega)}$ seem to be out of reach in the current setting. Further, it is not possible to derive first order optimality conditions in a general setting due to the fact that

$$Y_\infty := \left\{ v \in Y \;\middle|\; \int_\Omega \hat{W}(x,\nabla v(x))\,dx = \infty \right\}$$

is a dense subset in the space $W^{1,p}(\Omega)$. Under suitable assumptions, this issue can be overcome by choosing the space $W^{1,\infty}(\Omega)$. Still, results that apply to general cases have

not been derived yet. Also, the alternative first order optimality condition (3.5) turned out to be unsuited for the application in numerics so far. Thus, first order optimality conditions can only be derived formally, and a thorough analysis of this issue remains a subject of future research.

Chapter 4

Optimal Control of Nonlinear Elastic Contact Problems

Optimal control problems in nonlinear elasticity have recently received increased attention. Most notably, in [64, 66], an optimal control problem in hyperelasticity was used to compute designs of implants. In that context, an existence result was established, and the elaborated techniques serve as the starting point for our analysis. Further examinations were conducted in [39, 40]. There, the authors studied biological models, including the movement of heliotropic flowers. In addition, nonlinear elasticity is increasingly considered for shape optimization problems, see, e.g., [80, 87]. One application of this is the optimal design of structures.

The aim of this chapter is to extend existing results to optimal control of contact problems. In addition, suitable regularizations of the contact constraints are introduced to apply robust solution algorithms as developed in [67]. Correspondingly, convergence results are established for these new approaches.

This chapter is structured as follows. In Section 4.1, we derive an optimal control problem where a nonlinear elastic contact problem acts as the constraint. This yields a bilevel optimization problem that inherits all the theoretical difficulties from nonlinear elasticity. First and foremost, the bilevel structure does not admit a unique solution due to the lack of convexity. Thereafter, we apply the normal compliance method to remove the nonsmoothness resulting from the contact constraints. This leads to a regularized optimal control problem. The existence of solutions to both problems is discussed in detail. Section 4.2 addresses the convergence of the regularization approaches introduced in the prior section. At this, we modify the normal compliance method to derive convergence results. This new approach better reflects the structure of the entire optimal control problem, and as a result, strong structural assumptions are not required. Finally, Section 4.3 briefly discusses the derivation of KKT conditions.

Parts of this chapter have been published in [95].

4.1 Optimal control of contact problems

In the optimal control setting, we want to minimize an objective functional

$$J : Y \times U \to \mathbb{R}$$

of the form

$$J(y,u) := \frac{1}{2}\|y - y_{\mathrm{d}}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_U, \tag{4.1}$$

where $y_{\mathrm{d}} \in L^2(\Omega)$ represents the desired state and $\alpha > 0$ denotes the regularization parameter. This standard tracking type functional is obviously weakly lower semicontinuous and coercive w.r.t. its second argument. Here, we require $p \geq 2$ and $q = 2$ for the spaces $Y = W^{1,p}(\Omega)$ and $U = L^q(\Gamma_N)$. An alternative to choosing a tracking type functional was discussed in [39, 40]. There, the objective was shifted to approximate a desired direction instead of a desired deformation. This approach leads to more complicated objective functionals and requires a thorough examination. Such a study would be beyond the scope of this work, and therefore, we restrict the analysis here to tracking type functionals instead.
As the constraint, each optimal state y_* also needs to be a minimizer of the total energy functional:

$$y_* \in \operatorname*{argmin}_{v \in \mathcal{A}_c} I(v, u_*),$$

where u_* is the corresponding optimal control. With the objective functional at hand, the optimal control problem reads as follows:

$$\begin{aligned} &\min_{(y,u)\in Y\times U} J(y,u) \\ &s.t.\ y \in \operatorname*{argmin}_{v\in\mathcal{A}_c} I(v,u). \end{aligned} \tag{4.2}$$

4.1.1 Existence of optimal solutions

When proving the existence of optimal solutions to Problem (4.2), we encounter several difficulties. As already mentioned in Chapter 3, it is not possible to derive first order optimality conditions for the lower level problem without strong additional assumptions. Further, the total energy functional I is non-convex, and therefore, its minimizers do not have to be unique. In [64, 66], the existence of solutions to an optimal control problem in hyperelasticity without contact constraints has been proven. We can directly transfer the results from [64, 66] to our analysis. Before the existence of solutions is addressed, we introduce the following definition.

Definition 4.1 (Solution set)**.** The solution set $\mathcal{S}$ is defined as

$$\mathcal{S} := \left\{ (y,u) \in Y \times U \,\middle|\, y \in \operatorname*{argmin}_{v\in\mathcal{A}_c} I(v,u) \right\}.$$

The next theorem establishes the existence of optimal solutions to Problem (4.2).

Theorem 4.2. *The optimal control problem* (4.2) *has at least one optimal solution in* $\mathcal{S}$.

Proof. The proof follows the lines of [66, Proof of Theorem 3.1]. Let $(y_n, u_n) \subset \mathcal{S}$ be an infimizing sequence, whereby $J(y_n, u_n)$ is bounded. We know that such a sequence exists due to Assumption 2.31 and due to the definition of the tracking functional J. The coerciveness of J w.r.t. the second variable yields the boundedness of u_n. The boundedness of $I(y_n, u_n)$ follows from the same arguments as applied in the proof of Lemma 3.3. Accordingly, Lemma 2.37 implies the boundedness of y_n. From there, we can deduce the boundedness of the strain energy $I_{\text{strain}}(y_n)$.
Now, the reflexivity of $Y \times U$ yields the existence of a weakly converging subsequence, which we also denote by (y_n, u_n). Its weak limit is denoted by $(\overline{y}, \overline{u}) \in Y \times U$. Here, Lemma 2.38 and the weak closedness of $\mathcal{C}$ ensure that $\overline{y} \in \mathcal{A}_c$. Next, we have to verify that $(\overline{y}, \overline{u})$ satisfies again the constraint:

$$\overline{y} \in \operatorname*{argmin}_{v \in \mathcal{A}_c} I(v, \overline{u}).$$

Theorem 2.39 guarantees the existence of a state $\tilde{y} \in \mathcal{A}_c$ satisfying

$$\tilde{y} \in \operatorname*{argmin}_{v \in \mathcal{A}_c} I(v, \overline{u}).$$

Also, Lemma 2.36 yields the weak lower semi-continuity of I w.r.t. the sequence (y_n, u_n) and the weak continuity of the outer energy I_{out}. Then,

$$\limsup_{n\to\infty} I(y_n, u_n) \leq \limsup_{n\to\infty} I(\overline{y}, u_n) = I(\overline{y}, \overline{u}) \leq \liminf_{n\to\infty} I(y_n, u_n),$$

and consequently,

$$\lim_{n\to\infty} I(y_n, u_n) = I(\overline{y}, \overline{u}).$$

From there, we obtain

$$I(\tilde{y}, \overline{u}) \leq I(\overline{y}, \overline{u}) = \lim_{n\to\infty} I(y_n, u_n) \leq \lim_{n\to\infty} I(\tilde{y}, u_n) = I(\tilde{y}, \overline{u}).$$

Thus, $(\overline{y}, \overline{u})$ satisfies the constraints of the optimal control problem (4.2). Finally, the estimate

$$\inf_{(y,u)\in\mathcal{S}} J(y, u) \leq J(\overline{y}, \overline{u}) \leq \liminf_{n\to\infty} J(y_n, u_n) = \inf_{(y,u)\in\mathcal{S}} J(y, u)$$

concludes the proof. □

4.1.2 Regularized optimal control problem

Although the existence of optimal solutions has been established, the numerical computation of such solutions poses significant challenges due to the contact constraints and the resulting non-smoothness. In order to apply the specialized algorithm developed in

[67], we deploy the normal compliance method to relax the constraints. This leads to the following regularized problem:

$$\begin{aligned} &\min_{(y,u)\in Y\times U} J(y,u) \\ &s.t.\ y \in \operatorname*{argmin}_{v\in\mathcal{A}} I_\gamma(v,u), \end{aligned} \tag{4.3}$$

for some fixed parameter $\gamma > 0$. Two popular approaches for regularizing optimal control problems are the Lavrentiev regularization, cf. [48, 49, 70, 71, 86], and the Moreau-Yosida regularization, cf. [45, 46, 47, 52, 72]. Both approaches are designed to relax state constraints. The Moreau-Yosida method deploys a regularization for the characteristic function of the respective state constraints while the Lavrentiev approach aims at regularizing the state constraints directly. A comprehensive overview of regularization methods for optimal control problems can be found in [11].
The main reason why these established methods do not apply in our setting is the bilevel structure of the optimal control problem with no solution operator or first order optimality conditions for the lower level problem. As a result, the usual techniques to analyze regularized optimal control problems cannot be utilized in our case. Still, it is possible to prove that the regularized optimal control problem (4.3) is a suitable approximation of the original one (4.2), at least under sufficient assumptions. First, the existence of optimal solutions is shown.

Theorem 4.3. *Let $\gamma > 0$ be some fixed penalty parameter. Then, the optimal control problem* (4.3) *has at least one optimal solution.*

Proof. The regularization does not alter the two crucial properties of the total energy functional I which are coerciveness and weak lower semi-continuity w.r.t. infimizing sequences. Thus, the existence of optimal solutions can be proven analogously to the proof of Theorem 4.2. □

4.2 Convergence analysis for the regularized optimal control problem

The crucial part of every regularization scheme is to verify that solutions of the regularized problem approach solutions of the original one as the regularization parameter approaches its limit. However, the bilevel structure, the lack of first order optimality conditions, and the non-uniqueness of energy minimizers in hyperelasticity impede a convergence analysis along traditional lines. Thus, in order to obtain a satisfactory convergence result, we will need some additional structure or information on the problem. In this section, we discuss two alternative methods to show the desired convergence result. In the first approach, we utilize a structural assumption, namely that optimal solutions can be approximated by regularized solutions. In the second one, we modify the regularized energy functional by adding a small fraction of the objective functional to better couple optimality with the constraints. This allows us to obtain convergence of solutions without strong assumptions.

4.2.1 Convergence under a reachability assumption

In our setting so far, one critical case cannot be excluded. If no optimal solution pair of the original problem (4.2) can be approximated by a sequence of solutions of the regularized contact problem (3.2), we have no chance of proving any convergence result at all. In the general setting of hyperelastic contact problems, this case cannot be ruled out. Therefore, we have to require additional structure.
In this context, the following property ensures that solutions of the original contact problem (2.23) can be approximated by solutions of the regularized contact problem (3.2).

Definition 4.4 (Reachability)**.** A feasible solution $(y, u) \in \mathcal{S}$ is called reachable if for each sequence $\gamma_n \to \infty$, there exists a subsequence γ_{n_k} and a corresponding sequence $(\tilde{y}_{n_k}, \tilde{u}_{n_k}) \subset \mathcal{A} \times U$ satisfying $\tilde{y}_{n_k} \rightharpoonup y$, $\tilde{u}_{n_k} \to u$, and

$$\tilde{y}_{n_k} \in \operatorname*{argmin}_{v \in \mathcal{A}} I_{\gamma_{n_k}}(v, \tilde{u}_{n_k}).$$

The set of all reachable pairs is denoted by $\mathcal{R} \subset \mathcal{S}$.

Note that this definition is a slightly more general version of Remark 3.6. Although it admits more flexibility since the boundary force u is not fixed, verifying this property seems to be out of reach for general settings. At least in cases where the lower level problem admits a unique solution, Corollary 3.7 implies reachability. Since $\mathcal{R} \subset \mathcal{S}$, we obtain the estimate

$$\min_{(y,u)\in\mathcal{S}} J(y, u) \leq \inf_{(y,u)\in\mathcal{R}} J(y, u).$$

However, it is not clear whether both values coincide. In order to obtain a complete convergence result, the following assumption is necessary.

Assumption 4.5. *Assume that*

$$\min_{(y,u)\in\mathcal{S}} J(y, u) = \inf_{(y,u)\in\mathcal{R}} J(y, u).$$

This assumption is satisfied if at least one optimal solution of Problem (4.2) is also reachable. So far, there does not exist an approach to verify the reachability property for general problems. Such a verification is only possible in settings where the hyperelastic lower level problem admits a unique solution. This leaves us with a theoretical gap, and in the current setting, it remains unclear whether this gap can be closed. Next, we derive our first convergence result.

Theorem 4.6. *Let Assumption 4.5 hold. Further, let $\gamma_n \to \infty$ be a positive and monotonically increasing sequence of penalty parameters. In addition, let $(y_n, u_n) \subset \mathcal{A} \times U$ be a sequence of optimal solutions to the corresponding regularized optimal control problems* (4.3)*. Then,*

$$\lim_{n\to\infty} J(y_n, u_n) = \min_{(y,u)\in\mathcal{S}} J(y, u).$$

Furthermore, there exists a subsequence (y_{n_k}, u_{n_k}) *and a pair* $(\overline{y}, \overline{u}) \in \mathcal{A}_c \times U$ *such that* $y_{n_k} \rightharpoonup \overline{y}$ *in* Y *and* $u_{n_k} \to \overline{u}$ *in* U*. Additionally,* $(\overline{y}, \overline{u})$ *solves the original optimal control problem*

$$\min_{(y,u)\in Y\times U} J(y,u)$$
$$s.t.\ y \in \operatorname*{argmin}_{v\in\mathcal{A}_c} I(v,u).$$

Proof. We start by proving the boundedness of $J(y_n, u_n)$. Recalling the identity mapping id, it follows that $J(\mathrm{id}, u_z) < \infty$. Here, u_z denotes the zero boundary force on Γ_N. Due to Assumption 2.31(6), the pair $(\mathrm{id}, u_z) \in Y \times U$ satisfies the regularized constraint (3.2) for every parameter γ_n. This holds since the identity mapping is a natural state and $\mathrm{id} \in \mathcal{C}$. Therefore, the boundedness of $J(y_n, u_n)$ can be concluded so that $\limsup_{n\to\infty} J(y_n, u_n) < \infty$.
Let (y, u) be any reachable pair. Then, we can choose a subsequence (y_{n_k}, u_{n_k}) such that

$$\limsup_{n\to\infty} J(y_n, u_n) = \lim_{k\to\infty} J(y_{n_k}, u_{n_k}).$$

There also exists a sequence $(\tilde{y}_{n_k}, \tilde{u}_{n_k}) \subset \mathcal{A} \times U$ corresponding to γ_{n_k} with $\tilde{y}_{n_k} \rightharpoonup y$ in Y and $\tilde{u}_{n_k} \to u$ in U satisfying

$$\tilde{y}_{n_k} \in \operatorname*{argmin}_{v\in\mathcal{A}} I_{\gamma_{n_k}}(v, \tilde{u}_{n_k}).$$

The compact embedding $Y \hookrightarrow L^2(\Omega)$, cf. [2], implies $\tilde{y}_{n_k} \to y$ in $L^2(\Omega)$. Consequently, we conclude by optimality of (y_{n_k}, u_{n_k}) and strong continuity of J:

$$\limsup_{n\to\infty} J(y_n, u_n) = \lim_{k\to\infty} J(y_{n_k}, u_{n_k}) \le \lim_{k\to\infty} J(\tilde{y}_{n_k}, \tilde{u}_{n_k}) = J(y,u) \quad \text{for all } (y,u) \in \mathcal{R}.$$

Therefore,

$$\limsup_{n\to\infty} J(y_n, u_n) \le \inf_{(y,u)\in\mathcal{R}} J(y,u).$$

The coerciveness of the objective functional J w.r.t. the second variable yields the boundedness of u_n. Thus, Lemma 3.3 implies the boundedness of y_n. Hence, by reflexivity, there exists a subsequence of (y_{n_k}, u_{n_k}) such that simultaneously

$$\lim_{k\to\infty} J(y_{n_k}, u_{n_k}) = \liminf_{n\to\infty} J(y_n, u_n) \quad \text{and} \quad (y_{n_k}, u_{n_k}) \rightharpoonup (\overline{y}, \overline{u}).$$

Due to Lemma 3.4, we conclude that $(\overline{y}, \overline{u})$ satisfies the original constraint

$$\overline{y} \in \operatorname*{argmin}_{v\in\mathcal{A}_c} I(v, \overline{u}).$$

Further, by weak lower semi-continuity of J, we obtain

$$\begin{aligned}\min_{(y,u)\in\mathcal{S}} J(y,u) \le J(\overline{y},\overline{u}) &\le \lim_{k\to\infty} J(y_{n_k}, u_{n_k}) = \liminf_{n\to\infty} J(y_n, u_n) \\ &\le \limsup_{n\to\infty} J(y_n, u_n) \le \inf_{(y,u)\in\mathcal{R}} J(y,u).\end{aligned}$$

Invoking Assumption 4.5 leads to

$$\min_{(y,u)\in\mathcal{S}} J(y,u) = J(\overline{y},\overline{u}) = \lim_{n\to\infty} J(y_n,u_n).$$

Thus, $(\overline{y},\overline{u})$ is an optimal solution. Finally, we show strong convergence of the sequence u_{n_k}. The Sobolev embedding theorem yields the strong convergence of y_{n_k} in $L^2(\Omega)$. Therefore,

$$\frac{1}{2}\|y_{n_k} - y_\mathrm{d}\|^2_{L^2(\Omega)} \to \frac{1}{2}\|\overline{y} - y_\mathrm{d}\|^2_{L^2(\Omega)}.$$

By incorporating the convergence $J(y_{n_k},u_{n_k}) \to J(\overline{y},\overline{u})$, we can deduce that

$$\frac{\alpha}{2}\|u_{n_k}\|^2_{L^2(\Gamma_N)} \to \frac{\alpha}{2}\|\overline{u}\|^2_{L^2(\Gamma_N)}.$$

Since u_{n_k} is weakly converging in U, this implies the strong convergence in $L^2(\Gamma_N)$. □

In summary, we have been able to show a convergence result for the regularized optimal control problem (4.3). This has been possible only under the assumption of reachability. However, in applications, it is usually not possible to verify whether this assumption holds. The following critical case is conceivable: the original contact problem may have several solutions, some of which are in contact and some of which are not. In this case, our regularization scheme is biased towards those solutions that are in contact because violating the constraints allows reducing the energy. Consequently, we cannot show convergence along the lines of Theorem 4.6. The overarching problem seems to be that the pure normal compliance regularization is unsuited since it does not capture the entire structure of the optimal control problem. To compensate for this, we extend the normal compliance regularization to an approach more compatible with the optimal control problem (4.2).

4.2.2 A modified regularization

In order to improve the coupling of optimization and feasibility, we introduce an alternative regularized total energy function $\mathcal{E}_\gamma$, which contains an additional term from the objective functional. Roughly speaking, this introduces a bias of energy-minimizers towards optimality of the objective functional J. The modified regularized energy functional has the form

$$\mathcal{E}_\gamma(y,u) := I_\gamma(y,u) + \varphi(\gamma)\frac{1}{2}\|y - y_\mathrm{d}\|^2_{L^2(\Omega)}, \tag{4.4}$$

where $\varphi : [0,\infty[\to]0,\infty[$ is monotonically decreasing such that

$$\lim_{\gamma\to\infty} \varphi(\gamma) = 0.$$

As the following analysis will show, the latter property ensures that solutions of this new regularized problem can approach solutions of the original contact problem (2.23). First, we prove that the new regularized problem is well-posed.

Theorem 4.7. *Let $u \in U$ be a fixed boundary force and let $\gamma > 0$ be a fixed penalty parameter. Then, the minimization problem*

$$y \in \operatorname*{argmin}_{v \in \mathcal{A}} \mathcal{E}_\gamma(v, u) \tag{4.5}$$

has at least one solution.

Proof. We note that the functional $\mathcal{E}_\gamma$ is weakly continuous w.r.t. the second variable and weakly lower semi-continuous w.r.t. sequences that leave the strain energy I_{strain} bounded. Thus, the proof is completely analogous to the one of Theorem 2.39. □

Similarly, without giving the details of the proofs, we remark that the results of Lemma 2.37 and Lemma 3.3 also hold for $\mathcal{E}_\gamma$. Next, we establish that limits of the new regularized problem solve the original contact problem.

Lemma 4.8. *Let $\gamma_n \to \infty$ be a monotonically increasing sequence of penalty parameters. Furthermore, consider a weakly convergent sequence $(y_n, u_n) \rightharpoonup (\overline{y}, \overline{u})$ such that*

$$y_n \in \operatorname*{argmin}_{v \in \mathcal{A}} \mathcal{E}_{\gamma_n}(v, u_n).$$

Then, $(\overline{y}, \overline{u}) \in \mathcal{A}_c \times U$ with

$$\overline{y} \in \operatorname*{argmin}_{v \in \mathcal{A}_c} I(v, \overline{u})$$

and

$$\lim_{n \to \infty} \mathcal{E}_{\gamma_n}(y_n, u_n) = I(\overline{y}, \overline{u}).$$

Proof. Theorem 2.39 guarantees the existence of a state $\hat{y} \in \mathcal{A}_c$ with $I_{\text{strain}}(\hat{y}) < \infty$. From there, it follows that the sequence $\mathcal{E}_{\gamma_n}(\hat{y}, u_n)$ is bounded since u_n is bounded and

$$\gamma_n P(\hat{y}) = 0 \quad \text{for all } n \in \mathbb{N}.$$

Therefore, we deduce the boundedness of $\mathcal{E}_{\gamma_n}(y_n, u_n)$ due to

$$\mathcal{E}_{\gamma_n}(y_n, u_n) \leq \mathcal{E}_{\gamma_n}(\hat{y}, u_n).$$

We have to show that the pair $(\overline{y}, \overline{u})$ solves the contact problem (2.23). First, the boundedness of (y_n, u_n) and $\mathcal{E}_{\gamma_n}(y_n, u_n)$ implies the boundedness of $I_{\text{strain}}(y_n)$. Thus, Lemma 2.38 implies $\overline{y} \in \mathcal{A}$. By the same arguments as in the proof of Theorem 3.4, we obtain $\overline{y} \in \mathcal{A}_c$. Again, by using the techniques applied in the proof of Theorem 3.4 and by $\varphi(\gamma_n) \to 0$, we derive the estimates

$$\begin{aligned} \limsup_{n \to \infty} \mathcal{E}_{\gamma_n}(y_n, u_n) &\leq \limsup_{n \to \infty} \mathcal{E}_{\gamma_n}(\overline{y}, u_n) = \limsup_{n \to \infty} \left(I(\overline{y}, u_n) + \varphi(\gamma_n) \frac{1}{2} \|\overline{y} - y_{\text{d}}\|^2_{L^2(\Omega)} \right) \\ &= I(\overline{y}, \overline{u}) \leq \liminf_{n \to \infty} I(y_n, u_n) \leq \liminf_{n \to \infty} \mathcal{E}_{\gamma_n}(y_n, u_n). \end{aligned}$$

Accordingly,

$$\lim_{n \to \infty} \mathcal{E}_{\gamma_n}(y_n, u_n) = I(\overline{y}, \overline{u}).$$

Theorem 2.39 yields the existence of a state $\check{y}$ with

$$\check{y} \in \operatorname*{argmin}_{v \in \mathcal{A}_c} I(v, \overline{u}).$$

Therefore, it follows that

$$I(\check{y}, \overline{u}) \leq I(\overline{y}, \overline{u}) = \lim_{n \to \infty} \mathcal{E}_{\gamma_n}(y_n, u_n) \leq \lim_{n \to \infty} \mathcal{E}_{\gamma_n}(\check{y}, u_n) = I(\check{y}, \overline{u}).$$

□

Similarly to Theorem 3.19, we can derive a convergence rate for the new regularization.

Theorem 4.9. *Let $u \in U$ be a fixed boundary force, and let $\gamma_n \to \infty$ be a monotonically increasing sequence of penalty parameters. In addition, consider the setting of Theorem 3.19 with the derived convergence rate $\rho > 0$ such that*

$$\min_{v \in \mathcal{A}_c} I(v, u) - \min_{v \in \mathcal{A}} I_{\gamma_n}(v, u) \leq c \gamma_n^{-\frac{1}{\rho}}$$

for a suitable constant $c > 0$ and

$$\lim_{\gamma \to \infty} \frac{\gamma^{-\frac{1}{\rho}}}{\varphi(\gamma)} = 0. \tag{4.6}$$

Further, y_n denotes a sequence of corresponding minimizers satisfying

$$y_n \in \operatorname*{argmin}_{v \in \mathcal{A}} \mathcal{E}_{\gamma_n}(v, u).$$

Then, there exists a constant $C > 0$ such that

$$|\min_{v \in \mathcal{A}_c} I(v, u) - \mathcal{E}_{\gamma_n}(y_n, u)| \leq C \varphi(\gamma_n). \tag{4.7}$$

Proof. Let the sequence $\tilde{y}_n$ satisfy

$$\tilde{y}_n \in \operatorname*{argmin}_{v \in \mathcal{A}} I_{\gamma_n}(v, u).$$

Then, we obtain the estimate

$$I_{\gamma_n}(\tilde{y}_n, u) \leq I_{\gamma_n}(y_n, u) \leq I_{\gamma_n}(y_n, u) + \varphi(\gamma_n) \frac{1}{2} \|y_n - y_{\mathrm{d}}\|^2_{L^2(\Omega)} = \mathcal{E}_{\gamma_n}(y_n, u). \tag{4.8}$$

Additionally, Theorem 3.19 yields the following convergence rate

$$0 \leq \min_{v \in \mathcal{A}_c} I(v, u) - I_{\gamma_n}(\tilde{y}_n, u) \leq c \gamma_n^{-\frac{1}{\rho}}.$$

Next, w.l.o.g., it suffices to study the two cases

$$\mathcal{E}_{\gamma_n}(y_n, u) < \min_{v \in \mathcal{A}_c} I(v, u)$$

and

$$\mathcal{E}_{\gamma_n}(y_n, u) \geq \min_{v \in \mathcal{A}_c} I(v, u).$$

In the first case, we obtain

$$0 < \min_{v \in \mathcal{A}_c} I(v, u) - \mathcal{E}_{\gamma_n}(y_n, u) \overset{(4.8)}{\leq} \min_{v \in \mathcal{A}_c} I(v, u) - I_{\gamma_n}(\tilde{y}_n, u) \leq c\gamma_n^{-\frac{1}{\rho}}.$$

By definition, $\varphi(\gamma_n)$ approaches zero at a slower rate than $\gamma_n^{-\frac{1}{\rho}}$. This shows the first case. In second case, the same reasoning yields

$$\min_{v \in \mathcal{A}_c} I(v, u) \leq \mathcal{E}_{\gamma_n}(y_n, u) \leq \mathcal{E}_{\gamma_n}(\tilde{y}_n, u) \leq \min_{v \in \mathcal{A}_c} I(v, u) + \varphi(\gamma_n)\frac{1}{2}\|\tilde{y}_n - y_\mathrm{d}\|_{L^2(\Omega)}.$$

The boundedness of $\|\tilde{y}_n - y_\mathrm{d}\|_{L^2(\Omega)}$ follows from the boundedness of $\|\tilde{y}_n\|_Y$ due to Lemma 3.3, which concludes the proof. □

Summarizing this theorem, we have proven that the optimal regularized energy values of (4.5) approach the optimal energy value of the original contact problem (2.23) at least at the same rate as the regularization function φ approaches zero. The restriction to regularization functions satisfying (4.6) will be essential in the optimal control setting. Applying this new approach to the optimal control problem yields:

$$\begin{aligned} &\min_{(y,u) \in Y \times U} J(y, u) \\ &s.t.\ y \in \operatorname*{argmin}_{v \in \mathcal{A}} \mathcal{E}_\gamma(v, u). \end{aligned} \tag{4.9}$$

Next, we verify the existence of solutions to this problem.

Theorem 4.10. *Let $\gamma > 0$ be some fixed penalty parameter. Then, the optimal control problem* (4.9) *has at least one solution.*

Proof. The proof is analogous to the proof of Theorem 4.2. □

So far, no further restrictions of the regularization function φ have been necessary. However, in order derive a convergence result for the new regularized problem (4.9), additional structure is required. At this, we have to ensure that minimizing a part of the objective functional in the constraint is sufficiently weighted as the penalty parameter approaches infinity. Therefore, we introduce an additional condition for the function φ. Recall that for fixed u, the function $\gamma \to \min_{v \in \mathcal{A}} I_\gamma(v, u)$ is monotonically increasing and bounded. Moreover, by Proposition 3.5, we obtain

$$\lim_{\gamma \to \infty} \min_{v \in \mathcal{A}} I_\gamma(v, u) = \min_{v \in \mathcal{A}_c} I(v, u).$$

In the subsequent analysis, it is necessary that φ approaches zero at a sufficiently slow rate w.r.t. the elastic energy values. This property is specified in the following assumption.

Assumption 4.11. *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 at hand, we can state a convergence result without the assumption of reachability.

Theorem 4.12. *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 (4.2). In addition, let $(y_n, u_n) \subset \mathcal{A} \times U$ be a sequence of optimal solutions to the corresponding regularized problems (4.9), where the regularization function φ satisfies Assumption 4.11 w.r.t. u_*. Then,*

$$\lim_{n\to\infty} J(y_n, u_n) = J(y_*, u_*).$$

Further, there exists a subsequence (y_{n_k}, u_{n_k}) and a pair $(\overline{y}, \overline{u}) \in \mathcal{A}_c \times U$ such that $y_{n_k} \rightharpoonup \overline{y}$ in Y and $u_{n_k} \to \overline{u}$ in $L^2(\Gamma_N)$. Additionally, the pair $(\overline{y}, \overline{u})$ solves the original optimal control problem

$$\min_{(y,u)\in Y\times U} J(y,u)$$
$$s.t.\ y \in \operatorname*{argmin}_{v\in\mathcal{A}_c} I(v,u).$$

Proof. Let us construct a sequence $(\tilde{y}_n, u_*) \subset \mathcal{A} \times U$ that satisfies the regularized constraint (4.5) for each element of γ_n. Further, it should fulfill the condition

$$\limsup_{n\to\infty} J(\tilde{y}_n, u_*) \leq J(y_*, u_*).$$

To this end, let $\tilde{y}_n \subset \mathcal{A}$ be a sequence satisfying

$$\tilde{y}_n \in \operatorname*{argmin}_{v\in\mathcal{A}} \mathcal{E}_{\gamma_n}(v, u_*).$$

We know from Theorem 4.7 that such sequences exist. The minimization property of $\tilde{y}_n$ yields

$$\mathcal{E}_{\gamma_n}(\tilde{y}_n, u_*) - \mathcal{E}_{\gamma_n}(y_*, u_*) \leq 0 \quad \text{for all } n \in \mathbb{N}. \tag{4.10}$$

Then, we can derive the estimate

$$\begin{aligned}
\mathcal{E}_{\gamma_n}(\tilde{y}_n, u_*) - \mathcal{E}_{\gamma_n}(y_*, u_*) &= I_{\gamma_n}(\tilde{y}_n, u_*) - I(y_*, u_*) \\
&\quad + \varphi(\gamma_n)\left(\frac{1}{2}\|\tilde{y}_n - y_\mathrm{d}\|^2_{L^2(\Omega)} - \frac{1}{2}\|y_* - y_\mathrm{d}\|^2_{L^2(\Omega)}\right) \\
&\geq \min_{v\in\mathcal{A}} I_{\gamma_n}(v, u_*) - \min_{v\in\mathcal{A}_c} I(v, u_*) \\
&\quad + \varphi(\gamma_n)(J(\tilde{y}_n, u_*) - J(y_*, u_*)).
\end{aligned}$$

In combination with (4.10), this yields

$$J(\tilde{y}_n, u_*) \leq J(y_*, u_*) + \frac{\min_{v\in\mathcal{A}_c} I(v, u_*) - \min_{v\in\mathcal{A}} I_{\gamma_n}(v, u_*)}{\varphi(\gamma_n)}. \tag{4.11}$$

Since (y_n, u_n) is optimal and φ satisfies Assumption 4.11, we obtain

$$\limsup_{n\to\infty} J(y_n, u_n) \leq \limsup_{n\to\infty} J(\tilde{y}_n, u_*) \leq J(y_*, u_*).$$

By coerciveness of J in the second variable, u_n is bounded. Consequently, y_n is also bounded due to Lemma 3.3. Thus, we can choose a subsequence such that simultaneously

$$\lim_{k\to\infty} J(y_{n_k}, u_{n_k}) = \liminf_{n\to\infty} J(y_n, u_n) \quad \text{and} \quad (y_{n_k}, u_{n_k}) \rightharpoonup (\overline{y}, \overline{u}).$$

By Lemma 4.8, the pair $(\overline{y}, \overline{u})$ satisfies

$$\overline{y} \in \operatorname*{argmin}_{v\in\mathcal{A}_c} I(v, \overline{u}).$$

Due to the weak lower semi-continuity of J, we conclude

$$\begin{aligned} J(y_*, u_*) \leq J(\overline{y}, \overline{u}) &\leq \lim_{k\to\infty} J(y_{n_k}, u_{n_k}) \\ &= \liminf_{n\to\infty} J(y_n, u_n) \leq \limsup_{n\to\infty} J(y_n, u_n) \leq J(y_*, u_*). \end{aligned}$$

This estimate yields

$$\lim_{n\to\infty} J(y_n, u_n) = J(y_*, u_*) = J(\overline{y}, \overline{u}).$$

The strong convergence of u_n follows from the same arguments that have been applied in the proof of Theorem 4.6. □

In conclusion, if $\varphi(\gamma)$ tends to zero *sufficiently slowly*, then we can recover solutions of the original optimal control problem (4.2). To quantify a priori what *sufficiently slow* means, we can profit from the results elaborated in Chapter 3. Depending on the problem characteristics, Theorem 3.19 yields a convergence rate for the energy, and thus, a theoretically backed choice of $\varphi(\gamma)$.
Next, we derive qualitative estimates for the new regularized optimal control problem (4.9). In particular, it is of interest whether proving a convergence rate analogously to Theorem 3.19 is possible. However, this can only be achieved in special cases.

