An Optimal Control Approach to Implant Shape Design

Modeling, Analysis and Numerics

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

vorgelegt von
Lars Lubkoll aus Erlangen

Contents

Abstract

Facial trauma or congenital malformation of bones of the skull may degrade both skeletal integrity as well as the esthetic appearance. For the attending surgeon a prediction of the esthetic outcome of a bone replacement or augmentation implant insertion is challenging. Therefore, it would be advantageous if we were able to compute an implant shape from a given desired outcome. This task presents the main focus of this thesis. Besides the development of a model for the implant shape design problem, this work is concerned with the efficient solution and optimization of realistic models. This includes recent material laws for different soft tissue types as well as complex geometries attained from medical image data.

The implant shape design problem can be described as an optimal control problem with constraints given by the necessary optimality conditions in polyconvex hyperelasticity with nonlinear pressure-type boundary conditions. Important theoretical results, such as existence of solutions and higher regularity, are currently not available for such problems. Based on the existence result for polyconvex materials laws of Ball [15], existence of solutions of the nonconvex optimal control problem is proven for the case of a simpler Neumann boundary condition.

Due to the "impossible convexity" and the high nonlinearity of hyperelastic material laws the numerical solution of the arising problems is difficult. In this regard, an affine covariant composite step method for nonconvex, equality constrained optimization is presented. The corresponding globalization strategy is based on the affine covariant Newton method for underdetermined systems of Deuflhard [76] and cubic regularization methods for unconstrained optimization problems [277].

The linear systems arising from the discretization of constrained optimization problems are described by saddle point matrices. The efficient solution of these equality systems by conjugate gradient methods for convex and nonconvex problems is discussed. Moreover, an error estimator that fits into the affine covariant setting is presented.

The presented composite step method was implemented in the C++ finite element library KASKADE 7 [114]. The performance of the algorithm is demonstrated on several examples. Next to simple optimization problems, with admissible set given through models of linear and nonlinear heat transfer, we give four examples with nonconvex, hyperelastic constraints.

Zusammenfassung

Traumata und kongenitale Fehlbildungen der Schädelknochen können sowohl die Integrität des Skeletts also auch das ästhetische Erscheinungsbild beeinträchtigen. Für den behandelnden Chirurgen ist die Vorhersage der ästhetischen Folgen des Einsatzes eines Knochenersatz- oder Augmentationsimplantats schwierig. Aus diesem Grund wäre es von Vorteil Implantatformen auf Grundlage eines gewünschten Ergebnisses zu berechnen zu können. Diese Fragestellung steht im Fokus dieser Arbeit. Neben der Herleitung eines Modells für das Implantatdesignproblem wird die effiziente numerische Lösung und Optimierung für realistische Problemstellungen behandelt. Dazu gehören aktuelle Materialbeschreibungen sowie komplexe Geometrien welche aus medizinischen Bilddaten gewonnen wurden.

Das Implantatdesignproblem kann als Optimalsteuerungsproblem modelliert werden, mit Nebenbedingungen gegeben durch die notwendigen Optimalitätsbedingungen der polykonvexen Hyperelastizität mit Druckrandbedingungen. Für diese Probleme sind wichtige theoretische Ergebnisse, wie Existenz von Lösungen oder höhere Regularität, zur Zeit nicht verfügbar. Für den Fall einfacherer Neumannrandbedingungen wird, basierend auf Balls Existenzresultat für polykonvexe Materialgesetze [15], die Existenz von Lösungen des nichtkonvexen Optimalsteuerungsproblem gezeigt.

Auf Grund der "unmöglichen Konvexität" und der starken Nichtlinearität hyperelastischer Materialgesetze ist die numerische Lösung der auftretenden Probleme schwierig. Hierfür wird eine affin kovariante "composite step" Methode vorgestellt. Die zugehörige Globalisierungsstrategie basiert auf dem affin kovariante Newtonverfahren für unterbestimmte Systeme von Deuflhard [76] und kubischen Regularisierungsmethoden für unbeschränkte Optimierung [277].

Die linearen Gleichungssysteme, welche durch die Diskretisierung des beschränkten Optimierungsproblems entstehen, werden durch Sattelpunktmatrizen beschrieben. Die effiziente Lösung dieser Gleichungsysteme mittels konjugierter Gradientenverfahren für konvexe und nichtkonvexe Probleme wird diskutiert. Darüber hinaus wird ein Fehlerschätzer, der in den affin kovarianten Rahmen passt, vorgestellt.

Das vorgestellte "composite step"-Verfahren wurde in der C++-Finite-Elemente-Bibliothek Kaskade 7 [114] implementiert. Das Verhalten des Algorithmus wird anhand verschiedener Beispiele demonstriert. Neben einfachen Optimierungsproblemen, deren zulässige Menge wir durch Modelle der linearen und nichtlinearen Wärmeleitung beschreiben, werden vier Beispiele mit nichtkonvexen, hyperelastischen Nebenbedingungen vorgestellt.

Introduction

Computer-assisted therapy approaches are a valuable tool for improving quality and reducing costs of many therapeutic interventions. They offer new possibilities to physicians regarding education, training, communication with patients and in preoperative decision-making. Particularly in the field of patient-specific therapies there lies a high potential.

In this thesis we focus on implant shape design in the facial area where one is mainly concerned with two requirements. The first is restoration of functionality such as skeletal integrity. The second is an unobtrusive esthetic outcome, cf. [163, 282]. The latter is particularly difficult to realize manually. The aim of this thesis is to demonstrate the applicability of modern mathematics in the development of technical assistance tools that support the attending surgeons in the design of implants.

Currently therapy planning is largely based on general medical guidelines and statistical analysis. Recognizing individual patient-specific information on, amongst others, anatomy, physiology, metabolism and considering it for individual treatments is expected to strongly improve therapeutic outcomes [163, 282].

In order to avoid siloed solutions and redundancies a unified IT-infrastructure, such as the "Therapy Imaging and Model Management System" (TIMMS), as proposed by Lemke and Berliner [173] is mandatory. It brings together medical imaging devices and bio-sensors with modeling and simulation tools, visualization, technical intervention and validation. This requires the standardization of the information exchange between a large number of tools, which are partly new in therapy planning. Until now, such an infrastructure has not yet been put to practice. Nonetheless it is certainly required in the long-term and serves as a guideline for current research.

Probably the biggest step in this direction was the standardization of medical image data in the DICOM (Digital Imaging and Communications in Medicine) standard [1]. On its basis several technical assistance tools, in particular regarding patient-specific surgery planning, have been developed. Their utilization in clinical applications is termed Image Guided Therapy (IGT). For example, three dimensional visualizations of bones and soft tissues can facilitate the planning of surgical interventions [93, 71] and help in the communication with patients.

Workflow in the development of model-based therapies

Model-Based Therapies (MBT) are the consequent extensions of IGT. These are within the focus of TIMMS and additionally incorporate morphological, functional

and dynamic data to generate a patient-specific model [35, 184]. Researchers from several different fields are trying to overcome the manifold obstacles that arise in the development of technical tools assisting in therapy. For the particular case of surgery planning the most important steps are sketched in Fig. 1.

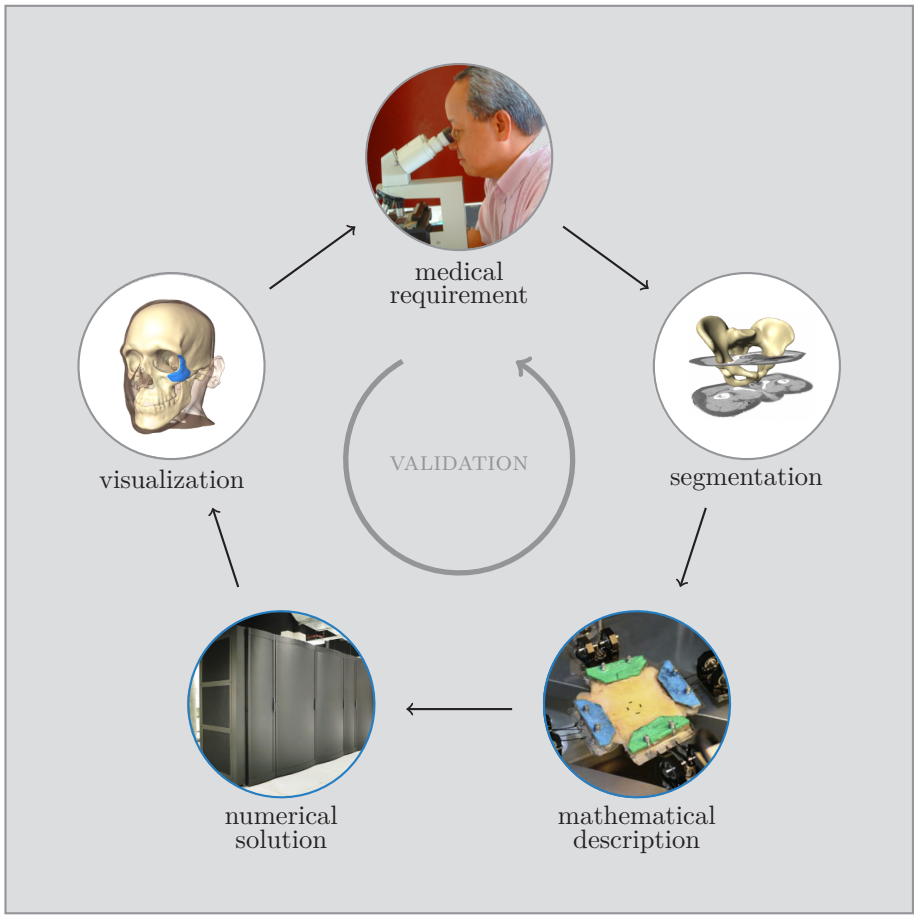


Figure 1.: Workflow for the development of model based therapies.
In the focus of this thesis are the steps that are marked by blue image borders: the derivation of mathematical models and their numerical solution.

1. A particular *medical requirement,* that could be solved automatically, must be identified and specified by the surgeon. In this phase close cooperation between physicians, biomechanics and applied mathematicians allows to balance clinical requirements with technical feasibility and to identify possibly occurring difficulties.

2. For reasonable patient-specific computations accurate descriptions of the considered geometry are required. Therefore three-dimensional models must be extracted from medical image data. This is referred to as the *segmentation* step. These descriptions should not only permit the separation of bony tissue from softer biological soft tissue, but also to distinguish the latter according

to mechanical properties and tissue types. Ideally information on fiber orientation, thickness and structure should also be extracted. These include muscle fibers as well as fibers embedded in the soft tissue's ground substance.

3. We shall need mathematical descriptions of the quantities of interest for different tissue types, and possibly the relation to their environment. Regarding the task of implant shape design the mechanical properties are of main interest[1]. Their description has to incorporate different types of tissues, depending on their microstructural characteristics, as well as varying material parameters in the tissue type itself. Thus a model for the specific medical task needs to be derived based on the available image data and material descriptions.
4. Methods for the *numerical solution* of this model need to be developed and implemented. Due to the geometric complexity of biological soft tissues relatively large problems are prone to arise. Highly efficient nonlinear solvers are needed for their solution. Moreover, for medical applications, reliability of the computation is mandatory.
5. In order to allow physicians, and possibly also patients, the interpretation of the numerical solution interactive *visualization* tools are required. Preferably these should allow the physician to modify the proposed solution, thus introducing his additional knowledge and experience into the planning process.
6. Eventually the most important phase is concerned with the *validation* of the models, cf. Lemke and Berliner [173]. This includes the validation of the models used for the description of the arising subproblems as well as the overall procedure and the therapeutical outcome after each medical intervention.

The segmentation and visualization steps mainly involve the application of well-developed mathematical tools. In contrast, the mathematical description of the quantities of interest and their numerical computation still lead to many open mathematical questions. In the context of implant shape design we are concerned with the deformation of the soft tissue as a consequence of implant insertion. Thus, we are concerned with the steps third and fourth step of the above depicted workflow. Before going into more detail the treatment of the other steps is addressed as well as an overview of previous work on computer assisted facial surgery.

Towards implant shape design

The identification of the design of implant shapes as medical requirement was realized prior to the work on this thesis within the context of the DFG research project Matheon A17. Regarding segmentation and visualization we rely on the expertise of the research group "Medical Planning" at the Zuse-Institute Berlin. The identification of bony tissue from computer tomography (CT) data is relatively well

[1] When implant shapes can be predicted with sufficient accuracy, another important feature is the growth of soft tissues [264, 288].

understood. Only in the presence of foreign material, such as bone screws, this is difficult. The differentiation of soft tissue types evokes the need of additional information attained via magnetic resonance tomography (MRT). In an ideal world both imaging devices would be applied simultaneously.

For the automated segmentation of different soft tissues statistical based methods are promising, cf. Kainmüller et al. [155, 156]. However, currently the segmentation of soft tissues is realized with the help of time-consuming manual intervention by experts [26]. Therefore patient-specific implant design is in general realized during the operation or preoperatively, either from three-dimensional models of the patients bone structure [71, 93, 211] or from three-dimensional models only distinguishing between bones and soft tissue [163]. Consequently the use of segmentation and visualization tools is mostly restricted to the setting of image guided therapy.

Related work

Recently first steps towards model-based therapy (planning) have been put to practice [163, 282, 283]. Regarding the prediction of the esthetic outcome of facial surgeries several approaches have been investigated. Partially these are already applied to assist in surgery planning [282]. A framework for the whole workflow from the problem specification by the surgeon over image segmentation and modeling to the computation of tissue displacements (including different facial expressions) and visualization has been proposed by Koch [163]. The mechanical behavior of soft tissues is described either by a mass-spring model or the model from linearized elasticity, Hooke's model. The tool-chain proposed by Schmidt et al. [228] follows a similar direction. There the focus is on assistance in osteotomy, which is the simulation of cutting and repositioning of bones. Also first attempts to incorporate nonlinear compressible neo-Hookean models have been realized by means of a homotopy approach. Osteotomy has also been investigated by Zachow [282]. Again the whole tool-chain from image segmentation to modeling, computation of tissue displacements and visualization is considered. This approach has been applied in more than 30 clinical cases [283]. In addition the modeling of facial expressions has been investigated, cf. Gladilin [108], Gladilin et al. [109]. Tissue growth was incorporated by Vandewalle et al. [264]. This growth can be triggered by implant insertions or bone relocations, in particular if the induced strains exceed the physiological limits [253, 288].

A highly accurate patient-specific model for the whole face has been created by Barbarino et al. [26], Mazza and Barbarino [183]. Their model includes parts of the mimic musculature as well as fat and skin tissue which were all described with nonlinear isotropic models.

From another point of view, not related to biomechanics, the optimization of elastic materials has recently been investigated in the PhD-thesis of Günnel [120].

Focus

The proposed tool-chains all focus on the solution of the forward problem of computing a soft tissue shape corresponding to a particular implant or bone repositioning. This approach asserts that reasonable implant shapes, that only need slight manual adjustments, are available. This may not always be the case. Especially in cases of large congenital deformations or severe traumata in the face, besides restoration of functionality, the esthetic outcome of an implant insertion is important [163, 282] but difficult to predict. Due to the complex mechanical behavior of biological soft tissues the estimation of its deformation is challenging. This is further complicated in the case of severe traumata. These often occur together with irreversible destruction of muscle tissue and scarring [59, 65], where both effects significantly alter a tissues mechanical properties.

The main focus of this thesis is to extend the previous approaches to the solution of the inverse problem of determining an implant shape from a given desired esthetic outcome. In addition we aim at incorporating recent state-of-the-art material laws. These laws must be nonconvex and mostly exhibit complex phenomena such as anisotropy and exponential growth of the elastic energy.

This requires to solve two bigger subproblems. First a suitable model for the implant shape design problem must be derived. For physically reasonable material descriptions only weak theoretical results are available [17]. Therefore the derivation of models may only be realized on a formal basis and rigorous theoretical results for these models are largely out of reach. Nonetheless, analyzing the problem from point-wise and function-space perspectives gives insights that help in modeling and the development of algorithms.

Second we need an algorithm that is able to solve these problems. In particular due to the complex models for biological soft tissues we need to develop an algorithm that captures significant parts of the underlying problem structure. For this an affine covariant composite step method for equality constrained optimization is developed, with particular focus on PDE-constraints and optimal control problems.

Outline

Regarding the mathematical description of the implant shape design problem we need some prerequisites from hyperelasticity. The basic setting will be introduced in Chap. 1. For a deformation φ of a domain Ω we denote the local stored energy density by W. The corresponding elastic energy stored in the material, the *strain energy*, is given via

$$\mathcal{E}^{\mathrm{str}}(\varphi) = \int_\Omega W(\nabla\varphi)\,\mathrm{d}\mu.$$

Then, hyperelastic theory leads to an optimization problem of the form

$$\min_{\substack{\varphi\in \mathrm{W}^{1,\mathrm{p}}(\Omega;\mathbb{R}^3),\\ \det(\nabla\varphi)>0}} \mathcal{E}(\varphi, g) := \mathcal{E}^{\mathrm{str}}(\varphi) + \mathcal{E}^{\mathrm{ext}}(\varphi, g), \tag{1}$$

where $\mathcal{E}^{\mathrm{ext}}$ is the energy associated with external forces g. Simple examples illustrate that the stored energy function W can not be convex (Sec. 1.1). Thus a more general setting is required. In order to elegantly introduce the suitable generalized convexity conditions we follow Pedregal [207, 208] and analyze (1) from the perspective of Young-measures (Sec. 1.2). While this it not necessary for proving existence of minimizers of (1), it nicely reveals the roles played by arguments coming from convex analysis and compactness arguments.

Equipped with a setting for the description of elastic materials a description of the implant design problem will be presented in Chap. 2. In this context, the implant can be interpreted as an obstacle to the elastic soft tissue (Sec. 2.1). If the setting is sufficiently regular the obstacle problem is related to pressure-type boundary conditions (Sec. 2.2). The latter seem to be better accessible numerically and analytically. However, it is unclear how to exactly incorporate these boundary conditions into the hyperelastic setting. If we relax the pressure-type boundary conditions to Neumann boundary conditions and measure the deviation between desired and computed solution with a cost functional J, the task of finding a reasonable shape of an implant can be formulated as bi-level optimization problem (Sec. 2.3):

$$\begin{aligned} &\min J(\varphi, g) \\ \text{subject to} \quad &\varphi \in \operatorname{argmin}_{\psi} \mathcal{E}(\psi, g). \end{aligned} \tag{2}$$

Despite the difficulties with the derivation of analytical results in elasticity theory existence of optimal solutions for (2) can be shown (Sec. 2.4).

If we want to incorporate pressure-type boundary conditions on $\Gamma_{\mathrm{c}} \subset \partial\Omega$ we have to replace the constraint by its first order optimality conditions which leads to an optimization problem with a partial differential equation (PDE) as constraint:

$$\begin{aligned} &\min J(\varphi, g) \\ \text{subject to} \quad &\frac{\partial}{\partial\varphi}\mathcal{E}^{\mathrm{str}}(\varphi, g)v - \int_{\Gamma_{\mathrm{c}}} g\operatorname{cof}(\nabla\varphi)nv \,\mathrm{ds} = 0 \quad \text{for all } v \in \mathrm{W}^{1,p}(\Omega;\mathbb{R}^3). \end{aligned} \tag{3}$$

This formulation has the advantage that insights gained regarding the numerical solution of PDEs and optimal control problems can be exploited in the development of algorithms. In the context of optimization algorithms an essential tool are the first order optimality conditions. We formally derive the Karush-Kuhn-Tucker (KKT) conditions for (3) and shortly discuss its validity (Sec. 2.5).

The solution of problems of the form (3) is still challenging. First because it is a (regularized) inverse problem. Second because the constraints from elasticity are

highly nonlinear. Their particular structure admits only weak theoretical results which interfere with the rigorous derivation of function space oriented numerical algorithms.

In Chap. 3 an *affine covariant composite step method* for the solution of equality constrained optimization problems is proposed under more regular conditions. To treat the competing aims of feasibility and optimality the Lagrange-Newton step is split into normal and tangential step (Sec. 3.1, Sec. 3.2). Globalization is based on the affine covariant Newton method for underdetermined systems of Deuflhard [76] and cubic regularization methods for unconstrained optimization problems, as suggested by Weiser et al. [277] (Sec. 3.3). For this scheme first theoretical results such as finite termination of the inner loops (Sec. 3.4) and transition to fast local convergence (Sec. 3.5) are established.

Following the description of our algorithm we turn to its practical realization in Chap. 4. Concerning the solution of the arising saddle point systems projected preconditioned conjugate gradient methods (PPCG) enjoy several advantageous properties (Sec. 4.1). In particular termination criteria that fit into the covariant setting are available, cf. [11, 247]. Moreover conjugate gradient methods yield descent directions, in contrast to other Krylov solvers such as MINRES or GMRES. PPCG methods are discussed for both convex (Sec. 4.2) and nonconvex (Sec. 4.3) problems. Furthermore, we introduce a hierarchical error estimator that fits into the chosen covariant setting (Sec. 4.4) and shortly discuss the approximation of the involved operators (Sec. 4.5).

Before turning to numerical examples we need specifications of the stored energy function for the particular tissue types. These are introduced in Chap. 5. state-of-the-art models are mostly derived within the framework of fiber-reinforced materials (Sec. 5.1). These are used to describe the mechanical behavior of biological soft tissues with respect to tensile forces (Sec. 5.2). Since most biological soft tissues are considered to be slightly compressible these models are augmented by suitable descriptions for volumetric deformations (Sec. 5.3). A particular difficulty in applications is the determination of patient-specific, spatially localized material parameters (Sec. 5.4).

In Chap. 6 numerical results for different test problems will be presented. We start with a simple two-dimensional model of nonlinear heat transfer. Then we will give two examples of complex anisotropic material laws on simple geometries. Finally two examples on real patient geometries are presented. The corresponding geometric data does neither contain information on fiber directions, necessary for the definition of anisotropic models, nor on different material types. For this reason a homogeneous, isotropic model will be employed in the last two examples.

This thesis closes with a discussion of the most relevant achievements and an outline some of the most important open theoretical, algorithmic and biomechanical questions, as well as some related to the establishment of TIMMS.

In order to increase the readability of this thesis different parts are, except for few exceptions, self-contained. The main blocks are the first two chapters, describing the mathematical framework and the employed model, and the third and fourth chapter, describing the used algorithm and its realization. The fifth chapter on biological soft tissues can be roughly understood without further knowledge, but is easier to understand with some background in elasticity theory. Eventually, to understand the numerical examples, that are presented in the sixth chapter, all previous chapters are relevant.

1. Elastic materials and the direct method

Descriptions of biological soft tissues are mostly based on the theory of nonlinear elasticity. The general setting, which is based on elementary physical considerations, is introduced in Sec. 1.1. It leads to descriptions of elastic deformations φ as minimizers of material specific energy functionals $\mathcal{E}$. These minimizers can not be unique, thus ruling out strict convexity of $\mathcal{E}$. Therefore, many important questions, such as well-definedness of the corresponding first order optimality conditions, are still open, cf. Ball [18, 19]. However, under dead load forces g acting on a measurable set $D \subset \Omega$, resp. $D \subset \partial\Omega$, with $|D| > 0$, existence of minimizers can be shown for a large class of problems, namely those whose energy functional is given via

$$\mathcal{E}(\varphi) = \int_\Omega W(\nabla\varphi(x))\,\mathrm{dx} - \int_D g(x)\varphi(x)\,\mathrm{dx},$$

where the *stored energy function* W can be written as convex function of the minors of $\nabla\varphi$. The latter property was introduced in Ball [15] under the name of *polyconvexity.* It can be motivated elegantly by the analysis of the optimization problem

$$\min_\varphi \mathcal{E}(\varphi)$$

with the *direct method* of the calculus of variations and *parametrized measures.* This will be the content of Sec. 1.2. Eventually, additional conditions under which the corresponding first order optimality conditions are well defined in $\mathrm{W}^{1,\infty}(\Omega)$ are discussed in Sec. 1.3.

Conventions. In order to increase readability and to not overload this presentation with technical details some commonly used conventions are adopted.

- **Bold** letters indicate a definition, whereas *italic* letters are used for emphasis.
- The soft tissue volume is denoted by $\Omega \subset \mathbb{R}^3$, which is assumed to be a bounded domain with Lipschitz boundary.
- The space of all $m \times n$-matrices is denoted by $\mathbb{M}^{m,n}$ and we set $\mathbb{M}^n := \mathbb{M}^{n,n}$. The space of symmetric $n \times n$-matrices is denoted by $\mathbb{S}^n$ and the space of orthogonal $n \times n$-matrices by $\mathbb{O}^n$. The subscript "+" denotes subsets of matrices with positive determinant, i.e.

$$\mathbb{K}^n_+ := \{F \in \mathbb{K}^n : \ \det(F) > 0\} \quad \text{for } \mathbb{K} = \mathbb{M}, \mathbb{S}, \mathbb{O}$$

and

$$\mathbb{R}_+ := \{t \in \mathbb{R} : \ t \geq 0\}\,.$$

- When extracting a subsequence out of a sequence $\{\varphi_j\}_j$ it will also be denoted by $\{\varphi_j\}_j$.
- Vector- and matrix-valued Sobolev spaces $\mathrm{W}^{1,\mathrm{p}}(\Omega;\mathbb{R}^m)$, resp. $\mathrm{W}^{1,\mathrm{p}}(\Omega;\mathbb{M}^{m,n})$, are written as $\mathrm{W}^{1,\mathrm{p}}(\Omega)$ if the image space can be easily deduced from context or is not relevant. The same applies for Lebesgue spaces $\mathrm{L}^{\mathrm{p}}(\Omega)$.
- With $\mathrm{W}^{1,\mathrm{p}}_0(\Omega)$ we denote *all* Sobolev spaces that incorporate homogeneous Dirichlet boundary conditions on a part of the boundary $\Gamma_{\mathrm{d}} \subset \partial\Omega$ with positive surface measure $|\Gamma_d| > 0$, not only those where $\Gamma_{\mathrm{d}} = \partial\Omega$. The case that Γ_{d} denotes only a part of the boundary is of main interest in this thesis. However, in order to keep focus on the relevant details, in theoretical results homogeneous Dirichlet boundary conditions are often assumed to hold on the whole boundary $\partial\Omega$.
- Subscripting of functions with one of its arguments denotes a partial derivative, i.e. for a function $\mathcal{L}(x,p)$ we have $\mathcal{L}_x(x,p) = \frac{\partial}{\partial x}\mathcal{L}(x,p)$.

1.1. Elasticity

We begin with introducing the basic notation of elasticity in Sec. 1.1.1. In continuum mechanics external forces and internal stresses are related via the *stress principle of Euler and Cauchy*, which is described in Sec. 1.1.2. It admits the formulation of static equilibrium conditions on the deformed domain $\Omega_{\mathrm{def}} = \varphi(\Omega)$. In order to express these equilibrium conditions on the undeformed domain the *Piola transform* is introduced in Sec. 1.1.3. Eventually, we need material specific *constitutive relations* to relate stresses with underlying deformations. In this thesis, we will focus on constitutive relations that can be described via an energy density, the *stored energy function*, see Sec. 1.1.4.

For a more detailed discussion of the mathematical theory of elastic materials the books of Ciarlet [56, 57], Holzapfel [135], Pedregal [208] are suggested to the interested reader. Also of interest are the books of Braess [41], Marsden and Hughes [181], Ogden [201], Sokolnikoff [239], Truesdell and Noll [261].

1.1.1. Kinematics

The domain $\Omega \subset \mathbb{R}^3$, occupied by a body in an equilibrium state, is called **reference configuration**. If forces act on this body it deforms to a new configuration Ω_{def}, determined by the **deformation**

$$\varphi : \Omega \ni x \mapsto \varphi(x) = (\mathrm{id} + u)(x) \in \Omega_{\mathrm{def}},$$

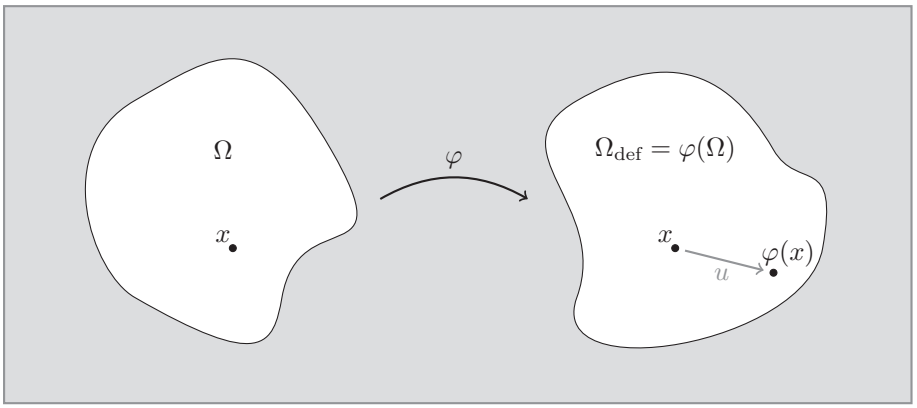


Figure 1.1.1.: Deformation of a domain Ω.

see Fig. 1.1.1. The deviation from the identity is the **displacement** $u = \varphi - \mathrm{id}$.

Next to adequate smoothness assumptions, the deformation φ must satisfy the additional **orientation preservation condition**

$$\det(\nabla\varphi(x)) > 0. \tag{1.1.1}$$

It rules out deformations that admit local self-penetration and guarantees local injectivity for sufficiently smooth deformation φ. Working in Sobolev spaces this argumentation is not valid any more. Nonetheless, the requirement that (1.1.1) holds almost everywhere in Ω will naturally arise in the context of hyperelastic compressible materials (Sec. 1.1.4).

Remark 1.1. For some materials, we can further restrict the orientation preservation condition to $\det(\nabla\varphi(x)) = 1$. In this case we speak of incompressible materials.

An important quantity in elasticity theory is the strain tensor, which describes the change in the length of line segments with respect to the Euclidean norm. Letting $x, x + h \in \Omega$ we compute with Taylor's formula for a smooth deformation φ

$$\|\varphi(x+h) - \varphi(x)\|^2 = \|\nabla\varphi(x)h\|^2 + o(\|h\|^2) = h^T\nabla\varphi(x)^T\nabla\varphi(x)h + o(\|h\|^2).$$

Thus, the change of length in line segments is dominated by the **(left) Cauchy-Green strain tensor**[1]

$$C(\nabla\varphi) := \nabla\varphi^T\nabla\varphi = I + \nabla u^T + \nabla u + \nabla u^T\nabla u.$$

Definition 1.2. The scaled deviation of the Cauchy-Green strain tensor from the identity

$$E(\nabla\varphi) := \frac{1}{2}(C(\nabla\varphi) - I) = \frac{1}{2}(\nabla u^T + \nabla u + \nabla u^T\nabla u)$$

is called **strain (tensor)**.

[1] As indicated by the naming there also exists a right Cauchy-Green strain tensor given through $\nabla\varphi\nabla\varphi^T$. Here, we mostly are concerned with the left Cauchy-green strain tensor. Recall that if $\nabla\varphi$ is unitarily diagonalizable it is normal and thus the left and right Cauchy-Green strain tensors coincide.

In the presence of small displacement gradients we may neglect the nonlinear part and replace the strain tensor by the **symmetric gradient**

$$\nabla^s u = \frac{1}{2}(\nabla u^T + \nabla u).$$

In contrast to the nonlinear strain tensor the symmetric gradient is not independent of the chosen coordinate system, i.e. it is not *frame-indifferent*, cf. Thm. 1.14. Thus it is already inadequate for small rotations. In this case the linearized theory yields non-physical "phantom" stresses [18]. The neglected term $\frac{1}{2}\nabla u^T \nabla u$ is the **geometric nonlinearity** and its incorporation is mandatory to derive reasonable models in the presence of large displacement gradients, regardless of the considered material. Material specific nonlinear behavior is captured by the *constitutive nonlinearity*, which will be introduced in Sec. 1.1.4.

1.1.2. Equilibrium conditions

Before describing further details of elasticity theory we recapitulate the framework for the description of *static equilibria*. It admits to relate external forces with the induced stresses in the material. A basic assumption in mechanics is that all acting forces can be partitioned into volume and surface forces. Honoring the main protagonists in the early study of static equilibria in the modern western world, static equilibria are defined in the **stress principle of Euler and Cauchy**.

Axiom 1.3 (Stress principle of Euler and Cauchy)**.** *Consider a body occupying a deformed region $\bar{\Omega}_{\text{def}}$, subjected to a body force*

$$f\colon \Omega_{\text{def}} \to \mathbb{R}^3$$

and a surface force

$$g\colon \Gamma_1^{\text{def}} \to \mathbb{R}^3,$$

where Γ_1^{def} is some measurable part of the boundary of Ω_{def}. Then there exists a vector field

$$t\colon \bar{\Omega}_{\text{def}} \times S \to \mathbb{R}^3 \qquad S := \{x \in \mathbb{R}^3 : |v| = 1\}$$

*called **Cauchy's stress vector** such that:*

1. For any subdomain $A \subseteq \bar{\Omega}_{\text{def}}$ and at any point $x \in \Gamma_1^{\text{def}} \cap \partial A$ where the unit outer normal vector n exists, $t(x, n) = g(x)$ holds.

   a) **Axiom of force balance**: For any subdomain $A \subseteq \bar{\Omega}_{\text{def}}$ holds

$$\int_A f(x)\,\mathrm{dx} + \int_{\partial A} t(x, n)\,\mathrm{ds} = 0 \tag{1.1.2}$$

b) **Axiom of moment balance**: For any subdomain $A \subseteq \bar{\Omega}_{\text{def}}$ holds

$$\int_A x \times f(x)\,\mathrm{dx} + \int_{\partial A} x \times t(x,n)\,\mathrm{ds} = 0 \tag{1.1.3}$$

where $\times$ denotes the vector/cross product.

As a consequence we get "one of the most important results in continuum mechanics" [56, p. 62], **Cauchy's theorem**.

Theorem 1.4 (Cauchy's theorem)**.** *Let the assumptions of Axiom 1.3 hold. Further assume that* $t(\cdot,n) \in \mathrm{C}^1(\Omega_{\text{def}};\mathbb{R}^3)$, $t(x,\cdot) \in \mathrm{C}(S;\mathbb{R}^3)$ *and* $f \in \mathrm{C}(\Omega_{\text{def}};\mathbb{R}^3)$*. Then there exists a symmetric tensor field* $T_{\text{def}} \in C^1(\Omega_{\text{def}};\mathbb{S}^3)$ *such that*

$$t(x,n) = T_{\text{def}}(x)n \qquad \text{for all } x \in \Omega_{\text{def}} \text{ and } n \in S, \tag{1.1.4}$$

$$\operatorname{div}(T_{\text{def}}(x)) + f(x) = 0 \qquad \text{for all } x \in \Omega_{\text{def}}, \tag{1.1.5}$$

$$T_{\text{def}}(x) = T_{\text{def}}^T(x) \qquad \text{for all } x \in \Omega_{\text{def}}. \tag{1.1.6}$$

The tensor $T_{\text{def}}(x)$ *is called* ***Cauchy stress tensor*** *at* $x \in \Omega_{\text{def}}$*.*

Proof. See [56, p. 63]. □

Remark 1.5. Here, the main point is that the stress vector t is linear in its second argument. Then, using Gauß' integral formula, *(1.1.2)* can be written as

$$\begin{aligned} \int_A f(x)\,\mathrm{dx} + \int_{\partial A} T_{\text{def}}(x)n\,\mathrm{ds} &= 0 \\ \Leftrightarrow \int_A \big[f(x) + \operatorname{div}(T_{\text{def}}(x))\big]\,\mathrm{dx} &= 0 \end{aligned}$$

Holding for every subdomain $A \subseteq \bar{\Omega}_{\text{def}}$ this leads to the differential equation *(1.1.5)*. The symmetry property *(1.1.6)* follows from *(1.1.3)*.

1.1.3. The Piola transform

The above equilibrium conditions are formulated on the unknown deformed domain Ω_{def}. Thus, we need a mapping that admits the expression of these equilibrium conditions on the undeformed domain Ω. This mapping and its properties will be discussed in this subsection.

First, recall the definition of the cofactor matrix from linear algebra.

Definition 1.6. Let $A = (a_{ij})_{i,j=1,\dots n} \in \mathbb{M}^n$, $n > 0$ be a $n \times n$-matrix and denote by $A_{ij}^{\#} \in \mathbb{M}^{n-1}$ the matrix that results when deleting the i-th row and j-th column from A. The scalars $(-1)^{i+j}\det(A_{ij}^{\#})$ are called the **cofactors** of A. The **cofactor matrix** is given by

$$\operatorname{cof}(A) = \Big((-1)^{i+j}\det(A_{ij}^{\#})\Big)_{i,j=1,\dots,n}.$$

Its transpose $\operatorname{adj}(A) = \operatorname{cof}(A)^T$ is called **adjugate matrix** of A.

Remark 1.7.

- Denoting the j-th column of a matrix A by A_j, the cofactors are related to the determinant via Laplace's formula as

$$\det(A) = \sum_{i=1}^{n} a_{ij}(-1)^{i+j} \det(A_{ij}^{\#}) = A_j^T \operatorname{cof}(A)_j,$$

  or in terms of the adjugate matrix, $\det(A) = \operatorname{adj}(A)_j A_j$. The derivative of the determinant in a direction δA is

$$\det{}'(A)\delta A = \operatorname{cof}(A) : \delta A = \operatorname{tr}\Big(\operatorname{cof}(A)^T \delta A\Big) = \operatorname{tr}\Big(\operatorname{adj}(A)\delta A\Big),$$

  where $\cdot : \cdot$ denotes the scalar product in $\mathbb{M}^{m,n}$, given through

$$A : B = \operatorname{tr}\left(A^T B\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}B_{ij}.$$

- If A is invertible, we have the identities

$$\operatorname{cof}(A) = A^{-T}\det(A), \quad \text{resp.} \quad \operatorname{adj}(A) = A^{-1}\det(A).$$

- The adjugate of the deformation gradient can be interpreted as a local measure for changes in the area of surfaces [56, 57].

Definition 1.8. The **Piola transform** $P_T : \bar{\Omega} \to \mathbb{M}^3$ of a 3×3-tensor $T : \bar{\Omega}_{\text{def}} \to \mathbb{M}^3$ is defined via

$$\begin{aligned} P_T(x) &:= \det(\nabla\varphi(x))T(\varphi(x))(\nabla\varphi(x))^{-T} \\ &= T(\varphi(x))\operatorname{cof}(\nabla\varphi(x)), \end{aligned}$$

for almost every $x \in \bar{\Omega}$ and all $\varphi \in \mathrm{W}^{1,\mathrm{p}}(\Omega)$, with $p > 1$, such that $\nabla\varphi$ is almost everywhere invertible.

We summarize the properties that are necessary to transform the relevant quantities.

Theorem 1.9. *For the Piola transform P_T holds:*

1. $\operatorname{div}\Big(P_T(x)\Big) = \det(\nabla\varphi)\operatorname{div}\Big(T(\varphi(x))\Big)$ for all $x \in \Omega$.
2. $P_T(x)n\,\mathrm{ds} = T(\varphi(x))n_{\text{def}}\,\mathrm{ds}_{\text{def}}$ for all $x \in \Omega$,
   where ds_{def} and ds are surface elements and n_{def} as well as n are the unit outer normals of $\partial\Omega_{\text{def}}$ resp. $\partial\Omega$.
3. The surface elements are related via

$$\det(\nabla\varphi(x))\Big|\nabla\varphi(x)^{-T}n\Big|\,\mathrm{ds} = \Big|\operatorname{cof}(\nabla\varphi(x))n\Big|\,\mathrm{ds} = \mathrm{ds}_{\text{def}}.$$

Proof. See [56, Thm. 1.7-1]. □

The Piola transform of the Cauchy stress tensor T_{def} at some point $x_\varphi = \varphi(x) \in \Omega_{\text{def}}$,

$$\hat{\sigma}(x) := \det(\nabla\varphi(x)) T_{\text{def}}(x_\varphi)(\nabla\varphi(x))^{-T} = T_{\text{def}}(x_\varphi)\text{cof}(\nabla\varphi(x)) \tag{1.1.7}$$

is called **first Piola-Kirchhoff stress tensor** and is, in general, not symmetric. It may be symmetrized which leads to the **second Piola-Kirchhoff stress tensor**

$$\hat{\Sigma}(x) := \det(\nabla\varphi(x))(\nabla\varphi(x))^{-1} T_{\text{def}}(x_\varphi)(\nabla\varphi(x))^{-T} = (\nabla\varphi(x))^{-1}\hat{\sigma}(x).$$

Hyperelastic materials, which will be introduced in the next subsection, naturally lead to the first Piola-Kirchhoff stress tensor. Thus we do not consider the second and refer the interested reader to [56].

1.1.4. Constitutive equations

The properties given in Cauchy's theorem are not sufficient for the determination the occurring stresses in the presence of body and/or surface forces. This is not surprising since up to now any material specific information is missing. This requires to relate the deformation with the corresponding stresses. More precisely, for elastic materials it is assumed that the Cauchy stress tensor only depends on the position x and the deformation gradient $\nabla\varphi(x)$.

Definition 1.10. A material is called **elastic** if there exists a mapping $\hat{T}\colon \Omega\times\mathbb{M}^3_+ \to \mathbb{S}^3_+$ such that

$$T_{\text{def}}(x_\varphi) = \hat{T}(x, \nabla\varphi(x)) \quad \text{for all } x_\varphi = \varphi(x) \in \Omega_{\text{def}}. \tag{1.1.8}$$

The mapping $\hat{T}$ is called the **response function** of T_{def}. The relation (1.1.8) is called **constitutive relation** or **material law**.

Remark 1.11. This definition is a simplifying assumption that facilitates the mathematical treatment with existing tools. Deriving effective stress tensors from descriptions of the underlying micro-structure by means of mathematical homogenization will in general lead to more complex relations, possibly involving nonlocal effects [251]. See also [56, Sec. 3.1] for comments on cases where the above definition is not adequate and references regarding *nonlocal elasticity*.

As a consequence we get the existence of a response function for the first Piola-Kirchhoff stress tensor

$$\hat{\sigma}(x) = \sigma(x, \nabla\varphi) \quad \text{with} \quad \sigma(x, \nabla\varphi) = \hat{T}(x, \nabla\varphi)\text{cof}(\nabla\varphi).$$

The explicit dependence on the deformation φ in formulations in terms of the response function are more expressive than the use of formulations based on the Piola-Kirchhoff stress tensors. Thus, in the following we only use the response function σ.

We further specify the relation between deformation and induced stresses. The second law of thermodynamics does not allow us to build a perpetual motion machine. Therefore, the work in closed processes should be non-negative and lengthy computations [122, Sec. 28] lead to a characterization of suitable stored energy functions.

Definition 1.12. An elastic material is called **hyperelastic** if there exists a function $W\colon \bar{\Omega} \times \mathbb{M}^3_+ \to \mathbb{R}$, differentiable in its second argument for each $x \in \bar{\Omega}$, such that

$$\sigma(x, F) = \frac{\partial W}{\partial F}(x, F) \quad \text{for all } x \in \Omega \text{ and all } F \in \mathbb{M}^3_+,$$

where σ is the response function of the first Piola-Kirchhoff stress tensor. W is an energy density and called **stored energy function**.

For hyperelastic materials we also call the relation $(x, \varphi) \mapsto W(x, \nabla\varphi(x))$ *constitutive relation* resp. *material law*, since this mapping uniquely determines (1.1.8).

Theorem 1.13. *An elastic material is hyperelastic if and only if the work is non-negative in closed processes.*

Proof. See [122, p. 186]. □

A particular feature of hyperelastic materials is the fact that in the presence of dead load forces g the associated deformation φ_g is a minimizer of the energy functional

$$\mathcal{E}(\varphi, g) = \mathcal{E}^{\text{str}}(\varphi) - \mathcal{E}^{\text{ext}}(\varphi, g).$$

For volume or Neumann boundary forces g the corresponding energy functional is $\mathcal{E}^{\text{ext}}(\varphi, g) = \int_{D_g} \varphi(x) g(x)\,\mathrm{dx}$, with $D_g = \Omega$, resp. $D_g = \Gamma_g \subset \partial\Omega$.

The specific form of the stored energy function is further restricted by the assumption of independence of the chosen coordinate system. As the following theorem shows, this is equivalent to the requirement that the stored energy function can be expressed in terms of the *Cauchy-Green strain tensor* instead of the deformation gradient.

Theorem 1.14. *The stored energy function $W\colon \bar{\Omega} \times \mathbb{M}^3_+ \to \mathbb{R}$ is called **frame-indifferent** if and only if one of the following equivalent conditions holds:*

- For all $x \in \bar{\Omega}$, all $F \in \mathbb{M}^3_+$ and all orientation preserving orthogonal matrices $Q \in \mathbb{O}^3_+$ holds
$$W(x, QF) = W(x, F).$$
- There exists a function $\tilde{W}\colon \bar{\Omega} \times \mathbb{S}^3_+ \to \mathbb{R}$ such that
$$W(x, F) = \tilde{W}(x, F^T F) \tag{1.1.9}$$
for all $x \in \bar{\Omega}$ and all $F \in \mathbb{M}^3_+$.

Proof. See [56, Thm. 4.2-1]. □

Remark 1.15.

- In order to get a proper splitting of the arising nonlinearities we call the nonlinearity of $\tilde{W}$ **constitutive nonlinearity**. The nonlinearity of W then comprises both the *geometric* and the *constitutive nonlinearity*.
- In order to guarantee frame-indifference, material laws are typically formulated in terms of *invariants*. For isotropic materials these are the principal or modified principal invariants (Def. 5.3 and Def. 5.6). Anisotropic material laws are often similar to isotropic laws, replacing or extending the (modified) principal invariants by (modified) mixed invariants (Def. 5.9). This strategy is referred to as *isotropization*. It is explained in more detail in Chap. 5.

An essential consequence of frame-indifference is the fact that it rules out convex stored energy functions W. Another requirement that admits an even simpler proof of the "impossible convexity" of W is related to the limit behavior for "extreme" strains, i.e. the case that for $x \in \Omega$ one of the eigenvalues $\lambda_i = \lambda_i(C)$ $i = 1, 2, 3$ of the (left) Cauchy-Green strain tensor $C = \nabla\varphi(x)^T \nabla\varphi(x)$ tends to 0 or ∞. W.l.o.g. let this eigenvalue be λ_1 and let $\lambda_2, \lambda_3 \in [c, d]$ for constants $c > 0$ and $d < \infty$. Then we have the equivalences

$$\lambda_1 \searrow 0 \quad \Leftrightarrow \quad \det(\nabla\varphi(x)) \searrow 0, \tag{1.1.10}$$

$$\lambda_1 \to \infty \quad \Leftrightarrow \quad \|\nabla\varphi(x)\| \to \infty, \tag{1.1.11}$$

$$\lambda_1 \to \infty \quad \Leftrightarrow \quad \|\mathrm{cof}(\nabla\varphi(x))\| \to \infty, \tag{1.1.12}$$

$$\lambda_1 \to \infty \quad \Leftrightarrow \quad \|\det(\nabla\varphi(x))\| \to \infty. \tag{1.1.13}$$

Assuming that infinite extensions require infinite energy, we deduce from the last three equivalences the necessity

$$(\|F\|, \|\mathrm{cof}(F)\|, |\det(F)|) \to (\infty, \infty, \infty) \Rightarrow W(x, F) \to \infty \quad \text{for } x \in \Omega \text{ and } F \in \mathbb{M}^3_+$$

as a reasonable condition for large strains. In the sharper form

$$W(x, F) \geq \alpha(\|F\|^p + \|\mathrm{cof}(F)\|^q + |\det(F)|^r) - \beta$$

with positive constants $\alpha > 0$, $p > 0$, $q > 0$, $r > 0$ and $\beta \in \mathbb{R}$ this assumption also provides the necessary coercivity inequality for the proof of existence of minimizers (see Thm. 1.42).

Condition (1.1.10) describes vanishing volumes. From a physical point of view it is reasonable that "an infinite pressure is required in order to annihilate volumes". Since infinite pressure yields an infinite stored energy [56, Sec. 4.6, Ex 4.9], this motivates the condition

$$\lim_{\det(F) \searrow 0} W(x, F) = \infty, \quad \text{for all } x \in \Omega, \tag{1.1.14}$$

for reasonable stored energy functions. Same as frame-indifference it rules out convexity of W.

Theorem 1.16. *Let $x \in \bar{\Omega}$ and*

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

be convex. Then:

1. Condition (1.1.14) cannot hold.

Proof. Noting that $\mathbb{M}^3_+$ is not convex we denote by $W_{\mathrm{co}} : \mathbb{M}^3 \to \mathbb{R}_\infty$ any convex extension of $W(x,\cdot)$ to the convex hull $\mathrm{co}(\mathbb{M}^3_+) = \mathbb{M}^3$. Let $I \in \mathbb{M}^3_+$ be the unit matrix and let

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \in \mathbb{M}^3_+, \; B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \notin \mathbb{M}^3_+.$$

As W_{co} is convex we have

$$\sup_{t\in[0,1]} W_{\mathrm{co}}(I + t(A-I)) = \max\{W_{\mathrm{co}}(I), W_{\mathrm{co}}(A)\} < \infty.$$

However, as

$$\lim_{t\to 1/2} I + t(A-I) = B \notin \mathbb{M}^3_+,$$

and assuming that W satisfies (1.1.14), we get

$$\lim_{t\to 1/2} W_{\mathrm{co}}(I + t(A-I)) = \infty.$$

Thus (1.1.14) cannot hold. □

Remark 1.17. Note that for $\varphi \in \mathrm{W}^{1,\mathrm{p}}(\Omega),\ p < \infty$, condition (1.1.14) implies that

$$\mathcal{E}^{\mathrm{str}}(\varphi) = \int_\Omega W(x, \nabla\varphi)\,\mathrm{dx} = \infty$$

on a dense subset of $\mathrm{W}^{1,\mathrm{p}}(\Omega)$.

1.2. The direct method and Young measures

We have identified convex functions as too restrictive to be of use in nonlinear elasticity. Thus, we need a more general framework in which we seek candidates for stored energy functions.

Inserting a traveling wave solution into the linearized equations of motion for an elastic material, we see that positive wave-speeds can only be guaranteed, if W is **rank-one convex**, i.e.

$$W(\lambda A + (1-\lambda)B) \leq \lambda W(A) + (1-\lambda)W(B),$$

for all $A, B \in \mathbb{M}^3$ such that $\operatorname{rank}(A-B) \leq 1$, cf. [233]. For twice differentiable stored energy functions W this is equivalent to the validity **Legendre-Hadamard condition** [72]

$$W''(A)\,(a \otimes b)^2 \geq 0, \quad \text{for all } a, b \in \mathbb{R}^3$$

for $A \in \mathbb{M}^3$.

Remark 1.18. Considering $A, B \in \mathbb{M}^{m,n}$ such that $A - B = e_i \otimes e_j$ for $1 \leq i \leq m,\ 1 \leq j \in n$ we recognize that a rank one convex function $W\colon \mathbb{M}^{m,n} \supseteq D \to \mathbb{R}_\infty$ is convex in each of its argument's entries and thus is locally Lipschitz continuous in $\operatorname{int}(D)$ [72, Chap. 2, Thm. 2.3].

For scalar problems, rank-one convexity coincides with convexity and the Legendre-Hadamard condition reduces to the positive semi-definiteness of the Hessian[2]. This is not the case for vectorial problems [72, 271]. In the latter case, accepting realistic traveling wave solutions as desirable, we seek a generalized convexity property of the stored energy function that is located somewhere between *convexity* and *rank-one convexity*.

Recall that hyperelastic problems with conservative loads are formulated as minimization problem

$$\bar{\varphi} \in \operatorname{argmin}_{\varphi \in \Phi} \mathcal{E}(\varphi, g),$$

where

$$\mathcal{E}(\varphi, g) = \mathcal{E}^{\mathrm{str}}(\varphi) - \mathcal{E}^{\mathrm{ext}}(\varphi, g)$$

and

$$\Phi := \left\{ \varphi \in \mathrm{W}^{1,\mathrm{p}}(\Omega) :\ \det(\nabla\varphi) \geq 0 \text{ a.e. in } \Omega,\ \varphi = 0 \text{ a.e. on } \Gamma_{\mathrm{d}} \right\}$$

is the admissible set. For volume forces the corresponding energy functional is $\mathcal{E}^{\mathrm{ext}}(\varphi, g) = \int_\Omega \varphi g \,\mathrm{d}\mu$, and for Neumann boundary forces $\mathcal{E}^{\mathrm{ext}}(\varphi, g) = \int_{\Gamma_{\mathrm{c}}} \varphi g \,\mathrm{ds}$, with $\Gamma_{\mathrm{c}} \subset \partial\Omega$.

The strategy of the *direct method of the calculus of variations* is to take an infimizing sequence $\lim_{j\to\infty} \mathcal{E}(\varphi_j, g) = \inf_{\psi\in\Phi} \mathcal{E}(\psi, g)$ and show that we can extract a weakly convergent subsequence $\varphi_j \rightharpoonup \bar{\varphi} \in \Phi$ such that $\mathcal{E}(\bar{\varphi}, g) = \inf_{\psi\in\Phi} \mathcal{E}(\psi, g)$. For both volume and boundary forces we have $\lim_{j\to\infty} \mathcal{E}^{\mathrm{ext}}(\varphi_j, g) = \mathcal{E}^{\mathrm{ext}}(\bar{\varphi}, g)$. For this reason the contribution from $\mathcal{E}^{\mathrm{ext}}$ is not relevant in the derivation of a mathematical setting that admits existence of minimizers for the problem

$$\min_{\varphi\in\Phi} \mathcal{E}(\varphi, g).$$

[2]For this reason, ellipticity, in the sense that the Legendre-Hadamard condition holds, and V-ellipticity, in sense of Tröltzsch [259], are sometimes not properly distinguished in literature.

Hence, in this section we restrict the discussion to $\mathcal{E}^{\mathrm{str}}$. First, we also neglect the orientation preservation condition $\det(\nabla\varphi) > 0$ as well as the boundary conditions and add them as soon as we have identified useful conditions for $\mathcal{E}^{\mathrm{str}}$. Therefore we focus on the problem

$$\min_{\varphi\in \mathrm{W}^{1,\mathrm{p}}(\Omega)} \mathcal{E}^{\mathrm{str}}(\varphi) := \int_\Omega W(x, \nabla\varphi(x))\,\mathrm{dx}. \tag{1.2.1}$$

Here, we mainly follow Pedregal [207] and start with a theorem that summarizes the idea behind this approach.

Theorem 1.19. *Consider the variational principle*

$$\min_{\varphi\in\mathcal{A}} \mathcal{E}^{\mathrm{str}}(\varphi)$$

where

- $\mathcal{A}$ *is a closed, convex subset of a reflexive Banach space* X*,*
- $\mathcal{E}^{\mathrm{str}}$ *is coercive, i.e.*

  $$\mathcal{E}(\varphi) \geq c\|\varphi\|_{\mathrm{X}}$$

  for some positive constant $c > 0$ *or*

  $$\lim_{\|\varphi\|_{\mathrm{X}}\to\infty} \mathcal{E}(\varphi) = \infty,$$

- $\mathcal{E}^{\mathrm{str}}$ *is lower semicontinuous with respect to the weak topology in* X*:*

  $$\mathcal{E}^{\mathrm{str}}(\bar\varphi) \leq \liminf \mathcal{E}^{\mathrm{str}}(\varphi_j) \quad \textit{for } \varphi_j \rightharpoonup \bar\varphi.$$

- *there exists at least one* $\bar\varphi \in \mathcal{A}$ *with* $\mathcal{E}^{\mathrm{str}}(\bar\varphi) < \infty$*.*

Then $\mathcal{E}^{\mathrm{str}}$ *has at least one minimizer in* $\mathcal{A}$*.*

Proof. See [208, Thm. 1.1]. □

Thus, the essential point to be analyzed is the weak lower semicontinuity of $\mathcal{E}$. We neglect the dependence of the stored energy function on the spatial variable and consider

$$\mathcal{E}^{\mathrm{str}}(\varphi) = \int_\Omega W(\nabla\varphi(x))\,\mathrm{dx},$$

where $\varphi \in \mathrm{W}^{1,p}(\Omega, \mathbb{R}^m)$ with $1 < p < \infty$ and

$$W\colon \mathbb{M}^{m,n} \to \mathbb{R}_\infty := \mathbb{R} \cup \{\infty\}\,,\ n \geq 1,\ m \geq 1$$

is continuous. The results of this section stay valid if W additionally depends measurably on the spatial variable x, i.e. if W is a Carathéodory function. As a first step to find suitable generalized convexity conditions for W, we need a representation result for weak limits in $\mathrm{L}^1(\Omega)$.

Theorem 1.20 (Existence of parametrized measures)**.**
Let $Z = \mathbb{R}^m$ or $Z = \mathbb{M}^{m,n}$ and let $z_j : \Omega \to Z$, $j \in \mathbb{N}$ be measurable functions that are bounded in $\mathrm{L}^p(\Omega)$ with $1 \leq p < \infty$, i.e.

$$\sup_j \|z_j\| \leq c_z$$

for some positive constant c_z. Then there exists a subsequence, again denoted by $\{z_j\}_j$, and a family of probability measures $\nu = \{\nu_x\}_{x\in\Omega}$, depending measurably on x, such that for any continuous function

$$W(\lambda)\colon Z \to \mathbb{R}_\infty$$

for which the sequence $\{W(z_j(x))\}_j$ is weakly convergent in $\mathrm{L}^1(\Omega)$ the weak limit is given by

$$W(z_j(x)) \rightharpoonup \bar{W}(x) = \int_Z W(\lambda)\,\mathrm{d}\nu_x(\lambda). \tag{1.2.2}$$

Proof. See [207, Thm. 6.2]. There a more general version is stated, that also admits its application in (Sobolev-)Orlicz spaces, cf. [3, 165]. □

Definition 1.21. Consider the notation of Thm. 1.20.

- The family of probability measures ν is called the **associated parametrized measure** of the (sub-)sequence $\{z_j\}_j$.
- If the sequence $\{z_j\}_j$ is a sequence of gradients of $\mathrm{W}^{1,p}$-functions, its associated parametrized measure is called **$\mathbf{W^{1,p}}$-parametrized measure** or **Young measure**.
- If ν is independent of the spatial variable x it is called **homogeneous ($\mathbf{W^{1,p}}$)-parametrized measure**.

In particular, if $z_j \rightharpoonup z$ in $\mathrm{L}^1(\Omega)$, then $z = \int_{\mathbb{R}^m} \lambda\,\mathrm{d}\nu_x(\lambda)$.

A consequence of Thm. 1.20 is the fact that every bounded sequence in $\mathrm{L}^p(\Omega)$ with $1 \leq p < \infty$ generates a parametrized measure ν. With its help we then can express the weak limits of superposition operators. Before continuing with the analysis of (1.2.1), two examples are given that illustrate how measures can be used for the description of weak limits.

Example 1.22. Consider the function

$$v_0(x) = \begin{cases} 1 & x \geq \frac{1}{2} \\ 0 & x < \frac{1}{2} \end{cases} \quad \text{on } [0,1]$$

extended 1-periodically to $\mathbb{R}$ and let $v_j(x) = v_0(jx)$ on $(0,1)$. Then we have for the oscillatory sequence $v_j \rightharpoonup \frac{1}{2} := v$ in $\mathrm{L}^p(0,1)$ for $p < \infty$, resp. $v_j \overset{*}{\rightharpoonup} v$ in $\mathrm{L}^\infty(0,1)$.

The corresponding homogeneous Young measure is $\nu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$, where δ_0, δ_1 are Dirac measures centered at 0, resp. at 1. For continuous functions W the sequence $\{W(v_j)\}_j$ converges weakly in $\mathrm{L}^1(0,1)$ to

$$\bar{W} = \int_{\mathbb{R}} \lambda \,\mathrm{d}\nu = \frac{1}{2}\left(W(0) + W(1)\right).$$

Note that in general, for a nonlinear function W, we have $\bar{W} \neq W\left(\frac{1}{2}\right)$. Thus, one of the big advantages of using parametrized measures is the fact we can describe weak limits of nonlinear functions.

Example 1.23. In Thm. 1.20 the sequence $\{W(z_j)_j\}$ was assumed to be weakly convergent in $\mathrm{L}^1(\Omega)$. This assumption will not always be satisfied. Since the space $\mathrm{L}^1(\Omega)$ is not reflexive we may not be able to extract weakly convergent subsequences from bounded sequences. The reason is the possible occurrence of *concentration effects*.

As a simple example, consider $v_j(x) = j\chi_{[0,1/j]}$. Then $\|v_j\|_{\mathrm{L}^1([0,1])} = 1$ for all j and for $g \in \mathrm{C}([0,1])$ we get

$$\begin{aligned}\int_0^1 v_j(x)g(x)\,\mathrm{dx} &= \int_0^{1/j} jg(x)\,\mathrm{dx} \\ &= \int_0^1 g(j^{-1}x)\,\mathrm{dx} \to g(0) = \int_0^1 g(x)\delta_0(x)\,\mathrm{dx}.\end{aligned}$$

Thus, $\{v_j\}_j$ is not weakly convergent in $\mathrm{L}^1([0,1])$. Instead it converges weakly-star in the space of regular Borel measures, which is isometrically isomorphic to the dual space of $\mathrm{C}([0,1])$, cf. Thm. A.3. This is referred to as *concentration* of $\{v_j\}_j$ (at $x = 0$). Concentration effects are the point that complicate the treatment of models from nonlinear elasticity.

The elementary operations in the analysis of $\mathrm{W}^{1,\mathrm{p}}$-parametrized measures are localization and homogenization, both admitting to extract information from particular homogeneous $\mathrm{W}^{1,p}$-parametrized measures. Localization is concerned with the focus on a particular measure $\nu_a \subset \nu$ for $a \in \Omega$, which is a homogeneous $\mathrm{W}^{1,p}$-parametrized measure, cf. Thm. A.5. In contrast, homogenization is concerned with averaging. In the scalar case, we get a homogeneous $\mathrm{W}^{1,\mathrm{p}}$-parametrized measure $\bar{\nu}$ via

$$\int_{\mathbb{R}^m} W(\lambda)\,\mathrm{d}\bar{\nu}(\lambda) = \frac{1}{|\Omega|}\int_\Omega \int_{\mathbb{R}^m} W(\lambda)\,\mathrm{d}\nu_x(\lambda)\,\mathrm{dx},$$

where W is continuous and bounded by a polynomial of order p, cf. Thm. A.4. A similar result for W defined on matrices will be given below, in Lem. 1.28, via

$$\int_{\mathbb{M}^{m,n}} W(F)\,\mathrm{d}\bar{\nu}(F) = \frac{1}{|\Omega|}\int_\Omega W(\nabla\varphi(x))\,\mathrm{dx}.$$

To keep this presentation short, we do not go into technical details. The interested reader is referred to [207].

1.2.1. Jensen's inequality

We know from Tonelli's theorem (Thm. A.7) that in the scalar case, i.e. the case when $\nabla\varphi(x)$ is vector-valued, weak lower semicontinuity of $\mathcal{E}$ is equivalent to the requirement of a convex integrand W. These are characterized by the classical Jensen inequality (1.2.3).

Theorem 1.24 (Jensen's inequality)**.** *Let ν be a positive Radon measure over a σ-algebra $\mathcal{M}$ on the set Ω such that $\nu(\Omega) = 1$ and let $f \in \mathrm{L}^1_\nu(\Omega)$, where $\mathrm{L}^1_\nu(\Omega)$ denotes the L^1-space with respect to a measure ν. Then every convex function W satisfies the inequality*

$$W\Big(\int_\Omega f \,\mathrm{d}\nu\Big) \leq \int_\Omega W(f)\,\mathrm{d}\nu. \tag{1.2.3}$$

Proof. See [208, Thm. 1.2]. □

In Sec. 1.1.4 we saw that convexity is not admissible in hyperelasticity. Luckily, for vector-valued functions, and matrix-valued gradients, the convexity requirement can be relaxed. A decisive role is played by a generalized Jensen inequality, similar to (1.2.3).

We start with considering weak lower semicontinuity of $\mathcal{E}^{\mathrm{str}}(\varphi) = \int_\Omega W(\nabla\varphi)\,\mathrm{d}\mu$ with respect to a particular sequence $\{\varphi_j\}_j$.

Theorem 1.25. *Consider a weakly convergent sequence $\varphi_j \rightharpoonup \varphi$ in $\mathrm{W}^{1,\mathrm{p}}(\Omega,\mathbb{R}^m)$, with $1 < p < \infty$ and $m \geq 1$. Let $\nu = \{\nu_x\}_{x\in\Omega}$ be its associated $\mathrm{W}^{1,\mathrm{p}}$-parametrized measure such that*

$$\nabla\varphi(x) = \int_{\mathbb{M}^{m,n}} F \,\mathrm{d}\nu_x(F) \quad \text{a.e. in } \Omega. \tag{1.2.4}$$

Let $W\colon \mathbb{M}^{m,n} \to \mathbb{R}_\infty$ be a continuous function, which is bounded from below. If

$$\liminf_{j\to\infty} \int_\Omega W(\nabla\varphi_j(x))\,\mathrm{dx} < \infty, \tag{1.2.5}$$

then

$$\int_{\mathcal{A}} W(\nabla\varphi(x))\,\mathrm{dx} \leq \liminf_{j\to\infty} \int_{\mathcal{A}} W(\nabla\varphi_j(x))\,\mathrm{dx}, \tag{1.2.6}$$

for all measurable subsets $\mathcal{A} \subset \Omega$, if and only if

$$W(\nabla\varphi(x)) \leq \int_{\mathbb{M}^{m,n}} W(F)\,\mathrm{d}\nu_x(F) \quad \text{a.e. in } \Omega. \tag{1.2.7}$$

Proof. Inserting (1.2.4) into (1.2.7) yields the generalized Jensen's inequality

$$W\left(\int_{\mathbb{M}^{m,n}} F\,\mathrm{d}\nu_x(F)\right) \leq \int_{\mathbb{M}^{m,n}} W(F)\,\mathrm{d}\nu_x(F).$$

Thus, the question of weak lower semicontinuity of $\mathcal{E}^{\mathrm{str}}(\varphi) = \int_{\mathcal{A}} W(\nabla\varphi_j(x))\,\mathrm{dx}$ is, in the given setting, equivalent to the question if this generalized Jensen inequality holds for W.
We first show that (1.2.7) implies (1.2.6). This follows directly using Thm. A.8, which is Fatou's lemma in the Young measure context, and (1.2.7)

$$\begin{aligned}\liminf_{j\to\infty} \int_{\mathcal{A}} W(\nabla\varphi_j(x))\,\mathrm{dx} &\geq \int_{\mathcal{A}} \int_{\mathbb{M}^{m,n}} W(F)\,\mathrm{d}\nu_x(F)\,\mathrm{dx} \\ &\geq \int_{\mathcal{A}} W(\nabla\varphi(x))\,\mathrm{dx}.\end{aligned}$$

For the other direction the interested reader is referred to [207, Thm. 3.1]. □

Remark 1.26. The above theorem can be extended to functionals that additionally depend on the state variable,

$$\mathcal{E}^{\mathrm{str}}(\varphi) = \int_{\Omega} W(\varphi(x), \nabla\varphi(x))\,\mathrm{dx},$$

cf. [207, Thm. 3.2]. From the embedding theorem of Rellich and Kondrachov [102, Chap. 1, Lem. 1.28] we deduce strong convergence of $\{\varphi_j\}_j$. Then the associated parametrized measure of $\{(\varphi_j, \nabla\varphi_j)\}_j$ is given by $\left\{\delta_{\varphi(x)} \otimes \nu_x\right\}_{x\in\Omega}$, where $\delta_{\varphi(x)}$ denotes Dirac's delta distribution centered at $\varphi(x)$, cf. [207, Sec. 6.6], and the same steps as above apply.

The question of weak lower semicontinuity of $\mathcal{E}^{\mathrm{str}}(\varphi) = \int_{\Omega} W\left(\nabla\varphi(x)\right)\,\mathrm{dx}$ with respect to a particular sequence $\{\varphi_j\}_j$ can be reduced to the point-wise generalized Jensen inequality (1.2.7), where ν_x is a homogeneous $\mathrm{W}^{1,p}$-parametrized measure (Thm. A.5). Consequently weak lower semicontinuity of $\mathcal{E}$ with respect to *any* weakly convergent sequence in $\mathrm{W}^{1,\mathrm{p}}(\Omega)$ requires

$$W\left(\int_{\mathbb{M}^{m,n}} F\,\mathrm{d}\nu(F)\right) \leq \int_{\mathbb{M}^{m,n}} W(F)\,\mathrm{d}\nu(F) \tag{1.2.8}$$

for *all* homogeneous $\mathrm{W}^{1,\mathrm{p}}$-parametrized measures ν. This inequality is referred to as **Jensen's inequality in the sense of Pedregal**.

Theorem 1.27. *The functional* $\mathcal{E}^{\mathrm{str}}(\varphi) = \int_{\Omega} W(\nabla\varphi(x))\,\mathrm{dx}$ *is weakly lower semicontinuous in* $\mathrm{W}^{1,\mathrm{p}}(\Omega)$*, with* $1 < p < \infty$*, if and only if* W *satisfies Jensen's inequality (in the sense of Pedregal), i.e.*

$$W\left(\int_{\mathbb{M}^{m,n}} F\,\mathrm{d}\nu(F)\right) \leq \int_{\mathbb{M}^{m,n}} W(F)\,\mathrm{d}\nu(F) \tag{1.2.9}$$

for all homogeneous $\mathrm{W}^{1,\mathrm{p}}$*-parametrized measures* ν*.*

Proof. See [207, Thm. 3.3]. □

From now on we only speak of Jensen's inequality and suppress the supplement "in the sense of Pedregal". If Jensen's inequality in the sense of Thm. 1.24 is employed we will refer to it as *classical* Jensen inequality.

1.2.2. Quasiconvexity

We begin with the motivation of a necessary condition for weak lower semicontinuity or equivalently, the validity of the point-wise condition (1.2.9). We recall a homogenization result, that will help us in the following. This and the following results are easier to understand if we assume that $G = \nabla\bar{\varphi}(x)$ for some function $\bar{\varphi}$, which will later be identified with the weak limit of certain sequences.

Lemma 1.28. *Let $G \in \mathbb{M}^{m,n}$, $\varphi_G(x) = Gx$ in Ω and $\varphi \in \mathrm{W}^{1,\mathrm{p}}(\Omega)$ with $\varphi - \varphi_G \in \mathrm{W}_0^{1,\mathrm{p}}(\Omega)$ and $1 \leq p \leq \infty$. Then, there exists a bounded sequence $\{\varphi_j\}_j$ in $\mathrm{W}^{1,\mathrm{p}}(\Omega)$, with $\varphi_j - \varphi_G \in \mathrm{W}_0^{1,\mathrm{p}}(\Omega)$ for all j, such that the associated $\mathrm{W}^{1,\mathrm{p}}$-parametrized measure $\bar{\nu}$ is homogeneous and given by*

$$\int_{\mathbb{M}^{m,n}} W(F)\,\mathrm{d}\bar{\nu}(F) = \frac{1}{|\Omega|}\int_\Omega W(\nabla\varphi(x))\,\mathrm{dx},$$

where

$$W \in \mathrm{X}^{\mathrm{p}} = \{W \in \mathrm{C}(\mathbb{M}^{m,n}) :\ |W(F)| \leq \text{const.}\,(1+|F|^p)\,,\ \textit{for all } F \in \mathbb{M}^{m,n}\}\,.$$

Proof. See [207, Lem. 8.2]. □

Consider the notation of Lem. 1.28. A subsequence of $\{\varphi_j\}_j$ converges weakly to some element $\varphi \in \mathrm{W}^{1,\mathrm{p}}(\Omega)$, with $\varphi - \varphi_G \in \mathrm{W}_0^{1,\mathrm{p}}(\Omega)$. As $\bar{\nu}$ is homogeneous and φ satisfies the affine boundary values, we have $\nabla\varphi(x) = G = \int_{\mathbb{M}^{m,n}} F\,\mathrm{d}\bar{\nu}(F)$ for almost all $x \in \Omega$. Then, Jensen's inequality (1.2.8) reduces to

$$W(G) = W\left(\int_{\mathbb{M}^{m,n}} F\,\mathrm{d}\bar{\nu}(F)\right) \leq \int_{\mathbb{M}^{m,n}} W(F)\,\mathrm{d}\bar{\nu}(F) = \frac{1}{|\Omega|}\int_\Omega W\left(\nabla\varphi(y)\right)\,\mathrm{dy}.$$

Consequently, the inequality

$$W(G) \leq \frac{1}{|\Omega|}\int_\Omega W\left(\nabla\varphi(y)\right)\,\mathrm{dy}$$

is a necessary condition for Jensen's inequality and thus for weak lower semicontinuity of $\mathcal{E}(\varphi)$.

Definition 1.29. Let $\varphi_G(x) = Gx$ for some $G \in \mathbb{M}^{m,n}$ and all $x \in \Omega$. A continuous function $W\colon \mathbb{M}^{m,n} \to \mathbb{R}$ is called (**$\mathbf{W^{1,p}}$-**)**quasiconvex at** G, for $1 \leq p \leq \infty$, if

$$W(G) \leq \frac{1}{|\mathcal{A}|}\int_{\mathcal{A}} W(\nabla\varphi(x))\,\mathrm{dx} \tag{1.2.10}$$

for all $\varphi \in \mathrm{W}^{1,\mathrm{p}}(\Omega)$ such that $\varphi - \varphi_G \in \mathrm{W}_0^{1,\mathrm{p}}(\Omega)$, and all measurable subsets $\mathcal{A} \subseteq \Omega$. If (1.2.10) holds for all $G \in \mathbb{M}^{m,n}$ we call W (**$\mathbf{W^{1,p}}$-**)**quasiconvex**.

Remark 1.30.

- The definition of $\mathrm{W}^{1,\infty}$-quasiconvexity corresponds to the original definition of quasiconvexity by Morrey [194]. The extension of quasiconvexity to $\mathrm{W}^{1,\mathrm{p}}$-spaces has been realized by Ball and Murat in [21], essentially based on results of Meyers [189] as well as Acerbi and Fusco [2], who derived conditions to deduce weak lower semicontinuity in $\mathrm{W}^{1,\mathrm{p}}(\Omega)$ from weak* lower semicontinuity in $\mathrm{W}^{1,\infty}(\Omega)$.
- The definitions of quasiconvexity actually only need to hold for one measurable and bounded domain $\mathcal{A} \subseteq \Omega$ in order to hold for all measurable and bounded domains $\mathcal{A} \subseteq \Omega$ [189, pp. 128-129]. Kristensen [166] showed that purely local definition of quasiconvexity, only involving W and a finite number of its partial derivatives, does not exist. For this reason, quasiconvexity is difficult to verify for a given function f, in particular in the case that f does not satisfy a stricter convexity property.
- There exists another definition of quasiconvexity in the fields of mathematical economics and operations research [70]. This definition describes a different concept than the one used here.

In the presence of suitable polynomial growth conditions, $\mathrm{W}^{1,\mathrm{p}}$-quasiconvexity is equivalent to $\mathrm{W}^{1,\infty}$-quasiconvexity.

Lemma 1.31. *For $1 \leq p < \infty$, an upper semicontinuous function W satisfying, for $c_1 > 0,\ c_2 > 0$ and all $F \in \mathbb{M}^{m,n}$,*

$$c_1 \leq W(F) \leq c_2(1 + |F|^p),$$

is $\mathrm{W}^{1,\mathrm{p}}$-quasiconvex if and only if it is $\mathrm{W}^{1,\infty}$-quasiconvex.

Proof. See Ball and Murat [21, Prop. 2.4] or Pedregal [207, Lem. 8.13]. □

To get a better idea about the relation of quasiconvexity to the validity of Jensen's inequality we consider a weakly convergent sequence $\varphi_j \rightharpoonup \bar{\varphi}$ such that $W(\nabla\varphi_j)$ is weakly convergent in $\mathrm{L}^1(\Omega)$ and let $G = \nabla\bar{\varphi}(x)$ for some $x \in \Omega$. Then the quasiconvexity condition requires

$$W(\nabla\bar{\varphi}(x)) \leq \frac{1}{|\Omega|}\int_\Omega W(\nabla\varphi_j(x))\,\mathrm{dx}$$

which implies

$$W(\nabla\bar{\varphi}(x)) \leq \liminf_{j\to\infty} \frac{1}{|\Omega|}\int_\Omega W(\nabla\varphi_j(x))\,\mathrm{dx}$$

and thus yields the desired weak lower semicontinuity. However, in the following we will see that the assumed weak convergence of $W(\nabla\varphi_j)$ in $\mathrm{L}^1(\Omega)$ is not clear a priori and must be enforced by additional growth conditions which are not compatible with the hyperelastic setting.

We first step back and recall that at the beginning of this subsection, we saw that quasiconvexity is a necessary condition for the validity of Jensen's inequality and thus weak lower semicontinuity. In general, the converse is not true. More precisely, (1.2.9) only holds if no concentration effects occur. To this end, let $\varphi_G(x) = Gx$ for some $G \in \mathbb{M}^{m,n}$ and consider, for $1 \leq p < \infty$ a bounded sequence $\{\psi_j\}_j$, with $\psi_j - \varphi_G \in \mathrm{W}_0^{1,\mathrm{p}}(\Omega)$ for all $j \in \mathbb{N}$, and associated $\mathrm{W}^{1,\mathrm{p}}$-parametrized measure $\nu = \{\nu_x\}_{x\in\Omega}$. Then, $\{\psi_j\}_j$ contains a weakly convergent subsequence with weak limit $\varphi(x) = \int_{\mathbb{M}^{m,n}} F \,\mathrm{d}\nu_x(F)$. According to Lem. 1.28 there exists another sequence $\{\varphi_j\}_j$, with $\varphi_j - \varphi_G \in \mathrm{W}_0^{1,\mathrm{p}}$ for all $j \in \mathbb{N}$, such that the associated $\mathrm{W}^{1,\mathrm{p}}$-parametrized measure $\bar{\nu}$ is homogeneous and given through

$$\int_{\mathbb{M}^{m,n}} W(F) \,\mathrm{d}\bar{\nu}(F) = \frac{1}{|\Omega|} \int_\Omega W(\nabla\varphi(x)) \,\mathrm{dx}.$$

The weak limit of $\{\varphi_j\}_j$ is affine and satisfies the same boundary conditions as φ_G. Consequently we have $\varphi_j \rightharpoonup \varphi_G$ and

$$G = \int_{\mathbb{M}^{m,n}} F \,\mathrm{d}\bar{\nu}(F). \tag{1.2.11}$$

Now, we consider a non-negative continuous function W, such that the sequence $\{W(\nabla\varphi_j)\}_j$ is bounded in $\mathrm{L}^1(\Omega)$. Since $\varphi_j - \varphi_G \in \mathrm{W}_0^{1,\mathrm{p}}(\Omega)$, we get from the $\mathrm{W}^{1,\mathrm{p}}$-quasiconvexity condition that

$$W(G) \leq \frac{1}{|\Omega|} \int_\Omega W\left(\nabla\varphi_j(x)\right) \,\mathrm{dx}$$

holds for all j. Consequently,

$$W(G) \leq \liminf_{j\to\infty} \frac{1}{|\Omega|} \int_\Omega W(\nabla\varphi_j(x)) \,\mathrm{dx}.$$

Depending on the behavior of $\{W(\nabla\varphi_j)\}_j$, we have to distinguish two cases. If $\{W(\nabla\varphi_j)\}_j$ does not develop concentrations we can extract a weakly convergent subsequence $W(\nabla\varphi_j) \rightharpoonup \bar{W}$ in $\mathrm{L}^1(\Omega)$. As the parametrized measure $\bar{\nu}$, that is generated by $\{\nabla\varphi_j\}_j$, is homogeneous, this limit is constant and can be expressed via

$$\bar{W} = \int_{\mathbb{M}^{m,n}} W(F) \,\mathrm{d}\bar{\nu}(F),$$

cf. Thm. 1.20. With (1.2.11) we get

$$W\left(\int_{\mathbb{M}^{m,n}} F \,\mathrm{d}\bar{\nu}(F)\right) = W(G) \leq \liminf_{j\to\infty} \frac{1}{|\Omega|} \int_\Omega W(\nabla\varphi_j(x)) \,\mathrm{dx} = \int_{\mathbb{M}^{m,n}} W(F) \,\mathrm{d}\bar{\nu}(F),$$

i.e. Jensen's inequality holds.

In contrast, if the sequence $\{W(\nabla\varphi_j)\}_j$ does develop concentrations we get the strict inequality

$$\liminf_{j\to\infty} \frac{1}{|\Omega|}\int_\Omega W(\nabla\varphi_j(x))\,\mathrm{dx} > \int_{\mathbb{M}^{m,n}} W(F)\;\mathrm{d}\bar{\nu}(F).$$

In order to prove Jensen's inequality, we would rather need that the limit on the left hand side is less or equal than the expression on the right hand side. Thus, in the presence of concentrations, we are not able to infer the validity of Jensen's inequality without further restrictions.