Proposition 4.13. *Let (y_*, u_*) denote an optimal solution to Problem (4.2). Furthermore, let $\gamma_n \to \infty$ be a monotonically increasing sequence of penalty parameters, where $(y_n, u_n) \subset \mathcal{A} \times U$ is a corresponding sequence of optimal solutions to the regularized optimal control problem (4.9). Denote by $\rho > 0$ the convergence rate derived in Theorem 3.19 such that*

$$\min_{v\in\mathcal{A}_c} I(v, u_*) - \min_{v\in\mathcal{A}} I_{\gamma_n}(v, u_*) \leq c\gamma_n^{-\frac{1}{\rho}}$$

for a suitable constant $c > 0$. In addition, the regularization function φ satisfies Assumption 4.11 w.r.t. u_*. If

$$J(y_*, u_*) \leq J(y_n, u_n)$$

holds for all $n \in \mathbb{N}$, then the convergence rate

$$0 \leq J(y_n, u_n) - J(y_*, u_*) \leq C \frac{\gamma_n^{-\frac{1}{\rho}}}{\varphi(\gamma_n)}$$

holds with $C > 0$.

Proof. Consider the sequence $\tilde{y}_n$ with

$$\tilde{y}_n \in \operatorname*{argmin}_{v \in \mathcal{A}} \mathcal{E}_{\gamma_n}(v, u_*).$$

Then, due to optimality of (y_n, u_n) and the estimate in (4.11), we derive

$$J(y_*, u_*) \leq J(y_n, u_n) \leq J(\tilde{y}_n, u_*) \leq J(y_*, u_*) + \frac{\min_{v \in \mathcal{A}_c} I(v, u_*) - \min_{v \in \mathcal{A}} I_{\gamma_n}(v, u_*)}{\varphi(\gamma_n)}.$$

Since

$$\lim_{n \to \infty} \frac{\gamma_n^{-\frac{1}{\rho}}}{\varphi(\gamma_n)} = 0$$

by definition, we obtain the desired convergence rate. □

Interestingly, the convergence rate of the optimal objective functional values directly depends on the convergence rate of the corresponding regularized energy values. In addition, we can prove that the regularized optimal control problem (4.9) leads to smaller values of the objective function.

Proposition 4.14. *Let $\gamma_n \to \infty$ be a monotonically increasing sequence of penalty parameters. Further, we denote by $(y_n, u_n) \subset \mathcal{A} \times U$ and $(\tilde{y}_n, \tilde{u}_n) \subset \mathcal{A} \times U$ sequences of optimal solutions to Problems (4.9) and (4.3), respectively. Then, the estimate*

$$J(y_n, u_n) \leq J(\tilde{y}_n, \tilde{u}_n)$$

holds.

Proof. We prove the statement by contradiction. Assume there exists an $n_0 \in \mathbb{N}$ such that

$$J(\tilde{y}_{n_0}, \tilde{u}_{n_0}) < J(y_{n_0}, u_{n_0}).$$

Let $\hat{y}_{n_0} \in \mathcal{A}$ satisfy

$$\hat{y}_{n_0} \in \operatorname*{argmin}_{v \in \mathcal{A}} \mathcal{E}_{\gamma_{n_0}}(v, \tilde{u}_{n_0}).$$

Then, due to optimality of (y_{n_0}, u_{n_0}), we obtain

$$J(\tilde{y}_{n_0}, \tilde{u}_{n_0}) < J(y_{n_0}, u_{n_0}) \leq J(\hat{y}_{n_0}, \tilde{u}_{n_0}),$$

and consequently,

$$\frac{1}{2}\|\tilde{y}_{n_0} - y_\mathrm{d}\|^2_{L^2(\Omega)} < \frac{1}{2}\|\hat{y}_{n_0} - y_\mathrm{d}\|^2_{L^2(\Omega)}.$$

However, this implies the estimate

$$\begin{aligned}\mathcal{E}_{\gamma_n}(\hat{y}_{n_0}, \tilde{u}_{n_0}) &> I_{\gamma_n}(\hat{y}_{n_0}, \tilde{u}_{n_0}) + \varphi(\gamma_{n_0})\frac{1}{2}\|\tilde{y}_{n_0} - y_\mathrm{d}\|^2_{L^2(\Omega)} \\ &\geq I_{\gamma_n}(\tilde{y}_{n_0}, \tilde{u}_{n_0}) + \varphi(\gamma_{n_0})\frac{1}{2}\|\tilde{y}_{n_0} - y_\mathrm{d}\|^2_{L^2(\Omega)} = \mathcal{E}_{\gamma_n}(\tilde{y}_{n_0}, \tilde{u}_{n_0}),\end{aligned}$$

which contradicts the minimization property of $\hat{y}_{n_0}$. □

So far, more general results seem to be out reach due to the inherent difficulty of the bilevel problem structure.

4.3 Formal KKT conditions for the optimal control problem

The inherent difficulties of nonlinear elasticity motivate the application of sophisticated algorithms as considered in [67]. However, to do so, the lower level energy minimizing problem has to be replaced by its first order optimality condition. Recalling the analysis from Section 3.2, it appears that so far, the only reasonable function space for differentiability of I_strain is $W^{1,\infty}(\Omega)$. To proceed towards KKT conditions in this setting, a local sensitivity of energy minimizers with respect to perturbations in the control would be necessary, e.g., by the application of an implicit function theorem.
A related result was briefly discussed in Theorem 2.34 and to a wider extent in [15, Chapter 6]. There, the analysis was conducted within a $W^{2,p}(\Omega)$-framework for $p > 3$. Unfortunately, this theory also requires very strong regularity assumptions on the problem data. To apply the implicit function theorem, we have to show the $W^{2,p}$-regularity of the solutions of the linearized elastic problems. Those assumptions are unlikely to be satisfied for many problems of interest. Particularly, the crucial case of mixed boundary conditions is generally ruled out.
Therefore, KKT conditions can only be derived in a formal way. Consider the common notation for optimal control problems $x := (y, u)$. In the context of formal first order optimality conditions for I_γ, we define:

$$c_\gamma(y,u)v = \partial_y I_\gamma(y,u)v - \gamma \int_{\Gamma_C} [y]_+^{k-1}\, v_3\, ds, \quad v \in Y.$$

Similarly, for the modified regularization (4.5), we obtain

$$c_\gamma(y,u)v = \partial_y \mathcal{E}_\gamma(y,u)v - \gamma \int_{\Gamma_C} [y]_+^{k-1}\, v_3\, ds, \quad v \in Y.$$

Then, the new optimal control problem reads as follows:

$$\begin{aligned} &\min_{(y,u)\in Y\times U} J(y,u) \\ &s.t.\ c_\gamma(y,u) = 0. \end{aligned} \tag{4.12}$$

Formally, the KKT conditions of (4.12) for given stationary point x_* state the existence of an adjoint state p such that:

$$\begin{aligned} J'(x_*) + c_\gamma'(x_*)^* p &= 0 \\ c_\gamma(x_*) &= 0. \end{aligned} \tag{4.13}$$

These equations serve as the starting point for the algorithmic analysis in Chapter 5.

Remark 4.15. *Replacing Problems* (3.2) *and* (4.5) *with their formal first order optimality conditions changes the fundamental structure of the respective optimal control problems* (4.3) *and* (4.9)*. Replacing the minimization problems allows stationary solutions that are no longer energy minimizers. This change has to be kept in mind when numerical results are analyzed.*

4.4 Summary

In summary, we have established a well-posed optimal control problem which incorporates the nonlinear elastic contact problem (2.23) as its constraint. This yields the bilevel optimization problem

$$\begin{aligned} &\min_{(y,u)\in Y\times U} J(y,u) \\ &s.t.\ y \in \operatorname*{argmin}_{v\in\mathcal{A}_c} I(v,u). \end{aligned}$$

Despite the inherent difficulties due to nonlinear elasticity and contact constraints, the existence result from [64, 66] has been successfully extended to our contact constrained problem. Via the application of the normal compliance method, we have obtained a numerically treatable problem:

$$\begin{aligned} &\min_{(y,u)\in Y\times U} J(y,u) \\ &s.t.\ y \in \operatorname*{argmin}_{v\in\mathcal{A}} I_\gamma(v,u). \end{aligned}$$

This transformation allows the application of the specialized algorithm developed in [64, 67]. However, for the pure normal compliance regularization, a corresponding convergence result has been achieved only under strong structural assumptions, and it remains unclear how to verify these assumptions for general settings. By considering a modified regularized energy functional of the form

$$\mathcal{E}_\gamma(y,u) := I_\gamma(y,u) + \varphi(\gamma)\frac{1}{2}\|y - y_\mathrm{d}\|^2_{L^2(\Omega)},$$

a convergence result was derived without relying on strong assumptions. Finally, we discussed the derivation of corresponding KKT conditions, which remains an open problem.

Chapter 5

Numerical Algorithms

This chapter is dedicated to an algorithmic discussion of problems involving nonlinear elasticity with contact constraints. Here, we only consider contact problems that have been relaxed by some regularization approach such as the normal compliance method described in Chapter 3. This significantly simplifies the numerical treatment since non-smoothness can be ruled out.
This chapter is structured as follows. In the first section, a cubic regularization method in the spirit of [107] is introduced to solve unconstrained minimization problems, such as (3.2). There, the non-convexity of the objective functional implies that the resulting linear systems are not positive definite in general. Two common approaches to overcome this are truncated and regularized CG methods. Truncated CG (TCG) methods usually yield cheaper steps considering the computation time. However, the computed search directions are often very irregular, leading to very small damping parameters and many outer iterations. This results from the termination of the algorithm at a point where things become particularly difficult, often resulting in irregular CG-iterates. For this reason, we opt for a regularized CG (RCG) method, which yields more regular iterates. The derived cubic regularization approach will be applied in Chapter 7 to test the convergence rates elaborated in Chapter 3. Additionally, we present a modified update strategy that incorporates the structural requirements of nonlinear elasticity.
Thereafter, Section 5.2 addresses the affine covariant composite step method introduced in [64, 67]. To apply this method, we have to replace the energy minimization in (4.3) and (4.9) with the respective formal first order conditions, leading to the optimal control problem

$$\begin{aligned} &\min_{(y,u)\in Y\times U} J(y,u) \\ &s.t.\ c_\gamma(y,u) = 0. \end{aligned}$$

The inherent difficulty of this problem, due to nonlinear elasticity, requires the application of robust solvers. In that context, an affine covariant composite method has been successfully applied in [64, 67] to solve optimal control problems of nonlinear elasticity. Finally, we construct a simple path-following approach in Section 5.3. This allows us to consider large normal compliance parameters γ to recover solutions of the original

optimal control problem (4.2).
Parts of this chapter have been published in [94, 95].

5.1 Cubic regularization approach

In [107], the authors addressed non-convex optimization by combining affine covariant techniques with a cubic regularization approach, based on the ideas of [37]. In particular, the elaborated algorithms have been successfully applied to hyperelastic problems. Here, we only introduce a simplified version of the algorithms developed in [107]. For an overview of affine covariant techniques in general, the reader is referred to the monograph [23].

5.1.1 A basic cubic regularization method

Let $(X, \langle\cdot,\cdot\rangle)$ be a Hilbert space and $f : X \to \mathbb{R}$ a sufficiently smooth but not necessarily convex function. We are interested in finding solutions to the optimization problem

$$x_* \in \operatorname*{argmin}_{x\in X} f(x).$$

To compute updates that improve optimality, we define a cubic model of f:

$$m_x(\delta x) := f(x) + f'(x)\delta x + \langle \delta x, f''(x)\delta x\rangle + \frac{\omega^{\mathrm{CR}}}{6}\|\delta x\|_{\mathrm{E}}^3,$$

with the positive regularization parameter ω^{CR} and a suitable norm $\|\cdot\|_{\mathrm{E}}$. Given a direction of descent δx of f, we compute a directional minimizer of m_x along δx, parametrized by the positive damping factor σ. We accept σ if the decrease predicted by the model m_x guarantees sufficient decrease of f. For a user-provided parameter $\eta_1^{\mathrm{CR}} \in]0, 1[$, the decrease is measured via the condition

$$\upsilon := \frac{f(x+\sigma\delta x) - f(x)}{m_x(\sigma\delta x) - m_x(0)} \geq \eta_1^{\mathrm{CR}}. \tag{5.1}$$

In the case that this condition is violated, ω^{CR} is increased by the factor $s_{\mathrm{I}}^{\mathrm{CR}} > 1$, and σ is recomputed. We repeat this process until the step is accepted. To avoid too small step sizes, we also allow a reduction of ω^{CR} if the decrease predicted by the model m_x ensures a certain decrease of f. This is the case if

$$\frac{f(x+\sigma\delta x) - f(x)}{m_x(\sigma\delta x) - m_x(0)} > \eta_2^{\mathrm{CR}},$$

with $\eta_2^{\mathrm{CR}} > \eta_1^{\mathrm{CR}} > 0$. To apply the reduction, ω^{CR} is multiplied with the factor $s_{\mathrm{D}}^{\mathrm{CR}} < 1$. For a suitable tolerance $\Lambda_{\mathrm{CR}} > 0$, the algorithm is considered converged if

$$\|\delta x\|_{\mathrm{E}} \leq \Lambda_{\mathrm{CR}} \tag{5.2}$$

and if f is convex at the current iterate. In this work, $\|\cdot\|_{\mathrm{E}}$ is chosen as the norm induced by linear elasticity. The resulting cubic regularization approach is summarized in Algorithm 1.

Algorithm 1 Cubic Regularization Method.

Solve:

$$x_* \in \operatorname*{argmin}_{x \in X} f(x).$$

Input: initial iterate x_0, parameter ω_0^{CR}, η_1^{CR}, η_2^{CR}, $s_{\mathrm{I}}^{\mathrm{CR}}$, $s_{\mathrm{D}}^{\mathrm{CR}}$, and Λ_{CR}.
Initialize: $k = 0$.
repeat
  $\delta x_k \leftarrow computeDirectionOfDescent()$
  repeat
    $\sigma_k \leftarrow computeDirectionalMinimizer()$
    $\upsilon_k \leftarrow \frac{f(x_k+\sigma_k\delta x_k)-f(x_k)}{m_{x_k}(\sigma_k\delta x_k)-m_{x_k}(0)}$
    if $\upsilon_k < \eta_1^{\mathrm{CR}}$ **then**
      $\omega_k^{\mathrm{CR}} \leftarrow \omega_k^{\mathrm{CR}} s_{\mathrm{I}}^{\mathrm{CR}}$
    else if $\upsilon_k > \eta_2^{\mathrm{CR}}$ **then**
      $\omega_k^{\mathrm{CR}} \leftarrow \omega_k^{\mathrm{CR}} s_{\mathrm{D}}^{\mathrm{CR}}$
    end if
  until step accepted (If Condition (5.1) is fulfilled.)
  $x_{k+1} \leftarrow x_k + \delta x_k \sigma_k$
  $\omega_{k+1}^{\mathrm{CR}} \leftarrow \omega_k^{\mathrm{CR}}$
  $k \leftarrow k + 1$
until convergent (If Condition (5.2) is fulfilled and if $f''(x_k)$ is positive definite.)

5.1.2 Computing a direction of descent

After applying a Galerkin-type discretization, the Newton equation

$$f''(x)\delta x = -f'(x)$$

represents a finite-dimensional linear system with the matrix $f''(x)$. As long as $f''(x)$ is positive definite, we can use a preconditioned conjugate gradient (PCG) method as illustrated in Algorithm 2 to compute a direction of descent.
Here, we choose a BPX-type preconditioner, cf. [13], to solve nonlinear elastic problems. However, if f is the total energy functional (2.21), positive definiteness is no longer guaranteed. This is due to the lack of convexity of nonlinear elastic problems. To overcome this, we add a positive definite matrix R_{E} to $f''(x)$:

$$f''(x) + \lambda R_{\mathrm{E}},$$

where $\lambda > 0$ is gradually increased until positive definiteness is reached. For the purpose of reducing the number of regularization steps, we choose a multiplicative update formula as described in Algorithm 3. One possible drawback of this approach is that the regularization parameter is scaled up too much. Consequently, the updates steps are unnecessarily damped, which can slow down the algorithm significantly.

Algorithm 2 Preconditioned Conjugate Gradient Method.

Solve: $Hx = b$.
Input: initial iterate x_0 and preconditioner Q.
Initialize: $r_0 = Hx_0 - b$, $d_0 = g_0 = Q^{-1}(-r_0)$, and $k = 0$.
repeat
$\alpha_k \leftarrow -\frac{r_k^T g_k}{d_k^T H d_k}$
$x_{k+1} \leftarrow x_k + \alpha_k d_k$
$r_{k+1} \leftarrow r_k + \alpha_k H d_k$
$g_{k+1} \leftarrow Q^{-1}(-r_{k+1})$
$\beta_{k+1} \leftarrow \frac{r_{k+1}^T g_{k+1}}{r_k^T g_k}$
$d_{k+1} \leftarrow g_{k+1} + \beta_{k+1} d_k$
$k \leftarrow k + 1$
until convergent

Algorithm 3 Update Formula for the Regularization Parameter.

Input: current regularization parameter λ, starting regularization parameter λ_{AInit}, and update parameter s_{A}.
if $\lambda = 0$ **then**
$\lambda \leftarrow \lambda_{\text{AInit}}$
else
$\lambda \leftarrow \lambda s_{\text{A}}$
end if

Remark 5.1. *A more sophisticated approach was considered in [64, Section 4.3]. There, an adaptive update formula for the regularization parameter was derived. Although adaptive methods are generally more technical to implement, the increased efficiency makes their use advisable.*

Combining the previous results yields a regularized preconditioned conjugate gradient (RPCG) method summarized in Algorithm 4. By recovering positive definiteness, Algorithm 4 can be applied in order to compute directions of descent of f.

For large values of λ, the Newton update δx approaches a gradient step w.r.t. the scalar product induced by the regularized matrix

$$f''(x) + \lambda R_{\text{E}}.$$

In the case of nonlinear elasticity, the natural choice for the regularization R_{E} is the Hessian matrix of linear elasticity. Besides CG methods, one can use the Chebyshev semi-iteration to solve linear systems.

Algorithm 4 Regularized Preconditioned Conjugate Gradient Method.

Solve: $Hx = b$.
Input: initial iterate x_0, preconditioner Q, regularization update parameter s_{A}, and starting regularization parameter λ_{AInit}.
Initialize: $r_0 = Hx_0 - b$, $d_0 = g_0 = Q^{-1}(-r_0)$, $\lambda = 0$, and $k = 0$.
repeat
  if $d_k^T H d_k < 0$ **then**
    $\lambda \leftarrow updateLambda(\lambda, \lambda_{\mathrm{AInit}}, s_{\mathrm{A}})$ (Algorithm 3)
    $H \leftarrow H + \lambda R_{\mathrm{A}}$
    restart (Restart the entire CG algorithm.)
  end if
  $\alpha_k \leftarrow -\frac{r_k^T g_k}{d_k^T A d_k}$
  $x_{k+1} \leftarrow x_k + \alpha_k d_k$
  $r_{k+1} \leftarrow r_k + \alpha_k H d_k$
  $g_{k+1} \leftarrow Q^{-1}(-r_{k+1})$
  $\beta_{k+1} \leftarrow \frac{r_{k+1}^T g_{k+1}}{r_k^T g_k}$
  $d_{k+1} \leftarrow g_{k+1} + \beta_{k+1} d_k$
  $k \leftarrow k + 1$
until convergent

Chebyshev semi-iteration

Here, only a short introduction of the Chebyshev semi-iteration is given. For a more detailed analysis, see [42]. Consider the linear system

$$Hx = b \tag{5.3}$$

and a suitable preconditioner Q. Moreover, let $\varsigma_{\min}$ and $\varsigma_{\max}$ denote the smallest and the largest eigenvalue of $Q^{-1}H$. Further, let Λ_{Cheb} denote the desired accuracy. Then, we can apply the Chebyshev semi-iteration illustrated in Algorithm 5 to solve (5.3). In contrast to CG methods, the Chebyshev semi-iteration is linear if the number of iterations is fixed. For the application of this method in our setting, it remains to derive estimates for $\varsigma_{\min}$ and $\varsigma_{\max}$. To do so, assume Algorithm 2 has been applied to solve a linear system with preconditioner Q and system matrix H. If the PCG method converges in m steps, it yields the parameters α_i, β_j with $i = 0,...,m-1$ and $j = 0,...,m-2$. Then, we can construct the tridiagonal matrix

$$T_m := \begin{pmatrix} \frac{1}{\alpha_0} & \frac{\sqrt{\beta_0}}{\alpha_0} & & & \\ \frac{\sqrt{\beta_0}}{\alpha_0} & \frac{1}{\alpha_1} + \frac{\beta_0}{\alpha_0} & \frac{\sqrt{\beta_1}}{\alpha_1} & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \frac{\sqrt{\beta_{m-2}}}{\alpha_{m-2}} \\ & & & \frac{\sqrt{\beta_{m-2}}}{\alpha_{m-2}} & \frac{1}{\alpha_{m-1}} + \frac{\beta_{m-2}}{\alpha_{m-2}} \end{pmatrix}.$$

Algorithm 5 Chebyshev Semi-Iteration.

Solve: $Hx = b$.
Input: preconditioner Q, smallest eigenvalue $\varsigma_{\min}$ and largest eigenvalue $\varsigma_{\max}$ of $Q^{-1}H$, and accuracy Λ_{Cheb}.
Initialize: $\kappa = \frac{\varsigma_{\max}}{\varsigma_{\min}}$, $\vartheta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$, $N = \lfloor \frac{\ln(\Lambda_{\text{Cheb}})}{\ln(\vartheta)} \rfloor + 1$, $c = \frac{\varsigma_{\max}-\varsigma_{\min}}{2}$, $\alpha = \frac{\varsigma_{\max}+\varsigma_{\min}}{2}$, $\mu_0 = -\alpha$, $\delta x_0 = Q^{-1}b$, $x_0 = x_{-1} = 0$, $\theta_0 = 0$.
for $k = 0,...,N$ **do**
  if $k = 1$ **then**
    $\theta_k \leftarrow \frac{-c^2}{(2\alpha)}$
  else if $k \geq 2$ **then**
    $\theta_k \leftarrow \frac{c^2}{(4\mu_k)}$
  end if
  if $k \geq 1$ **then**
    $\mu_k \leftarrow -(\alpha + \theta_k)$
  end if
  $x_{k+1} \leftarrow -\frac{\delta x_k + \alpha x_k + \theta_k x_{k-1}}{\mu_k}$
  $\delta x_{k+1} \leftarrow Q^{-1}(b - Hx_{k+1})$
end for

The eigenvalues of T_m yield adequate estimates for the eigenvalues of $Q^{-1}H$, cf. [88, Chapter 6]. Of course, such an approach is only reasonable if multiple systems involving the matrix H have to be solved, which is the case for preconditioners applied in the composite step method, see Section 6.3.

5.1.3 Nonlinear updates

At each iterate x, Algorithm 1 computes an update δx and adds it, with possible scaling σ, to the current iterate:

$$x + \sigma\delta x.$$

Instead of directly adding δx in this way, we want to take into account the required structural properties of the considered problem. For nonlinear elastic problems, x and $x + \delta x$ correspond to deformations of the domain $\overline{\Omega}$. Therefore, we are interested in maintaining the orientation-preserving condition:

$$\det(x + \delta x) > 0$$

at each point in the domain Ω. To achieve this, the update δx is transformed to incorporate this condition. A generalized concept of this approach is optimization on manifolds. Since a detailed analysis of this topic is beyond the scope of this work, the reader is instead referred to [1]. The modified update presented in this work was developed by Julián Ortiz in his PhD thesis which is in preparation.
Assume that the discretized domain $\overline{\Omega}_{\text{D}}$ consists of a set of tetrahedrons T_l such that $\overline{\Omega}_{\text{D}} = \bigcup_l T_l$, and intersections of the tetrahedrons are only allowed on their boundaries.

At a given deformation x, we obtain the corresponding set of deformed tetrahedrons denoted by T_{x_l}. Moreover, consider an arbitrary but fixed tetrahedron T_{x_k}, which can be represented by the points p_i with $i = 1,...,4$ and

$$p_i := \begin{pmatrix} p_{ix} \\ p_{iy} \\ p_{iz} \end{pmatrix}.$$

Note that each point p_i corresponds to a displaced vertex of the undeformed tetrahedron T_k. Further, i is the local index to identify the point p_i within the tetrahedron. Accordingly, there exists a global index, denoted by i_g, to identify each vertex in $\overline{\Omega}_{\mathrm{D}}$. For a given direction δx, we obtain the corresponding update:

$$\delta p_i := \begin{pmatrix} \delta p_{ix} \\ \delta p_{iy} \\ \delta p_{iz} \end{pmatrix}.$$

Adding this update to T_{x_k} leads to the deformed tetrahedron T'_{x_k} with the new coordinates:

$$p'_i := \begin{pmatrix} p'_{ix} \\ p'_{iy} \\ p'_{iz} \end{pmatrix} = \begin{pmatrix} p_{ix} \\ p_{iy} \\ p_{iz} \end{pmatrix} + \begin{pmatrix} \delta p_{ix} \\ \delta p_{iy} \\ \delta p_{iz} \end{pmatrix}.$$

This setting is illustrated in Figure 5.1.

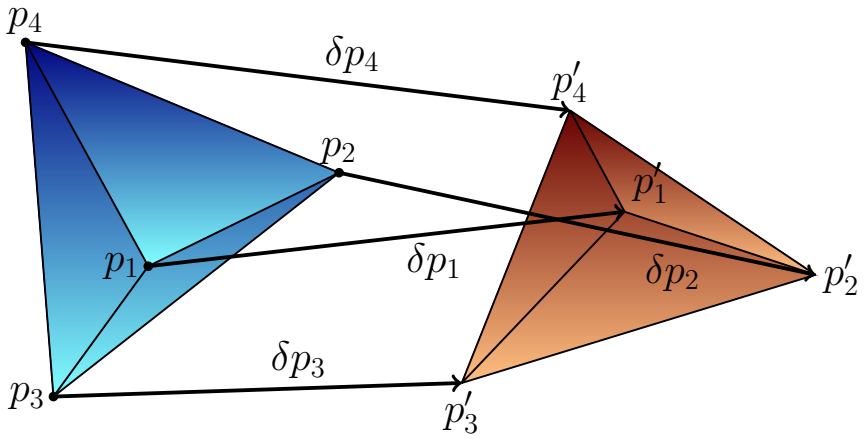

Figure 5.1: Deformation of a tetrahedron.

A heuristic way to ensure $\det(x+\delta x) > 0$ in a discrete setting is preserving the orientation of all tetrahedrons during the deformation. To achieve this, we define the matrices

$$M_x := \begin{pmatrix} p_{1x} & p_{2x} & p_{3x} & p_{4x} \\ p_{1y} & p_{2y} & p_{3y} & p_{4y} \\ p_{1z} & p_{2z} & p_{3z} & p_{4z} \\ 1 & 1 & 1 & 1 \end{pmatrix} \quad \text{and} \quad M_{dx} := \begin{pmatrix} \delta p_{1x} & \delta p_{2x} & \delta p_{3x} & \delta p_{4x} \\ \delta p_{1y} & \delta p_{2y} & \delta p_{3y} & \delta p_{4y} \\ \delta p_{1z} & \delta p_{2z} & \delta p_{3z} & \delta p_{4z} \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Writing $\mathrm{Exp}(\,\cdot\,)$ for the matrix exponential function, we compute

$$\tilde{M}_x := \mathrm{Exp}(M_{dx} M_x^{-1}) M_x \tag{5.4}$$

with the notation

$$\tilde{M}_x := \begin{pmatrix} \tilde{p}_{1x} & \tilde{p}_{2x} & \tilde{p}_{3x} & \tilde{p}_{4x} \\ \tilde{p}_{1y} & \tilde{p}_{2y} & \tilde{p}_{3y} & \tilde{p}_{4y} \\ \tilde{p}_{1z} & \tilde{p}_{2z} & \tilde{p}_{3z} & \tilde{p}_{4z} \\ \tilde{p}_{1w} & \tilde{p}_{2w} & \tilde{p}_{3w} & \tilde{p}_{4w} \end{pmatrix}.$$

Then,

$$\tilde{p}_i := \begin{pmatrix} \tilde{p}_{ix} \\ \tilde{p}_{iy} \\ \tilde{p}_{iz} \end{pmatrix}$$

corresponds to the vertex coordinates of the new deformed tetrahedron $\tilde{T}_{x_k}$ according to the nonlinear update formula described in (5.4). The new update is then defined by

$$\delta\tilde{p}_i := \tilde{p}_i - p_i.$$

Repeating this process for all tetrahedrons yields a new deformation update $\delta\tilde{x}$ for the entire domain. One drawback of this approach is that for each vertex p_j in $\overline{\Omega}_{\mathrm{D}}$, we compute multiple updates, one for each tetrahedron that contains the point p_j. To compensate for that, the respective displacement is averaged over the number of tetrahedrons containing p_j. The resulting update formula is summarized in Algorithm 6. u Since the updates for each tetrahedron are independent, the computation can be easily parallelized. We incorporate this new approach into our cubic regularization method and obtain Algorithm 7. There, this new nonlinear update is only applied if it yields a larger decrease of the objective function value than the linear one. By taking into account the underlying structure of nonlinear elastic problems, we expect to improve the performance of our cubic regularization approach. Corresponding numerical tests are conducted in Chapter 7.

Algorithm 6 Nonlinear Update.

Input: current iterate x and update δx.
for all vertices p_j **do**
  $\delta\overline{p}_j \leftarrow \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}^T$
end for
for all tetrahedrons T_k **do**
  $(p_1, p_2, p_3, p_4) \leftarrow computeCoordinates(x, T_k)$
  $(\delta p_1, \delta p_2, \delta p_3, \delta p_4) \leftarrow getLocalUpdate(dx, T_k)$
  $(\delta\tilde{p}_1, \delta\tilde{p}_2, \delta\tilde{p}_3, \delta\tilde{p}_4) \leftarrow computeNonlinearUpdate(p_1, p_2, p_3, p_4, \delta p_1, \delta p_2, \delta p_3, \delta p_4)$ (via (5.4))
  $\delta\overline{p}_{1_g} \leftarrow \delta\overline{p}_{1_g} + \delta\tilde{p}_1$
  $\delta\overline{p}_{2_g} \leftarrow \delta\overline{p}_{2_g} + \delta\tilde{p}_2$
  $\delta\overline{p}_{3_g} \leftarrow \delta\overline{p}_{3_g} + \delta\tilde{p}_3$
  $\delta\overline{p}_{4_g} \leftarrow \delta\overline{p}_{4_g} + \delta\tilde{p}_4$
end for
for all updates $\delta\overline{p}_j$ **do**
  $\delta\overline{p}_j \leftarrow (\delta\overline{p}_j)/numberOfAdjacentTetrahedrons$
  $\delta\tilde{x} \leftarrow transferUpdate(\delta\overline{p}_j)$
end for
return $\delta\tilde{x}$;

Algorithm 7 Cubic Regularization Method with Nonlinear Update.

Solve:

$$x_* \in \operatorname*{argmin}_{x \in X} f(x).$$

Input: initial iterate x_0, parameters ω_0^{CR}, η_1^{CR}, η_2^{CR}, $s_{\mathrm{I}}^{\mathrm{CR}}$, $s_{\mathrm{D}}^{\mathrm{CR}}$, and Λ_{CR}.
Initialize: $k = 0$.
repeat
  $\delta x_k \leftarrow computeDirectionOfDescent()$
  repeat
    $\sigma_k \leftarrow computeDirectionalMinimizer()$
    $\delta\tilde{x}_k \leftarrow computeNonlinearUpdate(x_k, \sigma_k \delta x_k)$ (Algorithm 6)
    if $f(x_k + \delta\tilde{x}_k) < f(x_k + \sigma_k \delta x_k)$ **then**
      $\delta\overline{x}_k \leftarrow \delta\tilde{x}_k$
    else
      $\delta\overline{x}_k \leftarrow \sigma_k \delta x_k$
    end if
    $\upsilon_k \leftarrow \frac{f(x_k + \delta\overline{x}_k) - f(x_k)}{m_{x_k}(\delta\overline{x}_k) - m_{x_k}(0)}$
    if $\upsilon_k < \eta_1^{\mathrm{CR}}$ **then**
      $\omega_k^{\mathrm{CR}} \leftarrow \omega_k^{\mathrm{CR}} s_{\mathrm{I}}^{\mathrm{CR}}$
    else if $\upsilon_k > \eta_2^{\mathrm{CR}}$ **then**
      $\omega_k^{\mathrm{CR}} \leftarrow \omega_k^{\mathrm{CR}} s_{\mathrm{D}}^{\mathrm{CR}}$
    end if
  until step accepted (If Condition (5.1) is fulfilled.)
  $x_{k+1} \leftarrow x_k + \delta x_k \sigma_k$
  $\omega_{k+1}^{\mathrm{CR}} \leftarrow \omega_k^{\mathrm{CR}}$
  $k \leftarrow k + 1$
until convergent (If Condition (5.2) is fulfilled and if $f''(x_k)$ is positive definite.)

5.2 Affine Covariant Composite Step Method

For a fixed penalty parameter γ, the optimal control problem (4.12) describes an equality constrained optimization problem of the form

$$\begin{aligned} &\min_{x \in X} f(x) \\ &s.t.\ c(x) = 0 \end{aligned}$$

that requires robust and efficient solution algorithms. Here, we choose the affine covariant composite step method considered in [64, 67]. The central idea of composite step methods is to split the Lagrange Newton step δx into a normal step $\delta n \in (\ker c'(x))^{\perp}$ and into a tangential step $\delta t \in \ker c'(x)$. The normal step approaches feasibility while the tangential step approaches optimality. This class of methods is widely applied in equality constrained optimization and optimal control, cf. [44, 78, 84, 105, 112].

Additionally, the algorithm proposed in [64, 67] applies a simplified normal step, denoted by δs, at the end of each iteration. This yields two major advantages. First, an affine covariant globalization scheme can be deployed. Second, the step δs also acts as a second order correction to avoid the well-known Maratos effect. The described splitting of the step δx is illustrated in Figure 5.2.

Finally, affine covariance ensures that norms are only evaluated in the domain space X but not in the image space of c. In the case of PDE-constraints, the image space is usually some dual space. Therefore, meaningful norms are hard to evaluate. In contrast, norms suited to the problem in the domain space can be evaluated much more easily.