To exclude these difficulties with Jensen's inequality, $\mathrm{W}^{1,\mathrm{p}}$-quasiconvexity typically turns up together with a polynomial growth condition of order p. Then, according to Lem. A.6, concentrations will not be relevant and $\{W(\nabla\varphi_j)\}_j$ converges weakly in $\mathrm{L}^1(\Omega)$. In contrast, "under no growth assumptions, there are still many open, delicate issues concerning $\mathrm{W}^{1,\mathrm{p}}$-quasiconvexity" Pedregal [207, p. 150].

We summarize the classical existence results for quasiconvex integrands into the following two theorems.

Theorem 1.32. *Let $W\colon \mathbb{M}^{m,n} \to \mathbb{R}$ be continuous. The functional $\mathcal{E}^{\mathrm{str}}(\varphi) = \int_\Omega W(\nabla\varphi(x))\,\mathrm{dx}$ is lower semicontinuous with respect to weak convergence in $\mathrm{W}^{1,\infty}(\Omega)$ if and only if W is $\mathrm{W}^{1,\infty}$-quasiconvex.*

Proof. See Pedregal [207, Thm. 3.4]. □

Theorem 1.33. *Let $W\colon \mathbb{M}^{m,n} \to \mathbb{R}_\infty$ be a continuous function such that for constants $c_0 \in \mathbb{R},\ c_1 > 0$ and all $F \in \mathbb{M}^{m,n}$*

$$c_0 \le W(F) \le c_1(1 + |F|^p),$$

with $1 < p < \infty$. Then the functional $\mathcal{E}(\varphi) = \int_\Omega W(\nabla\varphi(x))\,\mathrm{dx}$ is lower semicontinuous with respect to weak convergence of $\mathrm{W}^{1,\mathrm{p}}(\Omega)$ if and only if W is $\mathrm{W}^{1,\mathrm{p}}$-quasiconvex.

Proof. See Ball and Murat [21, Conj. 3.7]. □

The required polynomial growth condition does not fit into the setting of hyperelasticity. Consider an elastic stored energy function W that satisfies condition (1.1.1) and a weakly convergent sequence $\{\varphi_{\delta,j}\}_j$ such that on a subset $\mathcal{A}_\delta \subset \Omega$ with $0 < |\mathcal{A}_\delta| < \delta$ the sequence $\{\det(\nabla\varphi_{\delta,j})\}_j$ converges almost everywhere in $\mathcal{A}_\delta$ to 0. Then the sequence $\{W(\nabla\varphi_j)\}_j$ develops concentrations.. For this reason, the current understanding of quasiconvexity does not admit its application in hyperelasticity.

1.2.3. Polyconvexity

In the last section we saw that weak lower semicontinuity of quasiconvex energy functionals $\mathcal{E}^{\mathrm{str}}$ requires a polynomial upper bound. The problem is that we are not able

to guarantee the validity of Jensen's inequality if we don't have weak compactness of $\{W(\nabla\varphi_j)\}_j$. The latter property is enforced with the help of a polynomial growth condition, which is incompatible with the limit behavior $\lim_{\det(\nabla\varphi)\searrow 0} W(\nabla\varphi) = \infty$. For this reason, we have to stay closer to a convex setting. From a general perspective this is discussed in Sec. 1.2.3.1. More specific results for problems in hyperelasticity are the topic of Sec. 1.2.3.2.

1.2.3.1. The general setting

Recall that in $\mathrm{W}^{1,\mathrm{p}}(\Omega)$ with $1 \leq p < \infty$ growth conditions are not necessary for convex stored energy functions W, cf. Thm. A.7. In Sec. 1.1.4 we saw that convexity of W is not admissible. However, we can exploit this observation in the following way. Let M be a weakly continuous function, i.e. a function that maps weakly converging sequences $\nabla\varphi_j \rightharpoonup \nabla\bar{\varphi}$ into weakly converging sequences $M(\nabla\varphi_j) \rightharpoonup \bar{M}$. If $\bar{M} = M(\nabla\bar{\varphi})$, then we can identify candidates for stored energy functions via

$$W(\varphi) = V(M(\nabla\varphi)), \tag{1.2.12}$$

where $V(\cdot)$ is convex.

To make this more precise we need to identify the weakly continuous functions M of the deformation gradient $\nabla\varphi$. Keeping in mind the discussion of the last section, we observe that a necessary requirement for weak continuity of M is that both M and $-M$ are quasiconvex.

Definition 1.34. A function $M : \mathbb{M}^{m,n} \to \mathbb{R}$ is called **quasiaffine**[3] if M and $-M$ are quasiconvex.

Remark 1.35. Quasiaffine functions were considered in the work of John Ball under the name **null Lagrangians** [15, 20, 72]. For applications in the context of compensated compactness cf. [251, Cor. 17.2].

Theorem 1.36. *Let $M : \mathbb{M}^{m,n} \to \mathbb{R}$ be quasiaffine. Then the function $F \mapsto M(F)$ is affine in terms of the minors of F.*

Proof. For $m = n,\ n \leq 3$ the proof can be found in [15, Thm. 4.1]. For the general case see [72, Chap. 4, Thm. 1.5]. □

Since the minors naturally satisfy a polynomial growth condition they are the only weakly continuous function of the deformation gradient. Thus, we can further specify condition (1.2.12).

[3] Or equivalently, rank one affine or polyaffine [72, Chap. 4, Thm. 1.5].

Definition 1.37. A function $W\colon \mathbb{M} \to \mathbb{R}_\infty$ is called **polyconvex** if there exists a convex function $\mathbb{W}\colon \mathbb{M}^{\tau(m,n)} \to \mathbb{R}_\infty$ such that for all $F \in \mathbb{M}^{m,n}$

$$W(F) = \mathbb{W}(M(F)),$$

where

$$M : \mathbb{M}^{m,n} \to \mathbb{M}^{\tau(m,n)}, \quad F \mapsto (F, \operatorname{cof}_2(F), \ldots, \operatorname{cof}_{\min(m,n)}(F)),$$

$\operatorname{cof}_s(F)$ is the matrix of all $s \times s$-minors of F and

$$\tau(m,n) = \sum_{s=1}^{\min(m,n)} \frac{m!n!}{(s!)^2(m-s)!(n-s)!}.$$

The definition of polyconvexity, due to Ball [15], admits to prove the validity of Jensen's inequality, as necessary condition for weak lower semicontinuity of $\mathcal{E}^{\mathrm{str}}$. We only have to take care that the minors of the deformation gradient are well-behaved. In general, this requires that the integration order p satisfies $p \geq r := \max(m,n)$.

Theorem 1.38. *Let $W\colon \mathbb{M}^{m,n} \to \mathbb{R}_\infty$ be polyconvex and $p \geq r := \max(m,n)$. Then W satisfies Jensen's inequality for any homogeneous $\mathrm{W}^{1,\mathrm{p}}$-parametrized measure.*

Proof. Let ν be a homogeneous $\mathrm{W}^{1,\mathrm{p}}$-parametrized measure generated by a weakly convergent sequence $\{\nabla\varphi_j\}_j \subset \mathrm{W}^{1,\mathrm{p}}(\Omega)$. We denote its first moment by $G = \int_{\mathbb{M}^{m,n}} F \,\mathrm{d}\nu(F)$. According to Lem. A.6 we can w.l.o.g. assume that $\{|\nabla\varphi_j|^p\}_j$ is equi-integrable. As $M(\nabla\varphi_j) \rightharpoonup M(G)$ in $\mathrm{L}^{p/r}(\Omega)$ and $M(F) \leq c_1(1 + |F|^r)$, for some positive constant $c_1 > 0$, we have

$$M(\nabla\varphi_j) \rightharpoonup \int_{\mathbb{M}^{m,n}} M(F) \,\mathrm{d}\nu(F)$$

from Thm. A.4 and

$$\int_{\mathbb{M}^{m,n}} M(F) \,\mathrm{d}\nu(F) = M(G) = M\left(\int_{\mathbb{M}^{m,n}} F \,\mathrm{d}\nu(F)\right).$$

Using the convexity of $\mathbb{W}$ we get, with the classical Jensen inequality,

$$\begin{aligned}
\int_{\mathbb{M}^{m,n}} W(F) \,\mathrm{d}\nu(F) &= \int_{\mathbb{M}^{m,n}} \mathbb{W}(M(F)) \,\mathrm{d}\nu(F) \\
&\geq \mathbb{W}\left(\int_{\mathbb{M}^{m,n}} M(F) \,\mathrm{d}\nu(F)\right) \\
&= \mathbb{W}(M(G)) = W(G) = W\left(\int_{\mathbb{M}^{m,n}} F \,\mathrm{d}\nu(F)\right).
\end{aligned}$$

□

Corollary 1.39. *Let $W\colon \mathbb{M}^{m,n} \to \mathbb{R}_\infty$ be polyconvex and $p \geq r := \max(m,n)$. Then the functional*

$$\mathcal{E}^{\mathrm{str}}(\varphi) = \int_\Omega W(\nabla\varphi(x))\,\mathrm{dx}$$

is weakly lower semicontinuous in $\mathrm{W}^{1,\mathrm{p}}(\Omega)$.

Proof. This is a direct consequence of Thm. 1.38 and Thm. 1.27. □

We add a suitable coercivity condition to get

Theorem 1.40. *Let $W\colon \mathbb{M}^{m,n} \to \mathbb{R}_\infty$ be non-negative, polyconvex such that for $p \geq r := \max(m,n)$ the coercivity inequality*

$$c\,(|F|^p - 1) \leq W(F), \quad c > 0 \tag{1.2.13}$$

holds for all $F \in \mathbb{M}^{m,n}$. If there exists at least one $\varphi_0 \in \mathrm{W}^{1,\mathrm{p}}(\Omega)$ such that

$$\mathcal{E}^{\mathrm{str}}(\varphi_0) = \int_\Omega W(\nabla\varphi_0(x))\,\mathrm{dx} < \infty,$$

then there exists at least one minimizer $\bar{\varphi}$ of $\mathcal{E}^{\mathrm{str}}$.

Proof. From Thm. 1.38 we get that W satisfies inequality (1.2.9). Then weak lower semicontinuity of

$$\mathcal{E}^{\mathrm{str}}(\varphi) = \int_\Omega W(\nabla\varphi(x))\,\mathrm{dx}$$

follows from Thm. 1.27. We take a minimizing sequence $\{\varphi_j\}_j$, i.e.

$$\lim_{j\to\infty} \mathcal{E}^{\mathrm{str}}(\varphi_j) \to e = \inf_{\varphi\in\mathrm{W}^{1,\mathrm{p}}(\Omega)} \mathcal{E}^{\mathrm{str}}(\varphi) < \infty.$$

Thus we can extract a bounded subsequence of $\{\mathcal{E}^{\mathrm{str}}(\varphi_j)\}_j$ and, by (1.2.13), also of $\{\varphi_j\}_j$. Possibly again extracting a subsequence, we get $\varphi_j \rightharpoonup \bar{\varphi}$ for some $\bar{\varphi} \in \mathrm{W}^{1,\mathrm{p}}(\Omega)$. From the weak lower semicontinuity of $\mathcal{E}^{\mathrm{str}}$,

$$e = \liminf_{j\to\infty} \mathcal{E}^{\mathrm{str}}(\varphi_j) \geq \mathcal{E}^{\mathrm{str}}(\bar{\varphi}),$$

we deduce that $\bar{\varphi}$ is a minimizer of $\mathcal{E}^{\mathrm{str}}$. □

1.2.3.2. Polyconvex hyperelasticity

Having introduced the polyconvex framework, we now turn to its discussion in the context of hyperelasticity. In this subsection we assume that $n = m = 3$. Then the minors of $F \in \mathbb{M}^3$ are, besides F itself, the entries of the cofactor matrix $\operatorname{cof}(F)$ and $\det(F)$. Hence, a function $W\colon \mathbb{M}^3 \to \mathbb{R}_\infty$ is polyconvex if there exists a convex function $\mathbb{W}\colon \mathbb{M}^3 \times \mathbb{M}^3 \times \mathbb{R} \to \mathbb{R}_\infty$ such that

$$W(F) = \mathbb{W}(F, \operatorname{cof}(F), \det(F)). \tag{1.2.14}$$

Here polyconvexity admits a geometric interpretation. For some deformation φ the induced change of infinitesimal length elements is determined by $\nabla\varphi^T\nabla\varphi$, the change of area elements with normal n by $\operatorname{cof}(\nabla\varphi)n$ and the change of volumetric elements by $\det(\nabla\varphi)$. Thus $\mathbb{W}$ is convex in terms of the quantities that determine the changes of volumes of different codimension.

Recall that for $p > 3$, Sobolev's embedding theorem [102, Chap. 2, Thm. 1.2] yields continuity of the deformations. If we want to describe fracture or cavitation, frequent phenomena for rubber materials [170], discontinuous solutions in weaker Sobolev spaces are needed [16]. While these phenomena are not of interest here, many popular material laws are formulated in $\mathrm{W}^{1,2}(\Omega)$. For consistency with the models used in the numerical examples in Chap. 6, we also allow discontinuous deformations.

In order to get an existence result for three-dimensional problems in $\mathrm{W}^{1,2}(\Omega)$, we need a refined weak compactness result for the minors of the deformation gradient. Such a result has been established by Ball [15, Lem. 6.1, Thm. 6.2]. It is based on the Piola identity, $\operatorname{div}(\operatorname{cof}(\nabla\varphi)) = 0$, which admits to weaken the definition of $\det(\nabla\varphi)$ to the expansion

$$\det(\nabla\varphi) = \sum_{j=1}^{n} \frac{\partial\varphi_k}{\partial x_j}(\operatorname{cof}(\nabla\varphi))_{kj} \quad \text{for } k \in \{1, \dots, n\}\,.$$

Theorem 1.41. *Let $p \geq 2$, $q \geq 1$ such that $\frac{1}{s} := \frac{1}{p} + \frac{1}{q} \leq 1$ and $r \geq 1$. Then the mapping*

$$\mathrm{W}^{1,\mathrm{p}}(\Omega) \times \mathrm{L}^{\mathrm{q}}(\Omega) \ni (\varphi, \operatorname{cof}(\varphi)) \mapsto \det(\nabla\varphi) := \sum_{j=1}^{3} \frac{\partial\varphi_1}{\partial x_j}(\operatorname{cof}(\nabla\varphi))_{1j} \in \mathrm{L}^{\mathrm{s}}(\Omega)$$

is well defined and continuous. Furthermore

$$\left.\begin{aligned} \varphi_j &\rightharpoonup \varphi \text{ in } \mathrm{W}^{1,\mathrm{p}}(\Omega) \\ \operatorname{cof}(\nabla\varphi_j) &\rightharpoonup H \text{ in } \mathrm{L}^{\mathrm{q}}(\Omega) \\ \det(\nabla\varphi_j) &\rightharpoonup \delta \text{ in } \mathrm{L}^{\mathrm{r}}(\Omega) \end{aligned}\right\} \Rightarrow \left\{\begin{aligned} H &= \operatorname{cof}(\nabla\varphi) \\ \delta &= \det(\nabla\varphi). \end{aligned}\right. \tag{1.2.15}$$

Proof. See [56, Thm. 7.6-1], resp. [15, Lem. 6.1, Thm. 6.2]. □

Eventually we can establish an existence theorem for polyconvex hyperelasticity. Besides the relaxation from $p \geq 3$ to $p \geq 2$ this includes the incorporation of the orientation preservation condition $\det(\nabla\varphi) > 0$.

Theorem 1.42. *Let W, $\mathbb{W}$ be defined as in (1.2.14) and let*

$$\mathbb{W}(F, \operatorname{cof}(F), \det(F)) \geq \alpha(\|F\|^p + \|\operatorname{cof}(F)\|^q + |\det(F)|^r) - \beta, \tag{1.2.16}$$

where $\alpha > 0$, $\beta \in \mathbb{R}$ and $p \geq 2$, $\frac{1}{p} + \frac{1}{q} \leq 1$, $r \geq 1$. If there exists at least one element

$$\varphi_0 \in \Phi_0 := \left\{\varphi \in \mathrm{W}_0^{1,\mathrm{p}}(\Omega) : \ \det(\varphi) > 0 \ \textit{a.e. in } \Omega\right\}$$

such that $\mathcal{E}_l(\varphi_0) = \int_\Omega W(\nabla\varphi_0)\,\mathrm{dx} - l(\varphi_0) < \infty$, where $l : \Phi_0 \to \mathbb{R}$ is linear and continuous, then there exist at least one minimizer $\bar{\varphi}$ of $\mathcal{E}_l$ in Φ_0.

Proof. First we integrate equation (1.2.16) over Ω to deduce the necessary coercivity inequality in $\Phi_0 \times \mathrm{L}^{\mathrm{q}}(\Omega) \times \mathrm{L}^{\mathrm{r}}(\Omega)$:

$$\begin{aligned} \mathcal{E}_l(\varphi) &= \int_\Omega \mathbb{W}(\nabla\varphi, \operatorname{cof}(\nabla\varphi), \det(\nabla\varphi))\,\mathrm{dx} - l(\varphi) \\ &\geq \tilde{\alpha}(\|\varphi\|_{\mathrm{W}_0^{1,\mathrm{p}}(\Omega)} + \|\operatorname{cof}(\nabla\varphi)\|_{\mathrm{L}^{\mathrm{q}}(\Omega)} + \|\det(\nabla\varphi)\|_{\mathrm{L}^{\mathrm{r}}(\Omega)} - \tilde{\beta}, \end{aligned} \tag{1.2.17}$$

with $\tilde{\alpha} > 0$ and $\tilde{\beta} \in \mathbb{R}$. Now consider an infimizing sequence $\{\varphi_j\}_j$, i.e.

$$\lim_{j\to\infty} \mathcal{E}_l(\varphi_j) \to \inf_{\varphi\in\Omega} \mathcal{E}_l(\bar{\varphi}).$$

From equation (1.2.17) and Thm. 1.41 we infer the existence of a weakly convergent subsequence

$$\{(\nabla\varphi_j, \operatorname{cof}(\nabla\varphi_j), \det(\nabla\varphi_j))\}_j \rightharpoonup (\nabla\bar{\varphi}, \operatorname{cof}(\nabla\bar{\varphi}), \det(\nabla\bar{\varphi}))$$

in $\mathrm{W}_0^{1,\mathrm{p}}(\Omega) \times \mathrm{L}^{\mathrm{q}}(\Omega) \times \mathrm{L}^{\mathrm{r}}(\Omega)$. By definition of weak convergence we get $l(\varphi_j) \to l(\bar{\varphi})$. Consequently, in the following we only have to consider $\mathcal{E}^{\mathrm{str}}$. Due to Thm. 1.41, we still have weak compactness of M and get existence of a minimizer in $\mathrm{W}_0^{1,\mathrm{p}}(\Omega)$ analogously to Thm. 1.40.

It remains to establish that the weak limit $\bar{\varphi}$ is indeed contained in Φ_0. First the transition to strong convergence with Mazur's lemma (Lem. A.9) gives the existence of an almost everywhere pointwise convergent subsequence $\{\det(\nabla\varphi_j)\}_j$. Thus $\det(\nabla\bar{\varphi}) \geq 0$ almost everywhere in Ω. Now assume that there is a subset $\mathcal{A} \subset \Omega$ with $|\mathcal{A}| > 0$ such that $\det(\nabla\varphi) = 0$ almost everywhere in $\mathcal{A}$. Then

$$\infty = \int_{\mathcal{A}} W(\bar{\varphi})\,\mathrm{d}\mu = \int_\Omega \liminf_{k\to\infty} W(\varphi_k(x))\,\mathrm{d}\mu.$$

Since l is bounded this extends to

$$\infty = \mathcal{E}_l(\bar{\varphi}) = \lim_{j\to\infty} \mathcal{E}_l(\varphi_j),$$

which is in contradiction with the assumed existence of at least one element such that $\mathcal{E}_l$ is finite. Therefore $|\mathcal{A}| = 0$ and $\varphi \in \Phi_0$ must hold. □

Remark 1.43.

- The restriction to $\mathrm{W}^{1,\mathrm{p}}_0(\Omega)$ was done for simplicity. As usual, the above theorem can also be applied in the case of mixed boundary conditions, as long as Dirichlet boundary conditions are subscribed on a measurable part of the boundary $\Gamma_\mathrm{d} \subset \partial\Omega$ with positive surface measure $|\Gamma_\mathrm{d}| > 0$. Then the set of admissible deformations is

$$\Phi := \left\{ \varphi \in \mathrm{W}^{1,\mathrm{p}}(\Omega) : \ \det(\varphi) > 0 \text{ a.e. in } \Omega, \ \varphi = 0 \text{ a.e. on } \Gamma_\mathrm{d} \right\}.$$

- Observing $|\det(F)| \leq \|\mathrm{adj}(F)\|\|F\|$ we can drop the last summand in (1.2.16).
- For volume forces f or Neumann boundary forces g, defined on Γ_c, the functional l is given through

$$l(\varphi) = \mathcal{E}^{\mathrm{ext}}(\varphi, f) = \int_\Omega \varphi f \,\mathrm{d}\mu, \quad \text{resp.} \quad l(\varphi) = \mathcal{E}^{\mathrm{ext}}(\varphi, g) = \int_{\Gamma_\mathrm{c}} \varphi g \,\mathrm{ds}.$$

Recall, that the focus of this thesis is on implant shape design. An implant that occupies a domain Ω_{rigid} can be interpreted as an obstacle to the soft tissue. For this case we need another variant of the above existence theorem.

Theorem 1.44. *Let W, $\mathbb{W}$ be defined as in (1.2.14) and let*

$$\mathbb{W}(F, \mathrm{cof}(F), \det(F)) \geq \alpha(\|F\|^p + \|\mathrm{cof}(F)\|^q + |\det(F)|^r) - \beta,$$

where $\alpha > 0$, $\beta \in \mathbb{R}$ and $p \geq 2$, $\frac{1}{p} + \frac{1}{q} \leq 1$, $r \geq 1$. Let $\mathcal{A} \subset \mathbb{R}^3 \setminus \Omega_{\mathrm{rigid}}$ be closed. If there exists at least one element

$$\varphi_0 \in \Phi_{\mathcal{A}} := \left\{ \varphi \in \mathrm{W}^{1,\mathrm{p}}_0(\Omega) : \ \det(\varphi) > 0 \text{ a.e. in } \Omega \text{ and } \varphi \in \mathcal{A} \text{ a.e. on } \Gamma_\mathrm{c} \right\}$$

such that $\mathcal{E}^{\mathrm{str}}(\varphi_0) = \int_\Omega W(\nabla\varphi_0)\,\mathrm{dx} < \infty$, then there exists at least one minimizer $\bar{\varphi}$ of $\mathcal{E}^{\mathrm{str}}$ in $\Phi_{\mathcal{A}}$.

Proof. The proof is essentially the same as for Thm. 1.42. For this reason we only consider the additional requirement $\varphi(\Gamma_\mathrm{c}) \in \mathcal{A}$. Let $\{\varphi_j\}_j \subset \Phi_{\mathcal{A}}$ be an infimizing sequence and denote its weak limit by $\bar{\varphi}$. Since the trace operator $\mathrm{W}^{1,\mathrm{p}}(\Omega) \to \mathrm{L}^{\mathrm{p}}(\Gamma_\mathrm{c})$ is compact we can extract a subsequence that converges almost everywhere on Γ_c. Consequently, as $\mathcal{A}$ is closed, we have $\bar{\varphi} \in \mathcal{A}$ almost everywhere on Γ_c. □

Remark 1.45. Today the notion of polyconvexity is well established in the biomechanics community. Recognizing its relation to the Legendre-Hadamard condition and its implications [56, Sec. 5.10][161, 162, 185], material laws for all standard anisotropy classes have been derived [88, 91, 151, 190, 230, 231, 232, 233, 234, 244].

1.3. First order optimality conditions for a compressible Mooney-Rivlin material

In general, it is not clear whether for a given boundary force $g \in \mathrm{L}^2(\Gamma_c)$ a local minimizer $\varphi \in \Phi$ of the elastic energy functional $\mathcal{E} = \mathcal{E}^{\mathrm{str}} - \mathcal{E}^{\mathrm{ext}}$ satisfies the weak formulation

$$\mathcal{E}_\varphi(\varphi, g)h = 0 \quad \text{for all } h \in \mathrm{W}^{1,p}(\Omega).$$

In the context of compressible material laws the main difficulties are caused by the orientation preservation condition $\det(\nabla\varphi) > 0$. In particular, it implies for $1 \leq p < \infty$ that the set

$$\Phi_\infty := \left\{\varphi \in \mathrm{W}^{1,\mathrm{p}}(\Omega) : \ \mathcal{E}^{\mathrm{str}}(\varphi) = \infty\right\},$$

is a dense subset of $\mathrm{W}^{1,\mathrm{p}}(\Omega)$ for all $p < \infty$. Thus, differentiability cannot be expected in spaces weaker than $\mathrm{W}^{1,\infty}(\Omega)$.

To make the discussion concrete we consider a compressible Mooney-Rivlin material law in $\mathbb{R}^3$. This constitutive relation is polyconvex and isotropic. For $F \in \mathbb{M}^3$ its stored energy function is given via

$$W(F) = \frac{\alpha}{2}\|F\|^2 + \frac{\beta}{2}\|\mathrm{cof}\,(F)\|^2 + \Gamma\left(\det\left(F\right)\right),$$

with material parameters $\alpha > 0$, $\beta > 0$ and Γ denoting a volumetric penalty function. Here we exemplarily consider the penalty function suggested by Murnaghan [196],

$$\Gamma : \mathbb{R}_+ \to \mathbb{R}_+, \quad t \mapsto \frac{d}{2}t^2 + \frac{e}{2}t^{-k},$$

with material parameters $d, e, k > 0$. We study its derivatives at nonsingular $F \in \mathbb{M}^3_+$ in direction $\delta F \in \mathbb{M}^3$. Using $\det'(F)\delta F = \mathrm{adj}(F) : \delta F = \mathrm{cof}(F)^T : \delta F$ we get for the first directional derivative

$$W'(F)\delta F = \alpha F : \delta F + \beta \mathrm{cof}(F) : \mathrm{cof}'(F)\delta F + \Gamma'(\det(F))(\mathrm{cof}(F)^T : \delta F).$$

The second directional derivative is given via

$$\begin{aligned} W''(F)(\delta F_1, \delta F_2) = {} & \alpha \delta F_1 : \delta F_2 + \beta \mathrm{cof}'(F)\delta F_1 : \mathrm{cof}'(F)\delta F_2 \\ & + \beta \mathrm{cof}(F) : \mathrm{cof}''(F)(\delta F_1, \delta F_2) \\ & + \Gamma'(\det(F))(\mathrm{cof}'(F)^T \delta F_1^T : \delta F_2) \\ & + \Gamma''(\det(F))(\mathrm{cof}(F)^T : \delta F_1)(\mathrm{cof}(F)^T : \delta F_2). \end{aligned}$$

The validity of the above pointwise formulae follows, for $F \in \mathbb{M}^3_+$ and $\delta F_1, \delta F_2 \in \mathbb{M}^3$, directly from the definitions of the cofactor matrix and the determinant. Having considered properties of W as a nonlinear function of matrices $F \in \mathbb{M}^3$, we now turn to its study as superposition operators. In this regard, first consider cof.

Lemma 1.46. *Let $F \in \mathrm{L}^{\mathrm{p}}(\Omega)$ with $1 \leq p < \infty$. Then the mapping*

$$\mathrm{cof}'(F) : \mathrm{L}^{\mathrm{q}}(\Omega) \to \mathrm{L}^1(\Omega)$$

is linear and continuous for $p^{-1} + q^{-1} \leq 1$. The mapping

$$\mathrm{cof}''(F) : \mathrm{L}^{\mathrm{q}_1}(\Omega) \times \mathrm{L}^{\mathrm{q}_2}(\Omega) \to \mathrm{L}^1(\Omega)$$

is independent of F, bilinear and continuous for $q_1^{-1} + q_2^{-1} \leq 1$. For $N > 2$ we have $\mathrm{cof}^{(N)} = 0$.

Proof. The assertion follows from the observation that cof is a second order polynomial in the entries of F and Hölder's inequality. □

Definition 1.47. Let $\varphi \in \mathrm{W}^{1,\mathrm{p}}(\Omega)$ with $1 \leq p \leq \infty$. We call φ **nondegenerate** if there exists a constant $\varepsilon > 0$, such that

$$\det(\nabla\varphi) \geq \varepsilon \quad \text{almost everywhere in } \Omega.$$

Otherwise, we call φ **degenerate**. In the context of elasticity theory, we will also call displacements u nondegenerate if $\varphi = \mathrm{id} + u$ is nondegenerate.

Suppose there exists a degenerate local minimizer $\varphi \in \Phi$, i.e. there exists a sequence

$$\{x_k\}_k \subset \Omega,\ x_k \to x \in \Omega \quad \text{such that} \quad \det(\nabla\varphi(x_k)) \to 0.$$

This corresponds to a singularity of the stored energy function at x. This is reasonable if cutting or piercing processes are to be investigated. However, in this case other effects such as plasticity become dominant and elastic models, that only depend on the deformation gradient are no longer adequate. Regarding the design of reasonable implant shapes nondegeneracy is a natural assumption.

Lemma 1.48. *Assume that $F \in \mathrm{L}^{\mathrm{p}}(\Omega)$ is nondegenerate, $\mathrm{cof}(F) \in \mathrm{L}^{\mathrm{q}}(\Omega)$, and $\det(F) \in \mathrm{L}^{\mathrm{r}}(\Omega)$, where $1 \leq p, q, r < \infty$. Assume that the integrability indices $s_i \in [1, \infty]$, $i = 1, \ldots, N$, satisfy*

$$\begin{aligned}
N = 1 : &\qquad s_1^{-1} \leq 1 - \left(r^{-1} + q^{-1}\right) \\
N = 2 : &\qquad s_1^{-1} + s_2^{-1} \leq 1 - \max\left(r^{-1} + p^{-1}, 2q^{-1}\right) \\
N = 3 : &\qquad s_1^{-1} + s_2^{-1} + s_3^{-1} \leq 1 - \max\left(r^{-1}, p^{-1} + q^{-1}, 3q^{-1}\right).
\end{aligned}$$

Then, for $\delta F_i \in \mathrm{L}^{\mathrm{s_i}}(\Omega)$ we obtain

$$\frac{d^N}{(dF)^N}\Gamma(\det(F))(\delta F_1, \ldots, \delta F_N) \in \mathrm{L}^1(\Omega), \quad N = 1, 2, 3.$$

Proof. See [179, Lem. 4.2]. □

Now we turn to the study of derivatives of W.

Lemma 1.49. *Assume that* $F \in \mathrm{L}^{\mathrm{p}}(\Omega)$ *is nondegenerate,* $\operatorname{cof}(F) \in \mathrm{L}^{\mathrm{q}}(\Omega)$, *and* $\det(F) \in \mathrm{L}^{\mathrm{r}}(\Omega)$, *where* $1 \leq p, q, r < \infty$.

- If $0 \leq s_1^{-1} \leq 1 - (q^{-1} + \max(r^{-1}, p^{-1}))$, then

$$W'(F)\delta F \in \mathrm{L}^1(\Omega) \quad \text{for all } \delta F \in \mathrm{L}^{\mathrm{s}_1}(\Omega),$$

  is linear and continuous.
- If $0 \leq s_1^{-1} + s_2^{-1} \leq 1 - \max(2p^{-1}, r^{-1} + p^{-1}, 2q^{-1})$, then

$$W''(F)(\delta F_1, \delta F_2) \in \mathrm{L}^1(\Omega) \quad \text{for all } \delta F_i \in \mathrm{L}^{\mathrm{s}_i}(\Omega),\ i = 1, 2,$$

  is bilinear and continuous.
- If $0 \leq s_1^{-1} + s_2^{-1} + s_3^{-1} \leq 1 - \max(r^{-1}, p^{-1} + q^{-1}, 3q^{-1})$, then

$$W'''(F)(\delta F_1, \delta F_2, \delta F_3) \in \mathrm{L}^1(\Omega) \quad \text{for all } \delta F_i \in \mathrm{L}^{\mathrm{s}_i}(\Omega),\ i = 1, 2, 3,$$

  is trilinear and continuous.

Proof. See [179, Prop. 4.3]. □

Finally, we study conditions under which the formal directional derivatives of the strain energy

$$\mathcal{E}_{\varphi}^{\mathrm{str}}(\varphi)v_1 := \int_\Omega W'(\nabla\varphi)\nabla v \,\mathrm{d}\mu, \tag{1.3.1}$$

$$\mathcal{E}_{\varphi\varphi}^{\mathrm{str}}(\varphi)(v_1, v_2) := \int_\Omega W''(\nabla\varphi)(\nabla v_1, \nabla v_2) \,\mathrm{d}\mu, \tag{1.3.2}$$

$$\mathcal{E}_{\varphi\varphi\varphi}^{\mathrm{str}}(\varphi)(v_1, v_2, v_3) := \int_\Omega W'''(\nabla\varphi)(\nabla v_1, \nabla v_2, \nabla v_3) \,\mathrm{d}\mu, \tag{1.3.3}$$

are well defined. Moreover, we have to verify whether the remainder terms vanish.

Lemma 1.50. *Let* $\varphi = \mathrm{id} + u \in \Phi \cap \mathrm{W}^{1,\infty}(\Omega)$ *be nondegenerate. Then* $\mathcal{E}^{\mathrm{str}}$ *is twice directionally differentiable in* $\mathrm{W}^{1,\infty}(\Omega)$. *There exists a positive constant* $c_{r_2} > 0$, *such that corresponding remainder term can be estimated by*

$$r_2(\delta\varphi) \leq c_{r_2} \|\delta u\|_{\mathrm{W}^{1,\infty}(\Omega)} \|\delta u\|^2_{\mathrm{W}^{1,2}(\Omega)}.$$

Proof. See [179, Prop. 4.5]. □

With this result we get well-definedness of the necessary optimality conditions in $\mathrm{W}^{1,\infty}(\Omega)$.

Theorem 1.51. *Let $\varphi = \mathrm{id} + u \in \Phi \cap \mathrm{W}^{1,\infty}(\Omega)$ be a nondegenerate local minimizer of $\mathcal{E}^{\mathrm{str}}$ with $\mathcal{E}^{\mathrm{str}}(\varphi) < \infty$. Then it satisfies the weak formulation*

$$\mathcal{E}^{\mathrm{str}}_{\varphi}(\varphi)\delta\varphi = 0 \quad \text{for all } \delta\varphi \in \mathrm{W}^{1,\infty}(\Omega).$$

If $\mathcal{E}^{\mathrm{str}}_{\varphi\varphi}(\varphi) \geq \delta\|\delta u\|^2_{\mathrm{W}^{1,2}(\Omega)}$ with $\delta > 0$, then for sufficiently small $u \in \mathrm{W}^{1,\infty}(\Omega)$ and some constant $\varepsilon > 0$ we have the growth condition

$$\mathcal{E}^{\mathrm{str}}(\varphi + \delta\varphi) \geq \mathcal{E}^{\mathrm{str}}(\varphi) + \varepsilon\|\delta\varphi\|^2_{\mathrm{W}^{1,2}(\Omega)}.$$

In particular, φ is a $\mathrm{W}^{1,\infty}(\Omega)$-local minimizer of $\mathcal{E}^{\mathrm{str}}$.

Proof. To show that $\mathcal{E}^{\mathrm{str}}_{\varphi}(\varphi)\delta\varphi = 0$, we compute

$$\mathcal{E}^{\mathrm{str}}_{\varphi}(\varphi)(\pm\delta\varphi) = \lim_{t\to 0} \frac{\mathcal{E}^{\mathrm{str}}(\varphi + t\delta\varphi) - \mathcal{E}^{\mathrm{str}}(\varphi)}{t} \geq 0,$$

since φ is a local minimizer of $\mathcal{E}^{\mathrm{str}}$.

Regarding the second assertion, we note that

$$\begin{aligned}\mathcal{E}^{\mathrm{str}}(\varphi + \delta\varphi) - \mathcal{E}^{\mathrm{str}}(\varphi) &= \frac{1}{2}\mathcal{E}^{\mathrm{str}}_{\varphi\varphi}(\varphi)(\delta\varphi)^2 + r_2(\delta\varphi) \\ &\geq \frac{\delta}{2}\|\delta u\|^2_{\mathrm{W}^{1,2}(\Omega)} + r_2(\delta\varphi).\end{aligned}$$

Due to Lem. 1.50, we obtain, for $\|\delta u\|_{\mathrm{W}^{1,\infty}(\Omega)} \to 0$, the inequality

$$\mathcal{E}^{\mathrm{str}}(\varphi + \delta\varphi) - \mathcal{E}^{\mathrm{str}}(\varphi) \geq \left(\frac{\delta}{2} - c_{r_2}\|\delta u\|_{\mathrm{W}^{1,\infty}(\Omega)}\right)\|\delta u\|^2_{\mathrm{W}^{1,2}(\Omega)} \geq \varepsilon\|\delta u\|^2_{\mathrm{W}^{1,2}(\Omega)}.$$

□

1.4. Summary

Physical considerations yield the framework of hyperelastic material laws. For compressible materials the set of admissible functions is

$$\Phi := \left\{\varphi \in \mathrm{W}^{1,\mathrm{p}}(\Omega) : \ \det(\varphi) > 0 \text{ a.e. in } \Omega, \ \varphi = 0 \text{ a.e. on } \Gamma_{\mathrm{d}}\right\}.$$

In this regard, deformations Φ are given as possibly non-unique minimizers of hyperelastic energy functionals $\mathcal{E}(\varphi) = \mathcal{E}^{\mathrm{str}}(\varphi) - \mathcal{E}^{\mathrm{ext}}(\varphi, g)$, i.e.

$$\varphi \in \mathrm{argmin}_{\psi\in\Phi}\mathcal{E}(\psi),$$

where $\mathcal{E}^{\mathrm{ext}}$ describes the action of external forces g and $\mathcal{E}^{\mathrm{str}}(\varphi) = \int_\Omega W(\nabla\varphi)\,\mathrm{d}\mu$ is the strain energy determined by a material specific stored energy function W. In

particular the orientation preservation condition $\det(\nabla\varphi) > 0$ – and the corresponding limit behavior $\lim_{\det(F)\searrow 0} W(F) = \infty$ – imposes challenges. It implies that the singular set

$$\Phi_\infty = \left\{\varphi \in \mathrm{W}^{1,\mathrm{p}}(\Omega) :\ \mathcal{E}(\nabla\varphi) = \infty\right\}$$

is dense in $\mathrm{W}^{1,\mathrm{p}}(\Omega)$ and consequently $\mathcal{E}$ is neither bounded by a polynomial, nor convex.

Thus, the theoretical treatment of problems in nonlinear elasticity is challenging. In the chain of generalized convexity conditions

$$\text{convexity } \Rightarrow \textit{polyconvexity} \Rightarrow \text{ quasiconvexity } \Rightarrow \text{ rank-one convexity},$$

polyconvexity was identified as reasonable condition. To the current understanding of quasiconvexity, it requires a polynomial growth condition to prove existence of minimizers. This is in conflict with the limit behavior of W. Without such a growth condition polyconvexity is the only admissible candidate that fits into the setting of hyperelasticity and admits existence of minimizers. Moreover, polyconvex functions are sufficiently general to admit the development of complex, realistic models, as described in Chap. 5.

Another consequence of the orientation preservation condition is the fact that directional differentiability of $\mathcal{E}$ in $\mathrm{W}^{1,\mathrm{p}}(\Omega)$ can not be expected. Excluding the limit behavior for $\det(\nabla\varphi) \to \infty$ by the restriction of nondegeneracy on $\varphi \in \Phi$, i.e. $\det(\nabla\varphi) \geq \varepsilon_0 > 0$, the associated first order optimality conditions are at least well-defined in $\mathrm{W}^{1,\infty}(\Omega)$.

2. A mathematical model for implant shape design

The last chapter equipped us with a mathematical setting for the treatment of problems in hyperelasticity. Now we turn to the mathematical description of the implant shape design problem. We begin with the introduction of the forward problem in Sec. 2.1. Different formulations, incorporating the influence of the implant as obstacle or by pressure-type boundary conditions, are discussed in Sec. 2.2. The corresponding inverse problem for the determination of an implant shape from a desired displacement of the skin then is explained in Sec. 2.3. Replacing the pressure-type boundary conditions by dead load Neumann boundary conditions, which fit into the hyperelastic setting, a corresponding existence theorem is given in Sec. 2.4. Eventually, the first order optimality conditions for the chosen model are shortly discussed in Sec. 2.5.

The content of this chapter has essentially been published in [179].