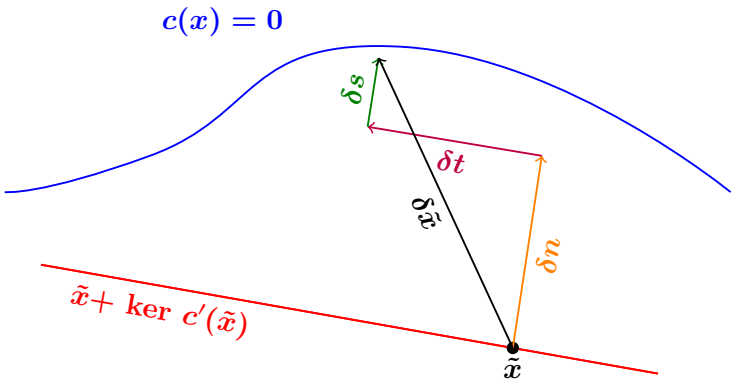

Figure 5.2: Splitting of the composite step.

This section gives a brief overview of the algorithm proposed in [64, 67].

5.2.1 Setting

Let $(X, \langle\cdot,\cdot\rangle)$ denote a Hilbert space and P a reflexive Banach space. Then, we can define the optimization problem

$$\begin{aligned} &\min_{x\in X} f(x) \\ &s.t.\ c(x) = 0, \end{aligned} \tag{5.5}$$

where $f : X \to \mathbb{R}$ and $c : X \to P^*$ are both twice continuously differentiable and $c'(x)$ is surjective. In optimal control settings, it is common to introduce the splitting $X = Y \times U$, $x = (y, u)$, and

$$c(x) = \hat{A}(y) - Bu, \tag{5.6}$$

where $\hat{A} : Y \to P^*$ is a nonlinear operator with continuous inverse and $B : U \to P^*$ a linear and compact operator. The above regularity assumptions imply that each stationary point x_* satisfies the KKT conditions

$$\begin{aligned} f'(x_*)v + pc'(x_*)v = 0 \quad &\text{for all } v \in X, \\ c(x_*) = 0 \quad & \end{aligned}$$

with the Lagrange multiplier $p \in P^{**}$. The Hilbert space structure allows the splitting

$$X = \ker c'(x_*) \oplus (\ker c'(x_*))^{\perp}. \tag{5.7}$$

Moreover, the equation

$$f'(x_*)v + pc'(x_*)v = 0 \quad \text{for all } v \in X$$

is equivalent to the system

$$\begin{aligned} f'(x_*)v = 0 \quad &\text{for all } v \in \ker c'(x_*), \\ (f'(x_*) + pc'(x_*))w = 0 \quad &\text{for all } w \in (\ker c'(x_*))^{\perp}. \end{aligned}$$

From the second equation, we obtain a formula to compute the Lagrange multiplier p:

$$\begin{pmatrix} M & c'(x)^* \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} v \\ p \end{pmatrix} = \begin{pmatrix} -f'(x) \\ 0 \end{pmatrix}, \tag{5.8}$$

cf. [64, Theorem 3.1]. Here, M denotes the Riesz isomorphism satisfying

$$(Mv)(w) = \langle v, w\rangle.$$

At a point $x \in X$, we denote the Lagrange multiplier computed via (5.8) by p_x.

5.2.2 Computation of the update steps

As stated in the introduction, the basic idea of the composite step method is to split the Lagrange Newton update step into a normal step $\delta n \in (\ker c'(x))^{\perp}$ and a tangential step $\delta t \in \ker c'(x)$. Additionally, we consider the positive damping factors ν and τ with the corresponding undamped steps Δn and Δt such that

$$\delta n = \nu \Delta n \quad \text{and} \quad \delta t = \tau \Delta t.$$

The two steps are added up to

$$\delta x = \delta n + \delta t.$$

Further, a second order correction $\delta s \in (\ker c'(x))^{\perp}$ is applied, yielding the final update

$$\delta \tilde{x} = \delta x + \delta s.$$

In the following, we discuss how to compute these updates and the corresponding Lagrange multiplier.

5.2.3 Computation of the Lagrange multiplier

Using the Lagrange formulation, we define

$$\mathcal{L}(x, p_x) := f(x) + p_x c(x).$$

Instead of computing p_x with (5.8) directly, a correction δp_x is computed via the system

$$\begin{pmatrix} M & c'(x)^* \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} v \\ \delta p_x \end{pmatrix} = \begin{pmatrix} -\mathcal{L}_x(x, p_{x_-}) \\ 0 \end{pmatrix}, \tag{5.9}$$

where p_{x_-} denotes the previous multiplier. Thereafter, the Lagrange multiplier p_x is updated:

$$p_x \leftarrow p_{x_-} + \delta p_x. \tag{5.10}$$

This update scheme is numerically more stable than the direct approach (5.8).

5.2.4 Computation of the normal step

To improve feasibility, the undamped normal step Δn has to satisfy

$$c(x) + c'(x)\Delta n = 0. \tag{5.11}$$

In general, Δn is not fully determined by this system. Therefore, Equation (5.11) is extended to a minimum norm correction problem:

$$\begin{aligned} &\min_{v \in X} \frac{1}{2}\langle v, v \rangle \\ &s.t.\ c'(x)v = -c(x). \end{aligned}$$

This leads to the system

$$\begin{pmatrix} M & c'(x)^* \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} \Delta n \\ q \end{pmatrix} = \begin{pmatrix} 0 \\ -c(x) \end{pmatrix}. \tag{5.12}$$

For general right hand sides

$$r = \begin{pmatrix} 0 \\ -g \end{pmatrix},$$

we use the short notation $c'(x)^-(g)$ to denote the primal component of solutions to (5.12).

5.2.5 Computation of the simplified normal step

The simplified normal step δs has to fulfill

$$c(x + \delta x) - c(x) - c'(x)\delta x + c'(x)\delta s = 0.$$

If $\nu = 1$, δs corresponds to the second step of a simplified Newton method for the equation $c(x) = 0$, where x serves as the starting point. Again, computing a minimum norm correction yields the system

$$\begin{pmatrix} M & c'(x)^* \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} \delta s \\ q \end{pmatrix} = \begin{pmatrix} 0 \\ -c(x + \delta x) + c(x) + c'(x)\delta x \end{pmatrix} \tag{5.13}$$

from which the step δs can be determined.

5.2.6 Computation of the tangential step

Consider a given normal step δn and Lagrange multiplier p_x. Then, we compute the tangential step δt such that $\delta x := \delta n + \delta t$ approximates the minimizer of a quadratic model q of the Lagrange function $\mathcal{L}$ on $\ker c'(x)$. Here, q is defined by

$$q(\delta x) := f(x) + f'(x)\delta x + \frac{1}{2}\mathcal{L}_{xx}(x, p_x)(\delta x)^2, \tag{5.14}$$

yielding the optimization problem

$$\begin{aligned} &\min_{\delta t \in X} q(\delta n + \delta t) \\ &s.t.\ c'(x)\delta t = 0. \end{aligned}$$

The condition

$$p_x c'(x)\delta t = 0$$

allows the equivalent reformulation to

$$\begin{aligned} &\min_{\delta t \in X} \left(\mathcal{L}_x(x, p_x) + \mathcal{L}_{xx}(x, p_x)\delta n\right)\delta t + \frac{1}{2}\mathcal{L}_{xx}(x, p_x)(\delta t)^2 \\ &s.t.\ c'(x)\delta t = 0. \end{aligned} \tag{5.15}$$

This reformulation increases the numerical stability of our approach since $\mathcal{L}_x(x, p_x) \to 0$ as $(x, p_x) \to (x_*, p_{x_*})$. If we neglect the damping parameter τ for now, Δt can be computed by solving the system

$$\begin{pmatrix} \mathcal{L}_{xx}(x, p_x) & c'(x)^* \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} \Delta t \\ \Delta p \end{pmatrix} = \begin{pmatrix} -\mathcal{L}_x(x, p_x) - \mathcal{L}_{xx}(x, p_x)\delta n \\ 0 \end{pmatrix}$$

if $\mathcal{L}_{xx}(x, p_x)$ is positive definite. With all the update formulas at hand, it remains to derive a suitable acceptance criterion for δx.

5.2.7 Acceptance criterion

When considering feasibility, it is reasonable to require that $\|\delta s\| \ll \|\delta x\|$. Thus, we choose the contraction condition

$$\Theta(\delta x) := \frac{\|\delta s\|}{\|\delta x\|} \leq \Theta_{\text{acc}} < 1 \tag{5.16}$$

for a user-provided parameter Θ_{acc}. To ensure that this conditioned is fulfilled, we deploy a simple predictor-corrector scheme. The basic idea is to construct a model $\tilde{\Theta}(\delta x)$ of $\Theta(\delta x)$ and compute the correction δx such that

$$\tilde{\Theta}(\delta x) \leq \Theta_{\text{d}}, \tag{5.17}$$

for a desired contraction factor $\Theta_{\text{d}} < \Theta_{\text{acc}}$. To this end, assume that there exist positive constants ω_{C}, ω_{L}, and $\omega_{\text{f}'}$ such that for all $v, w \in X$, the following inequalities hold:

$$\|c'(x)^-(c'(x+v) - c'(x))v\| \leq \omega_{\text{C}}\|v\|^2, \tag{5.18}$$

$$|(\mathcal{L}_{xx}(x+v, p) - \mathcal{L}_{xx}(x, p))(v, v)| \leq \omega_{\text{L}}\|v\|^3, \tag{5.19}$$

$$|(f'(x+v) - f'(x))w| \leq \omega_{\text{f}'}\|v\|\|w\|. \tag{5.20}$$

Moreover, we define the parametrized model for the contraction rate by

$$\tilde{\Theta}(\xi) := \frac{\tilde{\omega}_{\text{C}}}{2}\|\xi\|,$$

where $\tilde{\omega}_{\text{C}}$ is an estimate from below for ω_{C}. Then, the interpolation condition

$$\tilde{\Theta}(\delta x) = \Theta(\delta x)$$

yields the formula:

$$\tilde{\omega}_{\text{C}} = \frac{2\Theta(\delta x)}{\|\delta x\|} = 2\frac{\|\delta s\|}{\|\delta x\|^2}.$$

Inserting this model into (5.17) leads to the trust region constraint

$$\|\delta x\| \leq \frac{2\Theta_{\text{d}}}{\tilde{\omega}_{\text{C}}}. \tag{5.21}$$

From there, the corrections δx and δs are computed. If (5.16) is not satisfied, $\tilde{\omega}_{\text{C}}$ is updated again with the newly computed corrections δx and δs. The updated estimate $\tilde{\omega}_{\text{C}}$ is then used to compute new corrections δx and δs. This process is repeated until (5.16) is satisfied. In the following, we study how the trust region constraint (5.21) affects the computation of the damping parameters ν and τ.

Damping of the normal steps

Consider the direction Δn determined by (5.12). Then, a suitable damping parameter can be computed by

$$\nu = \min\Big\{1, \frac{2\Theta_{\mathrm{n}}}{\|\Delta n\|\tilde{\omega}_{\mathrm{C}}}\Big\},$$

guaranteeing that δn does not violate (5.21). The contraction factor $\Theta_{\mathrm{n}} \leq \Theta_{\mathrm{d}}$ permits some "elbow space" for the tangential step δt. Here, $\Theta_{\mathrm{n}} = \Theta_{\mathrm{d}}$ would imply $\delta t = 0$.

Damping of the tangential step

By definition, the steps δn and δn are orthogonal. Thus, we obtain the splitting

$$\|\delta n + \delta t\|^2 = \|\delta n\|^2 + \|\delta t\|^2.$$

Inserting this property into (5.21) yields

$$\|\delta t\| \leq \sqrt{\left(\frac{2\Theta_{\mathrm{d}}}{\tilde{\omega}_{\mathrm{C}}}\right)^2 - \|\delta n\|^2}\,,$$

and accordingly, the upper bound $\tau_{\max}$ for the damping parameter τ:

$$\tau \leq \tau_{\max} := \frac{\sqrt{\left(\frac{2\Theta_{\mathrm{d}}}{\tilde{\omega}_{\mathrm{C}}}\right)^2 - \nu^2\|\Delta n\|^2}}{\|\Delta t\|}.$$

An additional damping of Δt is required to approach optimality w.r.t. the objective functional f. Therefore, τ is chosen such that δt is a directional minimizer of the cubic model

$$m_{\tilde{\omega}_{\mathrm{f}}}(\delta x) := q(\delta x) + \frac{\tilde{\omega}_{\mathrm{f}}}{6}\|\delta x\|^3.$$

The parameter $\tilde{\omega}_{\mathrm{f}}$ is updated in each step as follows:

$$\tilde{\omega}_{\mathrm{f}} = \frac{6}{\|\delta x\|^3}(f(x + \delta x + \delta s) - q(\delta x)). \tag{5.22}$$

In summary, δt can be described as the solution to the following optimization problem:

$$\begin{gathered}\min_{\delta t \in X} m_{\tilde{\omega}_{\mathrm{f}}}(\delta n + \delta t)\\ s.t.\ \frac{\tilde{\omega}_{\mathrm{C}}}{2}\|\delta x\| \leq \Theta_{\mathrm{d}},\\ c'(x)\delta t = 0.\end{gathered}$$

To ensure that our algorithm also approaches optimality, we choose the natural acceptance criterion for the tangential step given by

$$\overline{\eta} \leq \eta := \frac{f(x + \delta x + \delta s) - m_{\tilde{\omega}_{\mathrm{f}}}(\delta n)}{m_{\tilde{\omega}_{\mathrm{f}}}(\delta x) - m_{\tilde{\omega}_{\mathrm{f}}}(\delta n)}, \tag{5.23}$$

for $\overline{\eta} \in]0,1[$. If a tangential step is rejected, $\tilde{\omega}_f$ is updated via (5.22). As a result, τ and δs are computed again. This approach is repeated until the tangential step δt is accepted. In the case that iterates are very far away from the feasible set $c(x) = 0$, a stagnation of the updates of $\tilde{\omega}_{\mathrm{f}}$ can occur. Then, the focus is set on feasibility first, and the tangential step δt is discarded, yielding $\delta x = \delta n$. The stagnation is measured via the condition

$$\tilde{\omega}_{\mathrm{f_{new}}} < \left(1 + \varrho \frac{1-\overline{\eta}}{2}\right) \tilde{\omega}_{\mathrm{f_{old}}}$$

for a user-provided parameter $0 < \varrho < 1$. For the sake of efficiency, we do not discard a tangential step if $\eta_{\min} < \overline{\eta}$ for a suitable parameter $\eta_{\min} > 0$.
So far, we have only given a brief summary of the composite step method, and the reader is referred to [64, 67] for a detailed discussion.

5.2.8 Convergence criterion

For given relative and absolute accuracies $\Lambda_{\mathrm{CSAb}} > 0$ and $\Lambda_{\mathrm{CSRel}} > 0$, we consider the composite step method convergent if the relative criterion

$$\|\delta x\| \leq \Lambda_{\mathrm{CSRel}} \|x\| \tag{5.24}$$

or the absolute criterion

$$\|x\| \leq \Lambda_{\mathrm{CSAb}} \quad \text{and} \quad \|\delta x\| \leq \Lambda_{\mathrm{CSAb}} \tag{5.25}$$

is fulfilled. A sketch of our derived composite step method is shown in Algorithm 8.

Algorithm 8 Affine Covariant Composite Step Method.

Solve:

$$\min_{x \in X} f(x)$$
$$s.t.\ c(x) = 0.$$

Input: initial iterate (x_0, p_0), initial Lipschitz constants $\tilde{\omega}_{\mathrm{C}_0}$, $\tilde{\omega}_{\mathrm{f}_0}$, parameters $\eta_{\min}$, $\overline{\eta}$, ϱ, accuracies Λ_{CSAb}, Λ_{CSRel}, and contraction factors Θ_{acc}, Θ_{n}, and Θ_{d}.
Initialize: $k = 0$.
repeat
  $p_{k+1} \leftarrow updateLagrangeMultiplier()$ (via (5.10))
  $(\Delta n_k, \Delta t_k) \leftarrow computeSteps(p_{k+1}, x_k)$
  repeat
    $(\nu_k, \tau_k, \delta s_k, \tilde{\omega}_{\mathrm{C}_k}, \tilde{\omega}_{\mathrm{f}_k}) \leftarrow computeUpdates()$
  until step accepted (If Conditions (5.16) and (5.23) are satisfied.)
  $x_{k+1} \leftarrow x_k + \delta x_k + \delta s_k$
  $k \leftarrow k + 1$
until convergent (If Condition (5.24) or (5.25) is fulfilled.)

It remains to verify that the composite step method converges to a solution of (5.5). Showing global convergence seems to be out of reach for the affine covariant approach

studied here. Affine covariant methods avoid the evaluation of $\|c(x)\|$ which is required for the usual globalization mechanisms. However, under suitable assumptions, fast local convergence can be shown. To do so, we assume sufficient smoothness and second order sufficient optimality conditions at a local minimizer x_*, which we call an SSC point.

Theorem 5.2. *Consider the optimization problem described in* (5.5). *Further, assume that the iterates of Algorithm 8 converge to the SSC point* x_*. *If Conditions* (5.18)-(5.20) *hold in a neighborhood of* x_*, *Algorithm 8 admits local quadratic convergence.*

Proof. See [64, Proof of Theorem 3.17]. □

5.2.9 Adaption to inexactness and nonlinear elasticity

So far, it has been assumed that all arising systems are solved exactly. However, for large scale problems, this becomes infeasible and iterative solvers have to be deployed. In our case, we rely on PCG methods to solve large scale linear systems. The implications of this inexactness for the composite step method have been analyzed in [92]. Additionally, a strategy for accuracy matching has been proposed and tested. This is important since very tight tolerances would render each step too expensive. Also, too loose tolerances might lead to the loss of robustness and increase the number of outer iterations. The results elaborated in [92] can be summarized as follows. First, the relative accuracy Λ_{CSNorm} to compute the (simplified) normal step and the Lagrange multiplier update can be constant. Second, the relative accuracy of the tangential step Λ_{CSTang} can be set to the contraction factor $\frac{\|\delta s\|}{\|\delta x\|}$ of the previous step. These two conditions ensure efficiency while maintaining at least local superlinear convergence of the composite step method.
In the case of nonlinear elastic problems, the orientation-preserving condition $\det y > 0$ has to be taken into account. If this condition is violated for a trial iterate $x + \delta x$, we apply the damping

$$\nu \leftarrow \frac{1}{2}\nu \quad \text{and} \quad \tau \leftarrow \frac{1}{2}\tau$$

until $x + \delta x$ is feasible w.r.t. the constraint $\det y > 0$. Afterwards, additional damping of δn and δt as described for Algorithm 8 is applied. This approach was also considered in [64].
Going back to optimal control of nonlinear elastic contact problems, we can solve (4.12) only for fixed $\gamma > 0$.

5.3 Path-Following

Since we want to approximate solutions of the original problem (4.2), we introduce a path-following method. It is reasonable to assume that the difficulty of (4.12) depends on the regularization parameter γ. Thus, we augment our optimization algorithm by a path-following method, where the composite step methods acts as the inner solver. At this, we apply a simple step-size strategy to gradually increase the parameter γ.

Due to the inherent lack of structure of optimization problems involving nonlinear elasticity, more sophisticated approaches seem to be out of reach in the current setting. Heuristic schemes based on the convergence speed of the path-following algorithm could be applied. However, this remains a subject of future research. There exists a large body of literature analyzing path-following approaches, and the reader is referred to [23, 45, 46, 47, 93] for a more in-depth discussion.
Recalling the notation $Z := X \times P$ and $z := (x, p)$, the KKT system (4.13) can be interpreted as a parameter-dependent nonlinear equation:

$$F(z, \gamma) = 0.$$

Path-following methods are widely applied to solve parameter-dependent problems. For convex settings, it can often be shown that a homotopy-path $\gamma \to z(\gamma)$ of zeros of $F(\cdot, \gamma)$ exists. Sometimes, even sensitivity and a priori length estimates can be derived, cf. [47]. In non-convex settings, such results can only be observed a posteriori by a numerical algorithm. Additionally, we cannot rule out the existence of several paths which may converge to local solutions or end prematurely. Therefore, it is essential to employ a robust correction method.
Starting at a point (z_0, γ_0) close to homotopy-path, we want to successively compute solutions on the path for a increasing sequence of parameters γ_k. Assume that (z_k, γ_k) is a solution close to the path. For the next iterate, γ_k is increased by some constant factor $s_\mathrm{p} > 1$:

$$\gamma_{k+1} = s_\mathrm{p} \gamma_k.$$

With the composite step method functioning as a robust corrector, this simple approach is sufficient. However, an adaptive choice of the update parameter is advisable in order to reduce computational costs. Next, we utilize the composite step method to obtain a corresponding solution pair (z_{k+1}, γ_{k+1}) close to the path, where z_k is used as starting point. This approach can interpreted as a classical continuation method for parameter-dependent systems. We repeat this process until a solution close to the path for the desired parameter $\gamma_{\max}$ is found. In our analysis, s_p is chosen constant. This might result in too rapid increases of γ. In those cases, we expect that the globalization mechanism of the composite step method can steer the iterate back to the path. Our basic path-following approach is illustrated in Algorithm 9.

Algorithm 9 Basic Path-Following Method.

Solve: $F(z, \gamma_{\max}) = 0$.
Input: starting value (z_0, γ_0), maximum path-parameter $\gamma_{\max}$, and update factor s_{p}.
$(z_0, converged) \leftarrow compositeStepMethod(z_0, \gamma_0)$
if not *converged* **then**
  return; (No initial solution on the path found.)
end if
do
  $z_{k+1} \leftarrow z_k$
  $\gamma_{k+1} \leftarrow s_{\mathrm{p}} \gamma_k$
  $\gamma_{k+1} \leftarrow \min(\gamma_{k+1}, \gamma_{\max})$
  $(z_{k+1}, converged) \leftarrow compositeStepMethod(z_{k+1}, \gamma_{k+1})$
  if not *converged* **then**
    return; (Algorithm did not converge for (z_{k+1}, γ_{k+1}).)
  else
    $k \leftarrow k + 1$
  end if
while $\gamma_k < \gamma_{\max}$
return z_{k+1}; (Algorithm converged.)

Chapter 6

A Corrected Inexact Projected Preconditioned Conjugate Gradient Method

In the previous chapter, we introduced the composite step method in a function space setting to solve optimal control problems. Since computations in infinite-dimensional spaces are infeasible, discretization schemes, e.g., finite elements, are applied to obtain finite-dimensional problems. At this, solving the resulting systems numerically in an efficient way becomes the main difficulty.
Assume that the infinite-dimensional optimal control problem (5.5) has been discretized by some Galerkin-type method. The discretized space for the primal variables is again denoted by $X := Y \times U$. Accordingly, we obtain the discretized space P. Discrete linear operators are represented by matrices and their adjoints by transposed matrices. Further, we have the splitting $x := (y, u)$ and

$$c(x) = \hat{A}(y) - Bu,$$

where c denotes the discretized version of (5.6) and B is linear. Consider a fixed iterate (y_k, u_k, p_k) of the composite step method with $A := \hat{A}'(y_k)$ and $\mathcal{L} := \mathcal{L}(y_k, u_k, p_k)$. Further, we assume that both $\hat{A}$ and $\hat{A}'$ are continuously invertible. Then, the formulas for the (simplified) normal step, the Lagrange multiplier, and the tangential step yield matrices of the form

$$H_n := \begin{pmatrix} M_y & 0 & A^T \\ 0 & M_u & -B^T \\ A & -B & 0 \end{pmatrix} \quad \text{and} \quad H_t := \begin{pmatrix} \mathcal{L}_{yy} & \mathcal{L}_{yu} & A^T \\ \mathcal{L}_{uy} & \mathcal{L}_{uu} & -B^T \\ A & -B & 0 \end{pmatrix}.$$

Here,

$$M := \begin{pmatrix} M_y & 0 \\ 0 & M_u \end{pmatrix}$$

corresponds to the Riesz isomorphism chosen for the composite step setting. In the context of nonlinear elasticity, B denotes the discrete operator corresponding to the

outer energy I_{out}, and A denotes the Hessian matrix of the regularized energy functional I_γ or $\mathcal{E}_\gamma$ without I_{out}.
For large systems, applying direct methods becomes infeasible due to both memory constraints and computation time. Proven methods to solve such systems are PCG methods. However, in our case, the matrices H_n and H_t are generally not positive definite. Since M is positive definite, this issue can be overcome by applying a projected preconditioned conjugate gradient (PPCG) method, at least for systems involving the matrix H_n. This approach restricts the space of iterates to a suitable subspace where we have positive definiteness, see, e.g., [33]. In [64], PPCG methods were deployed in the context of optimal control of nonlinear elasticity.
If a suitable regularization is applied to the tangential step matrix H_t, the PPCG method can be utilized to compute the tangential step as well. Regularizations are necessary since the block

$$\mathcal{L}_{xx} = \begin{pmatrix} \mathcal{L}_{yy} & \mathcal{L}_{yu} \\ \mathcal{L}_{uy} & \mathcal{L}_{uu} \end{pmatrix}$$

is not positive definite in general. In view of the optimal control problem (4.3), we can set $\mathcal{L}_{yu} = \mathcal{L}_{yu} = 0$. But the examination in this chapter also applies to more general cases.
One drawback of PPCG methods is that they require exact solvers for the arising block systems. Therefore, we will modify the PPCG method accordingly and derive a suitable solution algorithm which utilizes the special structure of the matrices H_n and H_t.
This chapter is organized as follows. Section 6.1 introduces PPCG methods and addresses their application in a composite step setting. In Section 6.2, we examine how to deal with the inexact solutions of subsystems and derive an inexact PPCG method. However, this new algorithm does no longer correspond to the original problem. Therefore, in Section 6.3, the inexact PPCG algorithm is extended by a correction mechanism which compensates for the inexactness. The resulting approach has been developed by Anton Schiela and Alexander Siegl in cooperation with the author, cf. [96]. Additionally, we discuss the implementation of our scheme.
Parts of this chapter have been published in [94].

6.1 PPCG methods

We start by analyzing PPCG methods for general settings. Consider the following linear system:

$$H \begin{pmatrix} x \\ p \end{pmatrix} = \begin{pmatrix} b_x \\ 0 \end{pmatrix}$$

with

$$H := \begin{pmatrix} M & C^T \\ C & 0 \end{pmatrix},$$

where M is symmetric and positive definite. The matrix C is surjective and usually describes equality constraints. As discussed above, the full system matrix H does not

need to be positive definite. PPCG methods extend classical PCG methods in the following way: to overcome non-convexity, they restrict the primal part x of each iterate to a subspace where the respective matrix H is positive definite. This is achieved by applying a suitable preconditioner. In our examination, we choose a preconditioner of the form

$$Q = \begin{pmatrix} \tilde{M} & C^T \\ C & 0 \end{pmatrix}. \tag{6.1}$$

Here, $\tilde{M}$ represents a preconditioner for M which is required to be symmetric and positive definite on $\ker C$. Recalling the PCG method defined in Algorithm 2, Q is applied to the negative residuum, yielding the equation

$$g_{k+1} = Q^{-1}(-r_{k+1}). \tag{6.2}$$

To increase readability, we omit the iteration index of the PCG method if it is clear from the context. Applying the preconditioner here is equivalent to the following minimization problem:

$$\begin{aligned} \min_{g_x \in X} \; & g_x^T r_x + \frac{1}{2}\langle g_x, g_x \rangle_{\tilde{M}} \\ s.t. \; & Cg_x = 0, \end{aligned} \tag{6.3}$$

with corresponding Lagrange multiplier g_p. Here, the indices x and p denote the primal and dual component, respectively. Writing

$$r = \begin{pmatrix} r_x \\ r_p \end{pmatrix},$$

Equation (6.2) decouples into the block system

$$\tilde{M} g_x + C^T g_p = -r_x, \tag{6.4}$$

$$C g_x = -r_p. \tag{6.5}$$

Inserting this approach into Algorithm 2 yields a projected preconditioned conjugate gradient method, see Algorithm 10. We verify that for proper right hand sides and initial conditions, the PPCG method solely operates on $\ker C$.

Proposition 6.1. *Consider Algorithm 10. If Equations* (6.4) *and* (6.5) *are solved exactly, the primal components x_k, $(d_k)_x$, and $(g_k)_x$ are again contained in the kernel of C in each iterate.*

Proof. Assume that in iteration k, x_k satisfies $Cx_k = 0$. Further, $(d_k)_x \in \ker C$, $(g_k)_x \in \ker C$, and $(r_k)_p$ is zero. As a result, after updating the residuum via

$$r_{k+1} \leftarrow r_k + \alpha_k H d_k,$$

the dual component $(r_{k+1})_p$ remains zero. Thus, applying the preconditioner Q to the negative residuum $-r_{k+1}$ yields $(g_{k+1})_x \in \ker C$. The new direction d_{k+1} results from simply adding g_{k+1} to the previous scaled direction $\beta_{k+1} d_k$. Accordingly, $(d_{k+1})_x \in \ker C$. Given the initial setting and formulas for x_0, r_0, d_0, and g_0, the statement follows via an induction argument. □

Algorithm 10 Projected Preconditioned Conjugate Gradient Method.

Solve: $H \begin{pmatrix} x \\ p \end{pmatrix} = \begin{pmatrix} b_x \\ 0 \end{pmatrix}$.

Input: initial iterate $z_0 = \begin{pmatrix} x_0 \\ p_0 \end{pmatrix}$ satisfying $Cx_0 = 0$.

Initialize: $r_0 = Hz_0 - b$, $d_0 = g_0 = Q^{-1}(-r_0)$, and $k = 0$.

repeat

$\alpha_k \leftarrow -\frac{r_k^T g_k}{d_k^T H d_k}$

$z_{k+1} \leftarrow z_k + \alpha_k d_k$

$r_{k+1} \leftarrow r_k + \alpha_k H d_k$

$g_{k+1} \leftarrow Q^{-1}(-r_{k+1})$

$\beta_{k+1} \leftarrow \frac{r_{k+1}^T g_{k+1}}{r_k^T g_k}$

$d_{k+1} \leftarrow g_{k+1} + \beta_{k+1} d_k$

$k \leftarrow k + 1$

until convergent

In addition to this, preconditioners should cluster the eigenvalues of the preconditioned matrix $Q^{-1}H$. Such a result was shown in [55, Theorem 2.1].

Theorem 6.2. *Assume the constraint matrix $C \in \mathbb{R}^{m \times n}$ has full rank and $\tilde{M} \neq M$. Furthermore, Z is an $n \times (n-m)$ basis for the nullspace of C. Then, the matrix $Q^{-1}H$ has*

1. *an eigenvalue at 1 with multiplicity $2m$, and*
2. *$n - m$ eigenvalues which are defined by the generalized eigenvalue problem*

$$Z^T M Z x_\lambda = \lambda Z^T \tilde{M} Z x_\lambda.$$

Proof. See [55, Proof of Theorem 2.1]. □

Moreover, an upper bound on the dimension of the corresponding Krylov subspace can be derived.

Theorem 6.3. *Let the constraint matrix $C \in \mathbb{R}^{m \times n}$ be of full rank with $m < n$. Furthermore, H is nonsingular, $\tilde{M} \neq M$, and b is an arbitrary right hand side. Next, let Z be an $n \times (n-m)$ basis for the nullspace of C such that*

$$(Z^T \tilde{M} Z)^{-1} (Z^T M Z)$$

has k ($1 \leq k \leq n - m$) distinct eigenvalues λ_i with ($1 \leq i \leq k$). Additionally, these eigenvalues have the multiplicity μ_i, where $\sum_{i=1}^{k} \mu_i = n - m$. Then, the dimension of the Krylov subspace $\mathcal{K}(Q^{-1}H, b)$ is at most $k + 2$.

Proof. See [55, Proof of Theorem 3.7]. □

In summary, we can conclude that the PPCG algorithm corresponds to a standard PCG method that operates on the linear subspace $\ker C$.

PPCG methods for optimal control problems

Throughout the subsequent analysis, we use the setting and the notation, as defined in the introduction of this chapter. We start by studying the computation of the (simplified) normal step and the Lagrange multiplier update.

Normal step system and Lagrange multiplier update

First, note that the right hand side for the (simplified) normal step (5.12) is of the form

$$b = \begin{pmatrix} 0 \\ 0 \\ b_p \end{pmatrix}.$$

Therefore, it has to be adjusted to fit into the setting of Algorithm 10. The respective system reads as follows:

$$H_n z = b. \tag{6.6}$$

Defining $C := (A, -B)$, $z := z_0 + \overline{z}$ with

$$z_0 := \begin{pmatrix} A^{-1} b_p \\ 0 \\ 0 \end{pmatrix},$$

and $\overline{z} \in C$, Equation (6.6) can be reformulated to

$$H_n \overline{z} = b - H_n z_0 = \begin{pmatrix} -M_y A^{-1} b_p \\ 0 \\ 0 \end{pmatrix}.$$

Taking into account the system for the Lagrange multiplier update (5.9), it suffices to study right hand sides of the form

$$b = \begin{pmatrix} b_y \\ b_u \\ 0 \end{pmatrix}.$$

Thus, we obtain systems of the form

$$\begin{pmatrix} M_y & 0 & A^T \\ 0 & M_u & -B^T \\ A & -B & 0 \end{pmatrix} \begin{pmatrix} y \\ u \\ p \end{pmatrix} = \begin{pmatrix} b_y \\ b_u \\ 0 \end{pmatrix}, \tag{6.7}$$

with

$$M := \begin{pmatrix} M_y & 0 \\ 0 & M_u \end{pmatrix}.$$

This system corresponds to solving the minimizing problem

$$\begin{aligned} \min_{(y,u)\in Y\times U} \frac{1}{2}y^T M_y y + \frac{1}{2}u^T M_u u - b_y^T y - b_u^T u \\ s.t.\ Ay - Bu = 0, \end{aligned} \tag{6.8}$$

with Lagrange multiplier p.