2.1. The forward problem as obstacle problem

In the following we consider a generic polyconvex, coercive stored energy function W in $\mathbb{R}^3$ with associated functional

$$\mathcal{E}^{\mathrm{str}}(\varphi) := \int_\Omega W(x, \nabla\varphi(x))\,\mathrm{dx},$$

comprising all involved soft tissue types. Implants and bones are much stiffer than the soft tissues and are considered to be rigid, such as in [26]. Thus, there is no need to distinguish between both and we denote their volume by $\Omega_{\mathrm{rigid}} = \Omega_{\mathrm{bone}} \cup \Omega_{\mathrm{implant}}$. We further denote the contact surface between elastic and rigid material by

$$\Gamma_{\mathrm{c}} = \bar{\Omega}_{\mathrm{rigid}} \cap \bar{\Omega}$$

and the skin surface by

$$\Gamma_{\mathrm{o}} \subset \partial\Omega.$$

The remaining part of the boundary is denoted by

$$\Gamma_{\mathrm{d}} = \partial\Omega \setminus (\Gamma_{\mathrm{c}} \cup \Gamma_{\mathrm{o}}).$$

We begin with the specification of the boundary conditions on Γ_{o} and Γ_{d}.

- The skin surface is not restricted by external sources. This yields natural, homogeneous Neumann boundary conditions on Γ_o.
- Expecting the Green's function of nonlinear elasticity to vanish quickly, the soft tissue domain may be restricted to a bounded region in the vicinity of the implant. This introduces an artificial boundary Γ_d at the virtual cutting line. Here, transparent boundary conditions [171] might be imposed. For simplicity, we just assume the tissue to be fixed on Γ_d. The Dirichlet boundary conditions on Γ_d will be incorporated in the search space

$$\Phi = \left\{ \varphi \in \mathrm{W}^{1,p}(\Omega) : \ \det(\nabla \varphi) > 0 \text{ a.e. in } \Omega \text{ and } \varphi = 0 \text{ a.e. on } \Gamma_d \right\}.$$

In the human body bones are often surrounded by skeletal muscle tissue, which is attached via tendons at its end. Besides this partial fixation the muscle is allowed to freely glide over the contact surface Γ_c [255]. We assume that friction on the contact surface is negligible. Then we can account for the presence of rigid materials by restricting the set of admissible functions to

$$\Phi_{\Omega_{\text{rigid}}} = \{ \varphi \in \Phi : \ \varphi(x) \notin \Omega_{\text{rigid}} \quad \text{a.e. in } \Omega \}.$$

The arising optimization problem

$$\min_{\varphi \in \Phi_{\Omega_{\text{rigid}}}} \mathcal{E}^{\text{str}}(\varphi) \tag{2.1.1}$$

is an obstacle problem. Existence of solutions for the latter formulation is due to Thm. 1.44.

2.2. The forward problem and pressure-type boundary conditions

The obstacle formulation combines a mechanical model for soft tissue with a purely geometric description for solids. Thus, during computations for the solution of the non-convex problem (2.1.1) we will retain Ω_{rigid}, but the soft tissue domain $\varphi_k(\Omega)$ will change in each iteration k. Consequently, to guarantee that $\varphi_k \in \Phi_{\Omega_{\text{rigid}}}$ we need to solve a contact problem at Γ_c. These are inherently non-smooth [164, 223]. In addition to the non-convexity of $\mathcal{E}^{\text{str}}$ this further complicates the numerical solutions of (2.1.1).

An alternative approach is motivated by a formal theorem, given in [56].

Theorem 2.1. *Let Γ_c, Γ_d be disjoint relatively open subsets of $\Gamma = \partial\Omega$ such that $|\Gamma \setminus \{\Gamma_c \cup \Gamma_d\}| = 0$ and $|\Gamma_d| > 0$. For a closed subset $\mathcal{A} \subset \mathbb{R}^3 \setminus \Omega_{\text{rigid}}$ we define*

$$\Phi_{\mathcal{A}} = \{ \varphi \in \Phi : \ \varphi(\Gamma_c) \subset \mathcal{A} \}$$

as the set of admissible solutions. Then a smooth enough solution φ of the minimization problem

$$\min_{\varphi\in\Phi_{\mathcal{A}}} \mathcal{E}^{\mathrm{str}}(\varphi)$$

is, at least formally, a solution of the boundary value problem

$$\begin{aligned}
-\mathrm{div}(\sigma(\nabla\varphi)) &= 0 && \text{in } \Omega, && \text{(2.2.1a)}\\
\varphi &= 0 && \text{on } \Gamma_{\mathrm{d}}, && \text{(2.2.1b)}\\
\varphi(\Gamma_{\mathrm{c}}) &\subset \mathcal{A}, && && \text{(2.2.1c)}\\
\sigma(\nabla\varphi(x))n(x) &= 0 && \text{if } x\in\Gamma_{\mathrm{c}} \text{ and } \varphi(x)\in\mathring{\mathcal{A}}, && \text{(2.2.1d)}\\
\sigma(\nabla\varphi(x))n(x) &= g(x)\mathrm{cof}(\nabla\varphi(x))n(x) && \text{if } x\in\Gamma_{\mathrm{c}} \text{ and } \varphi(x)\in\partial\mathcal{A}, && \text{(2.2.1e)}
\end{aligned}$$

with $g(x)\le 0$ for all $x\in\Gamma_{\mathrm{c}}$.

Proof. As indicated in the statement of the theorem computations for its "proof" can only be performed on a formal basis [56, Thm. 5.3-1].

For a rigorous proof we would need that $\varphi(\Gamma_{\mathrm{c}})$ and $\partial\mathcal{A}$ are smooth, at least where they are in contact. Moreover application of Green's and Taylor's formula on variations of the stored energy function is, again due to condition (1.1.1), only valid on a formal level. Heuristically the first point seems to be justifiable, as we expect reasonable implant shapes to be smooth where they are in contact with soft tissue. Analytically this is an extremely challenging problem. First the vectorial problem setting already suggests that we should not expect more than partial regularity. Secondly, condition (1.1.1) rules out standard and non-standard polynomial growth conditions, which play an important role in regularity theory [191]. Therefore, available partial regularity results only focus on polyconvex integrands that do not satisfy condition (1.1.1), cf. [50, 100, 117, 126]. □

The boundary condition on Γ_{c} can be expressed as boundary condition on the deformed boundary $\varphi(\Gamma_{\mathrm{c}})$:

$$\begin{aligned}
\sigma^{\varphi}(\nabla\varphi(x))n^{\varphi}(\varphi(x)) &= 0 && \text{if } x\in\mathring{\mathcal{A}}\\
\sigma^{\varphi}(\nabla\varphi(x))n^{\varphi}(\varphi(x)) &= g^{\varphi}(\varphi(x))n^{\varphi}(\varphi(x)) && \text{if } x\in\partial\mathcal{A} && \text{(2.2.2)}
\end{aligned}$$

with $g^{\varphi}(x)\le 0$. "The unilateral boundary condition of place on Γ_{c} constitutes a model of contact without friction with the obstacle's boundary $\mathcal{A}$. In this respect, the function $g^{\varphi}:\varphi(\Gamma_{\mathrm{c}})\to\mathbb{R}$, which measures the intensity of the contact load, is nothing but the Kuhn-Tucker multiplier associated with the constraint $\varphi(\Gamma_{\mathrm{c}})\subset\mathcal{A}$" [56, p. 214].

At least formally, we can replace the obstacle condition by the pressure-type boundary condition (2.2.2) on the deformed domain. With this reformulation we avoid the intrinsic non-smoothness of contact problems. Though, formulating this boundary

condition on the undeformed domain, as in (2.2.1e), we encounter a quadratic nonlinearity in the boundary conditions on Γ_c, the cofactor matrix $\mathrm{cof}(\nabla\varphi(x))$. This is due to the fact that the Piola-transform does not preserve angles.

Consequently these boundary conditions are in general not conservative [15, 46]. This means that, except for simple cases such as piece-wise constant pressure, we cannot expect existence of an associated energy functional $\mathcal{E}^{\mathrm{ext}}$ and we can not incorporate pressure-type forces in the way we incorporate undirected volume or boundary forces into the hyperelastic setting. At least on the level of first order optimality conditions we formally get

$$\mathcal{E}_\varphi^{\mathrm{str}}(\varphi)h = \int_{\Gamma_c} g\,\mathrm{cof}(\nabla\varphi)nh\,\mathrm{ds} \quad \text{for all } h \in \mathrm{W}_0^{1,\mathrm{p}}(\Omega), \tag{2.2.3}$$

In Chap. 1 we saw that the left hand side of (2.2.3) is in general not well-defined. Due to the occurrence of $\mathrm{cof}(\nabla\varphi)$ the same is true for the right hand side if $\varphi \in \mathrm{W}^{1,\mathrm{p}}(\Omega)$ with $p \leq 3$, i.e. if we do not get continuity of φ by the Sobolev embedding theorem. Though, in our context of implant shape design we expect smooth deformations and a smooth contact surface Γ_c. In such a setting equation (2.2.3) indeed describes a necessary equilibrium condition for the obstacle problem.

Using pressure-type boundary conditions instead of the obstacle formulation has two advantages. First, we avoid the treatment of a contact problem. Second, (2.2.3) is the weak formulation of a nonlinear PDE, with non-standard boundary conditions and we can exploit insights into the numerical treatment of PDE-problems.

2.3. The inverse problem

We now turn to the question of finding an implant shape for a given desired shape, such as the deformation φ_{ref} of the outer surface Γ_o. A priori we do not know if any seemingly reasonable surface shape can indeed be attained. Therefore we search for deformations φ that minimize a cost functional $J_0(\varphi)$ which measures the deviation from the desired deformation φ_{ref}. Since it is not yet clear which measures are used by humans to quantify abnormalities, or even "beauty", the choice of adequate cost functionals is left open. Any weakly lower semicontinuous functional, which is bounded from below, may be a candidate. Here we take the simplest reasonable one, the tracking-type functional

$$J_0(\varphi) = \frac{1}{2}\|\varphi - \varphi_{\mathrm{ref}}\|^2_{\mathrm{L}^2(\Gamma_o)},$$

resp. its Tikhonov-regularized counterpart

$$J(\varphi, g) = \frac{1}{2}\|\varphi - \varphi_{\mathrm{ref}}\|^2_{\mathrm{L}^2(\Gamma_o)} + \frac{\alpha}{2}\|g\|^2_{\mathrm{L}^2(\Gamma_c)},$$

where $\alpha > 0$ denotes the Tikhonov regularization parameter. Since we aim to avoid a contact problem, we describe the influence of the implant via (2.2.3). This leads to the optimal control problem

$$\min_{(\varphi,g)\in\Phi\times \mathrm{L}^2(\Gamma_\mathrm{c})} J(\varphi, g) \tag{2.3.1a}$$

$$\text{subject to} \qquad \int_\Omega W'(\nabla\varphi) : \nabla h \,\mathrm{d}\mu = \int_{\Gamma_\mathrm{c}} g\mathrm{cof}(\nabla\varphi) n h \,\mathrm{ds} \quad \text{for all } h \in \mathrm{W}_0^{1,\mathrm{p}}(\Omega). \tag{2.3.1b}$$

From an optimal soft tissue deformation φ, solving (2.3.1), an implant shape must be reconstructed. As depicted in Fig. 2.3.1, this can be easily done by filling the gap between undeformed and deformed contact boundary.

In this formulation, it is straightforward to satisfy additional medical requirements. For instance, no gaps should occur between soft tissue and implants since voids tend to be a source of infections. The chosen construction of the implant shape directly meets this requirement. In contrast, imposing this condition in the obstacle formulation is quite involved.

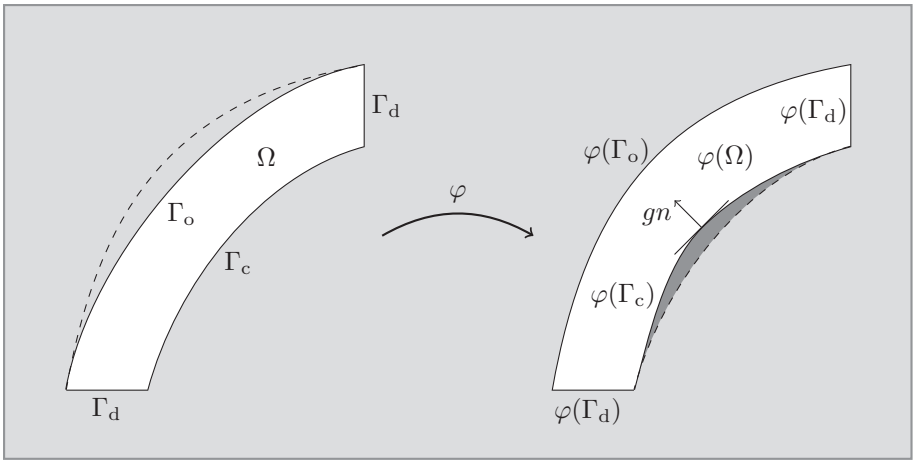


Figure 2.3.1.: Cross-section of the reference configuration (left) and the deformed state due to the normal force gn defining the implant volume in gray (right).

2.4. An existence result for dead load forces

Since the constraints of (2.3.1) are not well-defined we are not able to prove existence of solutions for this problem. Therefore, we will switch to a simplified setting and replace the pressure-type boundary condition $\sigma n = g\mathrm{cof}(\nabla\varphi)n$ by one of the following dead load boundary conditions:

$$\sigma n = \tilde{g} n \qquad \tilde{g} : \Gamma_\mathrm{c} \to \mathbb{R} \tag{2.4.1}$$

$$\sigma n = g \qquad g : \Gamma_\mathrm{c} \to \mathbb{R}^3. \tag{2.4.2}$$

Both conditions naturally enter linearly into the energy functional and can be augmented by a non-positivity constraint, such as $g \leq 0$ in the first and $\tilde{g} \leq 0$ in the second case. This simplification is reasonable if $\mathrm{cof}(\nabla\varphi) = \mathrm{cof}(I + \nabla u) \approx I$. From a practical point of view, one can expect that a solution of the simplified problem

$$\begin{aligned} &\min_{(\varphi,g)\in\Phi\times \mathrm{L}^2(\Gamma_\mathrm{c})} J(\varphi, g) \\ \text{subject to} \qquad &\varphi \in \mathrm{argmin}_{\psi\in\Phi}\mathcal{E}(\psi, g) := \mathcal{E}^{\mathrm{str}}(\psi) - \mathcal{E}^{\mathrm{ext}}(\psi, g) \end{aligned}$$

will yield an implant form that is sub-optimal with respect to the original problem, but still reasonable.

We begin with a lemma to establish that the minimizing property of deformations is retained by weak limits of suitable sequences $\{(\varphi_j, g_j)\}_j$.

Lemma 2.2. *Let*

$$\mathcal{E}^{\mathrm{str}}(\varphi) = \int_\Omega W(x, \nabla\varphi(x))\,\mathrm{dx}$$

be the energy functional associated to a polyconvex stored energy function W, satisfying a coercivity condition

$$W(x, \nabla\varphi(x)) \geq \alpha\left(\|\nabla\varphi(x)\|^p + \|\mathrm{cof}(\nabla\varphi(x))\|^q + |\det(\nabla\varphi(x))|^r\right) - \beta, \qquad (2.4.3)$$

with $\alpha > 0$, $,\beta \in \mathbb{R}$, $p \geq 2$, $\frac{1}{p} + \frac{1}{q} \leq 1$ and $r \geq 1$. Let

$$\mathcal{E}^{\mathrm{ext}}(\varphi, g) = \int_{\Gamma_\mathrm{c}} g(x)\varphi(x)\,\mathrm{dx}$$

be the functional associated with one of the dead load boundary conditions (2.4.1) or (2.4.2) and let $\mathcal{E}(\varphi, g) = \mathcal{E}^{\mathrm{str}}(\varphi) - \mathcal{E}^{\mathrm{ext}}(\varphi, g)$. Consider a weakly converging sequence $(\varphi_j, g_j) \rightharpoonup (\bar{\varphi}, \bar{g})$ in $\Phi \times \mathrm{L}^2(\Gamma_\mathrm{c})$, such that

$$\varphi_j \in \mathrm{argmin}_{\psi\in\Phi}\mathcal{E}\left(\psi, g_j\right),$$

and $\{\mathcal{E}(\varphi_j, g_j)\}_j$ is bounded from above. Then

$$\lim_{j\to\infty} \mathcal{E}\left(\varphi_j, g_j\right) = \mathcal{E}\left(\bar{\varphi}, \bar{g}\right) = \min_{\psi\in\Phi} \mathcal{E}\left(\psi, \bar{g}\right).$$

Proof. First, we show weak lower semicontinuity of $\mathcal{E}$. Weak lower semicontinuity of $\mathcal{E}^{\mathrm{str}}$ follows as in Thm. 1.42. The second part $\mathcal{E}^{\mathrm{ext}}$ is even weakly continuous. This follows via compactness of the trace mapping $\mathrm{W}^{1,\mathrm{p}}(\Omega) \to \mathrm{L}^2(\Gamma_\mathrm{c})$, which yields strong convergence $\varphi_j|_{\Gamma_\mathrm{c}} \to \bar{\varphi}|_{\Gamma_\mathrm{c}}$ in $\mathrm{L}^2(\Gamma_\mathrm{c})$, and weak convergence $g_j \rightharpoonup \bar{g}$ in $\mathrm{L}^2(\Gamma_\mathrm{c})$. Altogether, we conclude weak lower semicontinuity of $\mathcal{E}$, i.e.

$$\mathcal{E}(\bar{\varphi}, \bar{g}) \leq \liminf_{j\to\infty} \mathcal{E}(\varphi_j, g_j).$$

Next, by the minimizing property of φ_j we obtain $\mathcal{E}(\varphi_j, g_j) \leq \mathcal{E}(\bar{\varphi}, g_j)$ and

$$\limsup_{j\to\infty} \mathcal{E}(\varphi_j, g_j) \leq \limsup_{j\to\infty} \mathcal{E}(\bar{\varphi}, g_j) = \lim_{j\to\infty} \mathcal{E}(\bar{\varphi}, g_j) = \mathcal{E}(\bar{\varphi}, \bar{g}).$$

This implies

$$\limsup_{j\to\infty} \mathcal{E}(\varphi_j, g_j) \leq \mathcal{E}(\bar{\varphi}, \bar{g}) \leq \liminf_{j\to\infty} \mathcal{E}(\varphi_j, g_j)$$

and thus

$$\lim_{j\to\infty} \mathcal{E}(\varphi_j, g_j) = \mathcal{E}(\bar{\varphi}, \bar{g}).$$

The fact that $\bar{\varphi}$ is an energy minimizer of $\mathcal{E}(\cdot, \bar{g})$ follows from the minimizing property of φ_j and the established convergence result. To this end let $\tilde{\varphi}$ be a minimizer of $\mathcal{E}(\cdot, \bar{g})$. Then

$$\mathcal{E}(\tilde{\varphi}, \bar{g}) \leq \mathcal{E}(\bar{\varphi}, \bar{g}) = \lim_{j\to\infty} \mathcal{E}(\varphi_j, g_j) \leq \lim_{j\to\infty} \mathcal{E}(\tilde{\varphi}, g_j) = \mathcal{E}(\tilde{\varphi}, \bar{g}).$$

□

Theorem 2.3. *Let $J\colon \Phi \times \mathrm{L}^2(\Gamma_\mathrm{c}) \to \mathbb{R}$ be weakly lower semicontinuous satisfying*

$$J(\cdot, g) \geq c_J \|g\|^k_{\mathrm{L}^2(\Gamma_\mathrm{c})} \tag{2.4.4}$$

for all $g \in \mathrm{L}^2(\Gamma_\mathrm{c})$ and constants $c_J > 0$, $k > 0$. Let $\mathcal{E}, \mathcal{E}^{\mathrm{str}}, \mathcal{E}^{\mathrm{ext}}$ be defined as in Lem. 2.2. If, for each $g \in \mathrm{L}^2(\Gamma_\mathrm{c})$, there exists at least one $\varphi_g \in \Phi$ such that $\mathcal{E}^{\mathrm{str}}(\varphi_g, g) < \infty$, then the problem

$$\begin{aligned} &\min_{(\varphi,g)\in\Phi\times\mathrm{L}^2(\Gamma_\mathrm{c})} J(\varphi, g) \\ \textit{subject to} \qquad &\varphi \in \operatorname{argmin}_{\psi\in\Phi} \mathcal{E}(\psi, g) := \mathcal{E}^{\mathrm{str}}(\psi) - \mathcal{E}^{\mathrm{ext}}(\psi, g) \end{aligned}$$

has at least one solution.

Proof. We first note that inequality (2.4.3) implies for some constants $\tilde{\alpha} > 0$ and $\tilde{\beta} \in \mathbb{R}$

$$\mathcal{E}^{\mathrm{str}}(\varphi, \tilde{g}) \geq \tilde{\alpha} \left(\|\nabla\varphi\|^p_{\mathrm{L}^\mathrm{p}(\Omega)} + \|\mathrm{cof}(\nabla\varphi)\|^q_{\mathrm{L}^\mathrm{q}(\Omega)} + \|\det(\nabla\varphi)\|^r_{\mathrm{L}^\mathrm{r}(\Omega)} \right) - \tilde{\beta}. \tag{2.4.5}$$

According to Thm. 1.38 the functional $\mathcal{E}^{\mathrm{str}}$ is weakly lower semicontinuous in φ. Since $\mathcal{E}^{\mathrm{ext}}$ is weakly continuous in its second argument g the same holds for $\mathcal{E}$. Now let $\{(\varphi_j, g_j)\}_j$ be an infimizing sequence, i.e. with Thm. 1.42,

$$\varphi_j \in \operatorname{argmin}_{\psi\in\Phi} \mathcal{E}(\psi, g_j)$$

and

$$\lim_{j\to\infty} J(\varphi_j, g_j) = \inf_{(\varphi,g)\in\Phi\times\mathrm{L}^2(\Gamma_\mathrm{c})} J(\varphi, g) < \infty.$$

Due to (2.4.4) the sequence $\{g_j\}_j$ is bounded by some constant C_g, and as $\mathrm{L}^2(\Gamma_c)$ is reflexive we can extract a weakly convergent subsequence $g_j \rightharpoonup \bar{g}$ in $\mathrm{L}^2(\Gamma_c)$. We show that we can also extract a weakly converging subsequence from $\{\varphi_j\}_j$, using the linearity of the boundary conditions to prove the boundedness of $\{\varphi_j\}_j$.

First note that with w.l.o.g. $h \equiv 0$ and corresponding minimizer φ_h[1]

$$\begin{aligned}\mathcal{E}(\varphi_j, g_j) - \mathcal{E}(\varphi_h, h) &\leq \mathcal{E}(\varphi_h, g_j) - \mathcal{E}(\varphi_h, h) \\ &= -\int_{\Gamma_c} g_j \varphi_h \,\mathrm{ds} \leq \|g_j\|_{\mathrm{L}^2(\Gamma_c)} \|\varphi_h\|_{\mathrm{L}^2(\Gamma_c)} \leq C_g \|\varphi_h\|_{\mathrm{L}^2(\Gamma_c)},\end{aligned}$$

and consequently for all $j \in \mathbb{N}$

$$\tilde{\alpha} \|\nabla \varphi_j\|^p_{\mathrm{L}^p(\Omega)} \leq \mathcal{E}(\varphi_j, g_j) \leq \mathcal{E}(\varphi_h, h) + C_g \|\varphi_h\|_{\mathrm{L}^2(\Gamma_c)}.$$

We extract a weakly convergent subsequence $\varphi_j \rightharpoonup \bar{\varphi}$ in Φ and get with the weak lower semicontinuity of J:

$$J(\bar{\varphi}, \bar{g}) \leq \liminf_{j \to \infty} J(\varphi_j, g_j).$$

Analogously to Thm. 1.42 it follows that $\bar{\varphi} \in \Phi$ and according to Lem. 2.2 the weak limit $\bar{\varphi}$ is a minimizer of $\mathcal{E}(\cdot, \bar{g})$. □

2.5. Formal first order optimality conditions for the implant shape design problem

For the development of an optimization algorithm in the next chapter we will need the first order optimality conditions for the equality constrained optimization problem

$$\min_{(\varphi, g) \in \Phi \times \mathrm{L}^2(\Gamma_c)} J(\varphi, g)$$

$$\text{subject to} \qquad \int_\Omega W(\nabla \varphi) : \nabla h \,\mathrm{d}\mu = \int_{\Gamma_c} g \mathrm{cof}(\nabla \varphi) n h \,\mathrm{ds} \quad \text{for all } h \in \mathrm{W}^{1,\mathrm{p}}_0(\Omega).$$

In the chosen setting, it is not possible to rigorously derive these for the reduced equality constrained optimization problem. The main problem is the fact that, due to the orientation preservation condition (1.1.1), the justification of the formal first order conditions for polyconvex hyperelastic material laws is an open problem, cf. Ball [17]. In order to get differentiability of the elastic energy functional, we would need $\mathrm{W}^{1,\infty}$ as topological framework. However, due to the difficulties indicated in Chap. 1, a suitable existence result is not available. Therefore, we proceed in a formal way.

[1] Clearly $\varphi_h(x) = x$ should hold. However, this is not enforced in elasticity theory.

We begin with the formal Lagrangian function

$$\mathcal{L}(\varphi, g, p) = J(\varphi, g) + p\left(\mathcal{E}_\varphi\left(\varphi, g\right)\right).$$

The corresponding KKT-conditions are

$$\begin{aligned}
\frac{\partial}{\partial \varphi}\mathcal{L}(\varphi, g, p) &= J_\Phi(\varphi, g) + p\left(\mathcal{E}_{\varphi\varphi}(\varphi, g)\right) = 0,\\
\frac{\partial}{\partial g}\mathcal{L}(\varphi, g, p) &= J_g(\varphi, g) + p\left(\mathcal{E}_{\varphi g}(\varphi, g)\right) = 0,\\
\frac{\partial}{\partial p}\mathcal{L}(\varphi, g, p) &= \mathcal{E}_\varphi(\varphi, g) = 0.
\end{aligned}$$

As $\mathcal{E}_\varphi(\varphi, g)$ is in general not well-defined we cannot expect that the state and adjoint equation are well-defined. However, if the deformation gradients uniformly stay away from zero, if they are nondegenerate, the KKT-system is often well-defined in $\mathrm{W}^{1,\infty}$, cf. Sec. 1.3.

Considering our particular choice of boundary conditions and a Tikhonov-regularized, tracking type cost functional

$$J(\varphi, g) = \frac{1}{2}\|\varphi - \varphi_{\mathrm{ref}}\|^2_{\mathrm{L}^2(\Gamma_{\mathrm{o}})} + \frac{\alpha}{2}\|g\|^2_{\mathrm{L}^2(\Gamma_{\mathrm{c}}),}$$

with $\alpha > 0$, we can further specify our optimality conditions to

$$\varphi - \varphi_{\mathrm{ref}} + p\left(\mathcal{E}_{\varphi\varphi}(\varphi, g)\right) = 0, \tag{2.5.1a}$$

$$\alpha g + p\left(g\mathrm{cof}(\nabla\varphi)n\right) = 0, \tag{2.5.1b}$$

$$\mathcal{E}_\varphi(\varphi, g) = 0. \tag{2.5.1c}$$

Often (2.5.1b) is used to eliminate the control variable from the optimality system, cf. [178, 259]. This reduction is not compatible with our algorithmic approach of a composite step method, which will be introduced in the next chapter.

2.6. Summary

The influence of an implant on a hyperelastic soft tissue can be formulated as an obstacle problem, which requires the treatment of a contact problem. This can be avoided if instead of the implant shape we incorporate the pressure-type force $g^\varphi n^\varphi$ that is exerted on the soft tissue. The transformation of the corresponding equilibrium conditions onto the undeformed domain then leads to a nonlinear boundary condition $g\mathrm{cof}(\nabla\varphi)n$ with $g \leq 0$. In terms of this pressure-type formulation, the forward problem then is given through

$$\begin{aligned}
&\min_{\varphi\in\Phi} \mathcal{E}^{\mathrm{str}}(\varphi)\\
\text{subject to}\qquad &\sigma(\nabla\varphi)n = g\mathrm{cof}(\nabla\varphi)n \quad \text{on } \Gamma_{\mathrm{c}}.
\end{aligned}$$

The corresponding formal first order conditions are given via

$$\int_\Omega W(\nabla\varphi) : \nabla h \,\mathrm{d}\mu = \int_{\Gamma_c} g\mathrm{cof}(\nabla\varphi)nh \,\mathrm{ds} \quad \text{for all } h \in \mathrm{W}_0^{1,\mathrm{p}}(\Omega).$$

Regarding the determination of an implant shape, resp. its exerted force g, from a desired surface shape, leads to the Tikhonov-regularized equality constrained optimization problem

$$\begin{aligned} &\min_{(\varphi,g)\in\Phi\times \mathrm{L}^2(\Gamma_c)} J(\varphi,g) = J_0(\varphi) + \frac{\alpha}{2}\|g\|^2_{\mathrm{L}^2(\Gamma_c)} \\ \text{subject to} \qquad &\mathcal{E}^{\mathrm{str}}_\varphi(\varphi,g)v - \int_{\Gamma_c} g\mathrm{cof}(\nabla\varphi)nv \,\mathrm{ds} = 0 \quad \text{for all } v \in \mathrm{W}^{1,p}(\Omega;\mathbb{R}^3). \end{aligned}$$

The cost functional $J_0(\varphi)$ must be weakly lower semicontinuous. As the quantification of esthetics is not well understood we here employ the simple tracking-type functional $J_0(\varphi) = \frac{1}{2}\|\varphi - \varphi_{\mathrm{ref}}\|^2_{\mathrm{L}^2(\Gamma_o)}$, where φ_{ref} is the given deformation on the observation surface.

Theoretical results are challenging to attain for this problem. In particular, the nonlinear boundary, live load boundary conditions $\sigma(\nabla\varphi)n = g\mathrm{cof}(\nabla\varphi)n$ are difficult to incorporate. If these are replaced by simpler, dead load boundary conditions of the form $\sigma(\nabla\varphi)n = g$, resp. $\sigma(\nabla\varphi)n = gn$, we can prove existence of solutions for the regularized bi-level optimization problem, cf. Thm. 2.3.

For numerical computations we employ the formal first order conditions of the forward problem instead of the optimization problem and solve

$$\begin{aligned} &\min_{(\varphi,g)\in\Phi\times \mathrm{L}^2(\Gamma_c)} J(\varphi,g) := J_0(\varphi) + \frac{\alpha}{2}\|g\|^2_{\mathrm{L}^2(\Gamma_c)} \\ \text{subject to} \qquad &\int_\Omega W(\nabla\varphi) : \nabla h \,\mathrm{d}\mu = \int_{\Gamma_c} g\mathrm{cof}(\nabla\varphi)nh \,\mathrm{ds} \quad \text{for all } h \in \mathrm{W}_0^{1,\mathrm{p}}(\Omega). \end{aligned}$$

Formally, the corresponding Lagrangian is

$$\mathcal{L}(\varphi,g,p) = J(\varphi,g) + \int_\Omega W(\nabla\varphi) : \nabla p \,\mathrm{d}\mu - \int_{\Gamma_c} g\mathrm{cof}(\nabla\varphi)np \,\mathrm{ds}.$$

For a tracking-type cost functional the formal first order optimality conditions of the inverse problem are

$$\begin{aligned} \varphi - \varphi_{\mathrm{ref}} + p\,(\mathcal{E}_{\varphi\varphi}(\varphi,g)) &= 0, \\ \alpha g + p\,(g\mathrm{cof}(\nabla\varphi)n) &= 0, \\ \mathcal{E}_\varphi(\varphi,g) &= 0. \end{aligned}$$

3. An affine covariant composite step method

In this chapter an algorithm for nonlinear equality constrained optimization is presented. Particular focus is on the efficient solution of optimization problems with partial differential equations as constraints. These problems are originally posed in function space and become – after discretization – large scale problems with special structure, inherited from the infinite dimensional setting. Due to the mentioned difficulties associated with finite elasticity, a more regular problem setting is considered in the derivation of the algorithm. In this and the following chapter the usual notation of optimization and optimal control is used. Note that there is some ambiguity with elasticity where u denotes the displacement, a state variable, in contrast to its meaning in optimal control, where it denotes the control. After the introduction of the algorithm and practical details we turn back to the notation of elasticity. The content of this chapter has been published in Lubkoll, Schiela and Weiser [180], except for Sec. 3.5, which extends the results of [180].

To fix the problem setting, let $(\mathrm{X}, \langle\cdot,\cdot\rangle)$ be a Hilbert space and P a reflexive Banach space. In this setting, we consider the optimization problem

$$\begin{aligned} &\min_{x\in \mathrm{X}} f(x) \\ &\text{subject to} \qquad c(x) = 0, \end{aligned}$$

where $f : \mathrm{X} \to \mathbb{R}$ and $c : \mathrm{X} \to \mathrm{P}^*$ are twice continuously Fréchet differentiable. The constraint $c : \mathrm{X} \to \mathrm{P}^*$ maps into the dual space of P so that it can model a differential equation in weak form:

$$c(x) = 0 \ \text{ in } \mathrm{P}^* \quad \Leftrightarrow \quad c(x)v = 0 \ \text{ for all } v \in \mathrm{P}.$$

Here we use that P is a reflexive space, expressed a little sloppily by the relation $\mathrm{P} = \mathrm{P}^{**}$. In the context of optimal control it is common to split the variable x into two parts $\mathrm{X} = \mathrm{Y} \times \mathrm{U}$ and $x = (y, u)$, where y denotes the *state* and u the *control*. This splitting comes from the special structure of the equality constraints

$$c(x) = A(y) - Bu,$$

where $A : \mathrm{Y} \to \mathrm{P}^*$ is a nonlinear differential operator with continuous inverse, and $B : \mathrm{U} \to \mathrm{P}^*$ a linear, compact operator. Under these structural assumptions, it is

possible to show existence of minimizers and corresponding optimality conditions via the implicit function and the closed range theorem, cf. Sec. 3.1 and Sec. 3.2.

As algorithmic approach a *composite step method* is chosen. This class of methods is well established in nonlinear optimization. Its way to cope with the double aim of feasibility and optimality is to split the full Lagrange-Newton step δx into a *normal step* δn and a *tangential step* δt. More precisely, δn will be a minimum norm Gauss-Newton step for the solution of the underdetermined problem $c(x) = 0$, and δt aims to minimize f on the current null space of the linearized constraints. For globalization both are modified separately.

A couple of variants have been proposed in the literature, cf. [64, Sec. 15.4]. Our approach resembles the Vardi approach [265] in the sense that normal steps are computed as damped Newton steps for the underdetermined equation $c(x) = 0$ and thus always satisfy $\nu c(x) + c'(x)\delta n = 0$ for some damping factor $\nu \in]0, 1]$. Compared to the approach of Byrd-Omojokun [47, 48, 202], where normal steps are computed as minimizers of $\|c(x)\|$ in a trust region, Vardi methods need in addition surjectivity of $c'(x)$ as a prerequisite for the computation of steps. This is widely considered as a weakness of this class of methods as a basis for a general purpose solver.

Here, a Vardi-type method is an appropriate choice. Due to the above described structure of optimal control problems, one can usually exclude the case of non-surjective $c'(x)$. So the extra assumption imposed by Vardi-type methods is fulfilled.

The space P^* of residuals $c(x)$ is a dual space, which is often quite irregular and hard to compute. Therefore, we aim to avoid the computation of norms of residuals. Instead, an extension, due to Anton Schiela, of the *affine covariant Newton methods* for underdetermined problems, as described by Deuflhard [76, Sec. 4.4] is used here. In fact, if $f = \text{const.}$, the proposed composite step algorithm reduces to Deuflhard's method. In this context, a Vardi-type damping strategy is the natural result.

Regarding globalization of the tangential step trust region methods are widely used [64]. From an algorithmic perspective the choice of the trust region parameter is unsatisfactory, as its determination does not take into account problem specific information. An alternative is provided by *cubic regularization methods*, as used in [53, 54, 226, 277]. More precisely the globalization method used for the tangential step is motivated by Weiser et al. [277], where the regularization parameter is expressed in terms of an estimate of an affine invariant, in our context affine covariant, Lipschitz constant. In the absence of equality constraints, the proposed algorithm reduces to the cubic regularization method for unconstrained optimization of [277].

In this way, the globalization procedure, described in detail in Sec. 3.3, results in the following algorithmic behavior: *far away* from a feasible point priority is given to come close to a *feasible* solution. In this phase the method behaves like a damped Newton method for underdetermined systems. *Close* to the feasible manifold *optimality* is stressed, with the restriction that the iterates remain in the Kantorovich neighborhood of contraction around the feasible set. For this we use parametrized models for the nonlinearity of the functional and the constraints. Since our model

for the functional is *quadratic*, we use a *cubic* model for the error, while our *linear* model for the constraints is augmented by a *quadratic* model for the error.

In the following sections a practical algorithm is developed along these ideas. Some preliminary theoretical results, such as finite termination of the "inner loop" (Sec. 3.4) and fast local convergence (Sec. 3.5) are established. A proof of global convergence is not in the scope of this thesis, and will certainly require some modifications of the algorithm. In particular, it is known that affine covariant Newton methods, although very successful in practice, lack a rigorous proof of global convergence. Due to affine covariance the evaluation of $\|c(x)\|$ and thus the usual globalization mechanisms are not available.

The functional analytic framework for our algorithms forces us to distinguish precisely between primal and dual quantities. Emphasis is on the distinction between the linear functional $f'(x) \in \mathrm{X}^*$ and the gradient $\nabla f(x) \in \mathrm{X}$. Both are connected by the Riesz isomorphism $M : \mathrm{X} \to \mathrm{X}^*$, which maps $v \in \mathrm{X}$ to the linear functional $\langle v, \cdot \rangle \in \mathrm{X}^*$. In nonconvex problems M is usually a non-trivial mapping. Similarly, we use the derivative $c'(x) : \mathrm{X} \to \mathrm{P}^*$, instead of $\nabla c(x)$, which is widely seen in the literature, but not useful in a functional analytic setting. Concerning adjoint mappings, Banach space adjoints are used, i.e. $c'(x)^* : \mathrm{P} \to \mathrm{X}^*$ is defined by $(c'(x)^* p)(v) = pc'(x)v = p(c'(x)v)$. In this context, expressions like $c'(x)^* c'(x)$ are not well defined, since the range of $c'(x)$ is not related to the domain of $c'(x)^*$.

3.1. Lagrange multipliers and normal steps

Let us consider a generic equality constrained optimization problem on a Hilbert space X:

$$\min f(x) \tag{3.1.1a}$$

$$\text{subject to} \qquad c(x) = 0. \tag{3.1.1b}$$

Since f and c are twice Fréchet differentiable and $c'(x_*)$ is surjective at a stationary point x_* we can derive the KKT conditions at x_*. These conditions assert that there exists a Lagrange multiplier $p \in \mathrm{P}^{**} = \mathrm{P}$, such that

$$\mathcal{L}_x(x_*, p)v = f'(x_*)v + pc'(x_*)v = 0 \qquad \text{for all } v \in \mathrm{X} \tag{3.1.2}$$

$$c(x_*) = 0. \tag{3.1.3}$$

Here, (3.1.2) expresses the stationarity condition in $\ker c'(x_*)$:

$$pc'(x_*)v = 0 \Rightarrow f'(x_*)v = 0 \qquad \text{for all } v \in \ker c'(x_*). \tag{3.1.4}$$

Since X is a Hilbert space, equipped with scalar product $\langle \cdot, \cdot \rangle$, we can perform the splitting

$$\mathrm{X} = \ker c'(x_*) \oplus (\ker c'(x_*))^{\perp}$$

of X into $\ker c'(x_*)$ and its orthogonal complement $(\ker c'(x_*))^\perp$. Application of this splitting to (3.1.2) then yields the equivalence

$$\mathcal{L}_x(x_*,p)v = 0 \text{ for all } v \in \mathrm{X} \;\Leftrightarrow\; \begin{cases} f'(x_*)v = 0 \text{ for all } v \in \ker c'(x_*), \\ (f'(x_*) + pc'(x_*))w = 0 \text{ for all } w \in (\ker c'(x_*))^\perp . \end{cases}$$

The first condition on the right hand side characterizes stationarity of x_* and neither depends on p, nor on the scalar product. In contrast, the second condition

$$f'(x_*)w + pc'(x_*)w = 0 \quad \text{for all } w \in (\ker c'(x_*))^\perp , \tag{3.1.5}$$

depends on both the scalar product $\langle\cdot,\cdot\rangle$ and the Lagrange multiplier p. We will see that the validity of (3.1.5) has nothing to do with the stationarity of x_*. Instead, (3.1.5) holds for *arbitrary* $x \in \mathrm{X}$, as long as $c'(x)$ is surjective, and the corresponding Lagrange multiplier p can be computed by solving the linear system

$$\begin{pmatrix} M & c'(x)^* \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} v \\ p \end{pmatrix} + \begin{pmatrix} f'(x) \\ 0 \end{pmatrix} = 0, \tag{3.1.6}$$

where $M : \mathrm{X} \to \mathrm{X}^*$ is a Riesz isomorphism, characterized by $(Mv)(w) = \langle v, w\rangle$.

Theorem 3.1. *For $x \in \mathrm{X}$ assume that $c'(x) : \mathrm{X} \to \mathrm{P}^*$ is bounded and surjective. Then there is a unique element $p^x \in \mathrm{P}$ that solves (3.1.6) and satisfies*

$$f'(x)w + p^x c'(x)w = 0 \quad \textit{for all } w \in (\ker c'(x))^\perp . \tag{3.1.7}$$

Proof. Block operators of the form encountered in (3.1.6) are continuously invertible (in a Banach space context) as long as $c'(x)$ is bounded and surjective and the symmetric bilinear form $\langle v, w\rangle = (Mv)(w)$ is elliptic on $\ker(c'(x))$ and continuous. This is the result of the famous Brezzi splitting theorem [41, Thm. 4.3].

Now we test the first row of (3.1.6) with $w \in (\ker c'(x))^\perp$:

$$(Mv)(w) + p^x c'(x)w + f'(x)w = 0.$$

Since $w \in (\ker c'(x))^\perp$ and $v \in \ker c'(x)$, by the second row of (3.1.6), we conclude $(Mv)(w) = 0$ and thus (3.1.7). □

Definition 3.2. We call p^x, as defined in Thm. 3.1, the **Lagrange multiplier** of problem (3.1.1) at x.

Remark 3.3. In the context of the Euclidean scalar product

$$\langle v, w\rangle_2 := v^T w \quad \text{for } v, w \in \mathbb{R}^n$$

the multiplier p^x minimizes $\|f'(x)^T + c'(x)^T p\|_2$. For this reason a Lagrangian multiplier that is computed via (3.1.6) is known as a "least-squares estimate for p".

In the next section we will see that our special Lagrange multiplier enjoys a couple of favorable properties, also far away from an optimal solution. Before, we consider its relation to the current iterate x and what this means for the Lagrangian function of (3.1.1).

Lemma 3.4. *Let $x_0 \in \mathrm{X}$ and assume that f' and c' depend Lipschitz continuously on x. Further, assume that $c'(x) : \mathrm{X} \to \mathrm{P}^*$ is a bounded and surjective linear operator in an open neighborhood $\tilde{U}(x_0)$. Then the Lagrange multiplier p^x at x is given as Lipschitz continuous implicit function of x in some neighborhood around x_0, i.e. there exists a neighborhood $U(x_0)$ and a constant $\varepsilon_c(x_0) > 0$ such that for all $x \in U(x_0)$ holds*

$$\|p^x - p^{x_0}\| \leq \varepsilon_c(x_0)\|x - x_0\|.$$

Proof. Consider (3.1.6), i.e. for some $v \in \ker c'(x)$

$$a(x,p) := \begin{pmatrix} M & c'(x)^* \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} v \\ p \end{pmatrix} + \begin{pmatrix} f'(x) \\ 0 \end{pmatrix} = 0,$$

As $c'(x)$, and consequently $c'(x)^*$, are bounded linear operators, a is Fréchet differentiable in its second argument. Surjectivity of $c'(x)$ yields bijectivity of

$$a_p(x,p) \colon \mathrm{P} \to \mathcal{R}(a)$$

in $\tilde{U}(x_0)$. By assumption, both a and a_p are Lipschitz continuous at (x_0, p^{x_0}). Thus, we can apply the implicit function theorem (Thm. A.2) to $a \colon \mathrm{X} \times \mathrm{P} \to \mathcal{R}(a)$ to get the desired result. □

Lagrangian function. Let us discuss our result in terms of the Lagrangian function

$$\mathcal{L}(x,p) := f(x) + pc(x),$$

where $p = p^x$ is chosen as in Thm. 3.1. Our result implies that normal steps δn do not change the Lagrangian function up to first order:

$$\mathcal{L}_x(x,p^x)\delta n = f'(x)\delta n + p^x c'(x)\delta n = 0 \quad \text{for all } \delta n \in (\ker c'(x))^{\perp}.$$

Thus, our particular choice of the Lagrange multiplier p^x makes $\mathcal{L}_x(x,p^x)$ stationary in $(\ker c'(x))^{\perp}$. In contrast, for tangential steps δt, which are contained in $\ker c'(x)$, the relevant relation is given through

$$\mathcal{L}_x(x,p^x)\delta t = f'(x)\delta t + p^x c'(x)\delta t = f'(x)\delta t \quad \text{for all } \delta t \in \ker c'(x).$$

For the composite step $\delta x = \delta n + \delta t$ this yields

$$\mathcal{L}_x(x,p^x)\delta x = \mathcal{L}_x(x,p^x)(\delta n + \delta t) = f'(x)\delta t$$

If we look at a second order approximation of $\mathcal{L}$ along δx we obtain

$$\mathcal{L}(x+\delta x, p^x) = \mathcal{L}(x, p^x) + f'(x)\delta t + \frac{1}{2}\mathcal{L}_{xx}(x, p^x)(\delta x)^2 + o(\|\delta x\|^2).$$

Hence, p^x only enters in the second order approximation of $\mathcal{L}$. Below, in Sec. 3.2, we will construct a similar second order model for f, which avoids the well known Maratos effect [64, 206], i.e. our scheme will asymptotically fade into a Lagrange-Newton method.

3.2. Composite steps and their consistency

In this section we discuss some properties of composite steps and in particular their order of consistency, i.e. the asymptotic behavior of the difference between quadratic models and actual problem. Classically, composite steps are composed from a normal step δn and a tangential step δt. In our framework we add an additional *simplified normal step* δs that also plays the role of a *second order correction*.

For this purpose we introduce the following notation, which refers to a single step of our algorithm. Consider a fixed iterate x with Lagrange multiplier p^x, computed as in Thm. 3.1. We denote the (damped) normal step by $\delta n \in (\ker c'(x))^\perp$ and the tangential step by $\delta t \in \ker c'(x)$. The undamped normal step is denoted by Δn, so that $\delta n := \nu \Delta n$, where $\nu \in]0, 1]$ is a damping factor. A similar notation is conceivable for tangential steps. However, the computation of their directions and lengths may also be performed in one step.

The ordinary composite step is given by

$$\delta x := \delta n + \delta t, \tag{3.2.1}$$

and the extended composite step by

$$\delta \bar{x} := \delta x + \delta s.$$

The contributions to these steps have to fulfill the following equations:

$$c(x) + c'(x)\Delta n = 0 \quad \text{undamped normal step} \tag{3.2.2a}$$
$$c'(x)\delta t = 0 \quad \text{tangential step} \tag{3.2.2b}$$
$$[c(x+\delta x) - c(x) - c'(x)\delta x] + c'(x)\delta s = 0 \quad \text{simplified normal step.} \tag{3.2.2c}$$

Since in general $\ker c'(x)$ is non-trivial the steps are not fully determined by (3.2.2). We use our scalar product to uniquely determine Δn and δs as minimum norm corrections. The tangential step will be determined by approximately minimizing a quadratic model of $\mathcal{L}$ on $\ker c'(x)$, which corresponds to a quadratic model of f on the feasible set $c(x) = 0$.

3.2.1. Computation of steps via saddle point systems

We begin with the specification of conditions that determine the normal steps Δn, the Lagrange multiplier p^x, tangential steps δt, and the simplified normal step δs. The roles of the primal variables are illustrated in Fig. 3.2.1.

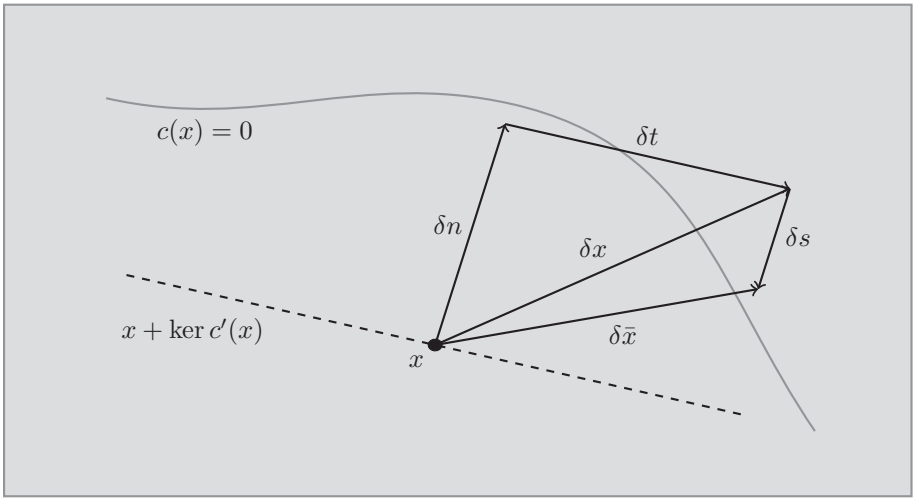


Figure 3.2.1.: Sketch of a composite step and corresponding second order corrected step.

3.2.1.1. Normal step

Since Δn and δs are both supposed to lie in $(\ker c'(x))^\perp$ we start with a short general discussion. First we note that the minimum norm problem

$$\min \ \frac{1}{2}\langle w, w\rangle \tag{3.2.3a}$$

$$\text{subject to} \quad c'(x)w + g = 0, \tag{3.2.3b}$$

is equivalent to finding $w \in (\ker c'(x))^\perp$ such that $c'(x)w + g = 0$. The optimality conditions for (3.2.3) motivate the following result.

Lemma 3.5. *Suppose that $w \in \mathrm{X}$ satisfies:*

$$\begin{pmatrix} M & c'(x)^* \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} w \\ q \end{pmatrix} + \begin{pmatrix} 0 \\ g \end{pmatrix} = 0 \tag{3.2.4}$$

for some $g \in \mathrm{P}^*$. Then $w \in (\ker c'(x))^\perp$.

Proof. This follows from the first row of (3.2.4), since

$$(Mw)(\xi) + q(c'(x)\xi) = 0 \text{ for all } \xi \in \mathrm{X} \ \Rightarrow \ (Mw)(\xi) = 0 \text{ for all } \xi \in \ker c'(x)$$

is equivalent to

$$w \in (\ker c'(x))^{\perp}\,.$$

□

At this point we again stress the fact that the choice of the Hilbert space scalar product $\langle \cdot, \cdot \rangle$ is crucial and depends on the function space context of the problem. Consequently, M, the Riesz-isomorphism of X, is usually a non-trivial linear operator. Further, we note that the normal step does not depend on the Lagrange multiplier p^x.

We denote the solution of (3.2.4) by

$$w := c'(x)^{-}g. \tag{3.2.5}$$

With this notation, we can define the normal step via:

$$\Delta n = -c'(x)^{-}c(x)$$

as the solution of (3.2.4) with $g = c(x)$.

3.2.1.2. Lagrange multiplier

We have already discussed the role of p^x and that it can be computed via (3.1.6) in Sec. 3.1. However, instead of computing p^x via (3.1.6), we obtain it via a correction δp to the previous multiplier p^{x_-}, i.e. $p^x = p^{x_-} + \delta p$, where x_- denotes the previous iterate. Recalling that $\mathcal{L}_x(x, p^{x_-}) = f'(x) + c'(x)^* p^{x_-}$ this is achieved by solving

$$\begin{pmatrix} M & c'(x)^* \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} w \\ \delta p \end{pmatrix} + \begin{pmatrix} \mathcal{L}_x(x, p^{x_-}) \\ 0 \end{pmatrix} = 0.$$

This formulation has the advantage that its right hand side tends to 0 when x tends to a local minimizer, which in turn improves numerical stability with respect to truncation and round-off errors. In exact arithmetic both alternatives yield the same result p^x, which therefore only depends on x, but not on previous Lagrange multiplier estimates.

3.2.1.3. Tangential step

Once we have computed the normal step δn, a prediction ν for its damping factor, and an adjoint state p^x we want to compute the tangential step δt. If the quadratic model

$$q(\delta x) := f'(x)(\delta x) + \frac{1}{2}\mathcal{L}_{xx}(x, p^x)(\delta x)^2$$

has a minimizer $\delta\bar{x}$ in $\ker c'(x)$, then we would like to have δt such that

$$\delta x := \delta n + \delta t$$

is an approximation of this minimizer. We call the exact tangential step $\Delta t := \delta\bar{x} - \delta n$. If q is nonconvex, δt should at least be a direction of descent. Thus, the quadratic problem we have to solve (for fixed δn) is

$$\min_{\delta x = \delta n + \delta t} \; q(\delta x) \tag{3.2.6a}$$

$$\text{subject to} \qquad c'(x)\delta t = 0. \tag{3.2.6b}$$

Omitting terms that are independent of δt and adding the term $p^x c'(x)\delta t = 0$ to the functional, this is equivalent to

$$\min_{\delta t} \; \Big(\mathcal{L}_x(x,p^x) + \mathcal{L}_{xx}(x,p^x)\delta n\Big)\delta t + \frac{1}{2}\mathcal{L}_{xx}(x,p^x)(\delta t)^2 \tag{3.2.7a}$$

$$\text{subject to} \qquad c'(x)\delta t = 0. \tag{3.2.7b}$$

This formulation, which only depends on the Lagrange function and its derivatives, reduces the influence of round-off errors close to the optimal solution, since $\mathcal{L}_x(x,p) \to 0$ for $x \to x_*$ and $p \to p_*$. If $\mathcal{L}_{xx}$ is positive definite on $\ker c'(x)$, then the exact minimizer Δt of problem (3.2.7) exists, and the corresponding first order optimality conditions are

$$\begin{pmatrix} \mathcal{L}_{xx}(x,p^x) & c'(x)^* \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} \Delta t \\ \Delta p \end{pmatrix} + \begin{pmatrix} \mathcal{L}_x(x,p^x) + \mathcal{L}_{xx}(x,p^x)\delta n \\ 0 \end{pmatrix} = 0. \tag{3.2.8}$$

If $\nu = 1$ we have $\delta n = \Delta n$. If further Δt solves (3.2.8) exactly, which asymptotically holds close to the optimal solution, we observe that $(\Delta x, \Delta p) = (\Delta n + \Delta t, \Delta p)$ satisfies the equations of the full Lagrange-Newton step:

$$\begin{pmatrix} \mathcal{L}_{xx}(x,p^x) & c'(x)^* \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta p \end{pmatrix} + \begin{pmatrix} \mathcal{L}_x(x,p^x) \\ c(x) \end{pmatrix} = 0. \tag{3.2.9}$$

Thus, Δp can be interpreted as a Newton update for the Lagrange multiplier respectively as the Lagrange multiplier at x with respect to the scalar product induced by $\mathcal{L}_{xx}(x,p^x)$.

3.2.1.4. Simplified normal step

After δx has been computed we can compute the simplified normal step, similar to the computation of the normal step, via a saddle point problem of the form (3.2.4), such that

$$\delta s := -c'(x)^-[c(x+\delta x) - c(x) - c'(x)\delta x]. \tag{3.2.10}$$

It follows from Lem. 3.5 that $\delta s \in (\ker c'(x_*))^\perp$, and thus

$$[f'(x) + p^x c'(x)]\delta s = 0. \tag{3.2.11}$$

Since the normal step satisfies $c'(x)\delta n + \nu c(x) = 0$, we can derive an alternative representation of the simplified normal step

$$\delta s := -c'(x)^-[c(x + \delta x) - (1 - \nu)c(x)].$$

In the undamped case $\nu = 1$ this relation reads $\delta s = -c'(x)^- c(x + \delta x)$, which is the second step of a simplified Newton method for the equation $c(x) = 0$, starting at x. This explains our naming of δs.

3.2.2. Order of consistency

A basic principle of equality constrained sequential quadratic programming (SQP) is to minimize a quadratic model of the functional subject to a linear model of the constraints. In this section we will study the order of consistency of these models, i.e. the order in which our local models approximate the true problem. This will provide the theoretical basis for the construction of our algorithm. Recalling that both f and c are twice Fréchet differentiable at x, the following quadratic model is used to approximate f on $\ker c'(x)$:

$$\begin{aligned} q(\delta x) &:= f(x) + f'(x)\delta x + \frac{1}{2}\mathcal{L}_{xx}(x, p^x)(\delta x)^2 \\ &= f(x) + f'(x)\delta x + \frac{1}{2}\left(f''(x) + \frac{1}{2}p^x c''(x)\right)(\delta x)^2. \end{aligned} \tag{3.2.12}$$

The last term, involving $c''(x)$ takes into account second order information of the equality constraints. This is necessary to achieve fast local convergence of the undamped Lagrange-Newton method. We will show that $q(\delta x)$ is second order consistent with $f(x+\delta x+\delta s)$, but only first order consistent with $f(x+\delta x)$. The latter is the reason for the well known Maratos effect, while the first result yields a possible remedy. For this reason we also refer to the simplified normal step as a *second order correction.*

Lemma 3.6. *Denote by $\delta x \in \mathrm{X}$ an arbitrary perturbation of $x \in \mathrm{X}$ and by δs the simplified normal step, determined through (3.2.10). Then we have the following consistency result:*

$$\|\delta s\| = o(\|\delta x\|), \tag{3.2.13}$$

$$f(x + \delta x) = q(\delta x) + O(\|\delta x\|^2), \tag{3.2.14}$$

$$f(x + \delta x + \delta s) = q(\delta x) + o(\|\delta x\|^2). \tag{3.2.15}$$

Proof. Estimate (3.2.13) follows directly from the definition of δs, using differentiability of c and invertibility of $c'(x)$ on $(\ker c'(x_*))^\perp$. Next, comparing the Taylor expansion for f at x with $q(\delta x)$, we get (3.2.14):

$$\begin{aligned} q(\delta x) - f(x+\delta x) &= q(\delta x) - \Big(f(x) + f'(x)\delta x + \frac{1}{2} f''(x)(\delta x)^2 + o(\|\delta x\|^2) \Big) \\ &= \frac{1}{2} p^x c''(x)(\delta x)^2 + o(\|\delta x\|^2) = O(\|\delta x\|^2). \end{aligned}$$

Testing the simplified step with p^x, and using Taylor's expansion for the equality constraints we compute

$$\begin{aligned} 0 &= p^x \Big([c(x+\delta x) - c(x) - c'(x)\delta x] + c'(x)\delta s \Big) \\ &= p^x \Big([c(x) + c'(x)\delta x + \frac{1}{2} c''(x)(\delta x)^2 + o(\|\delta x\|^2) - c(x) - c'(x)\delta x] + c'(x)\delta s \Big) \\ &= p^x \Big(\frac{1}{2} c''(x)(\delta x)^2 + c'(x)\delta s \Big) + o(\|\delta x\|^2). \end{aligned}$$

With (3.2.11) we obtain

$$f'(x)\delta s = -p^x c'(x)\delta s = \frac{1}{2} p^x c''(x)(\delta x)^2 + o(\|\delta x\|^2),$$

and with (3.2.12)

$$q(\delta x) = f(x) + f'(x)\delta x + \frac{1}{2} f''(x)(\delta x)^2 + f'(x)\delta s + o(\|\delta x\|^2). \tag{3.2.16}$$

Then, subtracting (3.2.16) from the Taylor expansion of f at x in direction $\delta x + \delta s$, we compute

$$\begin{aligned} &f(x+\delta x+\delta s) - q(\delta x) \\ &\quad = f(x) + f'(x)(\delta x + \delta s) + \frac{1}{2} f''(x)(\delta x + \delta s)^2 + o(\|\delta x + \delta s\|^2) - q(\delta x) \\ &\quad = \frac{1}{2} f''(x)(\delta s, 2\delta x + \delta s) + o(\|\delta x + \delta s\|^2) + o(\|\delta x\|^2). \end{aligned}$$

Eventually (3.2.13) implies $f''(x)(\delta s, 2\delta x + \delta s) = o(\|\delta x\|^2)$ and thus also the desired result (3.2.15). □

In our interpretation q is not a quadratic model of f on the linearization $c'(x)\delta t = 0$ of the feasible set. Rather it takes into account a better, quadratic, approximation of the true feasible set. To compare q and f we should not evaluate f at $x + \delta x$, but at a point that is closer to the true feasible set, at the second order corrected point $x + \delta x + \delta s$.

Quantitative estimates. After these qualitative considerations we discuss conditions under which the above estimates can be quantified more explicitly. Our conditions are mainly based on affine covariant Lipschitz conditions on $\mathcal{L}_{xx}$, f' and c'. These estimates provide the motivation for a couple of algorithmic choices in the following section and are the basis for finite termination and fast local convergence results for our algorithm. Recall that $v = c'(x)^- r$ denotes the minimum norm solution of the problem $c'(x)v = r$.

Lemma 3.7. *For $x, \delta x \in \mathrm{X}$ and corresponding simplified Newton step $\delta s \in \mathrm{X}$ we have the identity*

$$
\begin{aligned}
f(x+\delta x+\delta s) - q(\delta x) &= T_1 + T_2 \quad \textit{where} \\
T_1 &:= \mathcal{L}(x+\delta x, p) - \mathcal{L}(x,p) - \mathcal{L}_x(x,p)\delta x - \frac{1}{2}\mathcal{L}_{xx}(x,p)(\delta x, \delta x) \\
&= \int_0^1 \left(\mathcal{L}_x(x+\sigma\delta x, p) - \mathcal{L}_x(x,p) - \mathcal{L}_{xx}(x,p)\sigma\delta x\right)\delta x \mathrm{d}\sigma \\
&= \int_0^1 \int_0^1 \left(\mathcal{L}_{xx}(x+\tau\sigma\delta x, p) - \mathcal{L}_{xx}(x,p)\right)(\sigma\delta x, \delta x)\mathrm{d}\tau\mathrm{d}\sigma \\
T_2 &:= f(x+\delta x+\delta s) - f(x+\delta x) - f'(x)\delta s \\
&= \int_0^1 \left(f'(x+\delta x+\sigma\delta s) - f'(x)\right)\delta s \mathrm{d}\sigma \\
&= \int_0^1 \int_0^1 f''(x+\tau\delta x+\tau\sigma\delta s)(\delta x+\sigma\delta s, \delta s)\mathrm{d}\tau\mathrm{d}\sigma.
\end{aligned}
$$

Furthermore we have

$$
\delta s = \int_0^1 c'(x)^-(c'(x+\sigma\delta x) - c'(x))\delta x \mathrm{d}\sigma.
$$

Proof. The identities for T_1 and T_2 follow from the fundamental theorem of calculus. So it remains to show that

$$
f(x+\delta x+\delta s) - q(\delta x) = T_1 + T_2
$$

Indeed, using the identities $-c'(x)\delta s = c(x+\delta x) - c(x) - c'(x)\delta x$, and $(f'(x) + pc'(x))\delta s = 0$ we compute

$$
\begin{aligned}
T_1 + q(\delta x) &= \mathcal{L}(x+\delta x, p) - \mathcal{L}(x,p) - \mathcal{L}_x(x,p)\delta x - \frac{1}{2}\mathcal{L}_{xx}(x,p)(\delta x, \delta x) + q(\delta x) \\
&= f(x+\delta x) + (pc(x+\delta x) - pc(x) - pc'(x)\delta x) = f(x+\delta x) - pc'(x)\delta s \\
&= f(x+\delta x) + f'(x)\delta s = f(x+\delta x+\delta s) - T_2.
\end{aligned}
$$

The result on δs follows similarly from the fundamental theorem of calculus. □

Theorem 3.8. *Assume that there are positive constants ω_c, $\omega_{f'}$, and ω_L, such that for all $v, w \in \mathrm{X}$*

$$\|c'(x)^-(c'(x+v) - c'(x))v\| \le \omega_c \|v\|^2, \tag{3.2.17}$$

$$|(\mathcal{L}_{xx}(x+v,p) - \mathcal{L}_{xx}(x,p))(v,v)| \le \omega_L \|v\|^3, \tag{3.2.18}$$

$$|(f'(x+v) - f'(x))w| \le \omega_{f'} \|v\| \|w\|, \tag{3.2.19}$$

where (x,p) are taken among the iterates. Then for $\delta x \in \mathrm{X}$ and corresponding simplified normal step δs we have the estimates

$$\|\delta s\| \le \frac{\omega_c}{2} \|\delta x\|^2 \tag{3.2.20}$$

$$|f(x+\delta x+\delta s) - q(\delta x)| \le \frac{\omega_L}{6}\|\delta x\|^3 + \omega_{f'}\|\delta s\| \left(\|\delta x\| + \frac{1}{2}\|\delta s\|\right) \tag{3.2.21}$$

$$\le \left(\frac{\omega_L}{6} + \frac{\omega_{f'}\omega_c}{2}\left(1 + \frac{\omega_c}{4}\|\delta x\|\right)\right)\|\delta x\|^3. \tag{3.2.22}$$

Proof. Setting $v = \sigma\delta x$, we get

$$\|\delta s\| = \int_0^1 \frac{1}{\sigma}\|c'(x)^-(c'(x+\sigma\delta x) - c'(x))\|\sigma\delta x \mathrm{d}\sigma \le \frac{\omega_c}{2}\|\delta x\|^2,$$

which is (3.2.20). With respect to the Lipschitz constant for $\mathcal{L}_{xx}$ Lem. 3.7 yields

$$|f(x+\delta x+\delta s) - q(\delta x)| \le |T_1| + |T_2|.$$

With the assumed affine covariant Lipschitz conditions, we get, again with $v = \sigma\delta x$,

$$\begin{aligned} |T_1| &\le \int_0^1 \int_0^1 |\mathcal{L}_{xx}(x+\tau\sigma\delta x, p) - \mathcal{L}_{xx}(x,p)|\,(\sigma\delta x, \delta x)\mathrm{d}\tau\mathrm{d}\sigma \\ &\le \omega_L\|\delta x\|^3 \int_0^1\int_0^1 \tau\sigma^2 \mathrm{d}\tau\mathrm{d}\sigma = \frac{\omega_L}{6}\|\delta x\|^3, \end{aligned}$$

Setting $v = \delta x + \sigma\delta s$ and $w = \delta s$ we further obtain

$$\begin{aligned} |T_2| &\le \int_0^1 |(f'(x+\delta x+\sigma\delta s) - f'(x))\,\delta s|\,\mathrm{d}\sigma = \int_0^1 |(f'(x+\delta x+\sigma\delta s) - f'(x))\,\delta s|\,\mathrm{d}\sigma \\ &\le \int_0^1 \omega_{f'}\|\delta s\|\|\delta x + \sigma\delta s\|\mathrm{d}\sigma \le \frac{\omega_{f'}\omega_c}{2}\|\delta x\|^2\left(\|\delta x\| + \frac{\omega_c}{4}\|\delta x\|^2\right). \end{aligned}$$

Adding both estimates yields (3.2.21) and, inserting (3.2.20), yields (3.2.22). □

3.3. The globalization scheme

The globalization mechanism is a central part of any algorithm for nonlinear problems. The particular difficulty in equality constrained optimization is the simultaneous achievement of the potentially conflicting aims of feasibility and optimality.

As the determination of the feasible region is the prerequisite for finding an optimal solution, we attribute priority to feasibility. However, an algorithm that stresses this property too much is likely to be inefficient in finding an optimal point and may converge to a non-stationary feasible point. Thus, the main challenge is imposed by finding a the proper weighing of both aims. Roughly speaking a good algorithm should work as follows: *far away* from the feasible region, focus on getting close to it, *close to* the feasible region, focus on optimality, but take care to remain close to the feasible region. To render this vague idea useful we first have to quantify what *close* should mean.

Two popular techniques for balancing the conflicting aims are merit functions and filter methods [96]. Both combine monotonicity requirements on $f(x)$ and $\|c(x)\|$ to achieve $\|c(x)\| \to 0$ while minimizing f. However, both approaches are in conflict with our algorithmic paradigm that residual norms should not enter the algorithm.

Thus, we resort to different ideas, motivated by the affine covariant Newton methods for nonlinear equations of Deuflhard [76] and by cubic regularization methods [53, 54, 226, 277] for unconstrained optimization. In the context of (simplified) Newton's method one can argue that *close* to the solution means *safely within the region of local convergence*, so that we can find a feasible point easily within a few steps of Newton's method without damping. In Sec. 3.3.1 this idea is carried over to nonlinear optimization with equality constraints. Thus, an iterate is considered close to the feasible set, if a sequence of pure normal steps converges quickly to a feasible point.

To transform this idea into an algorithm, we have to *quantify* this region, at least by a rough heuristic estimate. Ways to construct such estimates are among the central topics in [76]. Here we only give a short motivation and refer to [76] for an in depth treatment for the case of nonlinear systems of equations.

In addition we need to control nonlinearities in the cost functional. For this, cubic regularization methods for unconstrained optimization are extended to the constrained case in Sec. 3.3.2.

3.3.1. Globalization with respect to feasibility

We begin with a discussion of the part of the damping strategy that deals with the non-linearity of the equality constraints $c(x) = 0$. The objective f does not yet play a role. It is considered in a second damping mechanism, described in the next subsection. A short sketch of the update-loop is given in Alg. 3.1.

Let us first recall the principal ideas of the affine covariant damping strategy.

Estimating the Newton contraction. Our main tool is the use of simplified Newton steps δs, which we already encountered in the last section. If $\nu = 1$, i.e. $\delta n = \Delta n$, then

$$c(x) + c'(x)\delta x = 0$$

Require: initial iterate x, Lipschitz constants $[\omega_c], [\omega_f]$
1: **repeat** // NLP loop
2: **repeat** // step computation loop
3: compute new trial correction δx
4: compute simplified correction δs
5: compute new Lipschitz constants $[\omega_c], [\omega_f]$
6: **until** trial correction δx accepted
7: $x \leftarrow x + \delta x + \delta s$
8: **until** converged

Algorithm 3.1: Outer and inner loop, strongly simplified

and δs satisfies

$$c(x + \delta x) + c'(x)\delta s = 0.$$

Thus, δx and δs can be interpreted as the first two steps of a simplified Newton method for the problem $c(\xi) = 0$. Consequently, if $\|\delta s\| \ll \|\delta x\|$ holds, we expect fast local convergence to a feasible point. Denoting the contraction factor by

$$\Theta(\delta x) := \frac{\|\delta s\|}{\|\delta x\|},$$

then $\Theta(\delta x) \ll 1$ is a good indicator that Newton contraction takes place.

Remark 3.9. In general, if $\nu < 1$, then δx and δs satisfy the equations

$$\begin{aligned} c(x) + c'(x)\delta x &= (1-\nu)c(x), \\ c(x + \delta x) + c'(x)\delta s &= (1-\nu)c(x). \end{aligned}$$

Thus, they form two steps of a simplified Newton method for the relaxed problem

$$c(\xi) = (1-\nu)c(x). \tag{3.3.1}$$

Again, $\|\delta s\| \ll \|\delta x\|$ indicates fast local convergence. The solutions of these problems locally define a path, the Newton path [76, Sec. 3.1.4], resp. in the context of underdetermined equations, the geodetic Gauss-Newton path [76, Sec. 4.4.2].

Acceptance criterion. The above considerations suggest to accept a trial correction δx whenever

$$\Theta(\delta x) := \frac{\|\delta s\|}{\|\delta x\|} \leq \Theta_{\text{acc}} < 1 \tag{3.3.2}$$

for a user-provided parameter Θ_{acc}. From Lem. 3.6 we know that $\lim_{\delta x \to 0} \Theta(\delta x) = 0$ and eventually δx will become acceptable for sufficiently small ν.

Adjustment of the error model. Along with the acceptance criterion we need a mechanism to compute acceptable corrections. To this end, we introduce a parametrized model for the contraction rate Θ,

$$[\Theta](\xi) := \frac{[\omega_c]}{2}\|\xi\|,$$

where $[\omega_c]$ is an estimate from below for the *affine covariant Lipschitz constant* ω_c, cf. Thm. 3.8. For a trial correction δx, and corresponding second order correction δs, we use the interpolation condition $[\Theta](\delta x) = \Theta(\delta x)$, to compute

$$[\omega_c] = \frac{2\Theta(\delta x)}{\|\delta x\|} = 2\frac{\|\delta s\|}{\|\delta x\|^2}, \tag{3.3.3}$$

If (3.3.2) fails, a new trial correction is computed such that

$$[\Theta](\delta x) \le \Theta_{\text{aim,x}}, \text{ i.e. } \frac{[\omega_c]}{2}\|\delta x\| \le \Theta_{\text{aim,x}}, \tag{3.3.4}$$

for another user provided contraction rate $\Theta_{\text{aim,x}} < \Theta_{\text{acc}}$. We can rewrite (3.3.4) as trust region constraint (cf. Fig. 3.3.1)

$$\|\delta x\| \le r_x := \frac{2\Theta_{\text{aim,x}}}{[\omega_c]}. \tag{3.3.5}$$

Successive updates of $[\omega_c]$ and δx yield a *predictor-corrector loop,* that terminates as soon as (3.3.2) is satisfied. The gap between $\Theta_{\text{aim,x}}$ and Θ_{acc} is necessary to guarantee finite termination, which will be discussed in Sec. 3.4.

As initial iterates for the outer iterations, depicted in Alg. 3.1, we use the estimates computed in the last iteration, resp. user provided estimates for the initial iteration.

Damping of normal steps. Up to now, this general scheme does not take into account the splitting $\delta x = \delta n + \delta t$ of the composite step. For this we need some slight adjustments. The situation is depicted in Fig. 3.3.1.

Recall that the normal step Δn is compute via (3.2.3) as minimum norm correction that satisfies $c(x)+c'(x)\Delta n = 0$. A damped normal step $\delta n = \nu\Delta n$ then is computed under the restriction $[\Theta](\delta n) \le \Theta_{\text{aim,n}} \le \Theta_{\text{aim,x}}$. Thus we compute the damping factor via

$$\nu := \min\left\{1, \frac{2\Theta_{\text{aim,n}}}{[\omega_c]\|\Delta n\|}\right\}. \tag{3.3.6}$$

The choice of the user specified contraction rate $\Theta_{\text{aim,n}} < \Theta_{\text{aim,x}}$ leaves some "elbow space" for the computation of the tangential step δt. Choosing $\Theta_{\text{aim,n}} = \Theta_{\text{aim,x}}$ would enforce $\delta t = 0$ as long as $\nu < 1$. Again we can interpret the restriction (3.3.6) as trust region constraint (cf. Fig. 3.3.1)

$$\|\delta n\| \le r_n := \frac{2\Theta_{\text{aim,n}}}{[\omega_c]}.$$

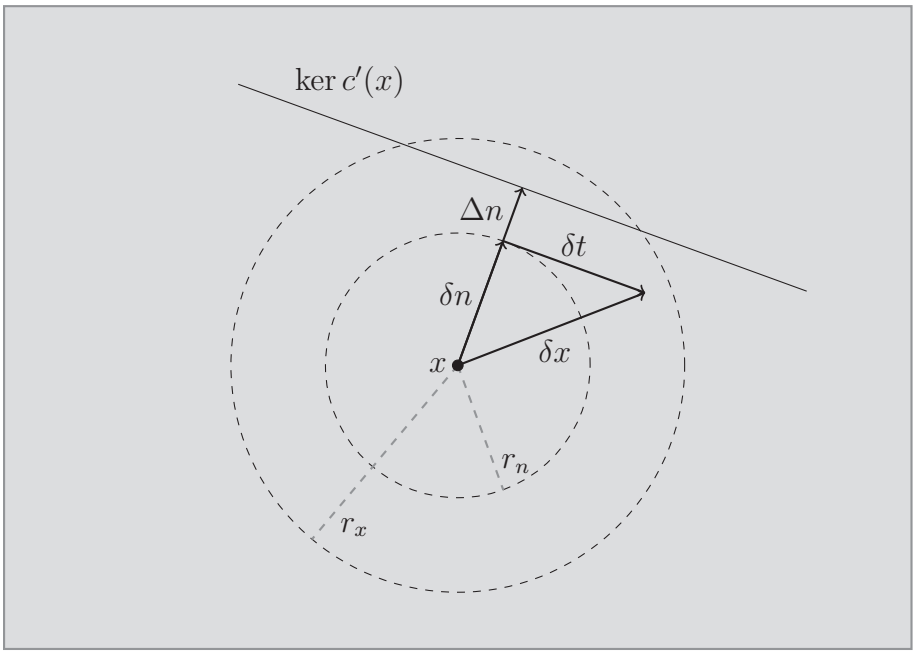


Figure 3.3.1.: Sketch of a damped composite step and trust regions.

Computation of the composite step. After the normal step δn, the tangential step will be computed. The latter also must be restricted such that (3.3.5) is fulfilled for the total step δx. By orthogonality of the tangential and normal step we have

$$\|\delta x\|^2 = \|\delta n\|^2 + \|\delta t\|^2.$$

Inserting into (3.3.5) and solving for δt we obtain

$$\|\delta t\| \leq \sqrt{\left(\frac{2\Theta_{\mathrm{aim,x}}}{[\omega_c]}\right)^2 - \|\delta n\|^2}. \tag{3.3.7}$$

In the case that δt is chosen from a one-dimensional subspace span $\{\Delta t\}$ we obtain for $\delta t := \tau \Delta t$ the restriction

$$\tau \leq \tau_{\max} := \frac{\sqrt{\left(\frac{2\Theta_{\mathrm{aim,x}}}{[\omega_c]}\right)^2 - \nu^2 \|\Delta n\|^2}}{\|\Delta t\|}.$$

If δt is computed from a higher dimensional subspace (cf. [277]), we have to take into account (3.3.7) directly as a trust region constraint.

3.3.2. Globalization with respect to optimality

While normal steps aim at feasibility and a criterion measuring the deviation from the constraint has been introduced, tangential steps are responsible for decrease in

the cost functional. Therefore we need a criterion that ensures decrease of the cost functional. Compared to the unconstrained case, the case with constraints contains additional difficulties.

First, we have to take into account the fact that the normal step may yield increase in the cost functional. In general, finding a feasible point may require an increase of the objective, relative to the current *infeasible* iterate. Thus, we cannot require decrease in the total step and decrease should only be measured for the tangential step. Thus, at first sight, a decrease condition of the form

$$f(x + \delta n + \delta t) < f(x + \delta n) \tag{3.3.8}$$

seems to be useful.

This leads us to the second difficulty, which arises most likely if acceptable normal steps are large, relative to the nonlinearity of the functional. Therefore, recall that tangential steps are computed with the help of a quadratic model at the current iterate x, but they are added to the normal step δn after its computation. For δt getting smaller and smaller during a globalization loop, (3.3.8) can only be guaranteed, if

$$f'(x + \delta n)\delta t < 0. \tag{3.3.9}$$

However, $f'(x+\delta n)$ does not enter the computation of δt, only $f'(x)$, so if δn is large there is no reason for (3.3.9) to hold. In this case, we might be forced to completely reject tangential steps until the iterates are close enough to the constraint.

Due to these two effects the design of a decrease based acceptance criterion is a delicate matter. Motivated by Weiser et al. [277] and Schiela [226] the proposed approach is based on cubic regularization. For this we define the cubic model

$$\begin{aligned} m_{[\omega_f]}(\delta x) &= q(\delta x) + \frac{[\omega_f]}{6}\|\delta x\|^3 \\ &= f(x) + f'(x)\delta x + \frac{1}{2}\mathcal{L}_{xx}(x,p)(\delta x)^2 + \frac{[\omega_f]}{6}\|\delta x\|^3 \end{aligned} \tag{3.3.10}$$

where $[\omega_f]$ is an estimate of the prefactor of the right hand side of (3.2.22). Then, recalling the trust region constraint (3.3.5), we compute tangential steps as solutions of the minimization problem

$$\begin{aligned} &\min_{\delta x = \delta n + \delta t \in \mathrm{X}} m_{[\omega_f]}(\delta x) \\ \text{subject to} \quad &\frac{[\omega_c]}{2}\|\delta x\| \le \Theta_{\mathrm{aim,x}}, \\ &c'(x)\delta t = 0. \end{aligned} \tag{3.3.11}$$

In this model $\Theta_{\mathrm{aim,x}} \in [\Theta_{\mathrm{aim,n}}, \Theta_{\mathrm{acc}}[$ is the contraction rate introduced in (3.3.4). Thus, tangential steps are computed as minimizers, or at least directional minimizers, along descent directions of (3.3.10).

Adjustment of the error model. We start with an extension of the strategy proposed for unconstrained optimization in [226]. Recall the definitions of the quadratic and cubic models

$$q(\delta x) = f(x) + f'(x)\delta x + \frac{1}{2}\mathcal{L}_{xx}(x,p)(\delta x)^2$$

and

$$m_{[\omega_f]}(\delta x) := q(\delta x) + \frac{[\omega_f]}{6}\|\delta x\|^3.$$

In Lem. 3.7 we observed that $q(\delta x)$ is a quadratic model for $f(x+\delta x+\delta s)$, but not for $f(x+\delta x)$. Therefore, we estimate $[\omega_f]$ via

$$[\omega_f] = \frac{6}{\|\delta x\|^3}\left(f(x+\delta x+\delta s) - q(\delta x)\right), \tag{3.3.12}$$

taking into account the restrictions

$$\rho_0[\omega_f]^{\text{old}} \leq [\omega_f]^{\text{new}} \leq \rho_1[\omega_f]^{\text{old}},$$

for $0 < \rho_0 < 1$ and $1 < \rho_1$. The first restriction guarantees $[\omega_f] > 0$, a necessary requirement for being able to determine finite tangential directions in the presence of non-convexities. The second dampens strong increases in the Lipschitz constant. This avoids the occurrence of oscillations of $[\omega_f]$. These restrictions can be relaxed along the lines of [226, Sec. 3.4].