Remark 6.4. *For the remainder of this chapter, we will denote the Lagrange multiplier of the composite method by p_c and the dual component of solutions to linear systems by p to avoid ambiguity.*

A common choice for the constraint preconditioner is

$$Q = \begin{pmatrix} 0 & 0 & A^T \\ 0 & M_u & -B^T \\ A & -B & 0 \end{pmatrix}, \tag{6.9}$$

see, e.g., [82, Chapter 5]. In the context of optimal control of nonlinear elasticity, this kind of preconditioner has been studied in [64]. For the optimal control setting presented above, we can derive the estimate

$$\begin{aligned} \langle u,u\rangle_U \leq \langle y,y\rangle_Y + \langle u,u\rangle_U &= \langle A^{-1}Bu, A^{-1}Bu\rangle_Y + \langle u,u\rangle_U \\ &\leq (1 + \|A^{-1}B\|^2_{U\to Y})\langle u,u\rangle_U. \end{aligned} \tag{6.10}$$

Thus, the matrix

$$\tilde{M} := \begin{pmatrix} 0 & 0 \\ 0 & M_u \end{pmatrix}$$

is spectrally equivalent to M on $\ker C$. This result ensures that the condition number of the preconditioned matrix $Q^{-1}H_n$ is independent of the discretization even if the block M_y is omitted in $\tilde{M}$. See also [64, Chapter 4]. When studying objective functionals of the form (4.1), the following norm choice is very common:

$$\|u\|_{\tilde{U}} := \sqrt{\alpha}\|u\|_U,$$

where $\alpha > 0$ denotes the Tikhonov regularization parameter. Then, we obtain the estimate

$$\langle u,u\rangle_{\tilde{U}} \leq (1 + \frac{1}{\alpha}\|A^{-1}B\|^2_{U\to Y})\langle u,u\rangle_{\tilde{U}},$$

which implies that the condition number of $Q^{-1}H_n$ can increase with $\frac{1}{\alpha}$.

Remark 6.5. *Note that the estimate in* (6.10) *depends on the inverse of A. In the case of nonlinear elasticity, this matrix, even if regularized, can be almost singular. Consequently, the condition number of $Q^{-1}H_n$ might deteriorate. One possible way to overcome this issue is to suitably regularize the matrix A.*

In the PCG method, the preconditioner is applied in the equation

$$Qg_{k+1} = -r_{k+1}.$$

Inserting the definition of Q yields

$$\begin{pmatrix} 0 & 0 & A^T \\ 0 & M_u & -B^T \\ A & -B & 0 \end{pmatrix} \begin{pmatrix} g_y \\ g_u \\ g_p \end{pmatrix} = - \begin{pmatrix} r_y \\ r_u \\ r_p \end{pmatrix}. \tag{6.11}$$

For initial iterates x_0 satisfying $Cx_0 = 0$, Algorithm 10 can be applied. Moreover, Proposition 6.1 ensures that $r_p = 0$ in each iterate. The application of the preconditioner Q can be split up into the subsystems

$$A^T g_p = -r_y, \tag{6.12}$$

$$M_u g_u - B^T g_p = -r_u, \tag{6.13}$$

$$Ag_y - Bg_u = 0. \tag{6.14}$$

This decoupling ensures that the preconditioner can be applied efficiently, provided that a factorization of A is at hand. However, for large scale systems, factorizing the matrix A is no longer possible. Thus, the linear systems involving A have to be solved inexactly, which is no longer consistent with the PPCG method. In Section 6.2, we will discuss techniques and methods to overcome this problem.

Remark 6.6. *In general settings, solving* (6.13) *can be challenging as well. In our study, the block M_u corresponds to a scaled mass matrix of the boundary control. Thus, it is very small compared to A and can be factorized at low computational cost. When this is not the case, an efficient solver for M_u has to be deployed as well. In many cases, M_u is a mass matrix, which can be replaced by its diagonal in the preconditioner Q.*

Tangential step

Here, we also apply the block preconditioner Q as defined in (6.9). In the tangential step matrix H_t, the block $\mathcal{L}_{xx}$ is not necessarily positive definite on $\ker C$, and (5.15) no longer corresponds to a quadratic minimization problem. Consequently, descent of tangential steps is not guaranteed, unless appropriate adjustments are made. Therefore, we apply a Hessian modification approach, where we add a positive definite regularization term R in the following way:

$$\begin{pmatrix} \mathcal{L}_{xx} + \lambda R & C^T \\ C & 0 \end{pmatrix} \begin{pmatrix} \delta t \\ q \end{pmatrix} = - \begin{pmatrix} \mathcal{L}_x + \mathcal{L}_{xx}\delta n \\ 0 \end{pmatrix}. \tag{6.15}$$

We set $R = M$ with the notation $R_y := M_y$ and $R_u := M_u$. Further, $\lambda \geq 0$ is an algorithmic parameter which is chosen sufficiently large to ensure that $\mathcal{L}_{xx} + \lambda R$ is positive definite on $\ker C$.

In our setting, each step is started with $\lambda = 0$. If non-convexities are encountered in the PPCG method, λ is updated according to Algorithm 3 with an initialization parameter λ_{tInit} and update parameter s_{t}. Thereafter, a new attempt to compute the tangential step is made. Subsequently, λ is increased until the PPCG method does not encounter directions of negative curvature anymore. This approach is summarized in Algorithm 11.

Algorithm 11 Regularized Projected Preconditioned Conjugate Gradient Method for the Tangential Step.

Solve: $H_t \begin{pmatrix} x \\ p \end{pmatrix} = \begin{pmatrix} b_x \\ 0 \end{pmatrix}$.

Input: initial iterate $z_0 = \begin{pmatrix} x_0 \\ p_0 \end{pmatrix}$ satisfying $Cx_0 = 0$, regularization update parameter s_{t}, and starting regularization parameter λ_{tInit}.

Initialize: $r_0 = H_t z_0 - b$, $d_0 = g_0 = Q^{-1}(-r_0)$, $\lambda = 0$, and $k = 0$.

repeat

  if $d_k^T H_t d_k < 0$ **then**

    $\lambda \leftarrow updateLambda(\lambda, \lambda_{\text{tInit}}, s_{\text{t}})$ (Algorithm 3)

    $H_t \leftarrow H_t + \lambda \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix}$

    restart (Restart the entire PPCG algorithm.)

  end if

  $\alpha_k \leftarrow -\frac{r_k^T g_k}{d_k^T H_t d_k}$

  $z_{k+1} \leftarrow z_k + \alpha_k d_k$

  $r_{k+1} \leftarrow r_k + \alpha_k H_t d_k$

  $g_{k+1} \leftarrow Q^{-1}(-r_{k+1})$

  $\beta_{k+1} \leftarrow \frac{r_{k+1}^T g_{k+1}}{r_k^T g_k}$

  $d_{k+1} \leftarrow g_{k+1} + \beta_{k+1} d_k$

  $k \leftarrow k + 1$

until convergent

The introduced PPCG method relies on the exact solutions of the systems involving the matrices A and A^T. However, for large linear systems, this becomes infeasible, and only iterative solvers can be used.

6.2 Inexact PPCG method

If not stated otherwise, we assume that A is positive definite and symmetric. Considering Equations (6.12) and (6.14), the default approach would be to apply a PCG method with high accuracy to ensure that iterates are again sufficiently close to the kernel of C. However, PCG methods are not linear, which rules out their application in the preconditioner Q. Also, solving systems with high accuracy is rather expensive. Therefore,

we have to fall back to linear iterative solvers and take into account the possible inexactness. The choice for the iterative solver will be discussed in Subsection 6.3.4. The corresponding inexact operators of A and A^{-1} are denoted by $\tilde{A}$ and $\tilde{A}^{-1}$, respectively. Note that $\tilde{A}$ cannot be evaluated explicitly since $\tilde{A}^{-1}$ is only available via an iterative method. Replacing A and A^{-1} with $\tilde{A}$ and $\tilde{A}^{-1}$ leads to the inexact preconditioner

$$\tilde{Q} := \begin{pmatrix} 0 & 0 & \tilde{A}^T \\ 0 & M_u & -B^T \\ \tilde{A} & -B & 0 \end{pmatrix}. \tag{6.16}$$

Then, the equation

$$\tilde{Q} g_{k+1} = -r_{k+1}$$

again splits into the systems

$$\tilde{A}^T g_p = -r_y, \tag{6.17}$$

$$M_u g_u - B^T g_p = -r_u, \tag{6.18}$$

$$\tilde{A} g_y - B g_u = 0. \tag{6.19}$$

By shifting the preconditioner to the inexact constraint $\tilde{C} := (\tilde{A}, -B)$, it is no longer guaranteed that the algorithm operates on the kernel of C. To adjust for this, we have to transfer the PPCG method to the kernel of $\tilde{C}$. This leads to the inexact normal step matrix

$$\tilde{H}_n := \begin{pmatrix} M_y & 0 & \tilde{A}^T \\ 0 & M_u & -B^T \\ \tilde{A} & -B & 0 \end{pmatrix}, \tag{6.20}$$

and consequently, to the new inexact system

$$\tilde{H}_n z = b.$$

Although this approach ensures positive definiteness, the evaluation of the product $\tilde{H}_n d_k$ seems to be infeasible since $\tilde{A}$ and $\tilde{A}^T$ cannot be directly applied. By a recursive computation, the application of these inexact operators can be avoided. Consider

$$\tilde{H}_n d_k = \begin{pmatrix} M_y & 0 & \tilde{A}^T \\ 0 & M_u & -B^T \\ \tilde{A} & -B & 0 \end{pmatrix} \begin{pmatrix} d_y \\ d_u \\ d_p \end{pmatrix} = \begin{pmatrix} M_y d_y \\ M_u d_u \\ 0 \end{pmatrix} + \begin{pmatrix} \tilde{A}^T d_p \\ -B^T d_p \\ \tilde{A} d_y - B d_u \end{pmatrix}.$$

While the first summand can be evaluated directly, the second one includes the operators $\tilde{A}$ and $\tilde{A}^T$. After defining

$$\xi_k := \begin{pmatrix} 0 & 0 & \tilde{A}^T \\ 0 & 0 & -B^T \\ \tilde{A} & -B & 0 \end{pmatrix} d_k,$$

we derive the recursive formula

$$\begin{pmatrix} 0 & 0 & \tilde{A}^T \\ 0 & 0 & -B^T \\ \tilde{A} & -B & 0 \end{pmatrix} d_k = \begin{pmatrix} 0 & 0 & \tilde{A}^T \\ 0 & 0 & -B^T \\ \tilde{A} & -B & 0 \end{pmatrix} g_k + \beta_k \xi_{k-1},$$

for $k \geq 0$. For the first summand, it holds that

$$\begin{pmatrix} 0 & 0 & \tilde{A}^T \\ 0 & 0 & -B^T \\ \tilde{A} & -B & 0 \end{pmatrix} g_k = \begin{pmatrix} -(r_k)_y \\ -B^T (g_k)_p \\ -(r_k)_p \end{pmatrix}.$$

By the same arguments as applied in the proof of Proposition 6.1, we conclude that $(r_k)_p = 0$ and that each iterate is contained in $\ker \tilde{C}$, for suitable starting values. Since d_0 is initialized with g_0, choosing

$$\xi_{-1} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

and $\beta_0 = 0$ allows the recursive computation of $\tilde{H}_n d_k$ without having to evaluate $\tilde{A}$. Still, for the initialization of the residuum, the product $\tilde{H}_n z_0$ is required. This can be overcome by choosing the initial iterate

$$z_0 = \begin{pmatrix} y_0 \\ u_0 \\ p_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Also this choice ensures that $\tilde{A} y_0 - B u_0 = 0$. Consequently, we obtain a PPCG method as defined in Algorithm 10. Further, it takes into account the application of inexact solvers for the subsystems. The resulting scheme is summarized in Algorithm 12.
The analysis for the tangential step follows analogously, with an additional regularization and

$$\tilde{H}_t := \begin{pmatrix} \mathcal{L}_{yy} & 0 & \tilde{A}^T \\ 0 & \mathcal{L}_{uu} & -B^T \\ \tilde{A} & -B & 0 \end{pmatrix}.$$

For the details, see Algorithm 13. In summary, we have derived an inexact PPCG method for the linear systems arising in the composite step method. At this, we applied the block preconditioner $\tilde{Q}$ without relying on solving the resulting equations exactly. However, in doing so, we shifted the problems to the kernel of the inexact constraints $\tilde{C}$. Therefore, it remains to transfer these results back to $\ker C$.

Algorithm 12 Inexact Projected Preconditioned Conjugate Gradient (IPPCG) Method for the (Simplified) Normal Step and the Lagrange Multiplier Update.

Solve: $\tilde{H}_n \begin{pmatrix} y \\ u \\ p \end{pmatrix} = \begin{pmatrix} b_y \\ b_u \\ 0 \end{pmatrix}$.

Initialize: $\xi_{-1} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$, $\beta_0 = 0$, initial iterate $z_0 = \begin{pmatrix} y_0 \\ u_0 \\ p_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$, $r_0 = \tilde{H}_n \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} b_y \\ b_u \\ 0 \end{pmatrix}$, $d_0 = g_0 = \tilde{Q}^{-1}(-r_0)$, and $k = 0$.

repeat

$\xi_k \leftarrow - \begin{pmatrix} (r_k)_y \\ B^T(g_k)_p \\ 0 \end{pmatrix} + \beta_k \xi_{k-1}$

$\tilde{H}_n d_k \leftarrow \begin{pmatrix} M_y(d_k)_y \\ M_u(d_k)_u \\ 0 \end{pmatrix} + \xi_k$

$\alpha_k \leftarrow -\frac{r_k^T g_k}{d_k^T \tilde{H}_n d_k}$

$z_{k+1} \leftarrow z_k + \alpha_k d_k$

$r_{k+1} \leftarrow r_k + \alpha_k \tilde{H}_n d_k$

$g_{k+1} \leftarrow \tilde{Q}^{-1}(-r_{k+1})$

$\beta_{k+1} \leftarrow \frac{r_{k+1}^T g_{k+1}}{r_k^T g_k}$

$d_{k+1} \leftarrow g_{k+1} + \beta_{k+1} d_k$

$k \leftarrow k + 1$

until convergent

Algorithm 13 Inexact Regularized Projected Preconditioned Conjugate Gradient Method for the Tangential Step.

Solve: $\tilde{H}_t \begin{pmatrix} y \\ u \\ p \end{pmatrix} = \begin{pmatrix} b_y \\ b_u \\ 0 \end{pmatrix}$.

Input: regularization update parameter s_{tI}, and starting regularization parameter $\lambda_{\mathrm{tIInit}}$.

Initialize: $\xi_{-1} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$, $\beta_0 = 0$, initial iterate $z_0 = \begin{pmatrix} y_0 \\ u_0 \\ p_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$, $r_0 = \tilde{H}_t \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} b_y \\ b_u \\ 0 \end{pmatrix}$, $d_0 = g_0 = \tilde{Q}^{-1}(-r_0)$, $\lambda = 0$, and $k = 0$.

repeat

$\xi_k \leftarrow -\begin{pmatrix} (r_k)_y \\ B^T(g_k)_p \\ 0 \end{pmatrix} + \beta_k \xi_{k-1}$

$\tilde{H}_t d_k \leftarrow \begin{pmatrix} (\mathcal{L}_{yy} + \lambda R_y)(d_k)_y \\ (\mathcal{L}_{uu} + \lambda R_u)(d_k)_u \\ 0 \end{pmatrix} + \xi_k$

if $d_k^T \tilde{H}_t d_k < 0$ **then**

$\lambda \leftarrow updateLambda(\lambda, \lambda_{\mathrm{tIInit}}, s_{\mathrm{tI}})$ (Algorithm 3)

restart (Restart the entire algorithm.)

end if

$\alpha_k \leftarrow -\frac{r_k^T g_k}{d_k^T \tilde{H}_t d_k}$

$z_{k+1} \leftarrow z_k + \alpha_k d_k$

$r_{k+1} \leftarrow r_k + \alpha_k \tilde{H}_t d_k$

$g_{k+1} \leftarrow \tilde{Q}^{-1}(-r_{k+1})$

$\beta_{k+1} \leftarrow \frac{r_{k+1}^T g_{k+1}}{r_k^T g_k}$

$d_{k+1} \leftarrow g_{k+1} + \beta_{k+1} d_k$

$k \leftarrow k + 1$

until convergent

6.3 An outer correction loop

The restriction to $\ker \tilde{C}$ allows the application of PPCG methods, as described in the previous section. However, the resulting solutions need to be corrected to satisfy the exact constraints $Cx = 0$. Thus, we add an outer correction loop to Algorithms 12 and 13 to recover solutions in $\ker C$. Again, we start with the (simplified) normal step system.

6.3.1 (Simplified) normal step system and Lagrange multiplier update

Roughly speaking, we construct an outer iterative method to solve the normal system

$$H_n z = b. \tag{6.21}$$

Given an initial iterate $z = \begin{pmatrix} x \\ p \end{pmatrix}$ with $Cx = 0$, a simple update δz can be derived via

$$H_n(z + \delta z) = b \Leftrightarrow H_n \delta z = b - H_n z. \tag{6.22}$$

An inexact update $\delta\tilde{z}$ for (6.22) can be computed with Algorithm 12. Since $(\delta\tilde{z})_x \in \ker \tilde{C}$, a projection back the original constraint is required to obtain the final update δz which is added to the initial iterate z. This process is repeated until we obtain a sufficiently accurate solution of (6.21). In the subsequent analysis, we will discuss this new method in detail.

Update of the dual component

Consider a given iterate $z = \begin{pmatrix} y \\ u \\ p \end{pmatrix}$ of our new outer iterative algorithm with

$$Ay - Bu = 0. \tag{6.23}$$

Motivated by the first line of Equation (6.6), the dual component p can be updated by solving

$$A^T p = b_y - M_y y.$$

A numerically more stable version is computing the update Δp via

$$A^T \Delta p = b_y - M_y y - A^T p \tag{6.24}$$

and adding it to p:

$$p \leftarrow p + \Delta p.$$

Remark 6.7. *For the theoretical analysis now, we assume that Equation* (6.24) *is solved exactly. In Subsection 6.3.4, we relax this condition and discuss the implications.*

Inexact update step

We recall the original problem

$$H_n \delta z = b - H_n z,$$

which corresponds to the system

$$\begin{pmatrix} M_y & 0 & A^T \\ 0 & M_u & -B^T \\ A & -B & 0 \end{pmatrix} \begin{pmatrix} \delta y \\ \delta u \\ \delta p \end{pmatrix} = \begin{pmatrix} b_y - M_y y - A^T p \\ b_u - M_u u + B^T p \\ -Ay + Bu \end{pmatrix}.$$

By inserting Conditions (6.23) and (6.24), the right hand side simplifies to

$$\hat{b} = \begin{pmatrix} 0 \\ b_u - M_u u + B^T p \\ 0 \end{pmatrix},$$

yielding the new system

$$H_n \delta z = \hat{b}.$$

To solve this equation, we apply a PPCG method, which requires a suitable preconditioner. As shown in the analysis of Section 6.1, a block preconditioner of the form (6.9) is a viable option. However, for large systems, only the inexact version as defined in (6.16) can be applied. Consequently, this choice implies that positive definiteness can only be guaranteed on $\ker \tilde{C}$. Thus, H_n has to be replaced by $\tilde{H}_n$, leading to the inexact system

$$\tilde{H}_n \delta \tilde{z} = \hat{b}. \tag{6.25}$$

This system can be solved with Algorithm 12 which yields the inexact update $(\delta \tilde{z})_x \in \ker \tilde{C}$. Since the original problem requires solutions in $\ker C$, the inexact update $\delta \tilde{z}$ has to be projected back to $\ker C$.

Projection

Writing $\delta \tilde{z} = \begin{pmatrix} \delta \tilde{y} \\ \delta \tilde{u} \\ \delta \tilde{p} \end{pmatrix}$, we project the inexact update back to $\ker C$ by solving

$$A \delta y = B \delta \tilde{u} \tag{6.26}$$

for δy, yielding $\begin{pmatrix} \delta y \\ \delta \tilde{u} \end{pmatrix} \in \ker C$.

Remark 6.8. *Again, we assume that Equation* (6.26) *is solved exactly, and we address the relaxation of this condition in Subsection 6.3.4.*

Before the primal variables y and u can be updated, a last modification is necessary.

Damping

Equation (6.21) corresponds to the minimization problem

$$\begin{aligned} \min_{(y,u)\in Y\times U} g(y,u) := \frac{1}{2}y^T M_y y + \frac{1}{2}u^T M_u u - b_y^T y - b_u^T u \\ s.t.\ Ay - Bu = 0. \end{aligned} \tag{6.27}$$

Therefore, we have to ensure that after the projection step, a descent of g is guaranteed. This can be achieved by computing a directional minimizer of g w.r.t. the direction $\begin{pmatrix}\delta y \\ \delta\tilde{u}\end{pmatrix}$. Consequently, we compute the optimal line search parameter

$$\omega := -\frac{\left(\begin{pmatrix}M_y & 0\\ 0 & M_u\end{pmatrix}\begin{pmatrix}y\\u\end{pmatrix} - \begin{pmatrix}b_y\\b_u\end{pmatrix}\right)^T \begin{pmatrix}\delta y\\ \delta\tilde{u}\end{pmatrix}}{\begin{pmatrix}\delta y^T & \delta\tilde{u}^T\end{pmatrix}\begin{pmatrix}M_y & 0\\ 0 & M_u\end{pmatrix}\begin{pmatrix}\delta y\\ \delta\tilde{u}\end{pmatrix}}. \tag{6.28}$$

Finally, the primal components are updated:

$$\begin{aligned} y &\leftarrow y + \omega\delta y \\ u &\leftarrow u + \omega\delta\tilde{u}. \end{aligned}$$

The entire method is described in Algorithm 14.

6.3.2 Tangential step

The analysis for the tangential step follows analogously. The only difference is the possible negative definiteness of $\mathcal{L}_{xx}$. Thus, a suitable regularization is incorporated analogously to Algorithm 13. Additionally, if a non-convexity of $\mathcal{L}_{xx}$ is encountered, the entire algorithm is restarted after the regularization. The resulting corrected inexact projected preconditioned conjugate gradient (CIPPCG) method for the tangential step is summarized in Algorithm 15. Next, it has to be verified that the derived CIPPCG method converges. Therefore, we show that it is equivalent to a gradient method, and as a result, inherits its convergence properties.

Algorithm 14 Corrected Inexact Projected Preconditioned Conjugate Gradient Method for the (Simplified) Normal Step and the Lagrange Multiplier Update.

Solve: $H_n \begin{pmatrix} y \\ u \\ p \end{pmatrix} = \begin{pmatrix} b_y \\ b_u \\ 0 \end{pmatrix}$.

Input: initial iterate $z_0 = \begin{pmatrix} y_0 \\ u_0 \\ p_0 \end{pmatrix}$ with $Ay_0 - Bu_0 = 0$.

Initialize: $k = 0$.

repeat

$\Delta p_k \leftarrow A^{-T}(b_y - M_y y_k - A^T p_k)$

$p_{k+1} \leftarrow p_k + \Delta p_k$

Solve: $\tilde{H}_n \begin{pmatrix} \delta\tilde{y}_k \\ \delta\tilde{u}_k \\ \delta\tilde{p}_k \end{pmatrix} = \begin{pmatrix} 0 \\ b_u - M_u u_k + B^T p_{k+1} \\ 0 \end{pmatrix}$ (Algorithm 12)

$\delta y_k \leftarrow A^{-1} B \delta\tilde{u}_k$

$$\omega_k \leftarrow -\frac{\left(\begin{pmatrix} M_y & 0 \\ 0 & M_u \end{pmatrix}\begin{pmatrix} y_k \\ u_k \end{pmatrix} - \begin{pmatrix} b_y \\ b_u \end{pmatrix}\right)^T \begin{pmatrix} \delta y_k \\ \delta\tilde{u}_k \end{pmatrix}}{\begin{pmatrix} \delta y_k{}^T & \delta\tilde{u}_k{}^T \end{pmatrix}\begin{pmatrix} M_y & 0 \\ 0 & M_u \end{pmatrix}\begin{pmatrix} \delta y_k \\ \delta\tilde{u}_k \end{pmatrix}}$$

$y_{k+1} \leftarrow y_k + \omega_k \delta y_k$

$u_{k+1} \leftarrow u_k + \omega_k \delta\tilde{u}_k$

$k \leftarrow k + 1$

until convergent

Algorithm 15 Corrected Inexact Projected Preconditioned Conjugate Gradient Method for the Tangential Step.

Solve: $H_t \begin{pmatrix} y \\ u \\ p \end{pmatrix} = \begin{pmatrix} b_y \\ b_u \\ 0 \end{pmatrix}$.

Input: initial iterate $z_0 = \begin{pmatrix} {y_0}^T & {u_0}^T & {p_0}^T \end{pmatrix}^T$ with $Ay_0 - Bu_0 = 0$, regularization update parameter s_{tIC}, and starting regularization parameter λ_{tICInit}.
Initialize: $k = 0$ and $\lambda = 0$.
repeat
$\Delta p_k \leftarrow A^{-T}(b_y - (\mathcal{L}_{yy} + \lambda R_y)y_k - A^T p_k)$
$p_{k+1} \leftarrow p_k + \Delta p_k$

$$\text{Solve: } \tilde{H}_t \begin{pmatrix} \delta\tilde{y}_k \\ \delta\tilde{u}_k \\ \delta\tilde{p}_k \end{pmatrix} = \begin{pmatrix} 0 \\ b_u - (M_u + \lambda R_u)u_k + B^T p_{k+1} \\ 0 \end{pmatrix} \begin{pmatrix} \text{Algorithm 13 with} \\ \text{regularization} \\ \text{parameters} \\ s_{\text{tIC}} \text{ and } \lambda_{\text{tICInit}} \end{pmatrix}$$

if $\tilde{H}_t$ was regularized during the computation of the previous step **then**
update λ accordingly.
$H_t \leftarrow H_t + \lambda \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix}$
$\tilde{H}_t \leftarrow \tilde{H}_t + \lambda \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix}$
restart (Restart the entire algorithm.)
end if
$\delta y_k \leftarrow A^{-1} B \delta\tilde{u}_k$

$$\omega_k \leftarrow -\frac{\left(\begin{pmatrix} \mathcal{L}_{yy} + \lambda R_y & 0 \\ 0 & \mathcal{L}_{uu} + \lambda R_u \end{pmatrix}\begin{pmatrix} y_k \\ u_k \end{pmatrix} - \begin{pmatrix} b_y \\ b_u \end{pmatrix}\right)^T \begin{pmatrix} \delta y_k \\ \delta\tilde{u}_k \end{pmatrix}}{\begin{pmatrix} {\delta y_k}^T & {\delta\tilde{u}_k}^T \end{pmatrix}\begin{pmatrix} \mathcal{L}_{yy} + \lambda R_y & 0 \\ 0 & \mathcal{L}_{uu} + \lambda R_u \end{pmatrix}\begin{pmatrix} \delta y_k \\ \delta\tilde{u}_k \end{pmatrix}}$$

$y_{k+1} \leftarrow y_k + \omega_k \delta y_k$
$u_{k+1} \leftarrow u_k + \omega_k \delta\tilde{u}_k$
if $\begin{pmatrix} {y_{k+1}}^T & {u_{k+1}}^T \end{pmatrix}(\mathcal{L}_{xx} + \lambda R)\begin{pmatrix} y_{k+1} \\ u_{k+1} \end{pmatrix} < 0$ **then**
$\lambda \leftarrow updateLambda(\lambda, \lambda_{\text{tICInit}}, s_{\text{tIC}})$ (Algorithm 3)
$H_t \leftarrow H_t + \lambda \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix}$
$\tilde{H}_t \leftarrow \tilde{H}_t + \lambda \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix}$
restart (Restart the entire algorithm.)
end if
$k \leftarrow k + 1$
until convergent

6.3.3 Equivalence to the gradient method

For the convergence analysis, we can w.l.o.g. combine the normal step and the tangential step systems to equations of the form

$$\begin{pmatrix} H_y & 0 & A^T \\ 0 & H_u & -B^T \\ A & -B & 0 \end{pmatrix} \begin{pmatrix} y \\ u \\ p \end{pmatrix} = \begin{pmatrix} b_y \\ b_u \\ 0 \end{pmatrix}, \tag{6.29}$$

where H_n and H_u are symmetric and positive definite. Accordingly, solving (6.29) is equivalent to the following minimization problem

$$\begin{aligned} \min_{(y,u)\in Y\times U} g(y,u) := \frac{1}{2} y^T H_y y + \frac{1}{2} u^T H_u u - b_y^T y - b_u^T u \\ s.t.\ Ay - Bu = 0. \end{aligned} \tag{6.30}$$

Assuming that A is positive definite, we define the solution operator

$$S := A^{-1}B.$$

Consequently,

$$y = Su,$$

and inserting it into (6.30) yields the unconstrained problem

$$\min_{u\in U} g(u) := \frac{1}{2} u^T S^T H_y S u + \frac{1}{2} u^T H_u u - b_y^T S u - b_u^T u.$$

A direction of descent of g can be computed via the minimization problem

$$\min_{\delta u\in U} \frac{1}{2} \delta u^T M_a \delta u + g'(u)\delta u,$$

for any positive definite preconditioner M_a. By inserting g', we obtain

$$\min_{\delta u\in U} \frac{1}{2} \delta u^T M_a \delta u + (u^T S^T H_y S + u^T H_u - b_y^T S - b_u^T)\delta u. \tag{6.31}$$

Thus, the solution δu of this minimization problem corresponds to the update of a gradient method with the preconditioner M_a. Analogously, we conduct this analysis for the CIPPCG method. In this setting, the computation of the inexact step via (6.25) is equivalent to

$$\begin{pmatrix} H_y & 0 & \tilde{A}^T \\ 0 & H_u & -B^T \\ \tilde{A} & -B & 0 \end{pmatrix} \begin{pmatrix} \delta\tilde{y} \\ \delta\tilde{u} \\ \delta\tilde{p} \end{pmatrix} = \begin{pmatrix} 0 \\ b_u - H_u u + B^T p \\ 0 \end{pmatrix} \tag{6.32}$$

and to the optimization problem

$$\begin{aligned} \min_{(\delta\tilde{y},\delta\tilde{u})\in Y\times U} \tilde{g}(\delta\tilde{y},\delta\tilde{u}) := \frac{1}{2}\delta\tilde{y}^T H_y \delta\tilde{y} + \frac{1}{2}\delta\tilde{u}^T H_u \delta\tilde{u} - ({b_u}^T - u^T H_u + p^T B)\delta\tilde{u} \\ s.t.\ \tilde{A}\delta\tilde{y} - B\delta\tilde{u} = 0. \end{aligned}$$

We define

$$\tilde{S} := \tilde{A}^{-1}B$$

and insert it into the above problem to obtain

$$\min_{\delta\tilde{u}\in U} \tilde{g}(\delta\tilde{u}) := \frac{1}{2}\delta\tilde{u}^T\tilde{S}^T H_y\tilde{S}\delta\tilde{u} + \frac{1}{2}\delta\tilde{u}^T H_u\delta\tilde{u} - ({b_u}^T - u^T H_u + p^T B)\delta\tilde{u}.$$

Combining this result with (6.24) leads to

$$\min_{\delta\tilde{u}\in U} \tilde{g}(\delta\tilde{u}) := \frac{1}{2}\delta\tilde{u}^T(\tilde{S}^T H_y\tilde{S} + H_u)\delta\tilde{u} - ({b_u}^T - u^T H_u + {b_y}^T S - u^T S^T H_y S)\delta\tilde{u}.$$

The comparison with (6.31) yields that $\delta\tilde{u}$ corresponds to a gradient step for the original function g w.r.t. the preconditioner $M_a := \tilde{S}^T H_y\tilde{S} + H_u$. Additionally, computing the projected direction δy via

$$A\delta y = B\delta\tilde{u}$$

corresponds to

$$\delta y = S\delta\tilde{u}.$$

As a result, computing a directional minimizer of g w.r.t. $\delta\tilde{u}$ via

$$\omega = -\frac{(u^T S^T H_y S + u^T H_u - b_y^T S - b_u^T)\delta\tilde{u}}{\delta\tilde{u}^T(S^T H_y S + H_u)\delta\tilde{u}}$$

is equivalent to the formulas for the optimal line search parameters in Algorithms 14 and 15. In summary, this shows that the CIPPCG algorithm is equivalent to a gradient method and has the same convergence properties.

6.3.4 Implementation and adjustments to nonlinear elasticity and inexactness

First, the implementation of the CIPPCG method is discussed.

Implementation

For the dual update (6.24) and the projection (6.26), we deploy a PCG method with a multigrid preconditioner Q_{bpx} of BPX-type. This preconditioner is combined with a block Jacobi smoother which utilizes the diagonal of 3×3 blocks of A, respecting the vector valued nature of the problem. This preconditioner Q_{bpx} will also be applied in the Chebyshev semi-iteration. Solving the dual update with a PCG method allows us to obtain an estimate for the eigenvalues of the preconditioned matrix $Q_{\text{bpx}}^{-1}A$ as discussed in Section 5.1. As a result, the linear Chebyshev semi-iteration, as defined in Algorithm 5, can be utilized to solve systems involving $\tilde{A}$ and $\tilde{A}^T$ in the application of the inexact preconditioner $\tilde{Q}$. The analysis in Subsection 6.3.3 only holds for solving (6.24) and (6.26) exactly. However, for the sake of efficiency, we do not impose high accuracies on the PCG methods involved, relying on a simple heuristic approach. As a

consequence, iterates are not guaranteed to be contained in $\ker C$. The starting accuracy for solving (6.24), (6.26), and (6.32) is chosen according to the required step accuracy Λ_{CS} in the composite step method as discussed in Subsection 5.2.9. For the Chebyshev semi-iteration, we apply a fixed accuracy Λ_{Cheb}.
A step is considered sufficiently accurate if the damping factor ω_k satisfies

$$0 < \omega_{\min} \leq \omega_k \leq \omega_{\max}$$

with user-provided parameters $\omega_{\min} < 1$ and $\omega_{\max} > 1$. This ensures that the final updates steps are not too small and do not decrease the convergence speed of the CIPPCG method too much. If this condition is not satisfied, we project the entire iterate back to $\ker C$ by applying a PCG method with high accuracy Λ_{CGMax} to solve

$$A\delta y_k = Bu_k - Ay_k. \tag{6.33}$$

Thereafter, only the state component is updated:

$$y_{k+1} \leftarrow y_k + \delta y_k.$$

During the computation of the subsequent iterates, the accuracies of (6.24), (6.26), and (6.32) are increased by the factor $\mu_{\mathrm{I}}^{\mathrm{CG}}$. Also, we increase the accuracy of the Chebyshev solver by $\mu_{\mathrm{I}}^{\mathrm{Cheb}}$. For high accuracies, (6.24), (6.26), and (6.32) are solved almost exactly. Accordingly, $\tilde{Q} \approx Q$. Therefore, ω_k should be close to 1, and fast convergence of the entire CIPPCG method can still be expected.