A modified decrease condition. To measure the quality of tangential steps we estimate the ratio between actual and predicted decrease via

$$\eta := \frac{f(x+\delta x+\delta s) - m_{[\omega_f]}(\delta n)}{m_{[\omega_f]}(\delta x) - m_{[\omega_f]}(\delta n)} \tag{3.3.13}$$

In this way we exclude the possible increase due to the normal step and avoid any additional function evaluations. Moreover, the denominator is guaranteed to be negative for $\|\delta t\| > 0$. Then the natural criterion for acceptance of the tangential step is

$$\eta \geq \underline{\eta} \tag{3.3.14}$$

for a user-defined lower bound $\underline{\eta} \in]0,1[$.

For $\delta n = 0$ this reduces to the standard decrease condition, which is widely used in trust region methods [64], and has been adapted in [226] to a cubic regularization method in unconstrained optimization. In the latter case, failure of (3.3.14) yields an increase in $[\omega_f]$ at least by a factor of $1 + \frac{1+\underline{\eta}}{2}$ [226]. Thus in the absence of normal steps, repeated failure of the acceptance test yields a quick increase of $[\omega_f]$.

For constrained problems the expected minimal increase depends on the relative contributions of damped normal resp. tangential step to the composite step, i.e. on the quantity

$$\rho = \frac{\|\delta n\|}{\|\delta x\|}.$$

More precisely, in Sec. 3.4 we will see that we only can guarantee an increase of $[\omega_f]$ by a factor $g(\rho) \in [1, 1 + \frac{1+\underline{\eta}}{2}]$. Thus, if the iterates are not sufficiently close to the constraint stagnating updates of the Lipschitz constant may occur. In this case, we should allow our algorithm to first improve feasibility before continuing optimization, i.e. we should discard the tangential step and accept the step $\delta x = \delta n$. To achieve this aim we monitor the increase in the Lipschitz constant after failure of (3.3.14). If

$$[\omega_f]^{\text{new}} < \left(1 + \rho_s \frac{1-\underline{\eta}}{2}\right)[\omega_f]^{\text{old}}, \tag{3.3.15}$$

for some algorithmic parameter $0 < \rho_s < 1$, then we accept the normal step but possibly discard the tangential step. In order to not waste computational resources we do not discard the tangential step if, for $\eta_{\min} < \underline{\eta}$, a relaxed acceptance test

$$\eta \geq \eta_{\min}, \tag{3.3.16}$$

is satisfied.

3.3.3. Avoiding interference of both schemes

The proposed acceptance test and update rules for the Lipschitz constants admit finite termination of the inner loops as long as not both acceptance criteria are violated in the same inner loop. If both acceptance criteria are violated, it may happen that the algorithm starts to cycle in the following scenario:

- i) A step is not acceptable in terms of (3.3.2), so $[\omega_c]$ is increased, but $[\omega_f]$ is decreased.
- ii) A step is not acceptable in terms of (3.3.14), so $[\omega_f]$ is increased, but $[\omega_c]$ is decreased.

In order to guarantee that this case cannot occur we additionally ensure monotonicity of the Lipschitz estimates after first failure of the corresponding acceptance test. For this, we slightly modify our update rules. In each inner loop, whenever

- i) (3.3.2) has failed at least once, we do not allow decrease in $[\omega_c]$ after failure of (3.3.14),
- ii) (3.3.14) has failed at least once, we do not allow decrease in $[\omega_f]$ after failure of (3.3.2).

In this way, if both (3.3.2) and (3.3.14) fail, we rule out cycling by strict monotonicity of the Lipschitz constants. In the more common cases of accepted steps or only one rejected acceptance criterion this modification is inactive.

3.3.4. Adjustments for nonlinear elasticity

If the constraint is given in terms of a problem from nonlinear elasticity our chosen setting is not fully admissible. Using the notation of elasticity, we recall that the orientation preservation condition

$$\det(\nabla\varphi) > 0$$

implies for hyperelastic energy functionals $\mathcal{E}$ the density of the set

$$\Phi_\infty = \left\{\varphi \in \mathrm{W}^{1,\mathrm{p}}(\Omega) :\ \mathcal{E}(\varphi) = \infty\right\}$$

in $\mathrm{W}^{1,\mathrm{p}}(\Omega)$. Besides the introduced inconveniences regarding the mathematical theory of nonlinear elasticity, this property may also cause failure of our algorithm. The reason is that we may compute corrections $\delta\varphi$ such that $\varphi + \delta\varphi \in \Phi_\infty$. Then the computation of the simplified normal step will fail.

Therefore, we monitor the determinant of the deformation gradient during our computations. If we encounter a violation of the orientation preservation condition, we decrease both damping factors. Since in this case no information is available to adjust the Lipschitz constants, we directly adjust the damping factors according to

$$\nu \leftarrow \frac{\nu}{2} \quad \text{resp.} \quad \tau \leftarrow \frac{\tau}{2}. \tag{3.3.17}$$

This reduction is repeated until the admissible domain is reached and we can switch back to the strategy described before.

In this way we guarantee that the computed iterates are nondegenerate. Thus, the KKT-systems of Sec. 2.5 are well-defined in $\mathrm{W}^{1,\infty}(\Omega)$. FE-solutions are computed in the same space and thus are indeed meaningful, despite the lacking regularity of problems from nonlinear elasticity, cf. Sec. 1.3.

3.3.5. Boundedness of algorithmic parameters

For proving finite termination of the inner loop, we need boundedness of the algorithmic parameters $[\omega_c]$ and $[\omega_f]$. This is essentially a consequence of Thm. 3.8.

Theorem 3.10. *Assume that there are positive constants ω_c, $\omega_{f'}$, and ω_L, such that for all $v, w \in \mathrm{X}$*

$$\|c'(x)^-(c'(x+v) - c'(x))v\| \le \omega_c\|v\|^2, \tag{3.3.18}$$

$$|(L_{xx}(x+v,p) - \mathcal{L}_{xx}(x,p))(v,v)| \le \omega_L\|v\|^3, \tag{3.3.19}$$

$$|(f'(x+v) - f'(x))w| \le \omega_{f'}\|v\|\|w\|, \tag{3.3.20}$$

where (x,p) are taken among the iterates. Further assume that the computed steps $\delta x \in \mathrm{X}$ are bounded. Then the algorithmic parameters $[\omega_c]$ and $[\omega_f]$ are bounded from above.

Proof. Both results follow directly from the update rules and the results of Thm. 3.8. Boundedness of $[\omega_c]$ is a consequence of (3.3.3) and (3.2.20) from Thm. 3.8:

$$[\omega_c] \leq \frac{2\Theta}{\|\delta x\|} = \frac{2\|\delta s\|}{\|\delta x\|^2} = \omega_c.$$

Similarly, boundedness of $[\omega_f]$ follows from (3.3.12) and (3.2.22) from Thm. 3.8:

$$\begin{aligned}[\omega_f] &\leq \frac{6}{\|\delta x\|^3}\left(\frac{\omega_L}{6}\|\delta x\|^3 + \frac{\omega_{f'}\omega_c}{2}\|\delta x\|^3 + \frac{\omega_{f'}\omega_c^2}{8}\|\delta x\|^4\right) \\ &= \omega_L + 3\omega_{f'}\omega_c + \frac{3}{4}\omega_{f'}\omega_c^2\|\delta x\|.\end{aligned}$$

□

3.4. Finite termination of inner loops

Throughout this section we restrict the discussion to one inner loop. In order to show that it terminates after a finite number of rejected steps we first consider the acceptance tests for $[\omega_c]$ and $[\omega_f]$ independently. Then we discuss the combination of both steps using the modification of Sec. 3.3.3.

3.4.1. Finite termination with respect to feasibility

Lemma 3.11. *If a trial correction is rejected due to failure of (3.3.2), then $[\omega_c]$ is increased at least by the fixed factor $\frac{\Theta_{\text{acc}}}{\Theta_{\text{aim,n}}} > 1$. Thus, as long as (3.3.14) does not fail, the inner loop terminates after a finite number of iterations.*

Proof. Using (3.3.5), i. e.

$$\frac{[\omega_c]^{\text{old}}}{2}\|\delta x\| \leq \Theta_{\text{aim,x}} \quad \Leftrightarrow \quad \frac{2}{\|\delta x\|} \geq \frac{[\omega_c]^{\text{old}}}{\Theta_{\text{aim,x}}},$$

failure of (3.3.2) yields

$$[\omega_c]^{\text{new}} = \frac{2\Theta(\delta x)}{\|\delta x\|} > \frac{2\Theta_{\text{acc}}}{\|\delta x\|} \geq \frac{\Theta_{\text{acc}}}{\Theta_{\text{aim,x}}}[\omega_c]^{\text{old}}.$$

□

3.4.2. Finite termination with respect to optimality

Before stating a similar result for the updates of $[\omega_f]$ we summarize some basic properties of minimizers of the cubic model m_ω.

Lemma 3.12. *A directional minimizer δt of m_ω satisfies*

$$(f'(x) + \mathcal{L}_{xx}(x,p)\delta n)\delta t \leq 0 \tag{3.4.1}$$

and

$$\begin{aligned} m_\omega(\delta x) - m_\omega(\delta n) &= \frac{1}{2}(f'(x) + \mathcal{L}_{xx}(x,p)\delta n)\delta t + \frac{\omega}{12}\Big(2\|\delta x\|^3 - 2\|\delta n\|^3 - 3\|\delta x\|\|\delta t\|^2\Big) \\ &\leq \frac{\omega}{12}\Big(2\|\delta x\|^3 - 2\|\delta n\|^3 - 3\|\delta x\|\|\delta t\|^2\Big). \end{aligned} \tag{3.4.2}$$

Proof. From the symmetry of $\frac{1}{2}\mathcal{L}_{xx}(x,p)(\delta t) + \frac{\omega}{6}\|\delta x\|^3$ in δt follows

$$m_\omega(-\delta t) < m_\omega(\delta t)$$

if

$$(f'(x) + \mathcal{L}_{xx}(x,p)\delta n)\delta t > 0.$$

Hence a directional minimizer δt satisfies (3.4.1). The Fréchet derivative of the regularization term is given via

$$\frac{\partial}{\partial \delta x}\left(\|\delta x\|^3\right)\delta h = 3\|\delta x\|\langle \delta x, \delta h\rangle \tag{3.4.3}$$

and thus the first order optimality conditions for m_ω read

$$0 = m'_\omega(\delta x)\delta h = (f'(x) + \mathcal{L}_{xx}(x,p)\delta n)\delta h + \mathcal{L}_{xx}(x,p)\delta t\delta h + \frac{\omega}{2}\|\delta x\|\langle \delta x, \delta h\rangle. \tag{3.4.4}$$

If $\delta h \in \ker c'(x)$ the derivative of $\|\delta x\|^3$ simplifies, due to the orthogonality of Δt and Δn, to

$$\frac{\partial}{\partial \delta x}\left(\|\delta x\|^3\right)\delta h = 3\|\delta x\|\langle \delta t, \delta h\rangle. \tag{3.4.5}$$

Inserting $m'_\omega(\delta t)\delta t = 0$ into the definition of m_ω yields (3.4.2). □

Now it is straightforward to prove

Lemma 3.13. *If a trial correction is rejected due to failure of (3.3.14), then either*

- *ω_L is increased at least by the fixed factor $1 + \rho_s\frac{1-\eta}{2} > 1$,*
- *or the trial correction is accepted, possibly discarding the tangential step.*

Thus, as long as (3.3.2) does not fail, the inner loop terminates after a finite number of iterations.

Proof. We first estimate the increase in $[\omega_f]$.

$$\begin{aligned}
[\omega_f]^{\text{new}} \geq [\omega_f] &= \frac{6}{\|\delta x\|^3}\Big(f(x+\delta x+\delta s) - q(\delta x)\Big) \\
&= \frac{6}{\|\delta x\|^3}\Big((f(x+\delta x+\delta s) - m_{[\omega_f]^{\text{old}}}(\delta n)) + m_{[\omega_f]^{\text{old}}}(\delta n) - q(\delta x)\Big) \\
&= \frac{6}{\|\delta x\|^3}\Big((\eta-1)\left(m_{[\omega_f]^{\text{old}}}(\delta n) - m_{[\omega_f]^{\text{old}}}(\delta n)\right) + \frac{[\omega_f]^{\text{old}}}{6}\|\delta x\|^3 \Big) \\
&= \frac{6}{\|\delta x\|^3}(\eta-1)(m_{[\omega_f]^{\text{old}}}(\delta x) - m_{[\omega_f]^{\text{old}}}(\delta n)) + [\omega_f]^{\text{old}} \\
&> \frac{6}{\|\delta x\|^3}(1-\underline{\eta})(m_{[\omega_f]^{\text{old}}}(\delta n) - m_{[\omega_f]^{\text{old}}}(\delta x)) + [\omega_f]^{\text{old}} \\
&\geq \frac{6}{\|\delta x\|^3}(1-\underline{\eta})\frac{[\omega_f]^{\text{old}}}{12}(3\|\delta x\|\|\delta t\|^2 + 2\|\delta n\|^3 - 2\|\delta x\|^3\Big) + [\omega_f]^{\text{old}} \\
&= \frac{6}{\|\delta x\|^3}(1-\underline{\eta})\frac{[\omega_f]^{\text{old}}}{12}(3\|\delta x\|(\|\delta x\|^2 - \|\delta n\|^2) + 2\|\delta n\|^3 - 2\|\delta x\|^3\Big) + [\omega_f]^{\text{old}} \\
&= (1-\underline{\eta})\frac{[\omega_f]^{\text{old}}}{2\|\delta x\|^3}\Big(\|\delta x\|^3 + \|\delta n\|^2(2\|\delta n\| - 3\|\delta x\|)\Big) + [\omega_f]^{\text{old}} \\
&= [\omega_f]^{\text{old}}\left(1 + \frac{1-\underline{\eta}}{2}\left(1 + \frac{\|\delta n\|^2(2\|\delta n\| - 3\|\delta x\|)}{\|\delta x\|^3}\right)\right).
\end{aligned}$$

Setting $\rho := \frac{\|\delta n\|}{\|\delta x\|}$ we obtain

$$\frac{[\omega_f]^{\text{new}}}{[\omega_f]^{\text{old}}} \geq g(\rho) := 1 + \frac{1-\underline{\eta}}{2}(1 + 2\rho^3 - 3\rho^2), \quad \rho \in [0,1].$$

The function g is monotonously decreasing on $[0,1]$ and bounded by its local extrema

$$1 = g(1) \leq g(\rho) \leq g(0) = 1 + \frac{1-\underline{\eta}}{2},$$

where the case $\rho = 0$ corresponds to the case of unconstrained optimization, i.e. $\delta n = 0$, cf. [226]. The other extreme $\rho = 1$ describes the case of a vanishing tangential step. In the latter case we are "far" away from the constraint and thus the computation of tangential steps may not make sense as the quadratic model is not an adequate model of the local constrained problem. Thus if $1+2\rho^3-3\rho^2 < \rho_s$, where $0 < \rho_s < 1$ is the algorithmic parameter from (3.3.15), the desired increase in $[\omega_f]$ can not be guaranteed any more. In this case we either have

$$[\omega_f]^{\text{new}} \geq \left(1 + \rho_s\frac{1-\underline{\eta}}{2}\right)[\omega_f]^{\text{old}}$$

or the trial correction is accepted due to (3.3.15), discarding the tangential step if (3.3.16) holds. □

We could also consider the quantity $\rho = \frac{\|\delta n\|}{\|\delta x\|}$ to prove increase of $[\omega_f]$ by a fixed factor. A priori it is not clear if δt is reasonable in this case. In particular for simpler problems we expect that tangential directions are meaningful even for significantly larger normal steps. For this reason we monitor the estimates $[\omega_f]$ instead of ρ. Then if globalization is required we only reject tangential steps if there is a real danger of not leaving the inner loop.

3.4.3. Finite termination of the combined scheme

In Lem. 3.11 and Lem. 3.13 we considered the case that only one of the two acceptance tests fails. If we allow both tests to fail, in the same inner loop, cycling might occur in the case of alternating failures of the acceptance tests for feasibility and optimality, see Sec. 3.3.3. For this reason we need the modification proposed in Sec. 3.3.3 to transfer the above results to the general case.

Theorem 3.14. *Assume that the affine covariant Lipschitz conditions (3.2.17)-(3.2.19) hold. Then the inner loop, as described in Alg. 3.2, terminates after a finite number of iterations.*

Proof. We assume that the inner loop does not terminate finitely and show that this implies either

$$[\omega_c] \to \infty \tag{3.4.6}$$

or

$$[\omega_f] \to \infty, \tag{3.4.7}$$

which is not consistent with Thm. 3.10.

If only one of the acceptance criteria (3.3.2) and (3.3.14) fails, we get from Lem. 3.11, resp. Lem. 3.13, that (3.4.6), resp. (3.4.7), holds. Thus we only have to consider the case that both criteria fail.

Let k be the first iteration where both criteria have failed before. Due to the modification of Sec. 3.3.3 none of the estimates for the Lipschitz constants is allowed to decrease during the following iterations in this inner loop. Then, if the inner loop does not terminate finitely, one of the two acceptance criteria is violated infinitely often after the k-th iteration and either (3.4.6) or (3.4.7) holds. □

We summarize the acceptance test in Alg. 3.2.

Require: Lipschitz constants $[\omega_c], [\omega_f]$, search directions $\Delta n, \Delta t$.
Accepted $\leftarrow$ false
ContractionFailedOnce $\leftarrow$ false
DecreaseFailedOnce $\leftarrow$ false
DiscardTangentialStep $\leftarrow$ false
repeat
  $\nu \leftarrow \frac{\rho_{\text{elbow}}\Theta_{\text{aim,n}}}{[\omega_c]\|\Delta n\|}$
  $\tau \leftarrow \tau \in \operatorname{argmin}_{\tau\in[0,\tau_{\max}]} m_{[\omega_f]}(\tau\Delta t)$
  if DiscardTangentialStep **then**
    $\delta x \leftarrow \nu\delta n$
  else
    $\delta x \leftarrow \nu\delta n + \tau\delta t$
  $\delta s \leftarrow$ via (3.2.10)
  compute new Lipschitz constants $[\omega_c]^{\text{new}}, [\omega_f]^{\text{new}}$ via (3.3.3) and (3.3.12)
  if ContractionFailedOnce **then**
    $[\omega_c] \leftarrow \max([\omega_c], [\omega_c]^{\text{new}})$
  else
    $[\omega_c] \leftarrow [\omega_c]^{\text{new}}$
  if DecreaseFailedOnce **then**
    $[\omega_f] \leftarrow \max([\omega_f], [\omega_f]^{\text{new}})$
  else
    $[\omega_f] \leftarrow [\omega_f]^{\text{new}}$
  if (3.3.2) fails **then**
    ContractionFailedOnce $\leftarrow$ true
  else
    if (3.3.14) fails **then**
      DecreaseFailedOnce $\leftarrow$ true
      if (3.3.15) fails **then**
        DiscardTangentialStep $\leftarrow$ true
    else
      Accepted $\leftarrow$ true
until Accepted

Algorithm 3.2: Globalization loop.

3.5. Transition to fast local convergence

We turn to the transition of our method to local quadratic convergence. Of particular interest is to show that the Maratos effect does not occur. As usual for local convergence results, we will assume sufficient smoothness and second order sufficient optimality conditions (SSC) at a local minimizer x_*. We call x_* an **SSC point**.

To keep the discussion concise we do not aim for the most general results, but remain in a rather simple setting. In particular, we only consider the case that normal and tangential steps can be computed exactly along Newton directions. This is in contrast to practical solvers, where at least the tangential steps are computed only inexactly up to a certain accuracy[1]. To retain fast local convergence in that setting an appropriate accuracy matching strategy has to be developed and analyzed. This is subject to ongoing work.

First, consider the classical, undamped Lagrange-Newton method

$$(x_{k+1}, p_{k+1}) = (x_k, p_k) - (\Delta x_k, \Delta p_k) = (x_k, p_k) - \mathcal{L}''(x_k, p_k)^{-1}\mathcal{L}'(x_k, p_k).$$

At an SSC point x_*, the Jacobian matrix $\mathcal{L}''(x_*, p^{x_*})$ is continuously invertible, and a perturbation argument yields that the same holds true in a neighborhood of (x_*, p^{x_*}). This implies that the undamped Lagrange-Newton method with iterates (x_k, p_k) locally converges quadratically towards (x_*, p^{x_*}), if, e.g., Lipschitz conditions like the ones used in Sec. 3.3.5 hold.

We will proof that local quadratic convergence follows for the variant with the adjoint update (3.1.6)

$$(x_{k+1}, w) = (x_k, p^{x_k}) - \mathcal{L}''(x_k, p^{x_k})^{-1}L'(x_k, p^{x_k}) \tag{3.5.1}$$

that is used here. To this end, it will first be shown that the undamped iteration locally admits quadratic convergence. Then we extend this result to the globalized variant.

As a preparatory step we show that small perturbations in p yield perturbations in the steps that are small *relative to the step length*.

Lemma 3.15. *Assume that $\mathcal{L}_{xx}(x, p^{x_*})$ is positive definite and $c''(x) : \mathrm{X} \times \mathrm{X} \to \mathrm{P}^*$ is bounded. Let p be a sufficiently small perturbation of p^{x_*}. Denote by Δx_* the solution of (3.2.6) with argument (x, p^{x_*}) and by Δx the solution of (3.2.6) with argument (x, p). Then there is a constant $c_* > 0$ such that*

$$\frac{\|\Delta x - \Delta x_*\|}{\|\Delta x\|} \leq c_* \|p - p^{x_*}\|. \tag{3.5.2}$$

[1] The maximal attainable accuracy of the discrete problems is typically significantly bigger than machine accuracy. Thus also with direct factorizations we cannot assume that we can solve the arising linear systems exactly or at least up to machine accuracy.

Proof. By assumption $\mathcal{L}_{xx}(x, p^{x_*})$ is positive definite on $\ker c'(x)$, i.e. there exists a positive constant $\gamma > 0$, such that for all $v \in \mathrm{X}$ we have

$$\gamma \|v\|^2 \leq \mathcal{L}_{xx}(x, p^{x_*})(v, v).$$

Hence, for a close-by Lagrange multiplier p we know that $\mathcal{L}_{xx}(x, p)$ is still positive definite on $\ker c'(x)$. Let Δx_* be the solution of (3.2.6) with (x, p^{x_*}), and Δx be the solution of (3.2.6) with p^{x_*} replaced by p. The corresponding first order optimality conditions read

$$0 = f'(x)v + \mathcal{L}_{xx}(x, p^{x_*})(\Delta x_*, v) \quad \text{for all } v \in \ker c'(x),$$

and

$$0 = f'(x)v + (\mathcal{L}_{xx}(x, p^{x_*}) + (p - p^{x_*})c''(x))(\Delta x, v) \quad \text{for all } v \in \ker c'(x).$$

Subtracting both equations yields

$$0 = \mathcal{L}_{xx}(x, p^{x_*})(\Delta x - \Delta x_*, v) + (p - p^{x_*})c''(x)(\Delta x, v).$$

Inserting $v = \Delta x - \Delta x_* \in \ker c'(x)$ (the normal components of the two steps do not differ) and using positive definiteness, we get

$$\gamma \|\Delta x - \Delta x_*\|^2 \leq \mathcal{L}_{xx}(x, p^{x_*})(\Delta x - \Delta x_*, \Delta x - \Delta x_*) = -(p - p^{x_*})c''(x)(\Delta x, \Delta x - \Delta x_*).$$

Taking norms, we obtain

$$\gamma \|\Delta x - \Delta x_*\|^2 \leq \|p - p^{x_*}\| \|c''(x)\| \|\Delta x\| \|\Delta x - \Delta x_*\|,$$

which yields, with $c_* := \gamma^{-1}\|c''(x)\|$, the desired inequality

$$\frac{\|\Delta x - \Delta x_*\|}{\|\Delta x\|} \leq c_* \|p - p^{x_*}\|.$$

□

Theorem 3.16. *Assume that the Lipschitz conditions (3.2.17),(3.2.18), and (3.2.19) hold in a neighborhood of $x_* \in \mathrm{X}$. Then, the iteration (3.5.1) locally admits quadratic convergence.*

Proof. For a pair $z = (x, p)$ let us introduce the notation $x := z_1$ to access the primal component of z. For given (x, p^x), we denote the next Newton iterate by (x_+, p_+). Since our update for p^x is not p_+, but p^{x_+}, computed via (3.1.6), we would like to estimate $\|x_+ - x_*\|$ in terms of $\|x - x_*\|$, namely we have to show that $\|x_+ - x_*\| = O(\|x - x_*\|^2)$. Using the Newton step, we compute

$$\begin{aligned}
x_+ - x_* &= (x_+ - x) + (x - x_*) = \Delta x - (x - x_*) \\
&= -\left[\mathcal{L}_{xx}\left(x, p^x\right)^{-1} \mathcal{L}_x\left(x, p^x\right)\right]_1 + [x - x_*, 0]_1 \\
&= -\left[\mathcal{L}_{xx}\left(x, p^x\right)^{-1} \mathcal{L}_x\left(x, p^x\right) - \mathcal{L}_{xx}\left(x, p^{x_*}\right)^{-1} \mathcal{L}_x\left(x, p^{x_*}\right)\right]_1 \qquad (3.5.3) \\
&\quad - \left[\mathcal{L}_{xx}\left(x, p^{x_*}\right)^{-1} \left[\mathcal{L}_x\left(x, p^{x_*}\right) - \mathcal{L}_x\left(x_*, p^{x_*}\right) + \mathcal{L}_{xx}\left(x, p^{x_*}\right)(x - x_*, 0)\right]\right]_1 \qquad (3.5.4)
\end{aligned}$$

From Lem. 3.15 and Lem. 3.4 we get

$$\begin{aligned}\left\|\left[\mathcal{L}_{xx}(x,p^x)^{-1}\mathcal{L}_x(x,p^x) - \mathcal{L}_{xx}(x,p^{x_*})^{-1}\mathcal{L}_x(x,p^{x_*})\right]_1\right\| &\leq c_*\|\Delta x\|\|p^x - p^{x_*}\| \\ &\leq c_*\varepsilon_c(x)\|\Delta x\|\|x - x_*\|.\end{aligned}$$

The second part (3.5.4) can be estimated via the fundamental theorem of calculus and the affine covariant Lipschitz condition

$$\left\|\mathcal{L}_{xx}(x,p^x)^{-1}\left(\mathcal{L}_{xx}(y,q)v - \mathcal{L}_{xx}(z,q)v\right)\right\| \leq \omega_L\|y-z\|\|v\|,$$

where $x, y, z, v \in \mathrm{X}$ and $p, q \in \mathrm{P}$.

$$\begin{aligned}&\left\|\left[\mathcal{L}_{xx}(x,p^{x_*})^{-1}[\mathcal{L}_x(x,p^{x_*}) - \mathcal{L}_x(x_*,p^{x_*}) + \mathcal{L}_{xx}(x,p^{x_*})(x-x_*,0)]\right]_1\right\| \\ =&\left\|\left[\int_0^1 \mathcal{L}_{xx}(x,p^{x_*})^{-1}[\mathcal{L}_{xx}(x+t(x_*-x),p^{x_*})(x_*-x,0) - \mathcal{L}_{xx}(x,p^{x_*})(x_*-x,0)]\,\mathrm{dt}\right]_1\right\| \\ \leq&\int_0^1 \left\|\left[\mathcal{L}_{xx}(x,p^{x_*})^{-1}[\mathcal{L}_{xx}(x+t(x_*-x),p^{x_*})(x_*-x,0) - \mathcal{L}_{xx}(x,p^{x_*})(x_*-x,0)]\right]_1\right\|\mathrm{dt} \\ \leq&\int_0^1 t\omega_L\|x_*-x\|^2\,\mathrm{dt} = \frac{\omega_L}{2}\|x_*-x\|^2.\end{aligned}$$

Combining both estimates we get

$$\|x_+ - x_*\| \leq c_*\varepsilon_c(x)\|\Delta x\|\|x-x_*\| + \frac{\omega_L}{2}\|x_*-x\|^2.$$

Next, we split $\|\Delta x\| = \|x_+ - x\| \leq \|x_+ - x_*\| + \|x - x_*\|$ and compute

$$\|x_+ - x_*\| \leq c_*\varepsilon_c(x)\|x_+-x_*\|\|x-x_*\| + c_*\varepsilon_c(x)\|x-x_*\|^2 + \frac{\omega_L}{2}\|x_*-x\|^2.$$

If $c_*\varepsilon_c(x)\|x-x_*\| \leq \varepsilon < 1$, this yields

$$\|x_+ - x_*\|(1-\varepsilon) \leq \left(c_*\varepsilon_c(x) + \frac{\omega_L}{2}\right)\|x-x_*\|^2.$$

□

Let us now study the influence of our globalization scheme close to an SSC point. For simplicity, we assume that close to the minimizer, where $\mathcal{L}_{xx}$ is positive definite on $\ker c'(x)$, tangential steps are computed in direction of the minimizer Δt of (3.2.7). Then we have $\delta t = \tau\Delta t$, where $\tau \in]0,1]$ is a damping factor, computed by solving (3.3.10) in the affine subspace $\delta n + \operatorname{span}\{\Delta t\}$. Thus, our damped composite step δx and the full Lagrange-Newton step Δx are related via

$$\begin{aligned}\delta x &= \delta n + \delta t = \nu\Delta n + \tau\Delta t, \\ \Delta x &= \Delta n + \Delta t.\end{aligned}$$

By orthogonality of Δn and Δt, as well as $\nu, \tau \in]0,1]$, this implies $\|\delta x\| \leq \|\Delta x\|$.

Theorem 3.17. *Assume that x_k converges to the SSC point x_*. Further assume that the Lipschitz conditions (3.2.17),(3.2.18), and (3.2.19) hold in a neighborhood of x_*. Then the globalized scheme admits local quadratic convergence.*

Proof. We have to show that our globalized scheme merges into the Lagrange-Newton method. First, we show that as $x_k \to x_*$ the corresponding damping factors ν_k and τ_k tend to 1. By our assumptions, the algorithmic parameters $[\omega_c]$ and $[\omega_f]$ remain bounded along x_k, while $\delta x_k \to 0$ and $\Delta x_k \to 0$. Thus, from (3.3.6) we get for k sufficiently large that $\nu_k = 1$.

Next, we show that $\tau_k \to 1$. Using the minimizing property of δx_k along the direction Δt_k and inserting $h = \Delta t_k$ into (3.4.4) we obtain

$$\begin{aligned}
&0 = m'_{[\omega_f]}(\delta x_k)\Delta t_k \\
&= (f'(x_k) + \mathcal{L}_{xx}(x_k, p_k)\delta n_k)\Delta t_k + \mathcal{L}_{xx}(x_k, p_k)(\delta t_k, \Delta t_k) + \frac{[\omega_f]}{2}\|\delta x_k\|\langle \delta x_k, \Delta t_k\rangle \\
&= (f'(x_k) + \mathcal{L}_{xx}(x_k, p_k)\delta n_k)\Delta t_k + \tau_k \left(\mathcal{L}_{xx}(x_k, p_k)(\Delta t_k, \Delta t_k) + \frac{[\omega_f]}{2}\|\delta x_k\|\,\|\Delta t_k\|^2\right).
\end{aligned}$$

A similar equation holds for the full tangential step Δt_k, which minimizes m_ω for $\omega = 0$:

$$\begin{aligned}
0 = m'_0(\delta x_k)\Delta t_k &= (f'(x_k) + \mathcal{L}_{xx}(x_k, p_k)\delta n_k)\Delta t_k + \mathcal{L}_{xx}(x_k, p_k)(\Delta t_k, \Delta t_k) \\
&= (f'(x_k) + \mathcal{L}_{xx}(x_k, p_k)\delta n_k)\Delta t_k + \mathcal{L}_{xx}(x_k, p_k)(\Delta t_k, \Delta t_k).
\end{aligned}$$

Subtracting both equations and solving for τ_k yields

$$\tau_k = \frac{\mathcal{L}_{xx}(x_k, p_k)(\Delta t_k, \Delta t_k)}{\mathcal{L}_{xx}(x_k, p_k)(\Delta t_k, \Delta t_k) + \frac{[\omega_f]}{2}\|\delta x_k\|\langle \Delta t_k, \Delta t_k\rangle}.$$

Since $\mathcal{L}_{xx}$ is positive definite near x_*, i.e. there exists a positive constant γ such that

$$\gamma\|v\|^2 \le \mathcal{L}_{xx}(x, p_*)(v, v)_{xx} \quad \text{for all } v \in \mathrm{X}.$$

With $[\omega_f]\|\delta x_k\| \to 0$ we get

$$\begin{aligned}
(1 - \tau_k) &\le \frac{[\omega_f]\|\delta x_k\|\langle \Delta t_k, \Delta t_k\rangle}{2\mathcal{L}_{xx}(x_k, p_k)(\Delta t_k, \Delta t_k)} \\
&\le \frac{[\omega_f]}{2\gamma}\|\delta x_k\| \\
&\le \frac{[\omega_f]}{2\gamma}\|\Delta x_k\|.
\end{aligned}$$

Recalling that $\nu = 1$ after a finite number of steps, this yields

$$\|\Delta x_k - \delta x_k\| = (1 - \tau_k)\,\|\Delta t_k\| \le \frac{[\omega_f]}{2\gamma}\,\|\Delta x_k\|^2.$$

Further, with the Lipschitz continuity of c', the definition of the second order correction (3.2.10) yields

$$\|\delta s_k\| \leq \text{const.} \, \|\Delta x_k\|^2 .$$

Consequently, the computed steps quadratically approach the full Lagrange-Newton steps and the iteration inherits local quadratic convergence from the latter. □

4. Computation of steps for optimal control problems

Up to here we described our composite step method mainly from the perspective of nonlinear optimization. Now we turn to the practical computation of steps in the context of optimal control problems.

In this section we assume that, by some Galerkin-type discretization, our infinite dimensional problem has been reduced to a finite dimensional one. Then, after choosing bases for the spaces X and P, which induces dual bases for X^* and P^*, the linear operators are represented by matrices and their adjoints by transpose matrices. The application of a linear functional $l \in \mathrm{X}^*$ to an element $x \in \mathrm{X}$ can be written in terms of their coefficient vectors as $l^T x$. The introduction as well as Sec. 4.2 and Sec. 4.3 have been published in Lubkoll, Schiela and Weiser [180].

To capture the structure of optimal control problems we split the primal variable into state and control, $x = (y, u)^T$, where $\mathrm{X} = \mathrm{Y} \times \mathrm{U}$ with $y \in \mathrm{Y}$ and $u \in \mathrm{U}$. We consider a problem of the form

$$\begin{aligned} &\min_{x=(y,u)} f(x) \\ \text{subject to} \qquad &c(x) = A'(y) - Bu = 0, \end{aligned}$$

where $A'(\cdot)$ is continuously invertible and B is linear. To simplify notation we will consider a fixed iterate $x_0 = (y_0, u_0)$ with corresponding Lagrange multiplier p_0 and let $A = A'(y_0)$ and $\mathcal{L} = \mathcal{L}(y_0, u_0, p_0)$.

Then, the saddle point matrices that occur in the computation of normal and tangential step read

$$H_n = \begin{pmatrix} M_y & & A^T \\ & M_u & -B^T \\ A^T & -B & \end{pmatrix}, \quad H_t = \begin{pmatrix} \mathcal{L}_{yy} & \mathcal{L}_{yu} & A^T \\ \mathcal{L}_{uy} & \mathcal{L}_{uu} & -B^T \\ A^T & -B & \end{pmatrix}.$$

In the following we only consider right hand sides of the form $(r_y, r_u, 0)^T$. Such a right hand side occurs in the computation of the tangential step and adjoint correction. For (simplified) normal steps this is not the case. Instead they satisfy, for some right hand side $r = (0, 0, r_p)^T$, the system

$$H_n z = r.$$

Equivalently, we can compute $z = z_0 + \tilde{z}$, with $z_0 = (A^{-1}r_p, 0, 0)^T$ and $\tilde{z} \in \ker c'(x_0)$ determined by

$$H_n \tilde{z} = r - H_n z_0 = \begin{pmatrix} -M_y A^{-1} r_p \\ 0 \\ 0 \end{pmatrix}.$$

Thus we can w.l.o.g. restrict the discussion to constraints $c'(x_0) = 0$. This admits to exploit the fact that the restriction of the search space to $\ker c'(x_0)$ yields a convex unconstrained optimization problem for problems involving H_n. Conjugate gradient methods (CG) that contain this restriction, *projected preconditioned conjugate gradient methods* (PPCG), are introduced in Sec. 4.1. Then different strategies for the computation of the normal step, resp. adjoint state or second order correction, are discussed in Sec. 4.2. We will restrict the discussion to the computation of $\tilde{z}$, since the same strategy will be applied for the computation of the adjoint state and the second order correction, the latter in a similar affine space as in the computation of the normal step.

Remark 4.1. We could also apply the Bramble-Pasciak conjugate gradient method [42] and its variants [229, 245], which employ an indefinite preconditioner Q and a non-standard inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ such that the preconditioned matrix $Q^{-1}H$ is symmetric and positive definite in the particular inner product. In the chosen setting, this approach is less effective, as it offers less possibilities for the reuse of assembled, and possibly processed, data in the computation of the tangential step.

Regarding the computation of the tangential step, we will also incorporate the restriction to $\ker c'(x_0)$ with the help of constraint preconditioners. However, since $\mathcal{L}_{xx}$ is in general not positive definite on $\ker c'(x_0)$ we shall need conjugate gradient methods for nonconvex problems. These are discussed in Sec. 4.3.

For an adaptive version of the affine covariant composite step method we introduce a hierarchical error estimator in Sec. 4.4. In the course of the algorithm we repeatedly have to solve equations involving the differential operators $A'(y_0)$, $(A'(y_0))^*$, M_u and $\mathcal{L}_{uu}(y_0, u_0, p_0)$. To do this efficiently we shortly introduce the employed approximation schemes in Sec. 4.5.

4.1. Projected preconditioned conjugate gradients

In the following, consider the solution of the linear saddle point system

$$H_n z = r$$

and denote the blocks of H_n as follows:

$$H_n = \begin{pmatrix} M & C^T \\ C & 0 \end{pmatrix}, \tag{4.1.1}$$

Require: initial iterate x, set $r = Hx - r$, $g = Q^{-1}r$ and $d = -g$.
1: **repeat**
2: $\sigma \leftarrow r^T g$
3: $\alpha \leftarrow \sigma / d^T H d$
4: $x \leftarrow x + \alpha d$
5: $r \leftarrow r + \alpha H d$
6: $g \leftarrow Q^{-1} r$
7: $\beta \leftarrow r^T g / \sigma$
8: $d \leftarrow -g + \beta d$
9: **until** convergent

Algorithm 4.1: Preconditioned conjugate gradient method.

where the components of the vectors are denoted by $z = (x, \lambda)^T$ and $r = (r_x, 0)^T$. Thus, C plays the role of $c'(x) = (A'(y_0),\ -B)$ and M is the Riesz isomorphism of X. In order to apply a conjugate gradient method (Alg. 4.1), we have to project the primal search directions $q = (q_x, q_\lambda)^T$ to $\ker C$.

This projection may be realized explicitly via computation of a basis $b_1, \ldots, b_{n-m}$ of $\ker C$. Setting $Z = [b_1, \ldots, b_{n-m}]$ a conjugate gradient method can be applied to the reduced, positive definite system $Z^T M Z$. However, the explicit computation of a basis of $\ker C$ is challenging, even more if we want a sparse basis with minimal number of entries, which turns out to be NP-hard [61].

More promising, for the class of problems we have in mind, is the implicit null space computation via *constraint preconditioners* which are of the form

$$Q_{\text{sc}} = \begin{pmatrix} \tilde{M} & C^T \\ C & 0 \end{pmatrix}, \tag{4.1.2}$$

where $\tilde{M}$ is a preconditioner for M and is assumed to be symmetric positive definite on $\ker C$. Thus $\tilde{M}$ induces a scalar product $\langle \cdot, \cdot \rangle_{\tilde{M}}$ on $\ker C$. Computation of $z = Q_{sc}^{-1} r$ (with $z = (x, \lambda)^T$) for an element $r = (r_x, 0)^T$ then is equivalent to solving the projected gradient problem

$$\begin{aligned} \min\ & r_x^T x + \frac{1}{2} \langle x, x \rangle_{\tilde{M}} \\ \text{subject to}\quad & Cx = 0, \end{aligned}$$

with Lagrangian multiplier λ. We see that this restricts the x-component of the search directions, at least in exact arithmetic, to $\ker C$. In particular, if M is positive definite on $\ker C$, the CG method only "sees" a positive definite part of H. Conjugate gradient methods that employ constraint preconditioners are known as **projected preconditioned conjugate gradient methods** (PPCG).

Interpreting the application of the constraint preconditioner as restriction of the search space yields the PPCG method in expanded form (Alg. 4.2). This variant is rather unstable with respect to round-off errors, thus leading to search directions that contain significant contributions from $(\ker C)^{\perp}$. The reason is, that this variant updates the primal component (the x-component) of the iteration vector only, but does not change the dual component p. The latter only converges to zero if the minimizer of

$$\frac{1}{2}x^T Mx + r_x^T x$$

is already contained in $\ker C$. This is easily seen from the first order necessary condition

$$r_x - M\bar{x} = C^T\bar{\lambda}.$$

Consequently, the residual $Mx - r_x$ does, in general, not tend to 0 during the CG-iteration. Instead it converges to an element in $\mathcal{R}(C^T) = (\ker C)^{\perp}$, i.e.

$$x \to 0 \qquad \text{and} \qquad \lambda \to \bar{\lambda}.$$

Denoting the exact primal solution by $\bar{x}$ and the condition number of C by $\kappa(C)$, the relative projection error in the application of the constraint preconditioner can be estimated by, cf. [115],

$$\frac{\|\bar{x} - x\|}{\|\bar{x}\|} \leq \text{const.}\kappa^2(C)\frac{\|\lambda\|}{\|x\|}.$$

If $\lambda \neq 0$, significant round-off and extinction errors are expected for $x \to 0$. To reduce this effect iterative refinement as well as a residual update strategy have been proposed in [115]. The latter consists of replacing, in each iteration, the primal residual x with $x - C^T q$, where q is a solution of

$$\min_v \|r_x - C^T v\|_2,$$

which reduces the ratio $\frac{\|\lambda\|}{\|x\|}$.

The residual update strategy cures the previously neglected influence of the dual variable. Thus, it is not expected to require such an update when working on the full saddle point system $Hz = r$ (Alg. 4.3). In this case the ordinary residual update of the CG method is

$$r \leftarrow r - \alpha Hd.$$

Considering primal and dual residual separately we get

$$\begin{pmatrix} r_x \\ r_\lambda \end{pmatrix} \leftarrow \begin{pmatrix} r_x - \alpha M d_x - \alpha C^T d_\lambda \\ r_\lambda - \alpha C^T d_x \end{pmatrix}.$$

Require: initial iterate x satisfying $Cx = 0$, set $r_x = Mx - r_x$, $r = (r_x, 0)$, $g =$ primal component of $Q_{\text{sc}}^{-1} r$ and $d = -g$.

repeat
  $\sigma \leftarrow r_x^T g$
  $\alpha \leftarrow \sigma / d^T M d$
  $x \leftarrow x + \alpha d$
  $r_x \leftarrow r_x + \alpha M d$
  $r \leftarrow (r_x, 0)$
  $g \leftarrow$ primal component of $Q_{\text{sc}}^{-1} r$
  $\beta \leftarrow r_x^T g / \sigma$
  $d \leftarrow -g + \beta d$
until convergent

Algorithm 4.2: Projected preconditioned conjugate gradient method in expanded form.

The first row contains a term corresponding to an update of the primal residual in the expanded form by $\alpha C^* d_p$ where d_p is a solution of

$$\min_v \|r_x - \alpha C^T v\|_2^2.$$

Thus, a local residual update is implicitly contained when working on the full system. Nonetheless iterative refinement or carefully chosen explicit recomputation of the residuals may be required if the conjugate gradient method converges slowly or convergence is delayed, cf. [177].

Require: initial iterate $z = (x, p)$ satisfying $Cx = 0$, set $r = Hz - r$, $g = Q_{sc}^{-1} r$ and $d = -g$.

repeat
  $\sigma \leftarrow r^T g$
  $\alpha \leftarrow \sigma / d^T H d$
  $z \leftarrow z + \alpha d$
  $r \leftarrow r + \alpha H d$
  $g \leftarrow Q_{sc}^{-1} r$
  $\beta \leftarrow r^T g / \sigma$
  $d \leftarrow -g + \beta d$
until convergent

Algorithm 4.3: Projected preconditioned conjugate gradient method, all-at-once form.

Besides the projection property, the preconditioner should cluster the eigenvalues of H in order to accelerate convergence. The spectrum of constraint preconditioners of the form (4.1.2) has been analyzed in [157]. In our notation their main results read

Theorem 4.2. *Denote the preconditioned matrix by $\mathcal{P} = Q_{sc}^{-1}H$. Then $\mathcal{P}$ has*

1. *the eigenvalue 1 with multiplicity $2m$,*
2. *$n-m$ eigenvalues which are defined by the generalized eigenvalue problem*

$$Z^T M Z y = \lambda Z^T \tilde{M} Z y, \quad y \in \mathbb{R}^{n-m}, \tag{4.1.3}$$

where Z denotes the projection on $\ker C$. Let $k \leq n-m$ be the number of distinct eigenvalues of the generalized eigenvalue problem. The Krylov space $\mathcal{K}(\mathcal{P}, d)$ is, for any right hand side d, at most of dimension $k+2$.

Proof. See [157, Thm. 2.1 and Thm. 3.7]. For the case of a non-vanishing lower right block cf. [85]. □

For a more detailed discussion of solution methods for saddle point matrices the interested reader is referred to the survey [34] as well as the references summarized in [229].

4.2. Computation of (simplified) normal steps and adjoint updates

Saddle point systems of the form

$$H_n z = r, \tag{4.2.1}$$

where the components of the vectors are denoted by $z = (x, \lambda)^T$ and $r = (r_x, 0)^T$, occur in the computation of the normal step Δn, the simplified normal step δs and in the computation of the adjoint state p. Depending on size and structure of the problem, different possibilities arise for its solutions. Recall that the saddle point matrix H_n corresponds to a strictly convex, equality constrained optimization problem and is invertible.

Problems of moderate size. If the problem size is moderate, the solution of (4.2.1) can be computed using a direct factorization of the saddle point matrix. The possibly high computational costs for the computation of the factorization are at least partially amortized by the multiple possibilities for its reuse. Besides the computation of p and δs this includes its application as a constraint preconditioner in the computation of the tangential step.

This works fine for moderately sized, stationary optimal control problems, usually in two spatial dimensions. However, for larger problems, such as time-dependent optimal control or larger three dimensional problems, sparse direct factorizations become prohibitively expensive, both in time and memory consumption.

Low dimensional control space. Let us consider the case that the space of controls is of low dimension (say, a couple of tens) and that $A = A'(y_0)$ can be factorized by a sparse direct solver. In this case it is possible to use a Schur-complement approach in order to solve (3.2.4) by factorization of A and a couple of back-solves. This can be interpreted as a direct solution of the system (4.2.1) with a special pivoting strategy, often not recognized by standard sparse solvers.

For this, consider the following block permutation of our system:

$$\begin{pmatrix} M_u & 0 & -B^T \\ 0 & M_y & A^T \\ -B & A & 0 \end{pmatrix} \begin{pmatrix} \delta n_u \\ \delta n_y \\ q \end{pmatrix} + r = 0. \tag{4.2.2}$$

The right lower 2x2 block

$$K = \begin{pmatrix} M_y & A^T \\ A & 0 \end{pmatrix}$$

is block triangular with invertible diagonal blocks, and thus is invertible by essentially inverting A and A^T. With that we can build a Schur complement with respect to M_u:

$$S = M_u - W^* K^{-1} W, \tag{4.2.3}$$

where W is defined as

$$W\colon \mathrm{U} \to \Phi^* \times \mathrm{P}^* \text{ via } W := \begin{pmatrix} 0 \\ -B \end{pmatrix}.$$

This strategy can be applied to fairly well resolved elliptic problems in two and three spatial dimensions. For a successful application of this approach we refer to [81], where an optimization problem from hyperthermia treatment was solved. There, the control consisted of 23 input parameters for the microwave antennas built into the hyperthermia applicator.

High dimensional control space. If neither direct factorizations nor the Schur complement reduction are applicable we have to use iterative solvers. In this regard we will employ the a PPCG method. We consider a constraint preconditioner of the form

$$Q_{\mathrm{sc}} = \begin{pmatrix} \tilde{M}_y & 0 & A^T \\ 0 & \tilde{M}_u & -B^T \\ A^T & -B & \end{pmatrix}.$$

The particular choice of $\tilde{M}$ affects both the convergence rate of the PPCG method and the computational effort for applying the preconditioner. A reasonable choice is the block triangular constraint preconditioner

$$Q_{\text{sc}} = \begin{pmatrix} 0 & 0 & A^T \\ 0 & M_u & -B^T \\ A & -B & \end{pmatrix}, \quad \text{i.e.} \quad \tilde{M} = \begin{pmatrix} 0 & 0 \\ 0 & M_u \end{pmatrix}. \tag{4.2.4}$$

Note that $\tilde{M}$ is spectrally equivalent to M on $\ker c'(x_0)$, since

$$\begin{aligned} \langle u, u\rangle_{\text{U}} &\leq \langle x, x\rangle_{\text{Y}\times\text{U}} = \langle y, y\rangle_{\text{Y}} + \langle u, u\rangle_{\text{U}} \\ &= \langle A^{-1}Bu, A^{-1}Bu\rangle_{\text{Y}} + \langle u, u\rangle_{\text{U}} \leq (1 + \|A^{-1}B\|^2_{\text{U}\to\text{Y}})\langle u, u\rangle_{\text{U}}. \end{aligned}$$

Often M_u is a scaled mass matrix and A an elliptic operator. For these efficient approximation schemes, that will be introduced in Sec. 4.5, are available. Note however, that since the constraint preconditioner has to project onto $\ker c'(x_0)$ we need highly accurate representations of A and, for symmetry reasons, also of A^T.

4.3. Computation of tangential steps

The standard PPCG method admits the solution of saddle point problems of the form (4.1.1) as long as M is positive definite on $\ker c'(x_0)$. Now we discuss the solution of

$$H_t z = r,$$

where $\mathcal{L}_{xx}$ is in general not positive definite on $\ker c'(x_0)$. In this case the conjugate gradient method is not directly applicable and must be modified. We will continue using the previously introduced notation, but mention that this section applies not only to constrained problems, but also to unconstrained ones.

Truncated conjugate gradient method. The most popular approach in this context is the truncated conjugate gradient method (TCG, Alg. 4.4), which terminates as soon as a direction of non-positive curvature is found. The used search directions span a subspace on which H_t is positive definite and no further modification of standard CG implementations are required. Working as long as possible on the original problem this approach seems to be quite effective in finding its way out of non-convexities, see Tab. 4.1. But we also observe that occasionally the TCG method does *not* lead us back into convex domains, at least not in a reasonable number of iterations. Here the problem is that the algorithm runs into a non-convexity which lead to termination of the TCG method after only few iterations. Thus only a very small subspace of the search space is covered and the computed direction may only very roughly lead us back into regions where the problem is convex. A popular

Require: initial iterate x, set $r = Hx - r$, $g = Q^{-1}r$ and $d = -g$.
1: **repeat**
2: **if** $d^T Hd < 0$ **do**
3: terminate
4: $\sigma \leftarrow r^T g$
5: $\alpha \leftarrow \sigma / d^T Hd$
6: $x \leftarrow x + \alpha d$
7: $r \leftarrow r + \alpha Hd$
8: $g \leftarrow Q^{-1}r$
9: $\beta \leftarrow r^T g / \sigma$
10: $d \leftarrow -g + \beta d$
11: **until** convergent

Algorithm 4.4: Truncated conjugate gradient method (TCG).

approach to improve the behavior of the TCG method is to also include the search direction for which the problem is nonconvex. This is motivated by the fact that this search direction still is a descent direction. However, in this case, the step size parameter α yields a local maximum of the energy norm. Therefore this last search direction is added "blindly" and may or may not increase the performance of the TCG method. The implementation used in this thesis incorporates this "blind" step.

Regularized conjugate gradient method. An alternative strategy is to modify H by adding multiples of the preconditioner Q. For some regularization parameter $\theta \geq 0$ such that $H + \theta Q$ is positive definite we can solve equations involving this regularized operator with a CG method. We refer to this strategy as regularized conjugate gradient method (RCG, Alg. 4.5).

Since a preconditioner is in general not given explicitly, but rather as an algorithm, it may not possible to directly compute the application of Q to a search direction d. However, starting from the observation that in the first CG iteration

$$q = Qd = -r$$

holds, we can update the quantity $q = Qd$, similar to the computation of the search directions d in the classical CG method, via

$$q \leftarrow -r + \beta q.$$

Lemma 4.3. *Consider the notation of Alg. 4.5, augmented by subscripts for the iterations numbers. Let $q_0 = r_0 - Hx_0$ and denote the sequence of generated conjugate search directions by $\{d_k\}_{k=1,\dots}$, the sequence of residuals by $\{r_k\}_{k=1,\dots}$ and the se-*

Require: initial iterate x, set $r = Hx - r$, $\theta = 0$, $g = Q^{-1}r$, $d = -g$ and $q = Qd = -r$.
repeat
  $z \leftarrow d^T H d + \theta d^T q$
  if $z < 0$ **do**
    increase θ
    restart
  $\sigma \leftarrow r^T g$
  $\alpha \leftarrow \sigma / z$
  $x \leftarrow x + \alpha p$
  $r \leftarrow r + \alpha(Hd + \theta q)$
  $g \leftarrow Q^{-1} r$
  $\beta \leftarrow r^T g / \sigma$
  $d \leftarrow -g + \beta d$
  $q \leftarrow -r + \beta q$
until convergent

Algorithm 4.5: Regularized conjugate gradient method (RCG).

quence of generated H-orthogonalization constants by $\{\beta_k\}_{k=1,\dots}$. Then the sequence $\{q_k\}_{k=1,\dots}$ generated via

$$q_k = -r_k + \beta_k q_{k-1}$$

satisfies

$$q_k = Qd_k.$$

Proof. For the initial iterate we have by definition $q_0 = -r_0 = Qd_0$. Let us assume that $q_k = Qd_k$. Then

$$d_{k+1} = -g_{k+1} + \beta d_k$$

and

$$q_{k+1} = -r_{k+1} + \beta_k q_k = -Qg_{k+1} + \beta_k Q d_k = Q(-g_{k+1} + \beta d_k) = Qd_{k+1}.$$

□

Consequently, our regularization only requires few additional arithmetic operations. The additional quantity $q = Qd$ may be required anyway, e.g. for termination criteria based on the Q-norm [133, 247]. We will not employ such a norm here. When considering the inexact solution of normal steps, the use of this norm may be advantageous, as it admits a proper matching of the inaccuracies.

For the choice of the regularization parameter θ, which, as usual, should be as small as possible and as big as necessary, we choose a simple heuristic. Starting the computation with $\theta = 0$ we discard the computed iterates as soon as we encounter a direction d of non-positive curvature $d^T H d < 0$ and update for some constant $c_d > 1$ the regularization parameter θ according to

$$\theta_{\text{new}} = \theta + \delta\theta \quad \text{with} \quad \delta\theta = c_d \frac{|d^T(H + \theta P)d|}{d^T P d}$$
$$\theta = \min\{\max\{\theta_{\text{new}}, \underline{c}_\theta \theta\}, \bar{c}_\theta \theta\},$$

with $1 < \underline{c}_\theta < \bar{c}_\theta$ such that the generated sequence of regularization parameters is strictly increasing each time a direction of non-positive curvature is encountered. The restriction $\theta \leq \bar{c}_\theta \theta_{\text{old}}$ is introduced as the update of θ according to $\delta\theta$ may be very large. From a purely theoretic perspective this is just what is required to guarantee convexity of the problem. In practice such large updates are not desirable. The reason is that high accuracy requirements are only imposed if we expect to be close to the solution. Far from it, where typically nonconvexities are encountered, the relative accuracy requirement δ_0 on the tangential direction will be low. Thus, instead of regularizing such that we can guarantee convexity of $H + \theta Q$ on the space spanned by the considered search directions, we rather aim at staying as close as possible to H, while increasing the subspace where $H + \theta Q$ is convex until the accuracy requirement are met.

After the update of θ we have to restart the CG iteration. One might assume that the previously computed solution could serve as new starting value. However, this is not the case, as even mild regularizations significantly alter the system matrix. Due to this restart the application of RCG is more expensive than one application of TCG, but for difficult problems this additional cost is often outweighed by a significantly reduced number of outer iterations.

We refer to Tab. 4.1 for a comparison of outer iteration numbers for different parameters. In all our computations we chose $\underline{c}_\theta = 2$ and $\bar{c}_\theta = 10$. We observe that the RCG method behaves more robustly than TCG, but occasionally it requires more outer iterations.

Remark 4.4.

- Note the analogy of the RCG method to hessian modification methods. We stress, however, that we do not add multiples of the identity matrix to the hessian, but rather add implicitly multiples of our preconditioner. Thus, we capture more of the underlying problem structure.
- The RCG method seems to be of particular interest if $H + \theta Q$ can be related to a physical model similar to H. This is the case in our numerical example problems from nonlinear elasticity where we use a simplified material model for preconditioning. Then we may interpret $H + \theta Q$ as the linearization of a model that corresponds to a more rigid material than the original one. Solutions of this problem enjoy better regularity properties compared to solutions

that have been computed by the TCG method. Even if such an interpretation is not admissible the RCG method seems, in the presence of reasonable preconditioners, to be more robust than the TCG method.

Require: initial iterate x, set $r = Hx - r$, $g = Q^{-1}r$, $\theta = 0$, $d = -g$ and $q = Qd = -r$.

repeat
  $z \leftarrow d^T H d + \theta d^T q$
  if $z < 0$ **do**
    if minimal decrease achieved **do**
      terminate
    else
      increase θ
      restart
  $\sigma \leftarrow r^T g$
  $\alpha \leftarrow \sigma / z$
  $x \leftarrow x + \alpha d$
  $r \leftarrow r + \alpha(Hd + \theta q)$
  $g \leftarrow Q^{-1} r$
  $\beta \leftarrow r^T g / \sigma$
  $d \leftarrow -g + \beta d$
  $q \leftarrow -r + \beta q$
until convergent

Algorithm 4.6: Hybrid conjugate gradient method (HCG).

Hybrid conjugate gradient method. We saw that TCG performs quite well except in the case that it runs into non-convexities early. Therefore, if we were able to make a reasonable choice when to prefer regularization and when to prefer truncation, then we would expect better performance of our composite step algorithm.

A suitable argument exists, if we take into account the fact we want to compute a tangential direction for our composite algorithm. It is well acknowledged that far from a local minimum we do not need to compute iterates, in our case tangential directions, overly accurate [76, 79]. Thus in this case we only need a moderate relative decrease $\delta = \delta_{\min}$ in the quantity underlying the used termination criterion of the employed Krylov solver. Only close to the solution, when some algorithmic quantities indicate fast local convergence, we adjust the accuracy requirements until it meets the desired relative accuracy $\delta = \varepsilon_{\text{tol}}$ of the outer iteration. This is explained in more detail at the end of this section.

As a consequence of the low accuracy that is imposed far from local minima only a small part of the actual problem structure is observable in the outer iteration.

Thus, indicators for fast local convergence may be too optimistic and may yield "false positives". In order to detect these, we will incorporate TCG as a fallback mechanism for RCG.

This means that we use RCG as default method. However, if a nonconvexity is encountered and additionally we observe a relative decrease of $\delta_{\min}$ or better, then we truncate. Regularization may have happened before truncation and thus we do not use a plain TCG method here, but apply the TCG approach to RCG. For consistency with [180] this combined (hybrid) method is called HCG (, Alg. 4.6)[1].

With HCG we can significantly reduce the computational costs when encountering such "false positives". Even more important, since we have to regularize less and thus stay closer to the original problem this admits to significantly reduce the number of required outer iterations, see Tab. 4.1. Thus, HCG performs like RCG would if we had a better indicator for fast local convergence.

Note that, when having to choose one of the conjugate gradient methods for nonconvex problems, Tab. 4.1 does not tell the full truth. Without additional structure we can never say which method should be expected to perform better. Thus, despite the attained promising results, the blind choice of one of the algorithms as black-box solvers for nonconvex problems is not recommended.

Alg.	TCG			RCG			HCG		
d \ c	10^2	10^3	10^4	10^2	10^3	10^4	10^2	10^3	10^4
10^{-5}	†	27	34	177	24	17	12	35	16
10^{-4}	24	34	29	21	36	17	24	22	14
10^{-3}	28	17	14	19	17	15	12	25	14
10^{-2}	10	19	16	18	14	18	13	18	17
10^{-1}	8	17	19	8	24	21	8	20	18
1	7	11	14	8	12	20	8	12	17

Table 4.1.: Required iterations for an example problem of nonlinear heat transfer (Sec. 6.1) for different parameters c and d on a fixed uniform grid with $h_{\max} = 2^{-7}$, $\alpha = 10^{-6}$ (†: not convergent within 500 iterations).

Termination criterion. It is well known that the widely used termination criteria for the dual norm of the preconditioned residual only yields a useful termination criterion in the case that $\kappa(Q^{-1}H) \approx 1$, i.e. if the preconditioner Q approximates H

[1] Calling this approach TRCG (for **t**runcated **r**egularized **c**onjugate gradient **m**ethod) is more intuitive and probably a better name.

very well, which we can not expect here. Instead, as already observed in the original paper of Hestenes and Stiefel [133], termination criteria for conjugate gradient methods should rely on the provable decrease of the energy error, or often better the relative energy error. Based on representation formulae given in [133] estimators for the absolute energy error $\|x - x_k\|_H$ and the relative energy error $\frac{\|x-x_k\|_H}{\|x\|_H}$ have been proposed in [11] resp. [247]. As all of the above presented conjugate gradient methods only work on subspaces where the, possibly regularized, problem is convex we can use the same termination criteria for non-convex problems. To control the algebraic errors we employ the estimate for the relative energy error proposed in [247]. Exploiting only local H-orthogonality the proposed estimate

$$\rho_{j,d} = \frac{\tilde{\rho}_{j,d}}{\xi_{j+d}} \tag{4.3.1}$$

with $\xi_{j+d} = \tilde{\rho}_{0,j+d} + r^T x_0 + r_0^T x_0$ and

$$\tilde{\rho}_{j,d} = \sum_{i=j}^{j+d-1} \alpha_i r_i^T g_i$$

is numerically stable. All quantities are available during computation, the only drawback lying in the fact that we need to perform $j + d$ iterations, for some look-ahead parameter d, in order to estimate the relative energy error in the j-th step. As the conjugate gradient method guarantees descent in the energy norm in each iteration,

$$\|x - x_{j+d}\|_H < \|x - x_j\|_H,$$

we accept the last iterate x_{j+d} if the estimate for $\frac{\|x-x_j\|_H}{\|x\|_H}$ is accepted.

In our computations we add another termination criterion to this estimate. This is required for the case that $\|x\|_H \approx \epsilon_{\max}$, where $\epsilon_{\max}$ is the maximal attainable accuracy of the linear system. Then, we cannot expect that the above termination criterion works. Therefore, we additionally terminate the CG iteration if the step length becomes negligible compared to the size of the current iterate, i.e. if

$$\|q\|_H \leq \epsilon_{\max} \|x\|_H .$$

Alternatively we may also add a criterion for the absolute energy error to achieve fast termination of the CG iteration for right hand sides that correspond to vanishing right hand sides plus numerical noise.

Accuracy matching. Far from the solution it does not make much sense to spend significant effort in the computation of highly accurate tangential directions. Therefore, following [76, Sec. 2.3.3] we content ourselves with a minimal relative accuracy of $\delta_0 = 0.25$. This guarantees that at least the leading two binary digits of the computed solution are correct. Though, close to the solution we should increase the

prescribed accuracy in order to profit from the local quadratic convergence of the Newton-Lagrange scheme. For constrained optimization problems this is not at all a trivial issue and under current investigation. Here we employ a heuristic argument. We decide being close to the solution if in the last, say k-th, step

- no damping occurred,
- no direction of non-positive curvature was encountered in the computation of the tangential direction,
- the estimate of the Kantorovich quantity satisfies $h_k = [\omega_c]\|\delta x_k\| < 1$.

In this case, we set the desired relative accuracy in the $(k+1)$-st step to

$$\delta_{k+1} = \min\left\{\delta_0, [\omega_f]\|\delta x_k\|\right\},$$

cf. [77, 78, 277].

The above choice $\delta_0 = 0.25$ implies that our algorithm will often overlook the presence of directions of negative curvature. To illustrate the differences between the different conjugate gradient methods in dealing with non-convexities, we employed $\delta_0 = 10^{-3}$ in the computations for Tab. 4.1. The look-ahead parameter is chosen to be $d = 10$ and the maximal attainable accuracy to $\epsilon_{\max} = 10^{-11}$. In the hybrid method, truncation is accepted as soon as at least the two most important digits are captured, i.e. if the relative energy error is decreased by a factor of 0.25. This is consistent with the introduced accuracy requirement.

4.4. Error estimation

In this section an error estimator for an adaptive version of the proposed composite step method is described. Error estimators for the discretization error in optimal control problems are based on strategies that have been applied to PDE problems. We shortly introduce some popular error estimation strategies for the simpler case of an elliptic PDE in Sec. 4.4.1 and discuss their applicability in the context of implant shape design. Then we turn to the description of an error estimator for the proposed affine covariant composite step method in Sec. 4.4.2.

4.4.1. Error estimation strategies

For the reliable computation of solutions to PDEs, here written in operator form

$$Ax = b \quad \text{in } \mathrm{X}^*,$$

we have to control the approximation error $\tilde{\epsilon}_h = \|x - \tilde{x}_h\|$ between the exact solution x and $\tilde{x}_h$, the computed solution in a finite dimensional subspace $\mathrm{X}_h \subset \mathrm{X}$. Let x_h be the exact solution in X_h. Exploiting Galerkin orthogonality, the approximation

error can be decomposed into algebraic error $\delta_h = \|x_h - \tilde{x}_h\|$ and discretization error $\epsilon_h = \|x - x_h\|$, i.e.

$$\tilde{\epsilon}_h^2 = \epsilon_h^2 + \delta_h^2. \tag{4.4.1}$$

This suggests that algebraic and discretization error should be in balance, a principle that is commonly followed in the derivation of error estimators. In practice neither x nor x_h are available. To get a computable estimate of the discretization error we concentrate on strategies that use $\tilde{x}_h$ as approximation of x_h and a better approximation $\hat{x}$ for the solution x. This yields an estimate for ϵ_h via $[\epsilon_h] = \|\hat{x} - \tilde{x}_h\|$.

Remark 4.5. As both $\hat{x}$ and $\tilde{x}_h$ are affected by algebraic errors the above splitting is not valid for computable quantities [12, 246]. Here, we assume that the error is dominated by the discretization error and neglect this issue. In a more detailed investigation of error estimators for optimal control problems this point should be taken into account.

The computation of $\hat{x}$ is in general too expensive to be realized with high accuracy. For this reason we are in general not in a position to accurately estimate the discretization error and a proper balancing of error sources is not possible. Instead one tries to achieve one of the following goals, cf. [31],

- either minimize the required number of degrees of freedom to guarantee a desired (relative) discretization accuracy,
- or minimize the (relative) discretization error for a given bound on the number of degrees of freedom.

Besides the fact that the first strategy may only be roughly be realized, due to the lack of accuracy in the discretization error, we may encounter the case that the desired (relative) discretization accuracy is chosen too small. Then we might exceed the available computational resources. These are known a priori, in contrast to the achievable (relative) accuracy. For this reason, at least for accurate computations on complex geometries, the only reasonable strategy is the second one.

In general the discretization error is not uniformly distributed on the computational domain. Instead, the error may be concentrated at problem specific features such as corner singularities or boundary layers [66, 67, 116]. For this reason, the estimate $[\epsilon_h]$ needs to be localized to admit efficient local mesh refinement. Here, localization to the cells $\{T\}$ of the spatial discretization is straightforward via $[\epsilon_h(T)] = \|(\hat{x} - \tilde{x}_h)\chi_T\|$, where χ_T is the indicator function of T, i.e.

$$\chi_T(x) = \begin{cases} 1 & \text{if } x \in T \\ 0 & \text{else} \end{cases}.$$

Some desirable properties of error estimators are collected in the following definition.

Definition 4.6. An error estimator $[\epsilon_h]$ is called

- **reliable** if there exists a constant $\kappa_1 \geq 1$ such that

$$\epsilon_h \leq \kappa_1 [\epsilon_h],$$

- **efficient** if there exists a constant $\kappa_2 \geq 1$ such that

$$[\epsilon_h] \leq \kappa_2 \epsilon_h,$$

- **asymptotically exact** if it is reliable and efficient and

$$\lim_{h \to 0} \kappa_i = 1 \quad \text{for } i = 1, 2.$$

Remark 4.7. Reliability and efficiency may be comprised in

$$\frac{1}{\kappa_1} \epsilon_h \leq [\epsilon_h] \leq \kappa_2 \epsilon_h.$$

The product $\kappa_1 \kappa_2 \geq 1$ is called **efficiency range** of the estimator. Both reliability and efficiency are properties that are satisfied by most error estimators for PDEs, at least up to higher order terms. In contrast, asymptotic exactness only is of value in theoretical considerations. In practice, for efficiency requirements, we can neither afford to generate grids that would admit to speak of asymptotic behavior nor to compute error estimates sufficiently accurate to guarantee this property. In scientific computing we are rather interested in estimators that are efficient to evaluate and are sufficiently accurate in the transient phase.

For the computation of $\hat{x}$ different strategies have been proposed. *Averaging error estimators*, also called *gradient recovery error estimators* are motivated by the observation that, in the case of smooth solutions, discrete stresses or fluxes $\tilde{x}_h$ are significantly less smooth than its continuous counterpart x. Thus,

$$\|x - A(x_h)\| \ll \|x - x_h\|,$$

where A is a projector into a smoother space. This suggests to base error estimators on the quantity $[\epsilon_h] = \|A(x_h) - x_h\|$, i.e. $\hat{x} = A(x_h)$. The simplicity of these error estimators is appealing. However, if the true stress or flux exhibits discontinuities, the performance of averaging error estimators can be poor [203]. Since the implant shape design problem is concerned with different soft tissue types such discontinuities can arise at interior tissue interfaces. Therefore the assumptions of averaging error estimators are not satisfied in our setting and we do not consider this strategy any further. For details, the interested reader is referred to the books [5, 79] as well as to the publications of Carstensen [51, 52].

An alternative to averaging error estimators are *hierarchical error estimators*. These directly compute an approximation $x_{\bar{h}}$ of x in a larger subspace $\mathrm{X}_{\bar{h}}$, satisfying $\mathrm{X}_h \subset$

$\mathrm{X}_{\bar{h}} \subset \mathrm{X}$. If $\mathrm{X}_{\bar{h}}$ is sufficiently large to capture significant parts of the discretization error, i.e. if the *saturation assumption*

$$\epsilon_{\bar{h}} \leq \beta \epsilon_h, \text{ with } 0 < \beta < 1,$$

holds, then accurate computation and localization of the error is to be expected. This inequality holds as long as oscillations in the right hand side of the operator equation in X are small [86].

As a disadvantage, this direct approach is computationally expensive. The increased computational cost for the computation of the error $[e_h] = x_{\bar{h}} - x_h$ can be significantly reduced if $\mathrm{X}_{\bar{h}}$ is constructed as a hierarchical extension from X_h, i.e. $\mathrm{X}_{\bar{h}} = \mathrm{X}_h \oplus \mathrm{X}_e$ [80, 289]. Then $[e_h] = x_{\bar{h}} - x_h \in \mathrm{X}_e = \mathrm{X}_{\bar{h}} \setminus \mathrm{X}_h$ is only computed in the extension space.

Regarding the energy error in elliptic PDEs the prototype of a hierarchical error estimator is the DLY estimator [80]. Using the hierarchical decomposition $\mathrm{X}_{\bar{h}} = \mathrm{X}_h \oplus \mathrm{X}_e$ the discretized system in $\mathrm{X}_{\bar{h}}$ is given through

$$\begin{pmatrix} A^{hh} & A^{eh} \\ A^{he} & A^{ee} \end{pmatrix} \begin{pmatrix} \eta_h \\ \eta_e \end{pmatrix} = \begin{pmatrix} b_h \\ b_e \end{pmatrix}.$$

The upper right block A^{eh} is assumed to be negligible, i.e. $\eta_h \approx {x_h}$[2]. Then, the above system is block triangular and we can compute the error estimate $[e_h] = \eta_e$ from the defect equation in X_e

$$A^{ee}\,[e_h] = b_e - A^{he} x_h.$$

Solving this equation in X_e is still expensive. Thus, to further reduce the computational costs, the operator A^{ee} is typically replaced by its lumped matrix $\hat{A}^{ee}$, which only contains the diagonal of A^{ee}. As is to be expected for elliptic PDEs, this does not affect efficiency and reliability but asymptotic exactness is lost [79, Sec. 6.1.4]. Thus we compute an estimate $[e_h]$ from the equation

$$\hat{A}^{ee}\,[e_h] = b_e - A^{he} x_h, \tag{4.4.2}$$

and localize it via $[\epsilon_h(T)] = \| [e_h]\, \chi_T \|$.

Remark 4.8. For sake of completeness we also mention *residual based error estimators*, which are the among the first proposed error estimators. Theoretically these are better understood than other a posteriori error estimators. However, in general these only provide a coarse global upper bound on the energy error. Moreover, they do not fit into an affine covariant algorithmic setting, where we avoid the evaluation

[2] This is supported by the gradient recovery result of Ovall [204]. However, this result relies on asymptotic exactness of the estimated function values and may not be valid for real world applications.

of residuals. We do not go into detail and refer the interested reader to the books of Ainsworth and Oden [5], Babuška et al. [14], Brenner and Scott [44], Deuflhard and Weiser [79], Verfürth [268]. We also suggest the report [127], where efficiency and reliability have been shown for a residual based error estimator without assuming that Galerkin orthogonality holds. Instead the authors exploit spectral equivalence of the BPX preconditioner [43] to estimate the $W^{-1,2}$-norm of the discrete residuals.

The above mentioned error estimation strategies all focus on the estimation of the (energy) error. In practical applications, this is not always the quantity of interest. Considering a more general setting for error measurement a duality-based method, the *dual weighted residual* method (DWR), cf. Bangerth and Rannacher [24], has been developed. There the discretization error is considered with respect to its influence on a functional J that measures the error in a specific quantity of interest. The essential idea behind this approach is the observation that solutions to a dual problem can be used to describe the actual influence of the primal residuals on the error in the quantity of interest J. For this we need approximations of the primal solution x as well as the corresponding Lagrange multiplier p. These are mostly computed using an averaging or a hierarchical extension approach. Using averaging strategies this method has been applied to a wide range of problems [30, 29]. Not directly obvious this approach incorporates a measure for the deviation from Galerkin orthogonality [12]. However, due its complicated setting several gaps remain to be filled [24] regarding rigorous theoretical backing.

Error estimation for optimal control problems. The extension of most of the strategies for error estimation in the PDE context to nonconvex optimal control problems is yet unclear. First, for these problems no energy norm exists to measure the error. Second, often it is a priori not clear how the quantity of interest should sensibly be chosen. Third, the coupling between the different variables complicates the attainment of theoretical results, such as reliability or efficiency. Therefore, these are harder to obtain than in the case of one variable. In particular, available results are typically based on relatively strong assumptions, such as convergence of the discrete solutions to the infinite dimensional solution [219].

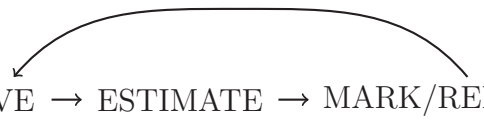
The DWR method has been extended to elliptic optimal control problems in [29, 30] and, additionally incorporating inequality constraints on the control variable, in [269]. Both approaches consider the cost functional J as quantity of interest. Then the accuracy of the discretization of the constraint is only considered with respect to its influence on the cost functional. Thus, admissibility of the iterates can only be guaranteed "in a very weak sense, possibly insufficient for the particular application" [30, p. 2]. To treat this issue one can combine the DWR approach with classical energy error estimation techniques for the PDE-constraint [30].

Another reasonable strategy is to base error estimates on the underlying (local) KKT-systems. Thus, one might choose to estimate the error in the state equation, the adjoint equation and possibly the variational equation. In academic examples,

the latter typically arises from a Tikhonov-regularized tracking type cost functional and is of minor interest. In contrast, both, the state and adjoint equation, contain significant information on the energy error in the primal variables. Treating both equations independently the whole machinery for error estimation in PDEs can be used. This idea has been pursued in [287], using averaging error estimators. There, also the consequences of inexact step computation, such as loss of Galerkin orthogonality, was investigated. Residual based error estimators, targeting at the error in the cost functional, have been proposed for optimal control problems with state and/or control constraints in [33, 219].

4.4.2. A hierarchical error estimator

We introduce a hierarchical error estimator for the Lagrange-Newton step δx in terms of the norm that is used in the composite step method. For nonlinear problems it is widely acknowledged that error estimation only is meaningful close to the exact solution x. Before turning to the practical realization of a hierarchical error estimator we discuss the question when error estimation is performed and how it should be interpreted. For this mainly two strategies are followed. In text books we most often encounter the *sequential* strategy



First, on a fixed grid, the discretized nonlinear problem is solved. The error of the nonlinear problem is estimated and the mesh is refined accordingly. Then, the solution process restarts, taking the previously computed solution as initial guess. Amongst others, this approach is followed in [29, 30, 33, 219, 269]. In this regard, it is usually assumed that the current discrete solution x_h is sufficiently close to x such that nonlinearities can be neglected in the error estimation procedure. For strongly nonlinear problems or coarse spatial discretization this assumption is in general not valid. In these cases one either has to consider the nonlinearities during error estimation or resort to other strategies.

An alternative is provided from the perspective of quasilinearization. Instead of estimating the error of the full nonlinear problem, we can estimate the error of the linearized problems, i.e. the error in the Lagrange-Newton steps. This admits the computation of meaningful error estimates without any need to justify the neglect of nonlinear contributions. Another advantage of this approach is the observation that we can start error estimation as soon as we observe that local convergence sets in. In this way we try to avoid the inefficient reduction of algebraic errors, when the discretization error is dominating, cf. [76, Sec. 8.3]. In contrast to the sequential strategy, this strategy requires to *integrate* error estimation into the optimization algorithm.

Here, we follow this integrated approach. Similar to the criteria used to control the relative accuracy of the tangential steps, we allow mesh refinement if

- no damping occurred,
- no direction of non-positive curvature was encountered in the computation of the tangential direction.

Since we only perform mesh refinement in the case that undamped steps are accepted and the tangential solver terminates without encountering nonconvexities, the step δx satisfies at some iterate (y_k, u_k, p_k) in X_h the local KKT-system

$$\begin{pmatrix} \mathcal{L}_{yy} & \mathcal{L}_{yu} & A^* \\ \mathcal{L}_{uy} & \mathcal{L}_{uu} & B^* \\ A & B & \end{pmatrix} \begin{pmatrix} \delta y \\ \delta u \\ q \end{pmatrix} = \begin{pmatrix} r_y \\ r_u \\ r_p \end{pmatrix}, \tag{4.4.3}$$

with $A = A'(y_k)$, $\mathcal{L} = \mathcal{L}(y_k, u_k, p_k)$ and right hand sides $r_y = -\mathcal{L}_y$, $r_u = -\mathcal{L}_u$, $r_p = -c$. Conceptually, a suitable hierarchical error estimator requires the solution of this in a hierarchically extended ansatz space

$$\mathrm{X}_{\bar{h}} = \mathrm{X}_h \oplus \mathrm{X}_e = \{\mathrm{Y}_h \times \mathrm{U}_h \times \mathrm{P}_h\} \oplus \{\mathrm{Y}_e \times \mathrm{U}_e \times \mathrm{P}_e\}.$$

The solution of this system is far too expensive to be of use in practice. Therefore, in order to efficiently compute an error indicator (e_y, e_u) the system (4.4.3) is simplified to block triangular form. This simplification requires to break the coupling between the different variables. How this should be done is not well understood. Here, we employ a heuristic that is motivated by the implant shape design problem. In this regard, we want to detect errors in the state variable which are caused by lack of precision in the control.