Remark 6.9. *Deriving an adaptive and more efficient way to determine the required accuracies and to show convergence for this inexact approach is beyond the scope of this work. Thus, it remains a subject of future research.*

Convergence criterion

The criterion derived here was proposed in [96]. At this, only the primal variable $x = (y, u)$ is taken into account. Let (x_*, p_*) denote the solution to (6.29) and (x_k, p_k) the current iterate of the CIPPCG method. For a given relative accuracy $\Lambda_{\mathrm{RelCIPPCG}}$, a desired convergence criterion would be

$$\|x_* - x_k\|_M \leq \Lambda_{\mathrm{RelCIPPCG}} \|x_*\|_M. \tag{6.34}$$

Recall that the matrix M corresponds to the Riesz isomorphism chosen for the composite step setting. Since the solution x_* is unknown, this condition cannot be evaluated exactly. Therefore, the goal here is to derive estimates for $\|x_*\|_M$ and $\|x_* - x_k\|_M$. Assume the contraction condition

$$\|x_* - x_k\|_M = \Theta^{k-m} \|x_* - x_m\|_M$$

holds for $0 < \Theta < 1$ and $m < k$. Then, we can derive the estimate

$$\|x_* - x_k\|_M \leq \Theta^{k-m} \left(\|x_* - x_k\|_M + \|x_k - x_m\|_M \right),$$

and thus,

$$\|x_* - x_k\|_M \leq \frac{\Theta^{k-m}}{1-\Theta^{k-m}}\|x_k - x_m\|_M.$$

Additionally, we obtain a lower bound of the norm of x_*:

$$\|x_*\|_M \geq \|x_k\|_M - \frac{\Theta^{k-m}}{1-\Theta^{k-m}}\|x_k - x_m\|_M.$$

For $k > m > i > 0$, a simple estimate for the contraction factor can be defined by

$$\tilde{\Theta} := \frac{\|x_k - x_m\|_M}{\|x_m - x_i\|_M}.$$

This yields the estimates

$$\|x_* - x_k\|_M \approx \frac{\tilde{\Theta}^{k-m}}{1-\tilde{\Theta}^{k-m}}\|x_k - x_m\|_M$$

and

$$\|x_*\|_M \approx \|x_k\|_M - \frac{\tilde{\Theta}^{k-m}}{1-\tilde{\Theta}^{k-m}}\|x_k - x_m\|_M,$$

which can be inserted into (6.34), leading to the approximated criterion

$$\frac{\tilde{\Theta}^{k-m}}{1-\tilde{\Theta}^{k-m}}\|x_k - x_m\|_M \leq \Lambda_{\text{RelCIPPCG}}\left(\|x_k\|_M - \frac{\tilde{\Theta}^{k-m}}{1-\tilde{\Theta}^{k-m}}\|x_k - x_m\|_M\right). \tag{6.35}$$

For the implementation, $m = k-1$ and $i = k-2$ were chosen. Consequently, at least two steps need to be computed in order to apply this convergence criterion. We also add the absolute acceptance condition

$$\|x_k - x_{k-1}\|_M \leq \Lambda_{\text{AbsCIPPCG}}, \tag{6.36}$$

for suitable $\Lambda_{\text{AbsCIPPCG}} > 0$. Moreover, $\Lambda_{\text{RelCIPPCG}}$ is set according to the step accuracy Λ_{CS} required by the composite step method as discussed in Subsection 5.2.9. The implementation is summarized in Algorithm 16 for the general setting described by (6.29). Note that applying this algorithm to compute the tangential step may require some regularization as described in Algorithm 15 for the regularization parameters s_{tIC} and λ_{tICInit}. For the sake of readability, we assume that the respective matrices are already sufficiently regularized and positive definite.

Non-convexity of the total energy functional

For nonlinear elastic problems, we know that the total energy functional I can be non-convex. Consequently, this also holds for I_γ and $\mathcal{E}_\gamma$. Therefore, at a given composite step iterate (y_k, u_k, p_{ck}), the operator $A := \hat{A}'(y_k)$ is not necessarily positive definite. During the computation of the composite step update $(\delta y_k, \delta u_k, \delta p_{ck}, \delta s_k)$, the non-convexity

Algorithm 16 Implementation of the Corrected Inexact Projected Preconditioned Conjugate Gradient Method.

Solve: $H \begin{pmatrix} y \\ u \\ p \end{pmatrix} = \begin{pmatrix} b_y \\ b_u \\ 0 \end{pmatrix}$.

Input: initial iterate $z_0 = \begin{pmatrix} y_0 \\ u_0 \\ p_0 \end{pmatrix}$ with $Ay_0 - Bu_0 = 0$, parameters: Λ_{CS} (accuracy required by the composite step method), $\Lambda_{\mathrm{AbsCIPPCG}}$, $\Lambda_{\mathrm{RelCIPPCG}} = \Lambda_{\mathrm{CS}}$, Λ_{CGMax}, Λ_{Cheb}, $\mu_{\mathrm{I}}^{\mathrm{CG}}$, $\mu_{\mathrm{I}}^{\mathrm{Cheb}}$, $\omega_{\min}$, and $\omega_{\max}$.

Initialize: $k = 0$.

repeat

Solve: $A^T \Delta p_k = b_y - M_y y_k - A^T p_k$ (Algorithm 2 with accuracy Λ_{CS})

$p_{k+1} \leftarrow p_k + \Delta p_k$

Solve: $\begin{pmatrix} H_y & 0 & \tilde{A}^T \\ 0 & H_u & -B^T \\ \tilde{A} & -B & 0 \end{pmatrix} \begin{pmatrix} \delta\tilde{y}_k \\ \delta\tilde{u}_k \\ \delta\tilde{p}_k \end{pmatrix} = \begin{pmatrix} 0 \\ b_u - H_u u_k + B^T p_{k+1} \\ 0 \end{pmatrix} \begin{pmatrix} \text{Algorithm 12} \\ \text{or 13 with } \Lambda_{\mathrm{CS}} \end{pmatrix}$

Solve: $A\delta y_k = B\delta\tilde{u}_k$ (Algorithm 2 with accuracy Λ_{CS})

$$\omega_k \leftarrow -\frac{\left(\begin{pmatrix} H_y & 0 \\ 0 & H_u \end{pmatrix}\begin{pmatrix} y_k \\ u_k \end{pmatrix} - \begin{pmatrix} b_y \\ b_u \end{pmatrix}\right)^T \begin{pmatrix} \delta y_k \\ \delta\tilde{u}_k \end{pmatrix}}{\begin{pmatrix} \delta y_k{}^T & \delta\tilde{u}_k{}^T \end{pmatrix}\begin{pmatrix} H_y & 0 \\ 0 & H_u \end{pmatrix}\begin{pmatrix} \delta y_k \\ \delta\tilde{u}_k \end{pmatrix}}$$

if $0 < \omega_{\min} \leq \omega_k \leq \omega_{\max}$ **then**

$y_{k+1} \leftarrow y_k + \omega_k \delta y_k$

$u_{k+1} \leftarrow u_k + \omega_k \delta\tilde{u}_k$

else

Solve: $A\delta y_k = Bu_k - Ay_k$ (Algorithm 2 with accuracy Λ_{CGMax})

$\Lambda_{\mathrm{CS}} \leftarrow \Lambda_{\mathrm{CS}} \mu_{\mathrm{I}}^{\mathrm{CG}}$

$\Lambda_{\mathrm{Cheb}} \leftarrow \Lambda_{\mathrm{Cheb}} \mu_{\mathrm{I}}^{\mathrm{Cheb}}$

$y_{k+1} \leftarrow y_k + \delta y_k$

$u_{k+1} \leftarrow u_k$

end if

$k \leftarrow k + 1$

until convergent (If Condition (6.35) or (6.36) is satisfied.)

can be detected by the CG methods applied to solve (6.24), (6.26), and (6.33). In that case, we choose a regularization factor $\lambda > 0$ and replace $\mathcal{E}_\gamma$ in (4.12) with

$$\hat{\mathcal{E}}_\gamma(y,u) := \mathcal{E}_\gamma(y,u) + \frac{\lambda}{2} q(y - y_k),$$

where q is a quadratic and positive definite energy. Note that the analysis for I_γ follows analogously. In this work, we choose

$$q(v) = \langle v, v \rangle_{\partial_{yy} I_{\text{strain}}(\text{id})},$$

which corresponds to the energy induced by linear elasticity. For this modification, we obtain

$$\partial_y \hat{\mathcal{E}}_\gamma(y,u) = \partial_y \mathcal{E}_\gamma(y,u) + \lambda q'(y - y_k)$$

such that

$$\partial_y \hat{\mathcal{E}}_\gamma(y_k,u_k) = \partial_y \mathcal{E}_\gamma(y_k,u_k).$$

Further, A is replaced by the regularized operator:

$$A_r := \partial^2_{yy} \hat{\mathcal{E}}_\gamma(y_k,u_k) = \partial^2_{yy} \mathcal{E}_\gamma(y_k,u_k) + \lambda q''(0).$$

This regularization has three main effects. First of all, for sufficiently large λ, A_r is positive definite. Thus, a PCG method can be applied. Due to the positive definiteness of A_r, normal steps are shifted towards descent of $\mathcal{E}_\gamma(y,u)$ at (y_k,u_k). This can be shown by considering the linearized constraint imposed on the normal step δn:

$$A_r \delta n_y - B \delta n_u + \partial_y \hat{\mathcal{E}}_\gamma(y_k,u_k) = 0.$$

In addition, $\mathcal{E}_\gamma(y,u)$ is linear in u, which yields

$$\partial_y \mathcal{E}_\gamma(y_k, u_k + \delta n_u)\delta n_y = (\partial_y \hat{\mathcal{E}}_\gamma(y_k,u_k) - B\delta n_u)\delta n_y = -(A_r \delta n_y)\delta n_y < 0.$$

Last, the regularization penalizes long steps, resulting in a more stable behavior of the optimization algorithm. This corresponds to the physical interpretation that the linear elastic regularization term adds an artificial stiffness to the material. To ensure the positive definiteness of A_r, we us the update formula of Algorithm 3 with the parameters λ_{EInit} and s_{E} until no directions of negative curvature are encountered anymore. After each regularization, all updates $(\delta y_k, \delta u_k, \delta p_{c_k}, \delta s_k)$ of the composite step method is recomputed with the new regularized energy functional $\hat{\mathcal{E}}_\gamma$. Moreover, when an update has been successfully computed, λ is set to zero again. This approach is summarized in Algorithm 17.

Remark 6.10. *An alternative method to regularization was also considered. If non-convexity of the energy was encountered at a composite step iterate (y_k,u_k), the control u_k was kept fixed, and Algorithm 1 was applied to compute a minimizer $\tilde{y}_k$ of $\mathcal{E}_\gamma(\,\cdot\,,u_k)$ or $I_\gamma(\,\cdot\,,u_k)$. Then, y_k was replaced by $\tilde{y}_k$ to ensure convexity of the energy. This heuristic scheme is no longer consistent with the composite step method elaborated in Section 5.2. So far, a rather unstable and erratic behavior has been observed, ruling out the application of this method. One possible explanation might be buckling which can occur during such an algorithm.*

Algorithm 17 Composite Step Method with Energy Regularization

Input: initial iterate (y_0, u_0, p_{c0}), regularization update parameter s_{E}, starting regularization parameter λ_{EInit}, initial Lipschitz constants $\tilde{\omega}_{\mathrm{C}_0}$, $\tilde{\omega}_{\mathrm{f}_0}$, parameters $\eta_{\min}$, $\overline{\eta}$, ϱ, accuracies Λ_{CSAb}, $\Lambda_{\mathrm{CSNorm}}$, and Λ_{CSRel}, and contraction factors Θ_{acc}, Θ_{n}, and Θ_{d}.
Initialize: $k = 0$ and $\lambda = 0$.
repeat
  repeat
    $\begin{pmatrix} y_{k+1} \\ u_{k+1} \\ p_{ck+1} \end{pmatrix} \leftarrow computeCompositeStep(y_k, u_k, p_{ck}, \lambda)$
    if $A + \lambda q''(0)$ is non-convex **then**
      $\lambda \leftarrow updateLambda(\lambda, \lambda_{\mathrm{EInit}}, s_{\mathrm{E}})$ (Algorithm 3)
    end if
  until $A + \lambda q''(0)$ is convex
  $\lambda \leftarrow 0$
  $k \leftarrow k + 1$
until convergent (If Condition (5.24) or (5.25) is fulfilled.)

Chapter 7

Numerical Examples

In the previous chapter, we have introduced three types of algorithms to solve problems involving nonlinear elasticity. These are a cubic regularization approach, a composite step method, and a path-following scheme. Here, we will test each algorithm by means of numerical examples. Before doing so, the general setting is defined. First, regarding nonlinear elasticity, a compressible Mooney-Rivlin model is chosen for the stored energy function:

$$\hat{W}(\nabla y) := a\|\nabla y\|^2 + b\|\operatorname{Cof}\nabla y\|^2 + c(\det\nabla y)^2 - d\ln\det\nabla y,$$

with the parameters:

$$a = 0.08625, \quad b = 0.08625, \quad c = 0.68875, \quad d = 1.895.$$

This choice describes a model for soft tissue, cf. [64, Chapter 6]. The derivation of these parameters for general cases is discussed in [15, Chapter 3-4]. In the optimal control setting, a tracking type functional defined by

$$J(y,u) := \frac{1}{2}\|y - y_{\mathrm{d}}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Gamma_N)}$$

serves as the objective functional. Hereby, we aim to approximate a reference deformation $y_{\mathrm{d}} \in L^2(\Omega)$ for a positive regularization parameter $\alpha > 0$. The function spaces are defined by

$$Y \times U := H^1(\Omega) \times L^2(\Gamma_N).$$

If not stated otherwise, the corresponding scalar product is set to

$$\langle (y,u),(y,u)\rangle := \frac{1}{2}\langle y,y\rangle_{H^1(\Omega)} + \frac{\alpha}{2}\langle u,u\rangle_{L^2(\Gamma_N)}.$$

Two geometries are considered here. The first one is described by a discretized plate $\overline{\Omega}_1 = [0,2]\times[0,2]\times[0,0.2]$. The respective starting grid is illustrated in Figure 7.1. On the side faces, homogeneous Dirichlet boundary conditions are enforced. Moreover, we choose the bottom face to be the Neumann boundary Γ_N, while the top face is the

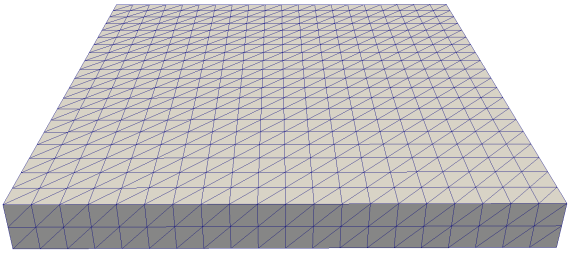

Figure 7.1: Initial grid for Problem 1.

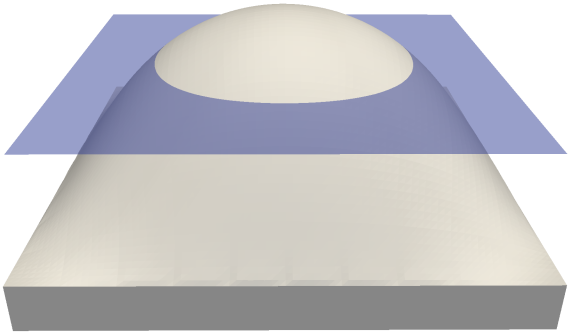

Figure 7.2: Desired deformation for Problem 1 with obstacle (transparent).

contact boundary Γ_C. Figure 7.2 shows the desired deformation y_d with the obstacle for three refinements. The desired deformations are precomputed for each refinement.
The second problem addresses a discretized cantilever $\overline{\Omega}_2 = [0,1] \times [0,0.5] \times [0,0.2]$, which is displayed in Figure 7.3. The back face is fixed at a wall, describing homogeneous Dirichlet boundary conditions. On the top face, the boundary force u is applied. Thus, it acts as the Neumann boundary Γ_N. The remaining faces function as the contact boundary Γ_C. Figure 7.4 illustrates the obstacle and the precomputed desired deformation y_d for three refinements. We expect the second problem to be much more challenging since the boundary conditions of the cantilever are less restrictive, allowing a wider range of deformations.
In both cases, we apply linear finite elements to obtain discretized variables. Here, only uniform refinements are considered, and the corresponding degrees of freedoms are displayed in Tables 7.1 and 7.2.

	1 Ref.	2 Ref.	3 Ref.	4 Ref.
Degrees of freedom y	6369	44415	330747	2550771
Degrees of freedom u	1323	5043	19683	77763
Degrees of freedom p	6369	44415	330747	2550771

Table 7.1: Degrees of freedom for Problem 1.

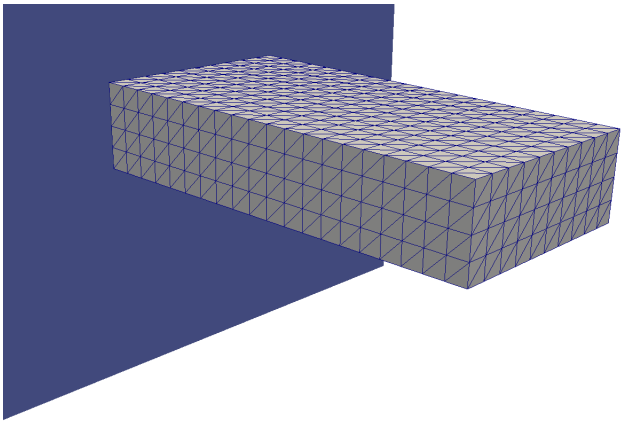

Figure 7.3: Initial grid for Problem 2.

	1 Ref.	2 Ref.	3 Ref.	4 Ref.
Degrees of freedom y	5865	42447	322971	2519859
Degrees of freedom u	693	2583	9963	39123
Degrees of freedom p	5865	42447	322971	2519859

Table 7.2: Degrees of freedom for Problem 2.

For the implementation, the finite element library KASKADE7 [32] was used. This library is based on the DUNE library [10] and has been developed at the Zuse Institute Berlin. Sparse linear systems that require direct solvers were solved by utilizing the software package UMFPACK [22]. Computing eigenvalues was conducted via the dstevr function from the software package LAPACK [3]. The C++ library Eigen, cf. [38], provided the matrix exponential function for the nonlinear update. Further, all optimization and path-following algorithms were implemented in the C++ library Spacy[1], which was developed for numerical optimization in a vector space setting. Regarding nonlinear elasticity, the library FunG [65] was applied to compute the derivatives of the stored energy function via automatic differentiation. All computation were carried out on a compute server with an Intel Xeon E7-2830 2.13GHz processor and 1TB RAM.
This chapter is structured as follows. In the first section, the cubic regularization approach (Algorithm 1) is applied to test the convergence rates derived in Chapters 3 and 4. Also, the performance of the nonlinear update (Algorithm 6) is analyzed in Section 7.2. Section 7.3 is dedicated to numerical examples for optimal control of nonlinear elasticity, testing the composite step method (Algorithm 17) combined with the CIPPCG algorithm (Algorithm 16). Finally, this chapter is concluded with numerical examples of path-following in Section 7.4. An analysis similar to the one presented in Subsection 7.3.2 has already been published in [94].

[1] https://spacy-dev.github.io/Spacy/

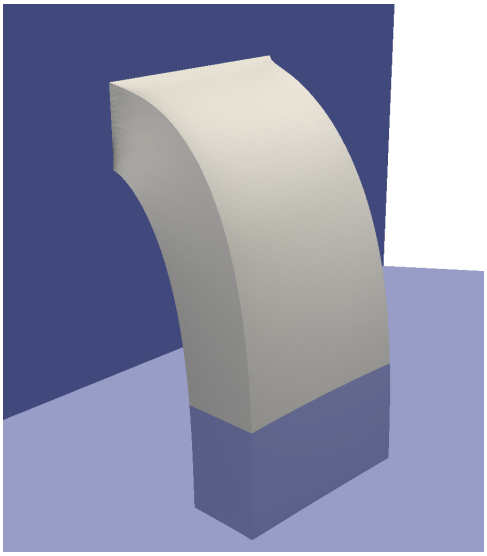

Figure 7.4: Desired deformation for Problem 2 with obstacle (transparent).

7.1 Numerical estimates for the convergence rates

Consider a monotonically increasing sequence of penalty parameters $\gamma_n \to \infty$ and the sequence of corresponding solutions y_n satisfying

$$y_n \in \operatorname*{argmin}_{v \in \mathcal{A}} I_{\gamma_n}(v, u). \tag{7.1}$$

We define

$$\Delta I(y_{n+1}, u) := I(y_{n+1}, u) - I(y_n, u)$$

and

$$\Delta I_{\gamma_{n+1}}(y_{n+1}, u) := I_{\gamma_{n+1}}(y_{n+1}, u) - I_{\gamma_n}(y_n, u).$$

As γ_n approaches infinity, the terms $\|[y_n]_+\|_{L^\infty(\Gamma_C)}$, $\Delta I(y_n, u)$, and $\Delta I_{\gamma_n}(y_n, u)$ should approach zero at a certain rate ρ^{e} as shown in Theorem 3.19. Analogously to the examples in [47], we can compute an estimate for ρ^{e} by

$$\tilde{\rho}_n^{\mathrm{e}} := \frac{\ln(\|[y_n]_+\|_{L^\infty(\Gamma_C)}) - \ln(\|[y_{n+1}]_+\|_{L^\infty(\Gamma_C)})}{\ln(\gamma_{n+1}) - \ln(\gamma_n)}. \tag{7.2}$$

The convergence rate estimates for the terms $\Delta I(y_n, u)$ and $\Delta I_{\gamma_n}(y_n, u)$ are defined analogously. Let

$$G := \{\gamma_1, \ldots, \gamma_l\}$$

be a set of strictly increasing and sufficiently large penalty parameters. Then, Algorithm 1 can be applied to obtain the corresponding solutions to Problem (7.1), yielding the set

$$L := \{y_1, \ldots, y_l\}.$$

These solutions allow the computation of estimates for ρ^{e} via (7.2). In the same way, we perform numerical experiments to test the results of Theorem 4.9. All computations for Problems 1 and 2 are conducted with three uniform refinements. The required parameters for Algorithm 1 and the applied boundary forces are summarized in Table 7.3. The boundary forces are chosen to be constant functions on Γ_N. As the initial iterate for Algorithm 1, we choose the identity mapping id.

	ω_0^{CR}	η_1^{CR}	η_2^{CR}	$s_{\mathrm{I}}^{\mathrm{CR}}$	$s_{\mathrm{D}}^{\mathrm{CR}}$	u	λ_{AInit}	s_{A}	Λ_{CR}
Problem 1	5	0.25	0.5	1.2	0.5	$(0, 0, 0.1)$	1e-8	5	1e-8
Problem 2	5	0.25	0.5	1.2	0.5	$(0, 0, -0.022)$	1e-8	5	1e-8

Table 7.3: Parameters for the cubic regularization approach (Algorithm 1).

7.1.1 Normal compliance method

Since the rates derived in Theorem 3.19 hinge on the exponent k in the penalty function

$$P(v) := \frac{1}{k} \int_{\Gamma_C} [v]_+^k \, ds, \quad k \in \mathbb{N},\ k > 1, \quad v \in Y,$$

any reasonable examination requires several tests for different exponents. Here, we choose $k = 2, 3, 4$. Recalling the setting of Theorem 3.19, $W^{1,p}(\Omega)$ is embedded into the space $C^{\beta}(\Omega)$ for $p > 3$ and $\beta \in]0.1[$. In the case $\gamma \to \infty$ and $\beta \to 1$, the estimates for ρ^{e} should approach the rate

$$\frac{1}{(k-1)+2}.$$

The resulting estimates for the convergence rates are depicted in Figures 7.5 to 7.10. For all results, we observe that the terms $\|[y]_+\|_{L^\infty(\Gamma_C)}$, $\Delta I(y, u)$, and $\Delta I_\gamma(y, u)$ exhibit the same asymptotic convergence rate, which is consistent with Theorem 3.19. Regarding Problem 2 (Figures 7.8 to 7.10), it takes larger parameters γ for the estimates to converge. This effect may be attributed to the increased difficulty of the problem. Additionally, the results indicate a faster rate

$$\rho^{\mathrm{e}} = \frac{1}{k-1},$$

which is significantly better than the rate predicted by Theorem 3.19. In summary, we can deduce that the theoretical estimates are not necessarily sharp, and further improvements may be possible.

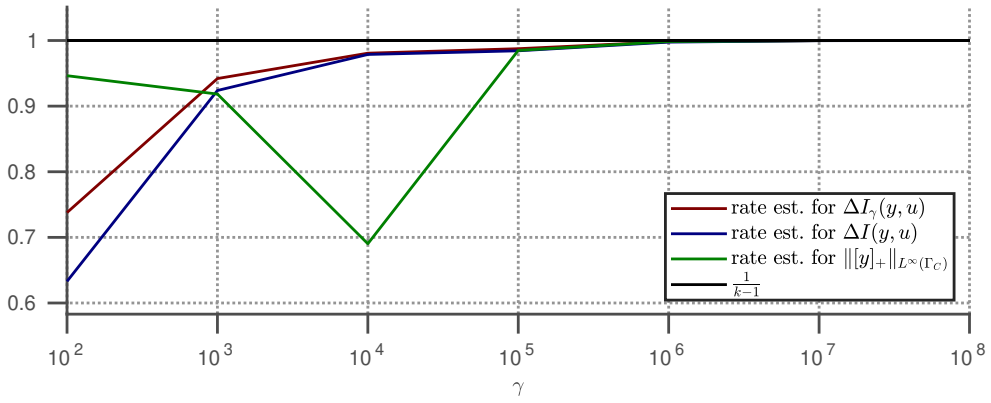

Figure 7.5: Estimated convergence rates for Problem 1 with $k = 2$.

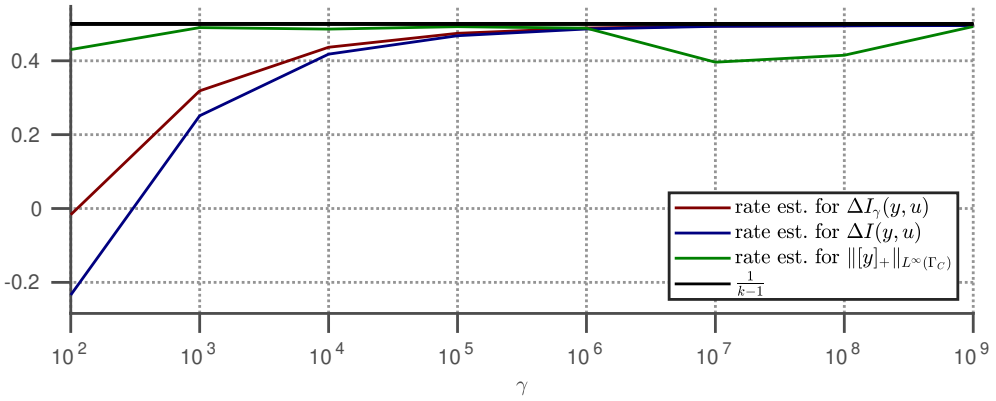

Figure 7.6: Estimated convergence rates for Problem 1 with $k = 3$.

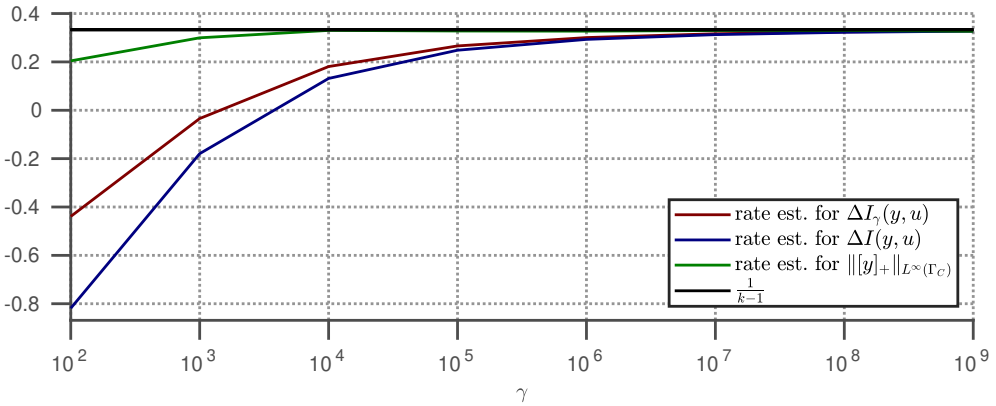

Figure 7.7: Estimated convergence rates for Problem 1 with $k = 4$.

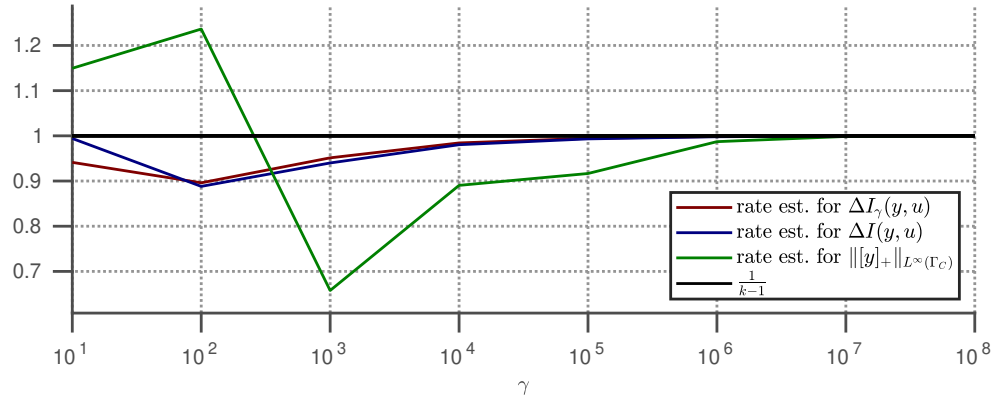

Figure 7.8: Estimated convergence rates for Problem 2 with $k = 2$.

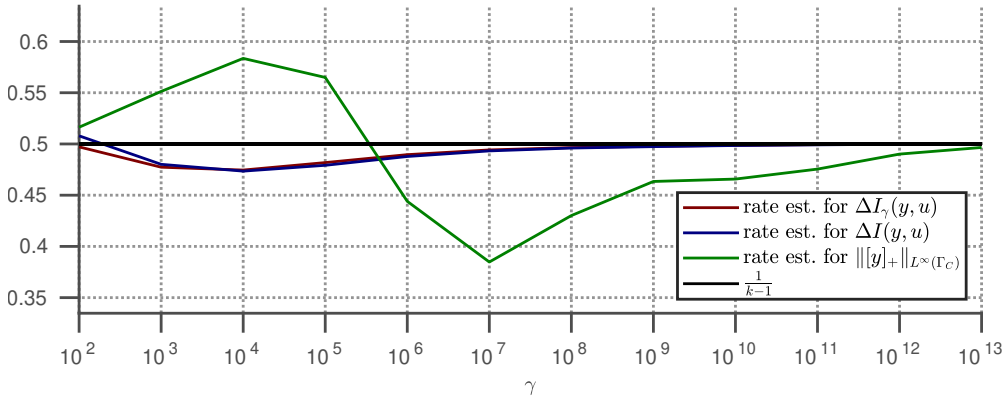

Figure 7.9: Estimated convergence rates for Problem 2 with $k = 3$.

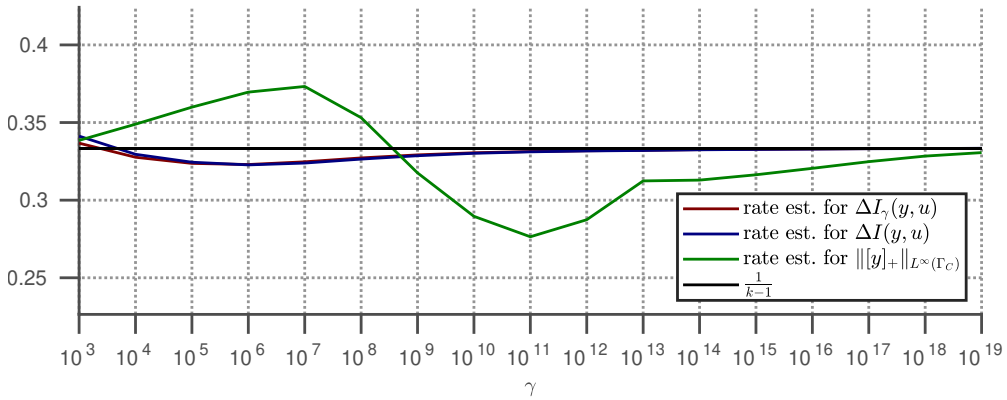

Figure 7.10: Estimated convergence rates for Problem 2 with $k = 4$.

7.1.2 Modified normal compliance regularization

In contrast to before, Theorem 4.9 only guarantees a convergence rate $\rho^{\mathcal{E}}$ for the modified functional

$$\mathcal{E}_\gamma(y,u) := I_\gamma(y,u) + \varphi(\gamma)\frac{1}{2}\|y - y_\mathrm{d}\|^2_{L^2(\Omega)}.$$

Again, we define

$$\Delta\mathcal{E}_{\gamma_{n+1}}(y_{n+1},u) := \mathcal{E}_{\gamma_{n+1}}(y_{n+1},u) - \mathcal{E}_{\gamma_n}(y_n,u).$$

Here, y_n denotes the sequence of minimizers of $\mathcal{E}_{\gamma_n}(\,\cdot\,,u)$ in $\mathcal{A}$. In particular, the derived rate in Theorem 4.9 only holds for the absolute value

$$|\Delta\mathcal{E}_{\gamma_{n+1}}(y_{n+1},u)|.$$

As γ approaches infinity, the function φ approaches zero at a user-provided rate ρ^φ, which leaves us with two parameters to choose. The parameters tested here are summarized in Table 7.4.

	$(k,\,\rho^\varphi)_1$	$(k,\,\rho^\varphi)_2$	$\varphi(\gamma)$
Problem 1	$(3,\frac{1}{4})$	$(4,\frac{1}{5})$	$\gamma^{-\rho^\varphi}$
Problem 2	$(2,\frac{1}{3})$	$(3,\frac{1}{5})$	$\gamma^{-\rho^\varphi}$

Table 7.4: Parameters for the modified regularization.

Analogously to (7.2), we compute estimates for $\rho^{\mathcal{E}}$. Theorem 4.9 states that $\rho^{\mathcal{E}}$ and ρ^φ coincide. Therefore, we expect the estimates for $\rho^{\mathcal{E}}$ to approach ρ^φ for sufficiently large γ. In the same way, convergence rates for

$$\Delta I(y_{n+1},u) := I(y_{n+1},u) - I(y_n,u)$$

and the maximum constraint violation $\|[y_n]_+\|_{L^\infty(\Gamma_C)}$ are investigated. The results are visualized in Figures 7.11 to 7.14.
The computed estimates support the rate derived in Theorem 4.9. They even indicate convergence rates for the terms $\|[y]_+\|_{L^\infty(\Gamma_C)}$ and $|\Delta I(y,u)|$, which are not backed up by theoretical results so far, encouraging further examinations. For the maximum constraint violation $\|[y]_+\|_{L^\infty(\Gamma_C)}$, we observe the rate $\frac{1}{k-1}$, coinciding with the results from the previous subsection.

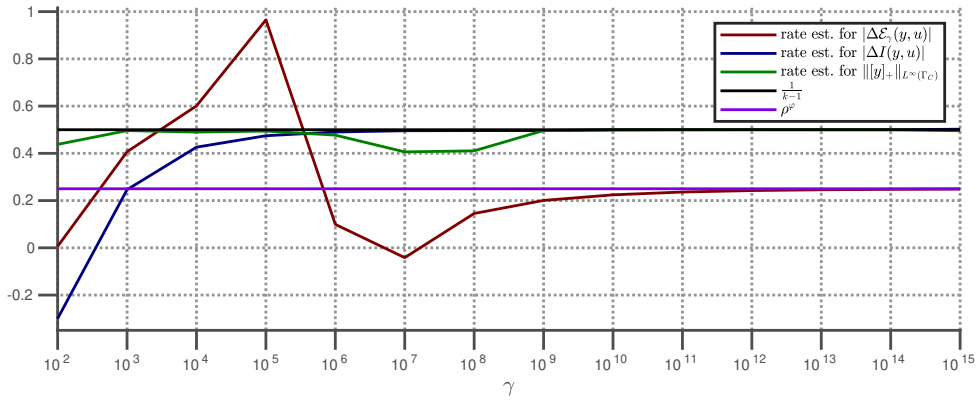

Figure 7.11: Estimated convergence rates for Problem 1 with $k = 3$ and $\rho^\varphi = \frac{1}{4}$.

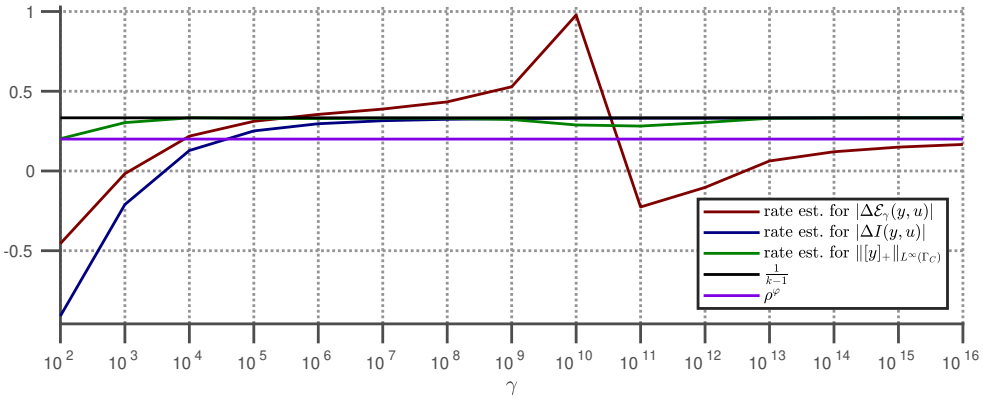

Figure 7.12: Estimated convergence rates for Problem 1 with $k = 4$ and $\rho^\varphi = \frac{1}{5}$.

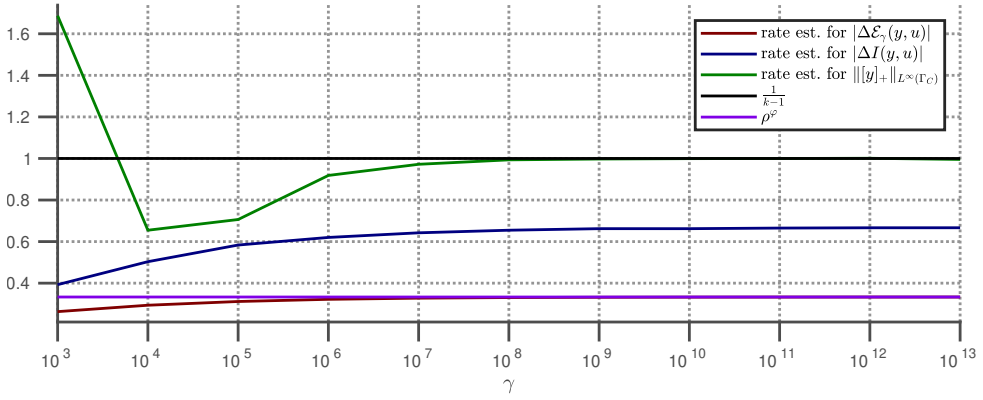

Figure 7.13: Estimated convergence rates for Problem 2 with $k = 2$ and $\rho^\varphi = \frac{1}{3}$.

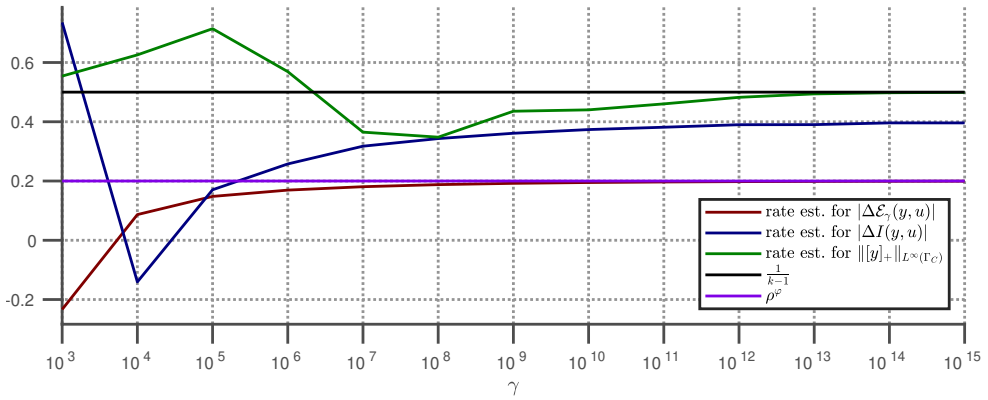

Figure 7.14: Estimated convergence rates for Problem 2 with $k = 3$ and $\rho^\varphi = \frac{1}{5}$.

7.2 Nonlinear updates

Here, Algorithms 1 and 7 are applied to the regularized problem

$$y \in \operatorname*{argmin}_{v \in \mathcal{A}} I_\gamma(v, u),$$

and their performances are compared. Related results are presented in the PhD thesis of Julián Ortiz, which is in preparation. The applied boundary forces and algorithmic parameters are shown in Tables 7.5 and 7.6. All computations are carried out for three uniform refinements.

	ω_0^{CR}	η_1^{CR}	η_2^{CR}	$s_{\mathrm{I}}^{\mathrm{CR}}$	$s_{\mathrm{D}}^{\mathrm{CR}}$	u	λ_{AInit}	s_{A}	k	Λ_{CR}
Prob. 1	5	0.25	0.5	1.2	0.5	$(0, 0, 0.1)$	1e-8	5	3	1e-8
Prob. 2	5	0.25	0.5	1.2	0.5	$(0, 0, -0.01)$	1e-8	5	3	1e-8

Table 7.5: Parameters for Algorithms 1 and 7 with $\gamma = 0$.

	ω_0^{CR}	η_1^{CR}	η_2^{CR}	$s_{\mathrm{I}}^{\mathrm{CR}}$	$s_{\mathrm{D}}^{\mathrm{CR}}$	u	λ_{AInit}	s_{A}	k	Λ_{CR}
Prob. 1	5	0.25	0.5	1.2	0.5	$(0, 0, 0.1)$	1e-8	5	3	1e-8
Prob. 2	5	0.25	0.5	1.2	0.5	$(0, 0, -0.015)$	1e-8	5	3	1e-8

Table 7.6: Parameters for Algorithms 1 and 7 with $\gamma = 100$.

Further, at an iterate y_k, we define the function value decrease

$$\Delta I_\gamma(y_k, u) := I_\gamma(y_{k-1}, u) - I_\gamma(y_k, u).$$

The required iterations and the corresponding function value decreases of the two algorithms are compared in Figures 7.15 to 7.18.

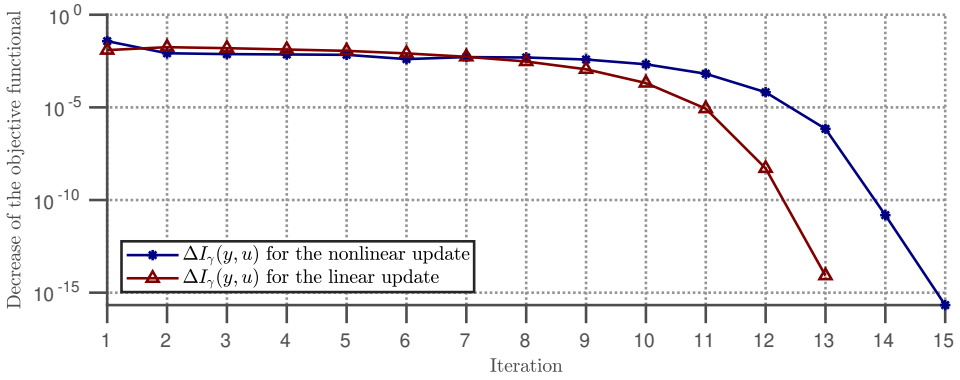

Figure 7.15: Function value decrease for Problem 1 with $\gamma = 0$.

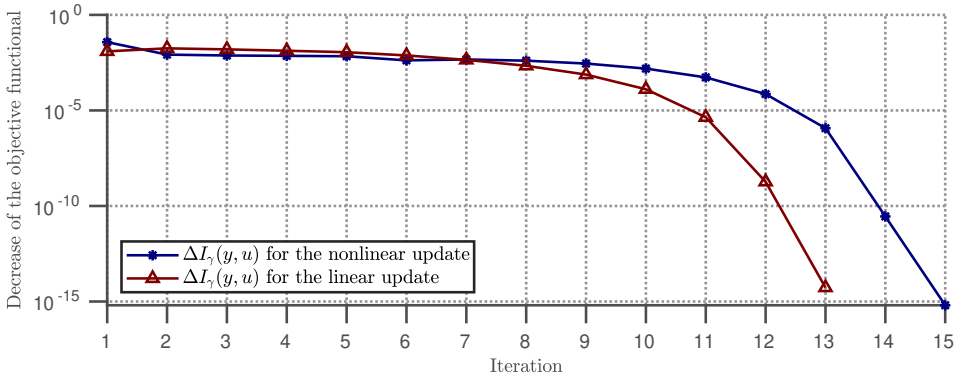

Figure 7.16: Function value decrease for Problem 1 with $\gamma = 100$.

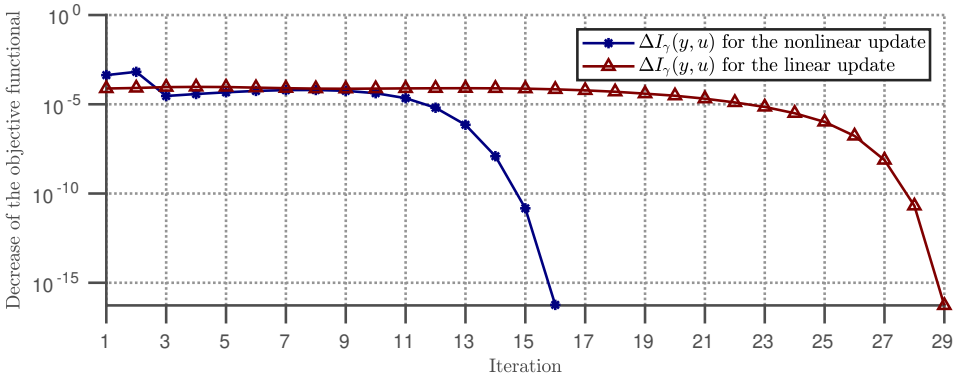

Figure 7.17: Function value decrease for Problem 2 with $\gamma = 0$.

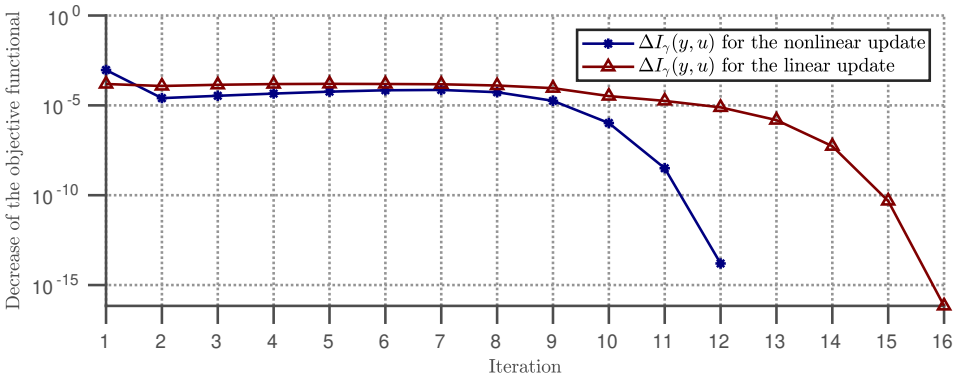

Figure 7.18: Function value decrease for Problem 2 with $\gamma = 100$.

For Problem 2 (Figures 7.17 and 7.18), the nonlinear update strategy significantly outperforms the standard approach. Although, this performance boost is smaller when γ is not zero. Figure 7.19 visualizes the difference between the two methods for the case $\gamma = 0$. There, we compare the deformation of the nonlinear update strategy and the linear one in the second iteration of Algorithms 1 and 7. We see that the nonlinear update yields larger deformations, and thus, faster convergence. In particular, rotational deformations are better approximated. Figure 7.20 illustrates the final deformation, which is the same for both approaches.

The performance increase cannot be observed for the first problem setting (Figures 7.15 and 7.16), where the nonlinear update strategy is even slightly slower. Additionally, the obstacle has barely an influence on the behavior of the algorithm. It seems that the nonlinear strategy only pays off in challenging settings where complex deformations such as rotations occur.

In summary, we conclude that taking into account the underlying structure of the problem can significantly improve the performance of the algorithm. The nonlinear updates, as presented here, yield promising results in the field of nonlinear elasticity. However, it remains the subject of future research to incorporate those updates into an optimal control setting.

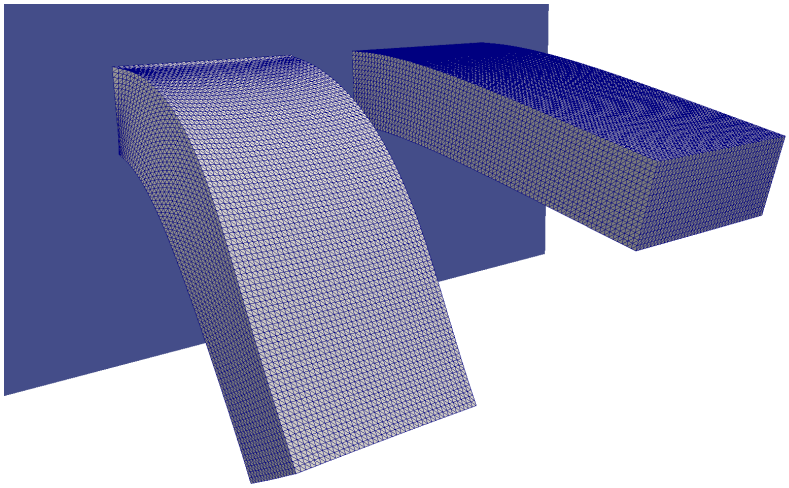

Figure 7.19: Left: second iterate of the nonlinear update strategy. Right: second iterate of the linear update strategy.

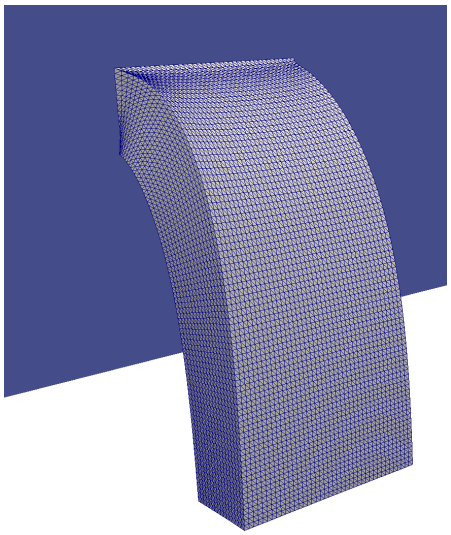

Figure 7.20: Final deformation for the linear and nonlinear update strategy.

7.3 Optimal Control

In this section, we will test the composite step method (Algorithm 17), whereby the arising linear systems are solved by Algorithm 16. As stated in the introduction of this chapter, we only consider the modified regularized optimal control problem

$$\min_{(y,u)\in Y\times U} J(y,u)$$
$$s.t.\ c_\gamma(y,u)=0,$$

where the energy minimization has been replaced by its formal first order optimality condition in the lower level problem. This is necessary to apply the composite step method. All algorithmic parameters are summarized in Tables 7.7 to 7.9.

	Λ_{CGMax}	Λ_{Cheb}	$\mu_{\mathrm{I}}^{\mathrm{CG}}$	$\mu_{\mathrm{I}}^{\mathrm{Cheb}}$	$\omega_{\min}$	$\omega_{\max}$	s_{tIC}	$\lambda_{\mathrm{tICInit}}$	$\Lambda_{\mathrm{AbsCIPPCG}}$
P. 1	1e-12	1e-2	0.1	0.25	0.1	10	100	1e-8	1e-12
P. 2	1e-12	1e-2	0.1	0.25	0.1	10	100	1e-8	1e-12

Table 7.7: Parameters for the CIPPCG method.

	Θ_{n}	Θ_{d}	Θ_{acc}	s_{E}	λ_{EInit}	Λ_{CSAb}	Λ_{CSRel}	$\Lambda_{\mathrm{CSNorm}}$
Problem 1	0.5	0.6	0.75	5	1e-8	1e-12	1e-6	1e-1
Problem 2	0.5	0.6	0.75	5	1e-8	1e-12	1e-6	1e-1

Table 7.8: Parameters for the composite step method.

	$\tilde{\omega}_{\mathrm{C}_0}$	$\tilde{\omega}_{\mathrm{f}_0}$	$\eta_{\min}$	$\overline{\eta}$	ϱ
Problem 1	1e-6	1e-6	1e-2	5e-2	0.25
Problem 2	1e-6	1e-6	1e-2	5e-2	0.25

Table 7.9: Parameters for the composite step method.

7.3.1 Performance of the CIPPCG method

First, we test the general performance of the composite step method combined with the CIPPCG algorithm. Therefore, we solve both problems with increasing refinements and compare the required computation time. Here, γ is set to zero. This simplifies the problems and allows more mesh refinements. The case $\gamma > 0$ will be discussed later. Further, we choose the regularization parameters $\alpha =$ 1e-1 for Problem 1 and $\alpha = 5$ for Problem 2. Table 7.10 shows the required computation time in hours, rounded to the second decimal place. For Problem 1, each refinement yields an increase of the computation time by about factor 10. The degrees of freedom are only increased by about

	1 Ref.	2 Ref.	3 Ref.	4 Ref.
Problem 1	0.11	1.00	9.95	97.39
Problem 2	0.24	2.92	62.56	204.57

Table 7.10: Computation time in hours.

factor 7.8. In absence of any grid-dependent behavior, the computation time would rise linearly with the degrees of freedom. However, we deploy a direct solver in the preconditioner, which does not yield linear scaling of the computation time w.r.t. the degrees of freedom. Also, in the case of nonlinear elasticity, grid-independent behavior cannot be guaranteed generally, as discussed in the following subsection. Still, the increase is within reasonable bounds, allowing us to solve large scale problems.
The second problem exhibits a more grid-dependent behavior since the increase in computation time cannot be described by a constant factor. In particular, the case of three refinements yields the largest increase by about factor 20. This issue will be discussed in detail when we study the two problems individually.
Next, the behavior of the composite step method is addressed. For the sake of brevity, we only study the results for three and four refinements since the larger problems are more relevant due to their increased accuracy.

Problem 1

The results for Problem 1 are depicted in Figures 7.21 to 7.32, where

$$\Delta J(x_k) := J(x_k) - J(x_{k-1}).$$

First, we notice that convergence was achieved within nine iterations in both cases, whereby no non-convexities are encountered. The damping factors ν and τ approach the value one very quickly (Figures 7.21 and 7.22), which suggests that the problem is not too difficult. Further, Figures 7.23 to 7.26 display that the norm of the updates and $|\Delta J(x_k)|$ approach zero. However, it seems that the region of superlinear convergence has not been reached yet. The number of required CIPPCG and IPPCG iterations has the same order of magnitude in both problems (Figures 7.27 to 7.30), indicating a grid-independent behavior of these two algorithms.
Finally, Figure 7.31 shows that the desired deformation is closely approximated. The optimal control is visualized in Figure 7.32.

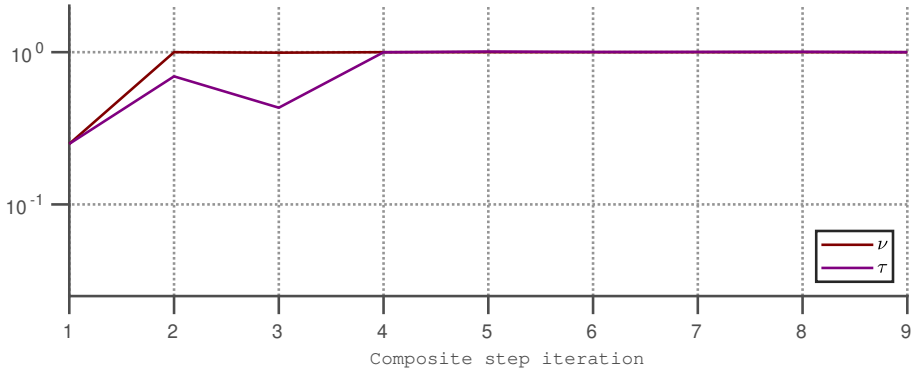

Figure 7.21: Damping factors for Problem 1 with three refinements.

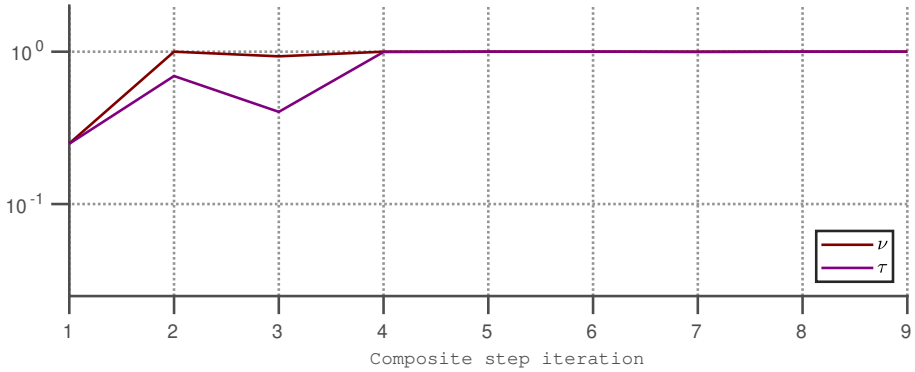

Figure 7.22: Damping factors for Problem 1 with four refinements.

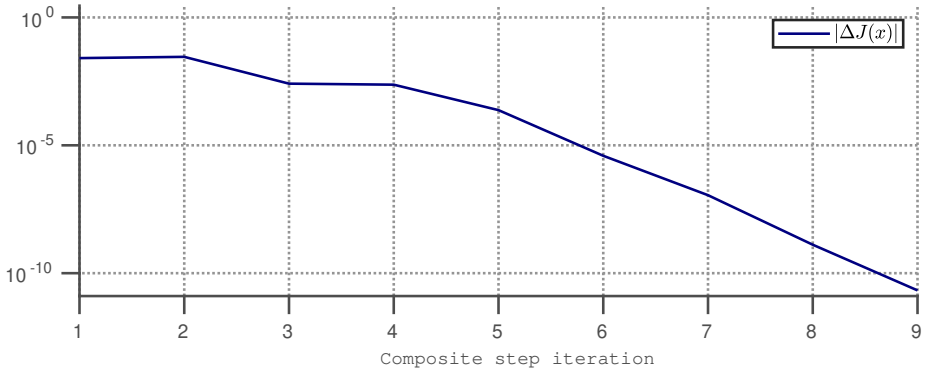

Figure 7.23: Change of the objective functional values for Problem 1 with three refinements.

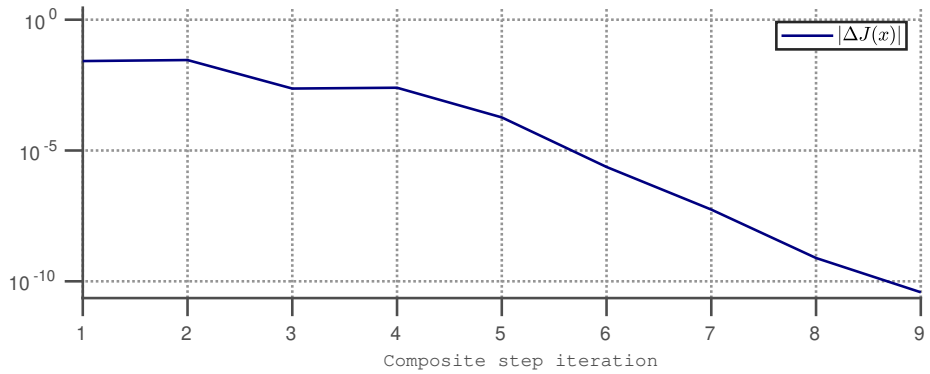

Figure 7.24: Change of the objective functional values for Problem 1 with four refinements.

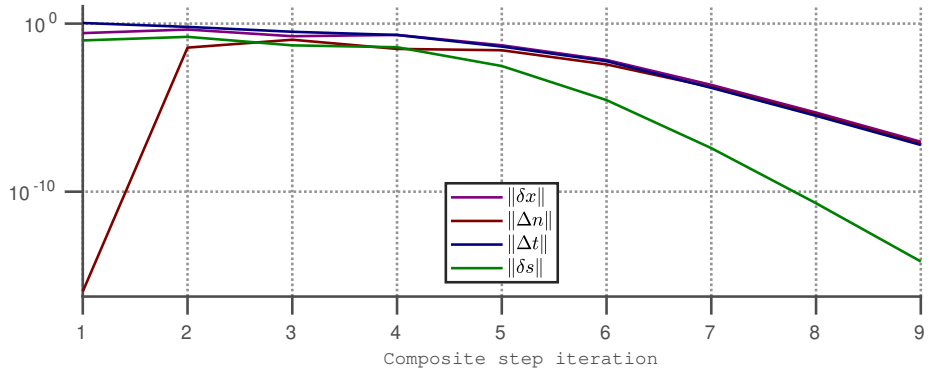

Figure 7.25: Norm of the updates for Problem 1 with three refinements.

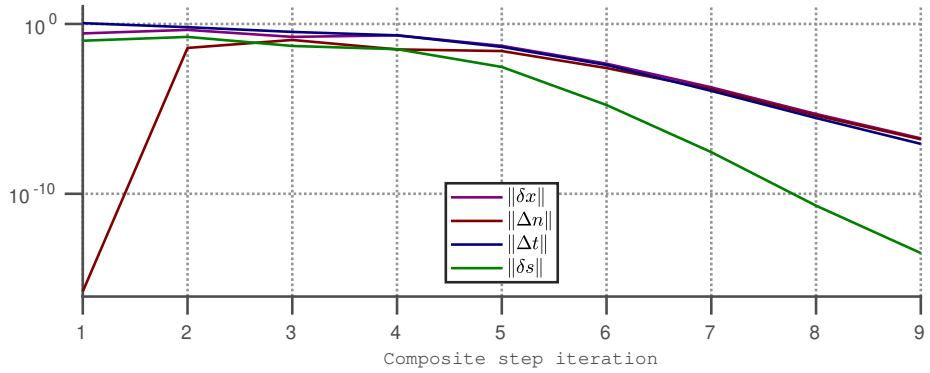

Figure 7.26: Norm of the updates for Problem 1 with four refinements.

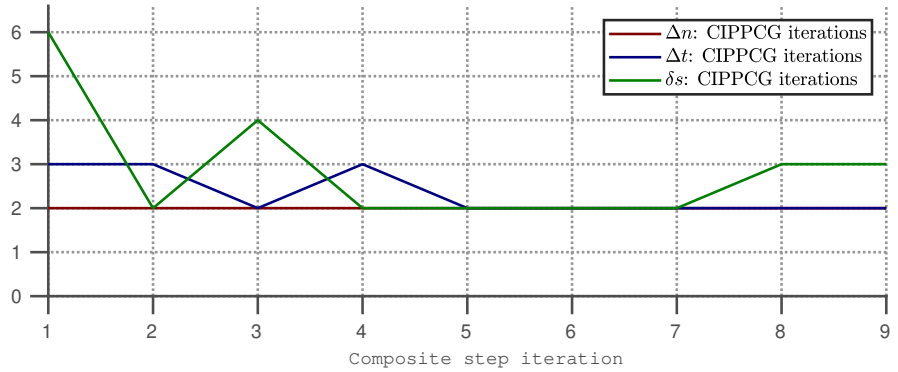

Figure 7.27: CIPPCG iterations for Problem 1 with three refinements.

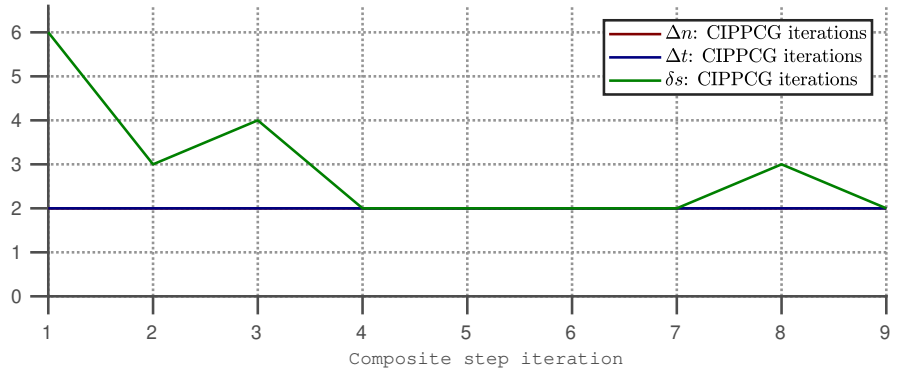

Figure 7.28: CIPPCG iterations for Problem 1 with four refinements.

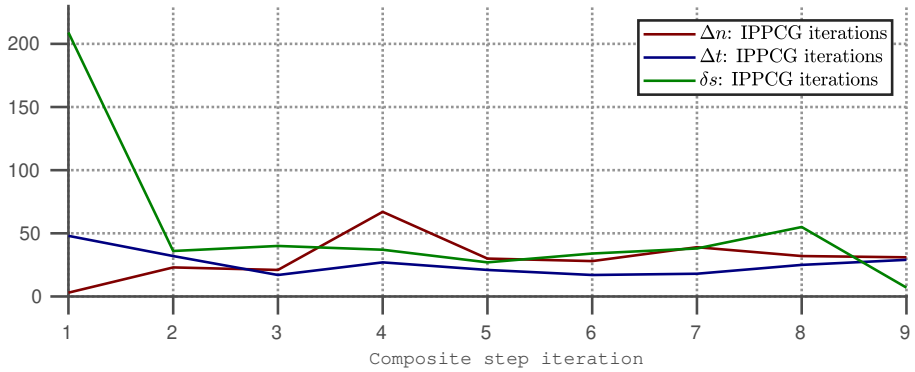

Figure 7.29: IPPCG iterations for Problem 1 with three refinements.

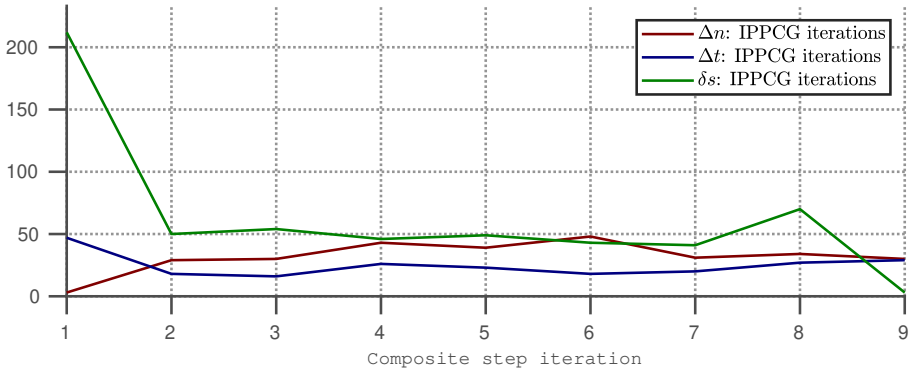

Figure 7.30: IPPCG iterations for Problem 1 with four refinements.

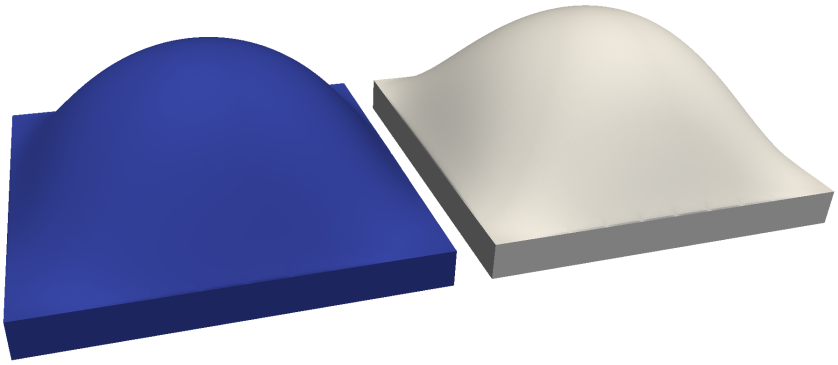

Figure 7.31: Optimal solution (left) and desired deformation (right) for Problem 1 with four refinements.

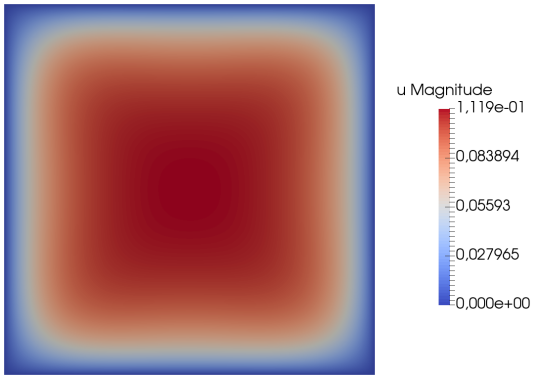

Figure 7.32: Optimal control (bottom face) for Problem 1 with four refinements. Color codes the intensity of the boundary forces.

Problem 2

As can be observed in Figures 7.33 to 7.43, Problem 2 is more difficult than the first one, which is reflected in the higher number of required composite step iterations. Here, the composite step method is faster in the case of four refinements. In contrast to before, the algorithm encounters non-convexities in the tangential step as well as in the total energy I. The iterations where the non-convexities occur are not the same, causing different behaviors for each refinement. Particularly, we observe more non-convex iterates in the case of three refinements, which might explain the higher number of composite step iterations. Nevertheless, the regularization mechanisms enable the composite step method to return to convex iterates again, and the algorithm finally converges in both cases. This is also reflected by the damping factors ν and τ, which are practically one in the last iterations (Figures 7.33 and 7.34).
The change in the objective functional value $|\Delta J(x)|$ and the update norm $\|\delta x\|$ both approach zero (Figures 7.35 to 7.38). Still, it seems that the algorithm has not entered the region of fast convergence yet.
Next, we note that the number of CIPPCG iterations is significantly higher for three refinements (Figures 7.39 and 7.40), in particular at iterates where the energy is non-convex. However, this may be caused by insufficient regularization. Considering the estimate in (6.10), we recall that the Hessian matrix of the elastic energy can still be almost singular, causing this irregular behavior. The same analysis holds for the IPPCG iterations (Figures 7.41 and 7.42).
Figure 7.43 depicts the resulting optimal solution compared to the desired deformation for four refinements. There, it can be seen that the desired deformation is closely approximated.
Finally, we conclude that our composite step algorithm combined with the CIPPCG method offers a robust approach to solve optimal control problems in nonlinear elasticity. It enables us to solve large scale problems and to deal with possible non-convexities. Still, the results suggest that a more sophisticated energy regularization could improve the performance significantly.

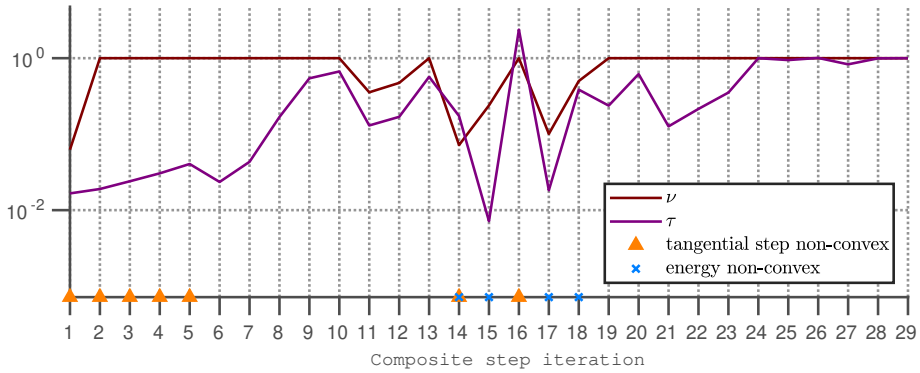

Figure 7.33: Damping factors for Problem 2 with three refinements.

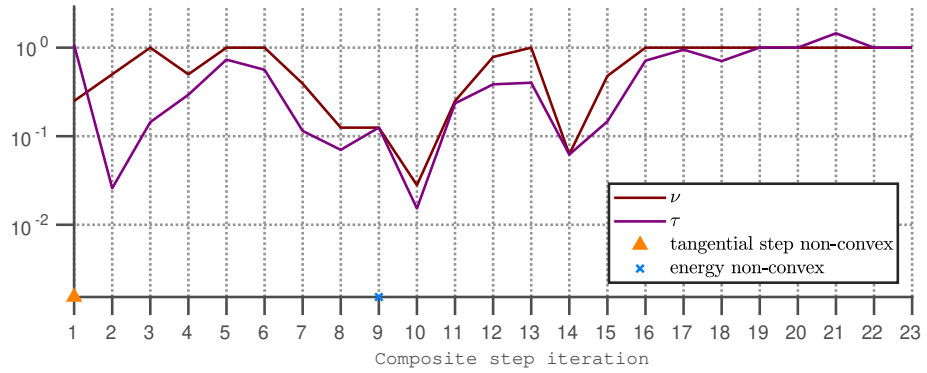

Figure 7.34: Damping factors for Problem 2 with four refinements.

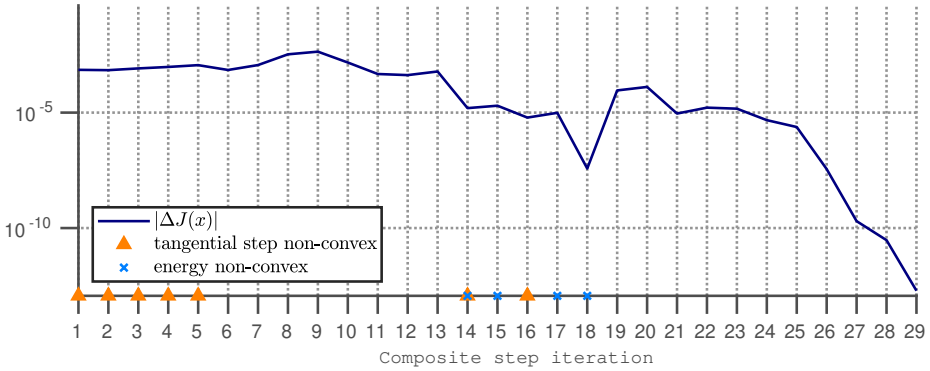

Figure 7.35: Change of the objective functional value for Problem 2 with three refinements.

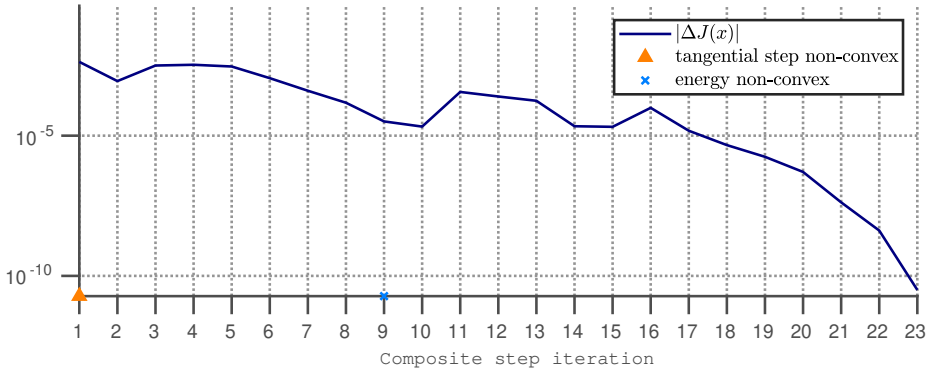

Figure 7.36: Change of the objective functional value for Problem 2 with four refinements.

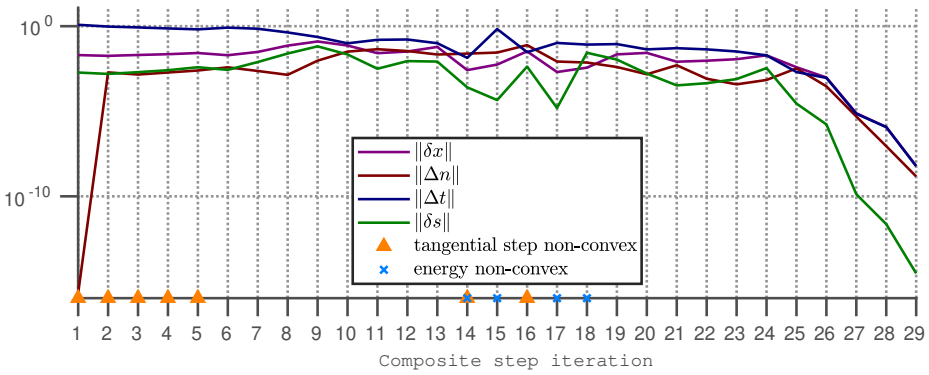

Figure 7.37: Norm of the updates for Problem 2 with three refinements.

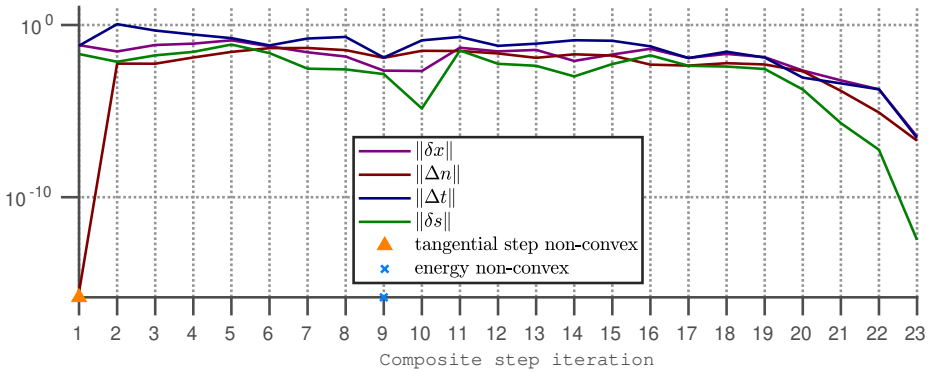

Figure 7.38: Norm of the updates for Problem 2 with four refinements.

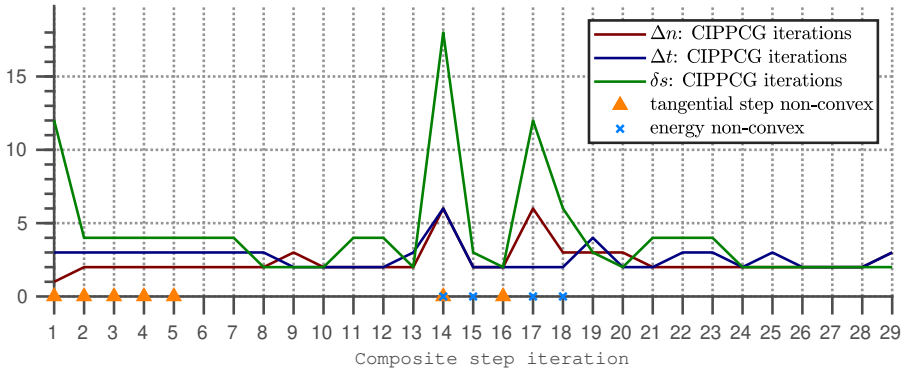

Figure 7.39: CIPPCG iterations for Problem 2 with three refinements.

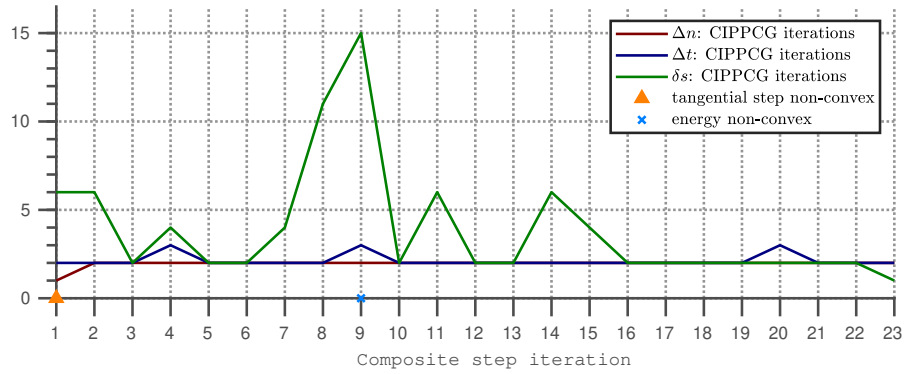

Figure 7.40: CIPPCG iterations for Problem 2 with four refinements.

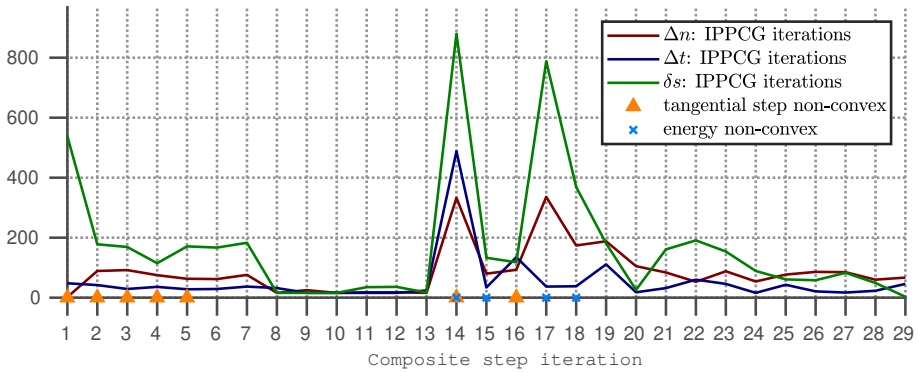

Figure 7.41: IPPCG iterations for Problem 2 with three refinements.

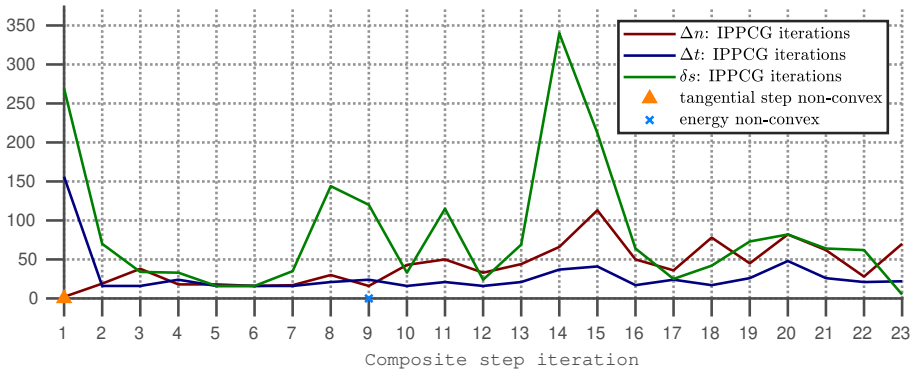

Figure 7.42: IPPCG iterations for Problem 2 with four refinements.

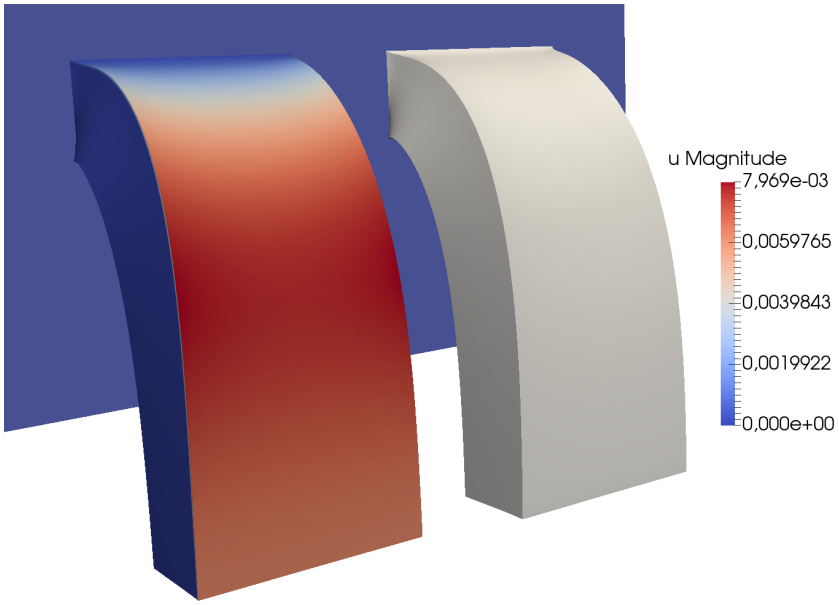

Figure 7.43: Optimal solution (left) and desired deformation (right) for Problem 2 with four refinements. Color codes the intensity of the boundary forces.

7.3.2 Choice of the functional analytic framework

For our optimal control problem, it is crucial to choose a suitable scalar product on $X = Y \times U$ since it induces orthogonality, the norm, and consequently, the measure of the step-length. Finding a scalar product such that (local) convergence theory in function space can be applied seems to be out of reach for nonlinear elasticity. The nonlinear elastic problem (3.2) alone is prone to a two-norm discrepancy. As discussed in Chapter 3, differentiability of the total energy functional I cannot be expected in spaces less regular than $W^{1,\infty}(\Omega)$. In contrast, the deformation space is only $W^{1,2}(\Omega)$, yielding a norm gap that is hard to bridge. Consequently, we have to expect grid-dependent behavior, at least for challenging problems where large strains occur. However, for simple problems, additional regularity of the steps can usually be observed. Such effects depend on the problem configuration and are difficult to verify a priori. In the following, we consider two different scalar products:

$$\langle (y,u),(y,u)\rangle_1 := \frac{1}{2}\langle y,y\rangle_{H^1(\Omega)} + \frac{\alpha}{2}\langle u,u\rangle_{L^2(\Gamma_N)}$$

and

$$\langle (y,u),(y,u)\rangle_2 := \frac{1}{2}\langle y,y\rangle_{L^2(\Omega)} + \frac{\alpha}{2}\langle u,u\rangle_{L^2(\Gamma_N)}.$$

The second scalar product is closer to the objective function J, but it does not take into account the regularity requirements of nonlinear elasticity. On the contrary, the first scalar product promotes smoother states. Although $W^{1,\infty}(\Omega)$-regularity is not guaranteed, it is significantly closer to the ideal situation.
We will test how the choice of the scalar product affects the performance for Problem 1 with three refinements and regularization parameter $\alpha =$ 5e-2. The results are analyzed by means of the damping factors ν and τ, see Figures 7.44 and 7.45. The numerical results confirm the prior considerations. While the first scalar product yields convergence within nine steps, the second one requires three times the number of iterations. Also, it exhibits very irregular behavior. Moreover, in the second case, non-convexities are

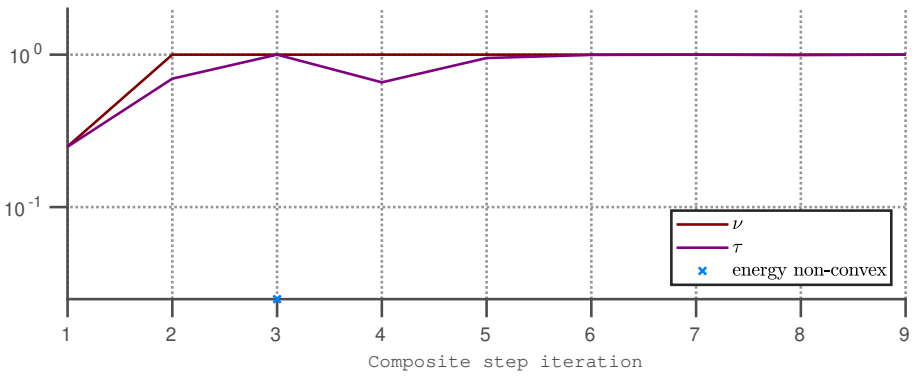

Figure 7.44: Damping factors for the first scalar product.

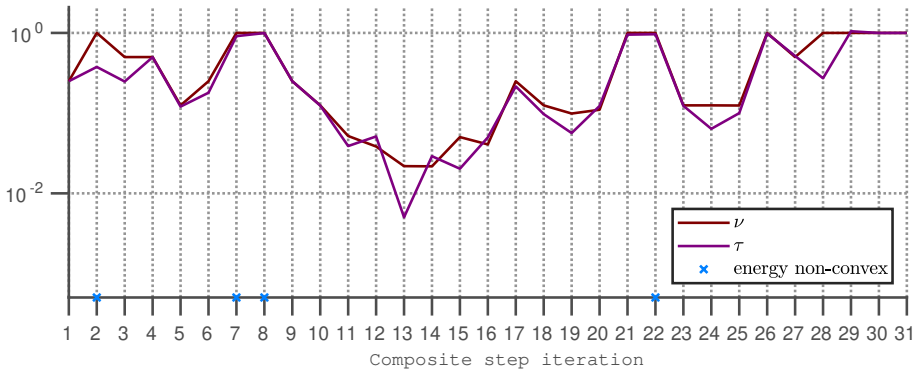

Figure 7.45: Damping factors for the second scalar product.

encountered four times compared to just one time when the first scalar product is chosen. A similar result was observed in [94].
In summary, we conclude that the functional analytical framework can be crucial for the performance of the applied algorithms. So far, utilizing an $H^1(\Omega)$-term in the scalar product produces satisfactory results. Still, finding more suitable choices is worth further investigations.

7.3.3 Optimal solutions that are not energy minimizers

Until now, we have not addressed how the underlying problem structure is changed if the energy minimizing condition in

$$\begin{aligned} &\min_{(y,u)\in Y\times U} J(y,u) \\ &s.t.\ y \in \operatorname*{argmin}_{v\in\mathcal{A}} I_\gamma(v,u) \end{aligned}$$

is replaced by its formal first order optimality condition, yielding

$$\begin{aligned} &\min_{(y,u)\in Y\times U} J(y,u) \\ &s.t.\ c_\gamma(y,u) = 0. \end{aligned} \tag{7.3}$$

While the first problem requires optimal states to be global minimizers of the regularized energy I_γ, the second problem admits solutions that are local minimizers or just arbitrary stationary points. Such solutions can be observed for special cases. An analogous observation follows for (4.9). Here, we solve (7.3) for Problem 2 and the parameters $\gamma = 100000$, $k = 4$, and $\alpha = 9.5$ with two refinements. In the composite step method, the linear systems are solved via a direct solver. Thus, the size of the problem is limited to two refinements. After obtaining a solution (y_*, u_*) to the optimal control problem (7.3), Algorithm 1 is applied to compute an energy minimizer of $I_\gamma(\,\cdot\,, u_*)$, denoted by y_e. At this, the identity mapping id is used as starting iterate. The resulting solutions are depicted in Figure 7.46, clearly showing that y_* and y_e are completely different defor-

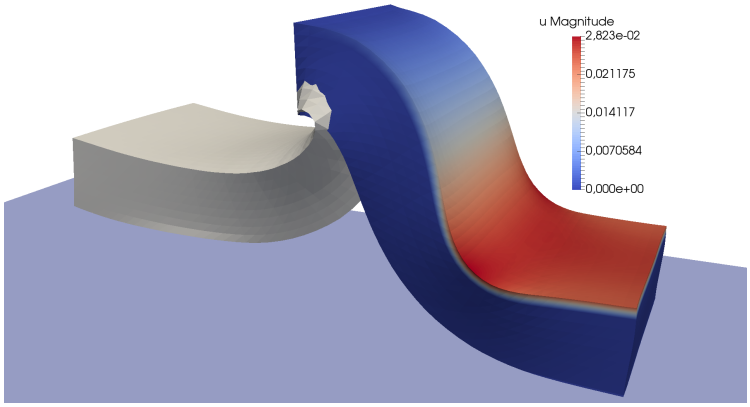

Figure 7.46: Optimal solution (right) with obstacle (transparent) and corresponding energy minimizer (gray and left) for Problem 2. Color codes the intensity of the boundary forces.

mations. Also, the respective function value $I_\gamma(y_e, u_*)$ is about -0.00029248, compared to 0.00426433 for $I_\gamma(y_*, u_*)$. This verifies that y_* cannot be a global energy minimizer, and thus, it does not satisfy the original constraint.
An interesting result is observed when the CIPPCG algorithm is applied instead of a direct solver. Then, the composite step method exhibits an oscillatory behavior, as illustrated in Figure 7.47 for the update norms. The algorithm was manually terminated after 64 iterations due to lack of progress. One explanation for this behavior might be

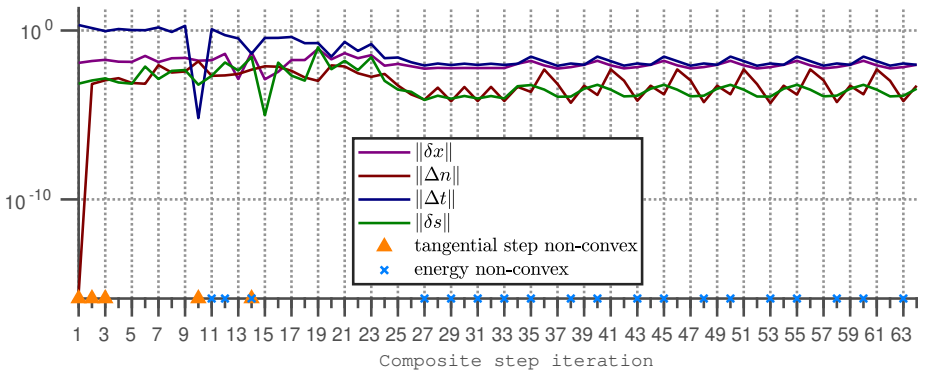

Figure 7.47: Norm of the updates for Problem 2 with two refinements.

the energy regularization required by the reoccurring non-convexity. By regularizing the elastic energy, as shown in Algorithm 17, the problem considered is fundamentally changed. Setting the regularization again to zero in the next iteration leaves us with an inconsistency between two successive steps, which could be responsible for the oscillation. By applying direct solvers, non-convexities are not detected. Consequently, the algorithm also accepts non-convex solutions, which cannot be energy minimizers. In general,

we want to avoid these cases all together and instead obtain solutions of the original regularized optimal control problem (4.3). So far, the methods derived here do not address this issue, leaving it as a subject for future research.

7.4 Path-Following

Here, we test Algorithm 9 for the original regularized optimal control problem

$$\min_{(y,u)\in Y\times U} J(y,u)$$
$$s.t.\ y \in \operatorname*{argmin}_{v\in\mathcal{A}} I_\gamma(v,u)$$

and the modified one

$$\min_{(y,u)\in Y\times U} J(y,u)$$
$$s.t.\ y \in \operatorname*{argmin}_{v\in\mathcal{A}} \mathcal{E}_\gamma(v,u),$$

where the energy minimizing conditions in the constraints are replaced by their formal first order conditions. The terms corresponding to the modified problem are marked with the subscript φ. Table 7.11 lists the chosen parameters. Additionally, the composite

	s_p	γ_0	γ_max	α	k	ρ^φ	$\varphi(\gamma)$	Refinements
Problem 1	10	10	1e13	0.2	4	0.18	$\gamma^{-\rho^\varphi}$	3
Problem 2	10	1000	1e12	15	4	0.18	$\gamma^{-\rho^\varphi}$	3

Table 7.11: Parameters for optimal control and path-following.

step method (Algorithm 17) acts as inner solver, and all linear systems are solved with the CIPPCG algorithm (Algorithm 16). The corresponding parameters are the same as displayed in Tables 7.7 to 7.9. At each iterate $z_k := (x_k, p_k)$ of the path-following approach with the respective parameter γ_k, the composite step method computes the next iterate z_{k+1} on the path for γ_{k+1}. To measure convergence, we define the update of the primal component

$$\Delta x_k := x_k - x_{k-1}.$$

For the path-following scheme to converge, $\|\Delta x\|$ has to approach zero. Analogously to the formula described in (7.2), we compute estimates for the convergence rates of the maximum constraint violation $\|[y]_+\|_{L^\infty(\Gamma_C)}$ and the change in the objective functional values

$$|\Delta J(x_k)| := |J(x_k) - J(x_{k-1})|.$$

The results are depicted in Figures 7.48 to 7.60. In Figures 7.48 and 7.49, we observe that the number of required composite step iterations is very high in the first iteration. After that, the path-following algorithm only requires few inner iterations. However, the number raises again at the end. This observation indicates that the problem is still

difficult or that the composite step method becomes instable for very large parameters γ.

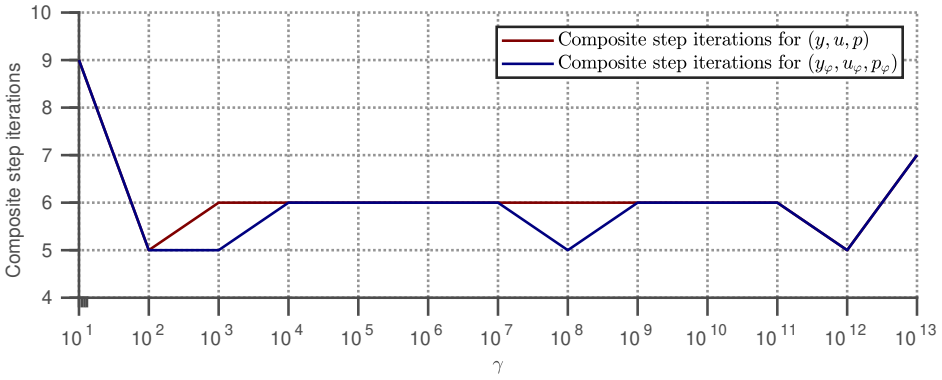

Figure 7.48: Inner iterations for Problem 1.

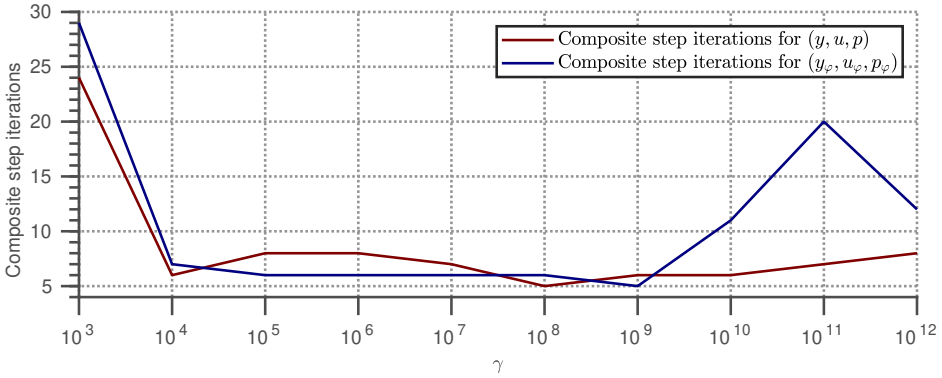

Figure 7.49: Inner iterations for Problem 2.

Figures 7.50 and 7.51 show that $\|\Delta x\|$ and $\|\Delta x_\varphi\|$ approach zero in both problem settings, indicating convergence of the entire path-following scheme. Moreover, the convergence speed only seems to be linear. However, due to the lack of theoretical results in the path-following setting, it is not clear whether faster rates can be expected. Further, the speed of convergence is slower for the modified regularization (4.9).
Next, Figures 7.52 and 7.53 illustrate that the objective functional values are monotonically increasing in all cases. This result is reasonable since relaxing the constraints allows for a larger choice of possible deformations, and thus, better approximations of the desired deformation y_d. Additionally, the modified regularization yields smaller function values, which is consistent with Proposition 4.14.
Regarding the term $|\Delta J(x)|$, Figures 7.54 and 7.55 show an estimated rate of about $\frac{1}{k-1}$. It is also indicated that $|\Delta J(x_\varphi)|$ approaches zero at the slower rate ρ^φ. This has to be expected since the regularization function φ, applied in (4.4), converges to zero at the rate ρ^φ, which is slower than the rate $\frac{1}{k-1}$ by construction.

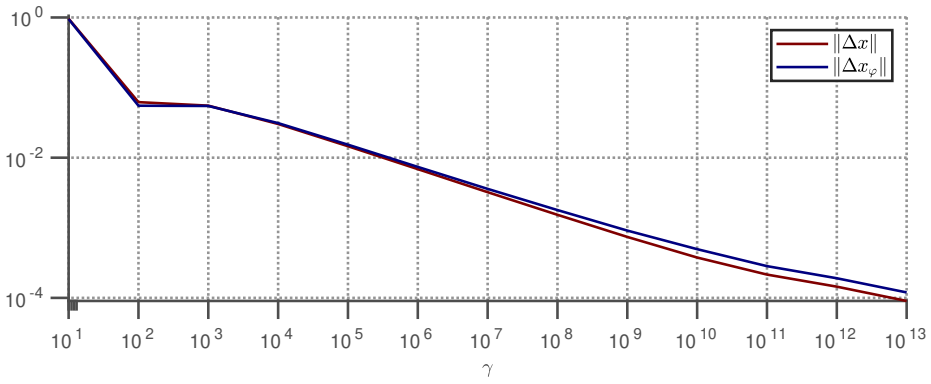

Figure 7.50: Norm of the updates for Problem 1.

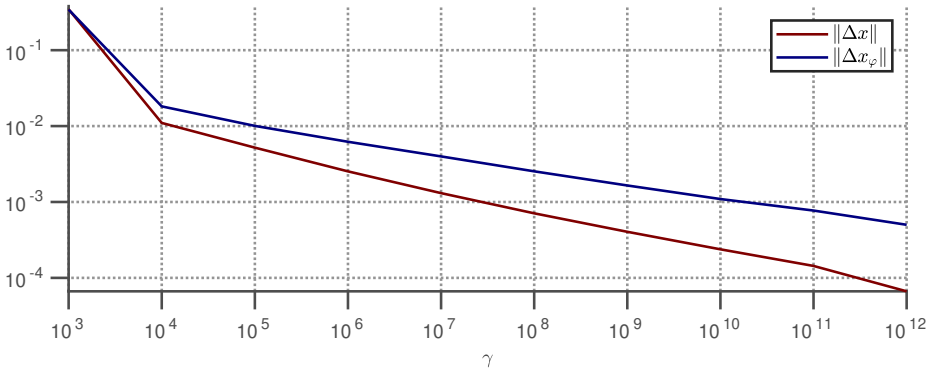

Figure 7.51: Norm of the updates for Problem 2.

The rate estimates for the maximum constraint violation indicate the convergence rate $\frac{1}{k-1}$ for the first problem (Figure 7.56). Here, the additional regularization does not seem to interfere with the convergence speed. Interestingly, the results coincide with those observed in Section 7.1. This raises the question whether the techniques from Section 3.3 can be transferred to the optimal control setting. For the second problem (Figure 7.57), the convergence rate of the maximum constraint violation appears to be slower. However, the results do not indicate an alternative value for the convergence rate, and further tests are required.
Figures 7.58 to 7.60 illustrate the optimal solutions for the largest parameter γ. Both approaches seem to converge to the same solution for Problem 1 and 2.
In summary, both regularization schemes appear to converge. Here, the additional regularization defined in (4.5) is not necessary to guarantee convergence, which may be caused by the simple problem setting. Also, the first approach exhibits faster convergence. Further, we observe convergence rates for the objective functional values and the maximum constraint violation. So far, neither the convergence of the entire path-following approach nor the observed rates have been verified by mathematical theory, motivating further examinations.

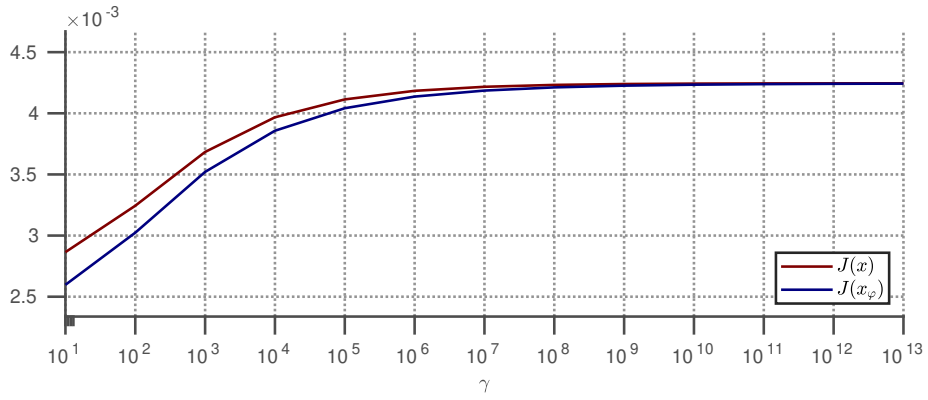

Figure 7.52: Objective functional values for Problem 1.

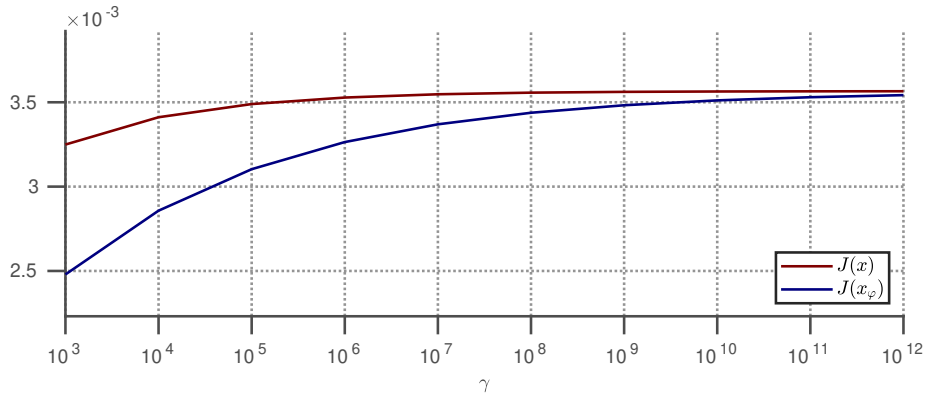

Figure 7.53: Objective functional values for Problem 2.

All in all, we have been able to solve optimal control problems with hyperelastic contact problems as constraints by combining a path-following method with a robust inner solver. So far a simple path-following scheme was sufficient. For the inner solver, an affine covariant composite step method provided the required robustness to solve problems involving nonlinear elasticity. The CIPPCG method was successfully deployed to solve the arising linear systems, allowing us to examine large scale problems.

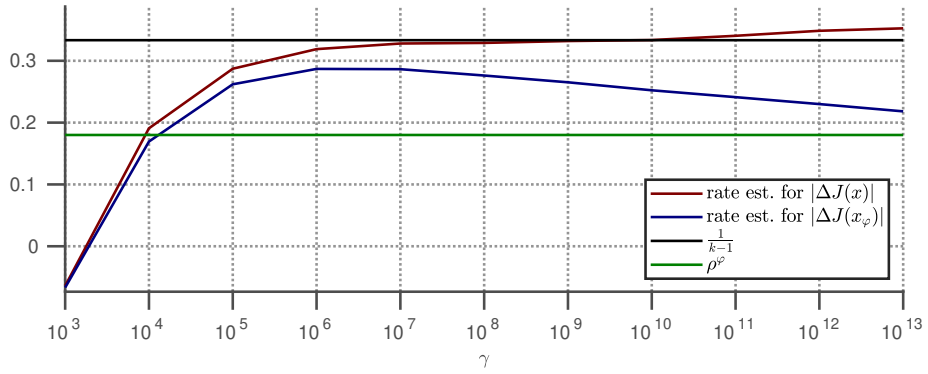

Figure 7.54: Estimated convergence rates for the objective functional values for Problem 1.

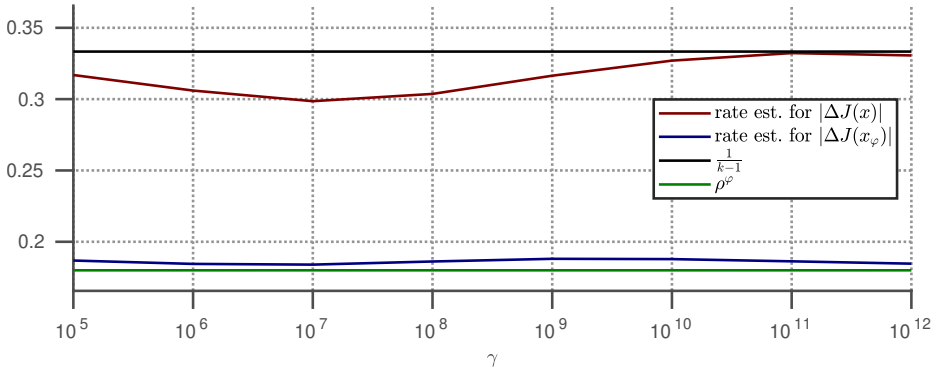

Figure 7.55: Estimated convergence rates for the objective functional values for Problem 2.

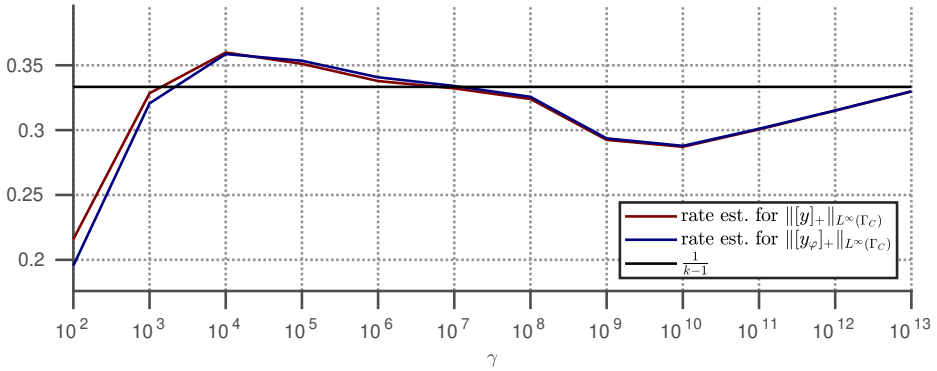

Figure 7.56: Estimated convergence rates for the maximum constraint violation for Problem 1.

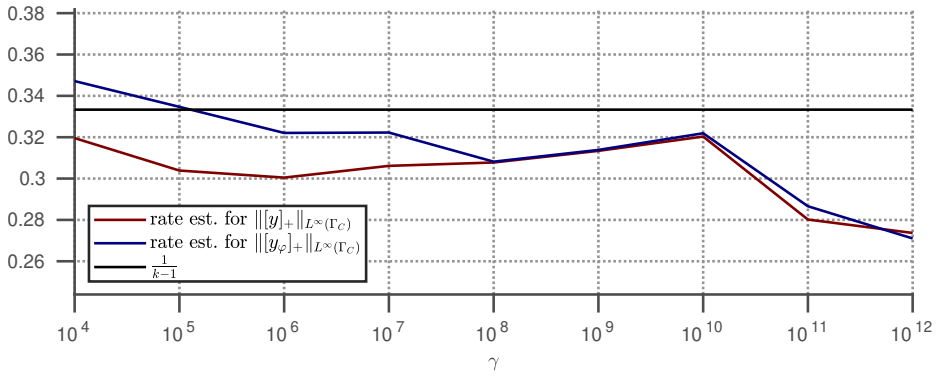

Figure 7.57: Estimated convergence rates for the maximum constraint violation for Problem 2.

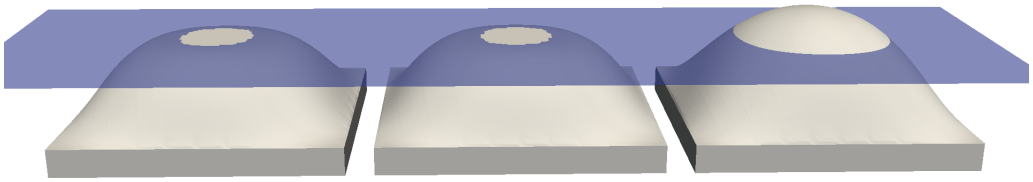

Figure 7.58: Optimal solutions for Problem 1 with $\gamma = 1e13$ and obstacle (transparent). Normal compliance regularization (left), modified regularization (middle), and reference deformation (right).

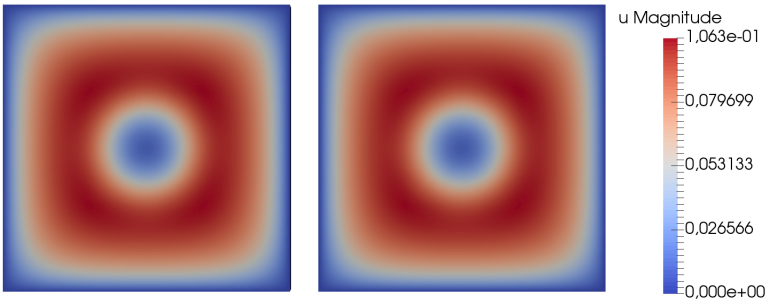

Figure 7.59: Optimal controls (bottom face) for Problem 1 with $\gamma = 1e13$. Normal compliance regularization (left) and modified regularization (right). Color codes the intensity of the boundary forces.

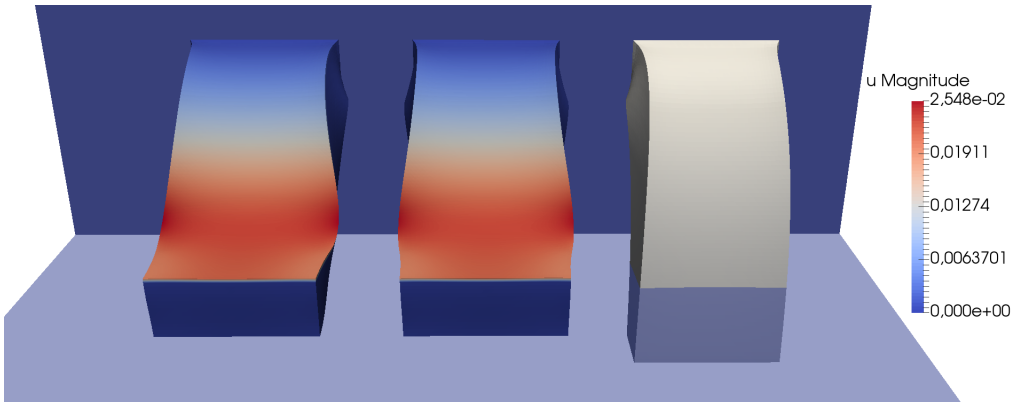

Figure 7.60: Optimal solutions for Problem 2 with $\gamma = 1e12$ and obstacle (transparent). Normal compliance regularization (left), modified regularization (middle), and reference deformation (right). Color codes the intensity of the boundary forces.

Chapter 8

Conclusion and Outlook

In this thesis, we have conducted a detailed theoretical and numerical investigation of optimal control of nonlinear elastic contact problems. Even without contact constraints, such optimal control problems are challenging in themselves due to the bilevel structure and the inherent difficulty of nonlinear elasticity. We extended the existing results, elaborated in [64, 66], by also considering contact constraints. These constraints add a non-smoothness to an already complex problem. A central part of the study was the relaxation of the contact constraints via the normal compliance method. In that context, convergence rates have been established. After applying the relaxation, we obtained regularized optimal control problems. The focus of the theoretical examination laid on deriving a corresponding convergence result. Showing such convergence has been achieved in two ways. First, strong structural assumptions yield the desired result. Second, we also obtained convergence by modifying the normal compliance regularization to better reflect the entire optimal control problem, allowing more general settings.

For the numerical simulations, we replaced the energy minimization property of hyperelasticity by its formal first order conditions. So far, these conditions are only valid for very restrictive settings. The replacement enabled us to apply the robust affine covariant composite step method put forward in [64, 67], which can deal with highly nonlinear problems. In addition, we introduced a specialized gradient-type method to solve the arising linear systems. This introduction was necessary since exiting algorithms are no longer suitable for large scale optimal control problems in nonlinear elasticity. Thorough testing was conducted, where we successfully solved complex optimal control problems in nonlinear elasticity.

At the end, these approaches were combined with path-following to solve regularized optimal control problems and approximate solutions to the original problem with contact constraints. However, it remains an open problem to back up these last results with mathematical theory. In summary, despite the challenging nature of optimal control and nonlinear elastic contact problems, new theoretical and numerical results have been established. Still, the examinations are by no means complete and leave us with a wide range of possible directions for future research.

The following extensions are conceivable. First, in the current implementation of the optimal control problem, solutions that are not energy minimizers are admissible. Such solutions have to be avoided since they are inconsistent with the original problem description of hyperelasticity. This is connected to the larger theoretical problem of deriving first order conditions for minimizers in a general setting. Also, the corresponding KKT conditions (4.13) only hold formally and require further investigations.
Second, since the long-term goal is to solve real-world applications, more complex geometries and contact constraints need to be considered. This also includes multi-body problems. Further, adaptive accuracy matching and regularization approaches are necessary to increase performance and solve more challenging problems. So far only simple heuristic approaches are deployed to determine the required accuracies of the applied algorithms. In that context, error estimates for adaptive mesh refinements are also essential.
Last, the nonlinear update strategy (Algorithm 6) yielded promising results in the field of nonlinear elasticity, encouraging further investigations. In particular, the embedding of this strategy into the optimal control setting could lead to significant performance gains.

List of Figures

List of Tables

Publications

[1] A. Schiela and M. Stöcklein. Optimal Control of Static Contact in Finite Strain Elasticity. *ESAIM: Control, Optimisation and Calculus of Variations*, 2020. `https://doi.org/10.1051/cocv/2020014`

[2] M. Schaller, A. Schiela, and M. Stöcklein. A composite step method with inexact step computations for PDE constrained optimization. Preprint SPP1962-098, October 2018

[3] A. Schiela and M. Stöcklein. Algorithms for Optimal Control of Elastic Contact Problems with Finite Strain. `https://eref.uni-bayreuth.de/52240/`, September 2019

Bibliography

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, USA, 2007.

[2] R. Adams. *Sobolev Spaces. Adams*. Pure and applied mathematics. Academic Press, 1975.

[3] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, 1999.

[4] L.-E. Andersson. A quasistatic frictional problem with normal compliance. *Nonlinear Analysis*, 16(4):347–369, 1991.

[5] L.-E. Andersson. A quasistatic frictional problem with a normal compliance penalization term. *Nonlinear Analysis: Theory, Methods & Applications*, 37(6):689–705, 1999.

[6] J. Ball. Convexity conditions and existence theorems in nonlinear elasticity. *Archive for Rational Mechanics and Analysis*, 63(4):337–403, 1977.

[7] J. Ball. Some open problems in elasticity. In *Geometry, mechanics, and dynamics*, pages 3–59. Springer, 2002.

[8] J. M. Ball. Global invertibility of Sobolev functions and the interpenetration of matter. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, 88(3-4):315–328, 1981.

[9] J. M. Ball and V. J. Mizel. One-dimensional variational problems whose minimizers do not satisfy the Euler-Lagrange equation. In *Analysis and Thermomechanics*, pages 285–348. Springer, 1987.

[10] P. Bastian, M. Blatt, C. Engwer, A. Dedner, R. Klöfkorn, S. Kuttanikkad, M. Ohlberger, and O. Sander. The distributed and unified numerics environment (DUNE). In *Proc. of the 19th Symposium on Simulation Technique in Hannover*, 2006.

[11] S. Bechmann and M. Frey. *Regularisierungsmethoden für Optimalsteuerungsprobleme*. Math. Inst. der Univ. Bayreuth, 2008.

[12] C. Bernardi, Y. Maday, and A. Patera. Domain decomposition by the mortar element method. *Asymptotic and Numerical Methods for Partial Differential Equations with Critical Parameters*, pages 269–286, 1993.

[13] J. H. Bramble, J. E. Pasciak, and J. Xu. Parallel multilevel preconditioners. *Mathematics of Computation*, 55(191):1–22, 1990.

[14] N. L. Carothers. *Real analysis*. Cambridge University Press, 2000.

[15] P. G. Ciarlet. *Mathematical Elasticity: Three-dimensional elasticity*. Number Bd. 1. North-Holland, 1994.

[16] P. G. Ciarlet and G. Geymonat. Sur les lois de comportement en élasticité non linéaire compressible. *CR Acad. Sci. Paris Sér. II*, 295:423–426, 1982.

[17] P. G. Ciarlet and J.-L. Lions. *Handbook of Numerical Analysis: Techniques of Scientific Computer (Part 1), Numerical Methods for Solids (Part 1), Solution of Equations in* $\mathbb{R}^n$ *(Part 2)*, volume 3. North-Holland, 1994.

[18] P. G. Ciarlet and J. Nečas. Unilateral problems in nonlinear, three-dimensional elasticity. *Archive for Rational Mechanics and Analysis*, 87(4):319–338, 1985.

[19] P. G. Ciarlet and J. Nečas. Injectivity and self-contact in nonlinear elasticity. *Archive for Rational Mechanics and Analysis*, 97(3):171–188, 1987.

[20] M. Cocu. Existence of solutions of signorini problems with friction. *International journal of engineering science*, 22(5):567–575, 1984.

[21] J. Davet. Sur les densités d'énergie en élasticité non linéaire: confrontation de modeles et de travaux expérimentaux. In *Annales des Ponts et Chaussées*, pages 2–33, 1985.

[22] T. Davis and I. Duff. An Unsymmetric-Pattern Multifrontal Method for Sparse LU Factorization. *SIAM Journal on Matrix Analysis and Applications*, 18(1):140–158, 1997.

[23] P. Deuflhard. *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms*. Springer Publishing Company, Incorporated, 2011.

[24] G. Eason and R. W. Ogden. *Elasticity: mathematical methods and applications: the Ian N. Sneddon 70th birthday volume*. Halsted Press, 1990.

[25] C. Eck, J. Jarusek, and M. Krbec. *Unilateral contact problems: variational methods and existence theorems*. CRC Press, 2005.

[26] C. Eck, J. Jarušek, and J. Stará. Normal compliance contact models with finite interpenetration. *Archive for Rational Mechanics and Analysis*, 208(1):25–57, 2013.

[27] J. Ericksen and R. Toupin. Implications of hadamard's conditions for elastic stability with respect to uniqueness theorems. *Canadian Journal of Mathematics*, 8:432–436, 1956.

[28] A. C. Eringen and D. Edelen. On nonlocal elasticity. *International Journal of Engineering Science*, 10(3):233–248, 1972.

[29] P. E. Farrell, A. Birkisson, and S. W. Funke. Deflation techniques for finding distinct solutions of nonlinear partial differential equations. *SIAM Journal on Scientific Computing*, 37(4):A2026–A2045, 2015.

[30] I. Fonseca and W. Gangbo. Local invertibility of sobolev functions. *SIAM journal on mathematical analysis*, 26(2):280–304, 1995.

[31] M. Giaquinta, G. Modica, et al. Cartesian currents, weak diffeomorphisms and existence theorems in nonlinear elasticity. *Archive for rational mechanics and analysis*, 106(2):97–159, 1989.

[32] S. Götschel, M. Weiser, and A. Schiela. Solving Optimal Control Problems with the Kaskade 7 Finite Element Toolbox. In A. Dedner, B. Flemisch, and R. Klöfkorn, editors, *Advances in DUNE*, pages 101 – 112. 2012.

[33] N. I. 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.

[34] P. L. Gould and Y. Feng. *Introduction to linear elasticity*. Springer, 1994.

[35] C. Gräser and R. Kornhuber. Multigrid methods for obstacle problems. *Journal of Computational Mathematics*, pages 1–44, 2009.

[36] C. Gräser, U. Sack, and O. Sander. Truncated nonsmooth Newton multigrid methods for convex minimization problems. In *Domain Decomposition Methods in Science and Engineering XVIII*, pages 129–136. Springer, 2009.

[37] A. Griewank. The modification of Newton's method for unconstrained optimization by bounding cubic terms. Technical Report NA/12, University of Cambridge, 1981.

[38] G. Guennebaud, B. Jacob, et al. Eigen v3. http://eigen.tuxfamily.org, 2010.

[39] A. Günnel. *Numerical Aspects in Optimal Control of Elasticity Models with Large Deformations*. PhD thesis, TU Chemnitz, 2014.

[40] A. Günnel and R. Herzog. Optimal control problems in finite-strain elasticity by inner pressure and fiber tension. *Frontiers in Applied Mathematics and Statistics*, 2:4, 2016.

[41] M. E. Gurtin. The linear theory of elasticity. In *Linear theories of elasticity and thermoelasticity*, pages 1–295. Springer, 1973.

[42] M. H. Gutknecht and S. Röllin. The Chebyshev iteration revisited. *Parallel Computing*, 28(2):263–283, 2002.

[43] W. Han and M. Sofonea. *Quasistatic contact problems in viscoelasticity and viscoplasticity*. American Mathematical Soc., 2002.

[44] M. Heinkenschloss and D. Ridzal. A matrix-free trust-region SQP method for equality constrained optimization. *SIAM J. Optim.*, 24(3):1507–1541, 2014.

[45] M. Hintermüller and K. Kunisch. Feasible and Noninterior Path–Following in Constrained Minimization with Low Multiplier Regularity. *SIAM Journal on Control and Optimization*, 45(4):1198–1221, 2006.

[46] M. Hintermüller and K. Kunisch. Path-following methods for a class of constrained minimization problems in function space. *SIAM Journal on Optimization*, 17(1):159–187, 2006.

[47] 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 Journal on Optimization*, 24(1):108–126, 2014.

[48] M. Hintermüller, F. Tröltzsch, and I. Yousept. Mesh-independence of semismooth Newton methods for Lavrentiev-regularized state constrained nonlinear optimal control problems. *Numerische Mathematik*, 108(4):571–603, 2008.

[49] M. Hinze and C. Meyer. Variational discretization of Lavrentiev-regularized state constrained elliptic optimal control problems. *Computational Optimization and Applications*, 46(3):487–510, 2010.

[50] S. Hüeber and B. I. Wohlmuth. An optimal a priori error estimate for nonlinear multibody contact problems. *SIAM Journal on Numerical Analysis*, 43(1):156–173, 2005.

[51] S. Hüeber and B. I. Wohlmuth. A primal–dual active set strategy for non-linear multibody contact problems. *Computer Methods in Applied Mechanics and Engineering*, 194(27-29):3147–3166, 2005.

[52] K. Ito and K. Kunisch. Semi–smooth newton methods for variational inequalities of the first kind. *ESAIM: Mathematical Modelling and Numerical Analysis*, 37(1):41–62, 2003.

[53] J. Jarušek. Dynamic contact problems with given friction for viscoelastic bodies. *Czechoslovak Mathematical Journal*, 46(3):475–487, 1996.

[54] K. Johnson. *Contact Mechanics*. Cambridge University Press, 1987.

[55] C. Keller, N. I. Gould, and A. J. Wathen. Constraint preconditioning for indefinite linear systems. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1300–1317, 2000.

[56] N. Kikuchi and J. T. Oden. *Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods*, volume 8. siam, 1988.

[57] A. Klarbring, A. Mikelić, and M. Shillor. Frictional contact problems with normal compliance. *International Journal of Engineering Science*, 26(8):811–832, 1988.

[58] A. Klarbring, A. Mikelić, and M. Shillor. A global existence result for the quasistatic frictional contact problem with normal compliance. In *Unilateral Problems in Structural Analysis IV*, pages 85–111. Springer, 1991.

[59] T. Laursen. Computational contact and impact mechanics, Springer. *Berlin, Heidelberg, New York (corr. 2nd printing)*, 2003.

[60] T. Laursen and J. Simo. A continuum-based finite element formulation for the implicit solution of multibody, large deformation-frictional contact problems. *International Journal for numerical methods in engineering*, 36(20):3451–3485, 1993.

[61] T. A. Laursen. *Computational contact and impact mechanics: fundamentals of modeling interfacial phenomena in nonlinear finite element analysis*. Springer Science & Business Media, 2013.

[62] E. H. Lieb and M. Loss. *Analysis, Volume 14*. 2001.

[63] J. L. Lions and G. Duvaut. *Inequalities in mechanics and physics*. Springer, 1976.

[64] L. Lubkoll. *An Optimal Control Approach to Implant Shape Design : Modeling, Analysis and Numerics*. PhD thesis, Bayreuth, 2015.

[65] L. Lubkoll. Fung-Invariant-based modeling. *Archive of Numerical Software*, 5(1), 2017.

[66] L. Lubkoll, A. Schiela, and M. Weiser. An optimal control problem in polyconvex hyperelasticity. *SIAM J. Control Opt.*, 52(3):1403 – 1422, 2014.

[67] L. Lubkoll, A. Schiela, and M. Weiser. An affine covariant composite step method for optimization with PDEs as equality constraints. *Optimization Methods and Software*, 32:1132–1161, 2017.

[68] J. Marsden and T. Hughes. *Mathematical Foundations of Elasticity*. Prentice-Hall Personal Computing Series. Prentice-Hall, 1983.

[69] J. Martins and J. 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.

[70] C. Meyer, U. Prüfert, and F. Tröltzsch. On two numerical methods for state-constrained elliptic control problems. *Optimization Methods and Software*, 22(6):871–899, 2007.

[71] C. Meyer, A. Rösch, and F. Tröltzsch. Optimal control of PDEs with regularized pointwise state constraints. *Computational Optimization and Applications*, 33(2-3):209–228, 2006.

[72] C. Meyer and I. Yousept. Regularization of state-constrained elliptic optimal control problems with nonlocal radiation interface conditions. *Computational Optimization and Applications*, 44(2):183–212, 2009.

[73] G. Müller. *Optimal control of time-discretized contact problems.* PhD thesis, Bayreuth, 2019.

[74] J. Nečas. *Direct Methods in the Theory of Elliptic Equations.* Springer Monographs in Mathematics. Springer Berlin Heidelberg, 2012.

[75] J. Oden. *Finite Elements of Nonlinear Continua.* McGraw-Hill, 1972.

[76] J. Oden and J. Martins. Models and computational methods for dynamic friction phenomena. *Computer Methods in Applied Mechanics and Engineering*, 52(1):527 – 634, 1985.

[77] R. Ogden. *Non-linear Elastic Deformations.* Dover Publications, 1997.

[78] 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.

[79] P. Panagitopoulos. Inequality problems in mechanics and applications. Convex and nonconvex energy functions, 1985.

[80] P. Penzler, M. Rumpf, and B. Wirth. A phase-field model for compliance shape optimization in nonlinear elasticity. *ESAIM: Control, Optimisation and Calculus of Variations*, 18(1):229–258, 2012.

[81] A. Popp. *Mortar methods for computational contact mechanics and general interface problems.* PhD thesis, Technische Universität München, 2012.

[82] T. Rees. *Preconditioning iterative methods for PDE constrained optimization.* PhD thesis, Oxford University, 2010.

[83] S. Reese and P. Wriggers. A finite element method for stability problems in finite elasticity. *International journal for numerical methods in engineering*, 38(7):1171–1200, 1995.

[84] 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.

[85] M. Rochdi, M. Shillor, and M. Sofonea. Quasistatic viscoelastic contact with normal compliance and friction. *Journal of Elasticity*, 51(2):105–126, 1998.

[86] A. Rösch and F. Tröltzsch. Existence of regular Lagrange multipliers for a nonlinear elliptic optimal control problem with pointwise control-state constraints. *SIAM Journal on Control and Optimization*, 45(2):548–564, 2006.

[87] M. Rumpf and B. Wirth. An elasticity-based covariance analysis of shapes. *International Journal of Computer Vision*, 92(3):281–295, 2011.

[88] Y. Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, USA, 2nd edition, 2003.

[89] A. S. Saada. *Elasticity: Theory and Applications*, volume 16. Elsevier, 2013.

[90] M. H. Sadd. *Elasticity: Theory, Applications, and Numerics*. Academic Press, 2009.

[91] O. Sander. *Multidimensional Coupling in a Human Knee Model*. PhD thesis, Freie Universität Berlin, 2008.

[92] M. Schaller, A. Schiela, and M. Stöcklein. A composite step method with inexact step computations for PDE constrained optimization. Preprint SPP1962-098, October 2018.

[93] A. Schiela and A. Günther. An interior point algorithm with inexact step computation in function space for state constrained optimal control. *Numerische Mathematik*, 119(2):373–407, 2011.

[94] A. Schiela and M. Stöcklein. Algorithms for Optimal Control of Elastic Contact Problems with Finite Strain. `https://eref.uni-bayreuth.de/52240/`, September 2019.

[95] A. Schiela and M. Stöcklein. Optimal Control of Static Contact in Finite Strain Elasticity. *ESAIM: Control, Optimisation and Calculus of Variations*, 2020. `https://doi.org/10.1051/cocv/2020014`.

[96] A. Siegl. Iterative Löser zur Schrittberechnung in einer Composite-Step Methode, Januar 2020. `https://eref.uni-bayreuth.de/55519/`.

[97] A. Signorini. Sopra alcune questioni di elastostatica. *Atti della Societa Italiana per il Progresso delle Scienze*, 1933.

[98] S. J. Spector. On uniqueness in finite elasticity with general loading. *Journal of Elasticity*, 10(2):145–161, 1980.

[99] S. J. Spector. On uniqueness for the traction problem in finite elasticity. *Journal of Elasticity*, 12(4):367–383, 1982.

[100] D. J. Steigmann. *Finite elasticity theory*. Oxford University Press, 2017.

[101] F. Stoppelli. Un teorema di esistenza e di unicita relativo alle equazioni dell'elastostatica isoterma per deformazioni finite. *Ricerche mat*, 3(247-267):59, 1954.

[102] F. Stoppelli. Sulla sviluppabilità in serie di potenze di un parametro delle soluzioni delle equazioni dell'elastostatica isoterma. *Ricerche Mat*, 4(58-73):59, 1955.

[103] C. A. Stuart. Special problems involving uniqueness and multiplicity in hyperelasticity. In *Nonlinear functional analysis and its applications*, pages 131–145. Springer, 1986.

[104] L. A. Taber. *Nonlinear theory of elasticity: Applications in Biomechanics*. World Scientific, 2004.

[105] A. Vardi. A trust region algorithm for equality constrained minimization: convergence properties and implementation. *SIAM J. Numer. Anal.*, 22(3):575–591, 1985.

[106] A. Weinstein. A global invertibility theorem for manifolds with boundary. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, 99(3-4):283–284, 1985.

[107] M. Weiser, P. Deuflhard, and B. Erdmann. Affine conjugate adaptive Newton methods for nonlinear elastomechanics. *Optimization Methods and Software*, 22(3):413–431, 2007.

[108] B. I. Wohlmuth. *Discretization Methods and Iterative Solvers Based on Domain Decomposition*, volume 17 of *Lectures Notes in Computational Science and Engineering*. Springer, Heidelberg, 2001.

[109] P. Wriggers. Computational contact mechanics. *Berlin, Heidelberg*, 2006.

[110] J. Youett, O. Sander, and R. Kornhuber. A globally convergent filter-trust-region method for large deformation contact problems. *SIAM Journal on Scientific Computing*, 41(1):B114–B138, 2019.

[111] J. W. Youett. *Dynamic large deformation contact problems and applications in virtual medicine*. PhD thesis, 2016.

[112] J. Ziems and S. Ulbrich. Adaptive multilevel inexact SQP methods for PDE-constrained optimization. *SIAM J. Optim.*, 21(1):1–40, 2011.

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet habe. Weiterhin erkläre ich, dass ich die Dissertation nicht bereits zur Erlangung eines akademischen Grades eingereicht habe und dass ich nicht bereits diese oder eine gleichartige Doktorprüfung endgültig nicht bestanden habe.

Zudem erkläre ich, dass ich die Hilfe von gewerblichen Promotionsberatern bzw. Promotionsvermittlern oder ähnlichen Dienstleistern weder bisher in Anspruch genommen habe noch künftig in Anspruch nehmen werde.

Weiterhin erkläre ich mich einverstanden, dass die elektronische Fassung der Dissertation unter Wahrung meiner Urheberrechte und des Datenschutzes einer gesonderten Überprüfung unterzogen werden kann, sowie dass bei Verdacht wissenschaftlichen Fehlverhaltens Ermittlungen durch universitätsinterne Organe der wissenschaftlichen Selbstkontrolle stattfinden können.

Ort, Datum	(Matthias Stöcklein)