We start with estimating the error $e_y = (e_y^h, e_y^e)$ in the state equation in $\mathrm{Y}_h \oplus \mathrm{Y}_e$. In order to incorporate information on the cost functional, this is propagated through the adjoint equation in $\mathrm{P}_h \oplus \mathrm{P}_e$ to estimate the corresponding dual variable $q = (q^h, q^e)$. Eventually the error in the control $e_u = (e_u^h, e_u^e)$ is estimated by inserting the error contributions from both variables into the variational equation in $\mathrm{U}_h \otimes \mathrm{U}_e$.

Let us state this more precisely. We compute an estimate $[e_y]$ for the error $e_y = (e_y^h, e_y^e)$ from the defect state equation:

$$\begin{pmatrix} A^{hh} & A^{eh} \\ A^{he} & A^{ee} \end{pmatrix} \begin{pmatrix} e_y^h \\ e_y^e \end{pmatrix} + \begin{pmatrix} B^{hh} & B^{eh} \\ B^{he} & B^{ee} \end{pmatrix} \begin{pmatrix} e_u^h \\ e_u^e \end{pmatrix} = \begin{pmatrix} r_p^h - A^{hh}\delta y - B^{hh}\delta u \\ r_p^e - A^{he}\delta y - B^{he}\delta u \end{pmatrix}.$$

At this point, the error in the control variable is not available and will be neglected, setting $e_u^h = 0$ and $e_u^e = 0$. Then the natural candidate for estimating the error in the state variable is the DLY estimator (4.4.2). Thus, we assume that for the estimates $\left[e_y^e\right]$ and $\left[e_y^h\right]$ we have

$$A^{eh}\left[e_y^e\right] = 0 \Leftrightarrow \left[e_y^h\right] = 0.$$

This yields for the error indicator in the extension space

$$\hat{A}^{ee}\left[e_y^e\right] = r_p^e - A^{he}\delta y - B^{he}\delta u,$$

where $\hat{A}^{ee}$ is the lumped matrix of A^{ee}.

To incorporate information from the cost functional we propagate the state error through the adjoint equation

$$\begin{pmatrix} \mathcal{L}_{yy}^{hh} & \mathcal{L}_{yy}^{eh} \\ \mathcal{L}_{yy}^{he} & \mathcal{L}_{yy}^{ee} \end{pmatrix} \begin{pmatrix} e_y^h \\ e_y^e \end{pmatrix} + \begin{pmatrix} \mathcal{L}_{yu}^{hh} & \mathcal{L}_{yu}^{eh} \\ \mathcal{L}_{yu}^{he} & \mathcal{L}_{yu}^{ee} \end{pmatrix} \begin{pmatrix} e_u^h \\ e_u^e \end{pmatrix} + \begin{pmatrix} (A^*)^{hh} & (A^*)^{eh} \\ (A^*)^{he} & (A^*)^{ee} \end{pmatrix} \begin{pmatrix} q^h \\ q^e \end{pmatrix}$$
$$= \begin{pmatrix} r_y^h - \mathcal{L}_{yy}^{hh}\delta y - \mathcal{L}_{yu}^{hh}\delta u \\ r_y^e - \mathcal{L}_{yy}^{he}\delta y - \mathcal{L}_{yu}^{he}\delta u \end{pmatrix}.$$

Again we neglect the contributions from the control variable, setting $e_u^h = 0$ and $e_u^e = 0$. Then, inserting the estimates $\left[e_y^h\right] = 0$ and $\left[e_y^e\right]$ for the error in the state variable, we get

$$\begin{pmatrix} (A^*)^{hh} & (A^*)^{eh} \\ (A^*)^{he} & (A^*)^{ee} \end{pmatrix} \begin{pmatrix} q^h \\ q^e \end{pmatrix} = \begin{pmatrix} r_y^h - \mathcal{L}_{yy}^{hh}\delta y - \mathcal{L}_{yu}^{hh}\delta u - \mathcal{L}_{yy}^{eh}\left[e_y^e\right] \\ r_y^e - \mathcal{L}_{yy}^{he}\delta y - \mathcal{L}_{yu}^{he}\delta u - \mathcal{L}_{yy}^{ee}\left[e_y^e\right] \end{pmatrix}.$$

Due to the summand $\mathcal{L}_{yy}^{eh}\left[e_y^e\right] \neq 0$ in the right hand side we can not assume that q^h coincides with the dual variable q of (4.4.3). Consequently, the consideration of the defect equation alone may neglect significant parts of the error in the adjoint equation. Still, for the efficient computation of estimates $\left[q^h\right]$ and $[q^e]$ we have to simplify the adjoint equation in the extended space $\mathrm{P}_h \oplus \mathrm{P}_e$ to block triangular form. At this point it is not clear whether $(A^*)^{eh}$ or $(A^*)^{he}$ should be neglected. Here we retain the upper right block $(A^*)^{eh}$. This is motivated by the idea that long-distance error transport can not be captured in the extension space. However, there is no strong argument for this choice and a better understanding of the consequences of the neglect of one of the off-diagonal blocks is desirable. Again replacing the differential operator $(A^*)^{ee}$ by its lumped form $(\hat{A}^*)^{ee}$ we get

$$\begin{pmatrix} (A^*)^{hh} & (A^*)^{eh} \\ & (\hat{A}^*)^{ee} \end{pmatrix} \begin{pmatrix} \left[q^h\right] \\ [q^e] \end{pmatrix} = \begin{pmatrix} r_y^h - \mathcal{L}_{yy}^{hh}\delta y - \mathcal{L}_{yu}^{hh}\delta u - \mathcal{L}_{yy}^{eh}\left[e_y^e\right] \\ r_y^e - \mathcal{L}_{yy}^{he}\delta y - \mathcal{L}_{yu}^{he}\delta u - \mathcal{L}_{yy}^{ee}\left[e_y^e\right] \end{pmatrix}.$$

Eventually, we propagate the error through the variational equation

$$\begin{pmatrix} \mathcal{L}_{uy}^{hh} & \mathcal{L}_{uy}^{eh} \\ \mathcal{L}_{uy}^{he} & \mathcal{L}_{uy}^{ee} \end{pmatrix} \begin{pmatrix} e_y^h \\ e_y^e \end{pmatrix} + \begin{pmatrix} \mathcal{L}_{uu}^{hh} & \mathcal{L}_{uu}^{eh} \\ \mathcal{L}_{uu}^{he} & \mathcal{L}_{uu}^{ee} \end{pmatrix} \begin{pmatrix} e_u^h \\ e_u^e \end{pmatrix} + \begin{pmatrix} (B^*)^{hh} & (B^*)^{eh} \\ (B^*)^{he} & (B^*)^{ee} \end{pmatrix} \begin{pmatrix} q^h \\ q^e \end{pmatrix}$$
$$= \begin{pmatrix} r_u^h - \mathcal{L}_{uy}^{hh}\delta y - \mathcal{L}_{uu}^{hh}\delta u \\ r_u^e - \mathcal{L}_{uy}^{he}\delta y - \mathcal{L}_{uu}^{he}\delta u \end{pmatrix}.$$

Recall that in $\mathrm{U}_h \oplus \mathrm{U}_e$ the operator $\mathcal{L}_{uu}$ yields a scaled mass matrix. Except for the scaling, we do not expect a large influence from this equation. In particular no long-distance transport phenomena will occur. Again it is not clear whether neglect of $\mathcal{L}_{uu}^{eh}$ or $\mathcal{L}_{uu}^{he}$ should be preferred. For the same reason as in the case of the adjoint equation we neglect the lower right part $\mathcal{L}_{uu}^{he}$. Again replacing the operator in U_e by its lumped form $\hat{\mathcal{L}}_{uu}^{ee}$, we get

$$\begin{pmatrix} \mathcal{L}_{uu}^{hh} & \mathcal{L}_{uu}^{eh} \\ & \hat{\mathcal{L}}_{uu}^{ee} \end{pmatrix} \begin{pmatrix} \left[e_u^h\right] \\ \left[e_u^e\right] \end{pmatrix}$$
$$= \begin{pmatrix} r_u^h - \mathcal{L}_{uy}^{hh}\delta y - \mathcal{L}_{uu}^{hh}\delta u - \mathcal{L}_{uy}^{eh}\left[e_y^e\right] - (B^*)^{hh}\left[q^h\right] - (B^*)^{eh}\left[q^e\right] \\ r_u^e - \mathcal{L}_{uy}^{he}\delta y - \mathcal{L}_{uu}^{he}\delta u - \mathcal{L}_{uy}^{ee}\left[e_y^e\right] - (B^*)^{he}\left[q^h\right] - (B^*)^{ee}\left[q^e\right] \end{pmatrix}.$$

The error is measured in the norm used in the composite step method:

$$[\varepsilon_h] = \| \left(\left[e_y^h\right] + \left[e_y^e\right], \left[e_u^h\right] + \left[e_u^e\right] \right) \| = \| \left(\left[e_y^e\right], \left[e_u^h\right] + \left[e_u^e\right] \right) \|.$$

Localization is straightforward via

$$[\varepsilon_h(T)] = \| \left(\left[e_y^e\right], \left[e_u^h\right] + \left[e_u^e\right] \right) \chi_T \|,$$

where χ_T is the indicator function of the grid cell T.

We will use this estimator in an adaptive version of the affine covariant composite step method. There we will measure the error in the Newton step if the local problems appear to be convex and no damping is required. As marking strategy the error equilibration strategy is the method of choice [79, p. 229]. Thus we mark all grid cells T for which the indicator satisfies

$$[\varepsilon_h(T)] \geq \frac{1}{n} [\varepsilon_h],$$

where n is the number of grid cells. The practical behavior of the derived error estimator is illustrated in Chap. 6.

4.5. Approximation of operators

In large scale problems the evaluation of the block triangular constraint preconditioner and the error estimator requires the solution of state, adjoint and variational equation. Depending on the structure of the involved operators different approximation schemes are favorable. In the variational equation we need to invert the scaled mass matrix $\mathcal{L}_{uu}$, which is discussed in Sec. 4.5.1. The treatment of the differential operators A and A^*, occurring in the state and the adjoint equations, is addressed in Sec. 4.5.2.

4.5.1. Approximation of the mass matrix

All three, the state constraint preconditioner for the normal step, the inexact state constraint preconditioner for the tangential step as well as the error indicator, involve the solution of the discretized variational equation

$$\alpha M u = b = r_u - B^T p,$$

where M is a mass matrix. In all three cases we will employ the Chebyshev semi-iteration [222, 110, 177, 95], named by Varga [266], as a polynomial preconditioner $Q_n(M)$. It is based on the Chebyshev polynomials

$$\mathcal{C}_k(x) = \cos(k \arccos(x)), \quad x \in [-1, 1],$$

which can be implicitly computed by the three-term recurrence

$$\mathcal{C}_0(x) = 1,\ \mathcal{C}_1(x) = x,\ \mathcal{C}_{k+1}(x) = 2x\mathcal{C}_k(x) - \mathcal{C}_{k-1}(x) \quad k > 0. \tag{4.5.1}$$

The scaled Chebyshev polynomials $p_k(x) = \frac{\mathcal{C}_k(x)}{\mathcal{C}_k(0)}$ are the unique solutions to the minimization problem

$$\min_{\substack{\varphi \in \Pi_k \\ \varphi(0)=0}} \max_{x \in [-1,1]} |\varphi(x)|, \tag{4.5.2}$$

and thus are the optimal polynomials p_k of order k with respect to the condition number of $p_k(A)A$ [177, 222]. This motivates, for an equation $Au = b$, the definition of a three-term recurrence, the Chebyshev semi-iteration, here with explicitly updated residuals [124].

Require: given u_0 and positive constants a, b, set $u_{-1} = 0,\ r_{-1} = 0,\ r_0 = b - Au_0,\ \beta_0 = -\frac{b^2}{2a}$ and $\gamma_0 = -a$.

1: **for** $k = 0, \ldots$ **do**
2: $\quad \beta_{k-1} \leftarrow \left(\frac{b}{2}\right)^2 \frac{1}{\gamma_{k-1}}$ if $k \geq 2$
3: $\quad \gamma_k \leftarrow -(a + \beta_k)$ if $k \geq 1$
4: $\quad u_{k+1} \leftarrow -\frac{1}{\gamma_k}(r_k + a x_k + \beta_k x_{k-1})$
5: $\quad r_{k+1} \leftarrow b - Au_{k+1}$

Algorithm 4.7: Chebyshev semi-iteration.

Realizing a fixed point iteration we need that the spectrum of A is contained in $[a - b, a + b]$. In particular for efficient application of the Chebyshev semi-iteration good estimates for spectral bounds $\lambda_{\min}$ and $\lambda_{\max}$ are required. For the case of a mass matrix that is preconditioned by a one-step Jacobi preconditioner, sufficiently good estimates for the spectral bounds have been obtained for many discretizations [273]. In this case, the Chebyshev semi-iteration provides a cheap and efficient

iterative solver and, if employed with a fixed number of iterations, it realizes a linear preconditioner.

Linearity of preconditioners is an important property. First because this is what we employ in theoretical considerations. Second, if nonlinear preconditioners such as a CG method – even with a fixed number of steps – are used as preconditioners for Krylov solvers this can significantly reduce their performance. This issue is nicely illustrated in [274].

4.5.2. Approximation of the stiffness matrix

Regarding the discretizations of the differential operators, occurring in the constraint and the adjoint equation, hierarchical multigrid solvers and preconditioners are among the most promising methods for adaptive algorithms. The two main reasons are the h-independent convergence rate and linear complexity with respect to the number of unknowns. Both properties are not only provable, but also can be verified for reasonable implementations of multigrid algorithms [260].

These mainly exploit the observation that the classification of high- and low-frequency error components in finite element (FE)-computations is dependent on the resolution of the spatial domain. High-frequency errors on a fine grid are not captured on coarser grids. In contrast low-frequency errors on fine grids may appear high-frequency on coarser ones, whereas smooth error components essentially can be represented on significantly coarser grids.

In order to exploit this insight, multigrid solvers use cheap smoothers, such as damped Jacobi- or Gauss-Seidel-iterations [79], to eliminate high-frequency error components. Repeated application on grid with different spatial resolution then admits to significantly reduce the oscillatory components of the algebraic error. Eventually on the coarsest grid the remaining error can be eliminated by a direct solver.

To express this formally we consider the operator equation $Au = b$ in X^* and a sequence of hierarchical grids, resp. nested FE-spaces

$$\mathrm{S}_0 \subset \cdots \subset \mathrm{S}_j \subset \mathrm{X}$$

with corresponding projected differential operators $A_k = A|_{\mathrm{S}_k}$. We denote the projection operators by $I_{k-1}^k : \mathrm{S}_{k-1} \to \mathrm{S}_k$ and the corresponding restriction operators by $I_k^{k-1} : \mathrm{S}_k \to \mathrm{S}_{k-1}$. The essential ingredients of the multigrid method are captured in Alg. 4.8.

For adequate choices of the algorithmic parameters, ν_1 applications of the smoother eliminate the high-frequency error components in S_k. Then, the remaining error is projected to a coarser grid and eliminated there. Assuming that the remaining error is "smooth" with respect to S_k we expect relatively small loss due to restriction and interpolation operators. After the computation of the coarse grid correction one

Require: given smoothing parameters ν_1, ν_2, grid level k, operator A_k and right hand side b_k, initial value u_k.

mgCycle:
if $k = 0$ **do**
  use direct solver to solve $A_0 \delta u_0 = b_0$
  $u_0 \leftarrow u_0 + \delta u_0$
else do
  apply ν_1 steps of a smoother to the system $A_k u_k = b_k$
  compute the restriction of the residual to the coarse space $r_{k-1} \leftarrow I_k^{k-1} r_k = I_k^{k-1}(b_k - A_k u_k)$
  compute $A_{k-1} = I_{k-1}^k A_k I_k^{k-1}$, set $\delta u_{k-1} = 0$
  apply **mgCycle**$(\nu_1, \nu_2, k-1, A_{k-1}, r_{k-1}, \delta u_{k-1})$
  correct the fine grid solution $u_k \leftarrow u_k + I_{k-1}^k \delta u_{k-1}$
  apply ν_2 steps of a smoother to the system $A_k u_k = b_k$

Algorithm 4.8: Two-grid correction scheme.

further applies a smoother ν_2 times to eliminate possibly remaining high-frequency error contributions. On the coarsest grid a direct solver can be employed to compute a highly accurate coarse grid correction. Alg. 4.9 is essentially the V-cycle multigrid method. For the realization of a linear solver, that can satisfy prescribed accuracy requirements, we repeatedly apply the above algorithm on the defect equation. See Briggs et al. [45], Trottenberg et al. [260] for more details and the W-cycle as well as the full multigrid (FMG) scheme. Same as the Chebyshev semi-iteration the multigrid method provides a linear preconditioner if employed with a fixed number of iterations.

Require: given smoothing parameters ν_1, ν_2, operator A and right hand side b.

while convergence test failed **do**
  set $r \leftarrow b - Au$ and $\delta u = 0$
  apply **mgCycle**$(\nu_1, \nu_2, 0, A, r, \delta u)$
  set $u \leftarrow u + \delta u$

Algorithm 4.9: V-cycle multigrid method.

On optimized academic examples multigrid solvers are extremely fast. This is illustrated in Tab. 4.2 where the iteration numbers are given for a simple example of linear heat transfer on the unit cube, with right hand side $f = 1$. On real-world problems, where already small and coarse geometries may contain larger numbers of degrees of freedoms in a linear finite element space, their performance is less outstanding, but still impressive. This is illustrated in Tab. 4.3 for equations of linear elasticity with the same right hand side. On the left of this table the iterations

numbers and computational time on the unit cube are given. On the right we give the corresponding numbers on the geometry used in Sec. 6.2.2.1. Thus, we will have to apply several V-cycles to get a reasonable preconditioner for our KKT-systems. If used to approximate the differential operators in the constraint preconditioner (4.2.4) we even need to actually solve the constraint and adjoint equation. For these reason the repeated application of multigrid solvers and preconditioners for problems in constraint preconditioners for PPCG methods still is quite expensive.

Laplace (2D, unit cube)			Laplace (3D, unit cube)		
dof	#iter.	time	dof	#iter.	time
545	7	2.2 ms	369	8	2.6 ms
8 321	7	32 ms	2 465	9	21 ms
33 025	7	0.13 s	17 985	10	0.18 s
131 585	7	0.59 s	137 345	10	1.7 s
2 099 201	7	11 s	1 073 409	11	16 s

Table 4.2.: Computation times for simple test problems (rel. acc.: 10^{-9}, smoothing steps: 10).

Linearized elasticity (unit cube)			Linearized elasticity (real world geometry)		
dof	#iter.	time	dof	#iter.	time
53 955	43	3.6 s	213 687	25	18 s
412 035	49	35 s	1 523 604	48	4.4 min
3 220 227	53	5 min	11 461 518	72	49 min

Table 4.3.: Computation times for simple test problems from linearized elasticity (both 3D, rel. acc.: 10^{-9}, smoothing steps: 10, resp. 20 for the second problem)

We note that in the context of pressure-type boundary conditions the lower left and upper right block in the KKT-system not only contain the differential operator $W_{\varphi\varphi}$, but also non-symmetric contributions from the Piola-transformed pressure-type boundary conditions $g\mathrm{cof}(\nabla\varphi)n$. Moreover, due to the polyconvexity of the stored energy function, the corresponding differential operator may not be elliptic. Consequently, in our setting, the application of multigrid solvers and preconditioners is not backed by a solid theoretical basis. However, in our computations they seemed to work well.

4.6. Summary

Let us summarize the step computations within the different settings. For the computations of δn, p^x and δs we can always assume positive definiteness of M on $\ker c'(x)$, and thus unique solvability of the corresponding system. For moderately sized problems, or a low dimensional control space, the solution can be found by direct elimination methods. Otherwise a PPCG method, i.e. a conjugate gradient method combined with a constraint preconditioner, can be used.

The situation is different for tangential steps δt. As $\mathcal{L}_{xx}$ is in general indefinite we have to use one of the modifications from Sec. 4.3, the HCG method, to compute descent directions for the cost functional. The restriction to $\ker c'(x)$ is again incorporated with the help of a constraint preconditioner. For problems of moderate size or low dimensional control space we can reuse the direct factorization which was computed for the determination of the normal step as preconditioner. If this approach is not admissible we will use the same block triangular preconditioner

$$Q_{\mathrm{sc}} = \begin{pmatrix} 0 & 0 & A^T \\ 0 & M_u & -B^T \\ A & -B & \end{pmatrix}$$

as in the computation of the normal step. To efficiently evaluate this preconditioner we have to efficiently solve the state, variational and adjoint equation. For this we replace the inversion of both A and A^T by a multigrid solver. Since the constraint preconditioner has to project onto $\ker c'(x_0)$, and in the absence of further analysis, it is necessary to solve the arising systems $A'(y_0)\delta n_y = r_p + B\delta n_u$ to high accuracy. Relaxing this condition on Q_{sc} is subject to ongoing work. In contrast M_u^{-1} can be replaced by a fixed number of Chebyshev semi-iterations [110, 124, 274], which needs not to be overly accurate. The required spectral bounds for the Chebyshev semi-iterations for the preconditioned matrix $Q_{\mathrm{jac}}M_u$, where Q_{jac} represents one step of the Jacobi iteration, are taken from [273].

For the purpose of error estimation and adaptive mesh refinement in an affine covariant setting, a hierarchical estimator for the error of the primal variables in the Lagrange-Newton step was proposed. Both the block triangular preconditioner and the error indicator require the solution of equations involving the mass matrix resp. the differential operators of the state and adjoint equation. For the first again the Chebyshev semi-iteration is employed, whereas the differential operators are treated with multigrid preconditioners (25 V-cycles and 20 pre- and post-smoothing steps).

5. Mechanical behavior of biological soft tissues

In the last chapters a model for the description of the implant shape design problem and an algorithm for its solution have been presented. It leaves to specify descriptions for the occurring soft tissue types. In literature often models of linearized elasticity are employed. This is on the one hand due to their simple formulation that are straightforward to implement efficiently. On the other hand realistic models are challenging for numerical solvers. Thus, in order to perform relevant computations, we need to understand state-of-the-art models for biological soft tissues.

Compared with industrially manufactured materials, biological soft tissues exhibit a far more complex mechanical behavior. This includes anisotropy [168], thermo- and viscoelastic behavior [220, 280] as well as complex metabolic interactions [99, 135, 136, 139, 147, 255] and self-regulating mechanisms such as tissue growth [212, 253, 288].

Due to the various different observed phenomena, general accurate mathematical descriptions of the mechanical properties of biological soft tissues are not available. Nonetheless in the described setting of polyconvex hyperelasticity reasonable descriptions of the static mechanical behavior have been proposed for many soft tissue types. These descriptions are mainly based on phenomenological continuum models[1].

In most biological soft tissues the mechanical properties are essentially determined through the properties of the extracellular matrix (ECM) between the cells. The ECM itself mainly consists of ground substance, connecting the cells and its chemical processes, and of three different fiber types (elastic, collagen and reticulin fibers).

- *Elastic fibers* mainly consist of elastin surrounded by fibrillin and are responsible for the tensile elastic properties of a tissue, i.e. the ability to return to its initial configuration when being stretched and then released.

[1] In the context of modeling it is often distinguished between three different approaches. *Continuum models* are derived from general considerations on the macroscopic material structure. *Phenomenological models* rely on the fitting of heuristically chosen expressions to experimental data. *Structural models* use probabilistic descriptions of insights in the microscopic material structure for the description of macroscopic properties. We will that state-of-the-art models for biological soft tissues are in general based on a combination of all three of the mentioned strategies. Insight in a tissue's micro-structure and functional properties are extensively used to derive structural models and as guidance for the development of phenomenological models, both incorporating basic requirements from continuum mechanics.

- *Collagen fibers*, often occurring as parallely aligned bundles, can sustain large tensile forces thus giving strength to biological tissues. “Its importance to man may be compared to the importance of steel to our civilization [...]” Fung [99, p. 251].

- *Reticulin fibers* consist of collagen surrounded by a glycoprotein and form nets from fine collagen bundles. Similar to collagen fibers they are responsible for the ability to sustain forces in many tissues such as blood vessels or smooth muscle tissue.

The presence of elastic and collagen fibers is illustrated in Fig. 5.1 for the case of subcutaneous tissue such as found in the papillary dermis, a deeper layer of skin tissue.

Eventually muscles themselves are assembled in a fibrous structure. Fibers largely determine the tensile mechanical behavior of biological soft tissues. Under compression they buckle and do not contribute directly to the mechanical response [136]. But they contribute indirectly, due to their influence on the permeability of liquids. This indirect relation is usually neglected in biomechanical models.

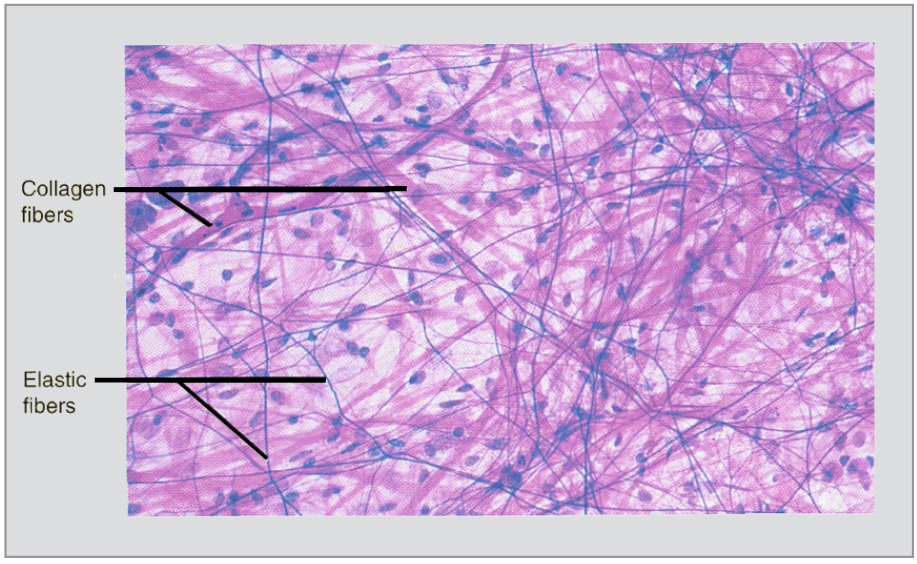


Figure 5.1.: Fiber structure of subcutaneous tissue.

We begin with the introduction of general strategies for the description of constitutive relations in the hyperelastic setting and the commonly used setting for fiber-reinforced materials in Sec. 5.1. Then, specifications for the behavior of different tissue types with respect to tensile (Sec. 5.2) and compressive forces (Sec. 5.3) are discussed. Eventually we shortly address the attainment of patient-specific material parameters in Sec. 5.4.

5.1. Modeling framework

In order to reduce the complexity in the development of constitutive laws for biological soft tissues it is commonly assumed that the polyconvex stored energy function W can be split into independent contributions[2]. Most of these splittings are of the form

$$W = W_{\mathrm{t}} + W_{\mathrm{c}}, \tag{5.1.1}$$

where W_{t} describes the behavior with respect to isochoric (volume-preserving) deformations , and W_{c} the behavior with respect to purely volumetric deformations. The latter is assumed to be isotropic, whereas

$$W_{\mathrm{t}} = W_{\mathrm{t,iso}} + W_{\mathrm{t,aniso}}, \tag{5.1.2}$$

can be split into isotropic and anisotropic contributions $W_{\mathrm{t,iso}}$, resp. $W_{\mathrm{t,aniso}}$. The stored energy function $W_{\mathrm{t,aniso}}$ describes the influence of fibers. Models of this form are called *fiber-reinforced* models. Since fibers buckle under compression they are considered to be relevant only in the presence of tensile forces. The contrary holds for liquids that are dominate the elastic behavior under compressive forces. In biomechanical models, the splitting (5.1.1) often corresponds to different models for the description of tension and compression experiments.

5.1.1. Isotropic materials

Isotropic materials do not contain any directional information on a macroscopic level.

Definition 5.1. A response function $\hat{T}\colon \bar{\Omega} \times \mathbb{M}^3_+ \to \mathbb{S}^3$ is **isotropic** at $x \in \bar{\Omega}$ if it satisfies

$$\hat{T}(x, FQ) = \hat{T}(x, F) \quad \text{for all } F \in \mathbb{M}^3_+ \text{ and all } Q \in \mathbb{O}^3_+.$$

We call an elastic material isotropic if the response function of the corresponding stress tensor is isotropic at every $x \in \bar{\Omega}$.

Isotropy implies that we can express the stress tensor in terms of the strain tensor instead of the deformation gradient.

Theorem 5.2. *A response function $\hat{T}\colon \bar{\Omega} \times \mathbb{M}^3_+ \to \mathbb{S}^3$ is isotropic at $x \in \bar{\Omega}$ if and only if there exists a mapping $\tilde{T}(x, \cdot)\colon \mathbb{S}^3_+ \to \mathbb{S}^3$ such that for all $F \in \mathbb{M}^3_+$ holds*

$$\hat{T}(x, F) = \tilde{T}(x, FF^T).$$

[2] In fact all common splittings do not yield independent contributions. In particular for complex models it is rather unclear how the interaction between the different summands affects the overall model.

Proof. See Ciarlet [56, Thm. 3.4-1]. □

Incorporating also the property of frame-indifference admits to specify the form of stress tensors for isotropic materials. First recall the invariants of a matrix.

Definition 5.3. Let $A \in \mathbb{M}^3$ with eigenvalues $\lambda_i,\ i = 1, 2, 3$. Then its **principal invariants** $\iota_A = (\iota_1, \iota_2, \iota_3)$ are

$$\begin{aligned} \iota_1 &= \operatorname{tr}(A) = \lambda_1 + \lambda_2 + \lambda_3, \\ \iota_2 &= \frac{1}{2}\Big(\operatorname{tr}(A)^2 - \operatorname{tr}(A^2)\Big) \\ &= \operatorname{tr}(\operatorname{cof}(A)) \\ &= \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_1, \\ \iota_3 &= \det(A) = \lambda_1\lambda_2\lambda_3. \end{aligned}$$

Remark 5.4. If $A = F^T F \in \mathbb{S}^3_+$, for $F \in \mathbb{M}^3_+$, is a left Cauchy-Green strain tensor, then its eigenvalues are also called **principal stretches** [201]. The corresponding invariants ι_1, ι_2 and ι_3 are called **principal strain invariants**. Note that in this case we have

$$\iota_3 = \det(A) = \det(F^T F) = \det(F)^2.$$

Thus for some deformation φ and $F = \nabla\varphi$, the third principal strain invariant describes volume changes.

Next to the famous existence result of Ball [15], which is not restricted to isotropic materials, the Rivlin-Ericksen theorem is the most important theorem in the theory of isotropic elasticity.

Theorem 5.5 (Rivlin-Ericksen)**.** *A mapping $\hat{T}\colon \mathbb{M}^3_+ \to \mathbb{S}^3$ satisfies*

$$\hat{T}(QF) = Q\hat{T}(F)Q^T \quad \text{and} \quad \hat{T}(FQ) = \hat{T}(F) \tag{5.1.3}$$

for all $F \in \mathbb{M}^3_+$ and all $Q \in \mathbb{O}^3_+$, if and only if for all $F \in \mathbb{M}^3_+$ holds

$$\hat{T}(F) = \tilde{T}(FF^T),$$

where $\tilde{T} : \mathbb{S}^3_+ \to \mathbb{S}^3$ is of the form

$$\tilde{T}(C) = f_0(\iota_C)I + f_1(\iota_C)C + f_2(\iota_C)C^2,$$

and $f_i,\ i = 1, 2, 3$ are real-valued functions of the principal strain invariants $\iota_C = (\iota_1, \iota_2, \iota_3)$ of the strain tensor $C = F^T F \in \mathbb{S}^3_+$.

Proof. See [56, Thm. 3.6-1]. Note that no regularity assumptions on $\tilde{T}$ are required. Instead symmetry of the strain and stress tensors, the axiom of frame-indifference as well as the assumption of isotropy (5.1.3) yield simultaneous diagonalizability of C and $\tilde{T}(C)$ as essential tool in the proof of the Rivlin-Ericksen theorem. □

The Rivlin-Ericksen theorem tells us that for hyperelastic materials the stress tensor, and consequently the stored energy function W, can be written in terms of the principal strain invariants. Since the invariant ι_3 describes volume changes the splitting(5.1.1) is mostly assumed to be of the form

$$W = W_{\mathrm{t}}(\iota_1, \iota_2) + W_{\mathrm{c}}(\iota_3).$$

However, the first and second principal strain invariant are not isochoric. For this reason first expressions for the influence of the shear parts of the first and second principal invariant on the volumetric part had already been derived in [209]. However, for sake of simplicity, it is commonly assumed that W_{c} only depends on the third strain invariant.

For a proper splitting into isochoric and volumetric contributions Penn [210] introduced a modified set of invariants $\bar{\iota}_C$.

Definition 5.6. Let $A \in \mathbb{M}^3$ with eigenvalues $\lambda_i,\ i = 1, 2, 3$. Then its **modified principal invariants** $\bar{\iota}_A = (\bar{\iota}_1, \bar{\iota}_2, \bar{\iota}_3)$ are

$$\begin{aligned} \bar{\iota}_1 &= \iota_1 \iota_3^{-\frac{1}{3}}, \\ \bar{\iota}_2 &= \iota_2 \iota_3^{-\frac{2}{3}}, \\ \bar{\iota}_3 &= \iota_3. \end{aligned}$$

Remark 5.7. Since we can recover the principal invariants from the modified principal invariants we may also use the latter in the Rivlin-Ericksen theorem.

Penn then investigated models of the form

$$W = W_{\mathrm{inc}}(\bar{\iota}_1, \bar{\iota}_2) + W_{\mathrm{vol}}(\bar{\iota}_3).$$

Assuming constant compressibility, Penn concluded that for large strains the relation between volume changes and stresses is no more physically reasonable and rejected this approach. Similar results in numerical experiments have been presented in [90].

Nonetheless this strategy is appealing as it facilitates the interpretation of the contributions of tensile and compressive responses and its fitting to experimental results, and is widely used [128, 135, 140, 137]. Moreover, premature rejection of the splitting based on modified invariants is challenged by Hartmann and Neff [128]. Stressing the fact that actually the interplay of complex nonlinear models for applied tensile and compressive forces is not well understood, an expression for the volumetric part is proposed that does not admit the previously observed deficiencies, see Sec. 5.3.

The isotropic parts of the models presented in Sec. 5.2 and Sec. 5.3 will all be of the form

$$W = W_{\mathrm{t}}(\iota_1, \iota_2) + W_{\mathrm{c}}(\iota_3) \quad \text{resp.} \quad W = W_{\mathrm{t}}(\bar{\iota}_1, \bar{\iota}_2) + W_{\mathrm{c}}(\bar{\iota}_3),$$

depending of the employed set of invariants. The fact that the model based on the principal invariants does not properly separate isochoric and volumetric contributions is often neglected in the modeling process. In many cases this is not a problem since the subsequent fitting to measurement data may at least partially correct these inaccuracies in the modeling process.

Remark 5.8.

- Another idea was followed in the development of Ogden's models, cf. Ogden [199, 200, 201]. These are expressed in terms of the principal stretches, the eigenvalues of the strain tensor. The most popular Ogden-type models, the neo-Hookean and the Mooney-Rivlin material law, can also be expressed in terms of the principal strain invariants. Models which do not allow a formulation in terms of its invariants require the additional solution of eigenvalue problems and thus are not widely used.
- An interesting alternative approach was followed in Criscione et al. [68, 69]. There, use of the "K-invariants", based on the Hencky, "true", strain tensor $\ln(\nabla\Phi^T\nabla\Phi)$ and orthogonality requirements, was proposed. The use of orthogonal, physically meaningful, invariants admits the construction of constitutive relations that are more robust with respect to measurement errors than formulations based on (modified) principal invariants. Yet further examinations have to reveal merits and shortcomings of this approach.

5.1.2. Fiber-reinforced materials

The presence of fibers leads to anisotropic mechanical behavior of biological soft tissues. In view of the splitting

$$W_{\mathrm{t}} = W_{\mathrm{t,iso}} + W_{\mathrm{t,aniso}},$$

these properties are incorporated into hyperelastic models by augmenting an isotropic model with an anisotropic one.

The standard strategy for the derivation of anisotropic model is *isotropization*. In order to incorporate directional information we need a projection of the strain tensor onto the fiber direction $v \in \mathbb{R}^3$, $|v| = 1$. For $d \in \mathbb{R}^3$ the orthogonal projection on $\mathrm{span}(v)$ is given through

$$v\,(v,d) = v\left(v^T d\right) = (v \otimes v)\,d,$$

where $(\cdot,\cdot)$ denotes the Euclidean scalar product in $\mathbb{R}^3$ and $\otimes$ the Kronecker product. Thus, the projection is characterized by the structural tensor

$$M := v \otimes v. \tag{5.1.4}$$

This tensor must be incorporated in the definition of strain invariants that describe anisotropic materials. This leads to a new set of invariants.

Definition 5.9. Let $A \in \mathbb{M}^3$ and $M \in \mathbb{S}^3$. The corresponding **mixed invariants** are

$$\begin{aligned} \iota_4 &= \operatorname{tr}(AM), \\ \iota_5 &= \operatorname{tr}(A^2M), \\ \iota_6 &= \operatorname{tr}(AM^2). \end{aligned}$$

Remark 5.10.

- For a detailed discussion of these and more matrix invariants in the context of biomechanical modeling we refer to [88] and the references therein.
- Same as for the principal invariants we can define modified mixed invariants via

$$\begin{aligned} \bar{\iota}_4 &= \iota_4 \iota_3^{-\frac{1}{3}}, \\ \bar{\iota}_5 &= \iota_5 \iota_3^{-\frac{2}{3}}, \\ \bar{\iota}_6 &= \iota_6 \iota_3^{-\frac{1}{3}}. \end{aligned}$$

- The definition applies for the more general setting $M \in \mathbb{S}^3$. Here, we are only interested in structural tensors M of the form given in (5.1.4). Then we have

$$M^2 = \left(vv^T\right)\left(vv^T\right) = v \underbrace{\left(v^T v\right)}_{=1} v^T = vv^T = M. \tag{5.1.5}$$

  In particular this implies that $\iota_4 = \iota_6$.

For the description of anisotropic materials we can use the same models as for isotropic materials, just replacing the (modified) principal invariants by (modified) mixed invariants, i.e.

$$W_{\text{aniso}} = W_{\text{aniso}}(\iota_4, \iota_5, \iota_6).$$

In this way we restrict the strain invariants to particular fiber directions. For more complex patterns of the embedded fibers $W_{\text{aniso}} = \sum_{\text{fiber}} W_{\text{aniso,fiber}}$ is composed of independent models for each fiber direction.

The mixed invariants admit geometric interpretations. Considering the deformation gradient $F = \nabla\varphi$, the corresponding strain tensor $C = F^T F$, we get for the first mixed invariant

$$\begin{aligned} \iota_4(C) &= \operatorname{tr}(CM) = \operatorname{tr}(F^T F v v^T) \\ &= \langle F^T F v, v \rangle = \|Fv\|^2. \end{aligned} \tag{5.1.6}$$

Thus, the first mixed invariant ι_4 locally measures the change of fiber length associated with a deformation. Due to its simple interpretation it is the mixed invariant

that is used in most proposed anisotropic constitutive relations that are based on a reinforcement model.

For the second invariant we get

$$\iota_5(C) = \mathrm{tr}(C^2 M) = \langle C^2 v, v\rangle = \|Cv\|^2. \tag{5.1.7}$$

Its interpretation is less obvious. It incorporates the fiber stretch, reaction to shear deformations as well as the deformations of surface area elements normal to the fiber direction, cf. Merodio [185].

Remark 5.11.

- Fiber reinforcements can be further refined. As fibers buckle under compression we may include a linear dependance on $\chi_{\iota_4 \geq 1}$. Additionally fibers are in general not exactly aligned along specific directions. This can be incorporated using statistical descriptions of fiber dispersion [9, 104, 137].
- In our framework the chosen stored energy functions are always required to be polyconvex. Thus they are also rank-one convex, implying that the Legendre-Hadamard condition holds and the corresponding acoustic tensor is elliptic [233].
  If material failure is of interest, loss of ellipticity can be related to the onset of fiber failure or instabilities. If fiber reinforcements only depend on the first mixed invariant ι_4, measuring the change in fiber length, we intuitively may relate the onset of loss of ellipticity with the failure mechanism of fiber kinking [153, 187, 188] if $\iota_4 < 1$, i.e. if fibers are compressed. If $\iota_4 > 1$ the corresponding failure mechanism is de-bonding of fibers. Reinforcements depending on the second mixed invariant ι_5 additionally admit the description of more complex failure mechanisms such as fiber splitting and matrix failure. These failure mechanisms and corresponding conditions on the stored energy functions are discussed in more detail in [186].

5.2. Elastic response with respect to isochoric deformations

Having roughly specified the setting in which most models for biological soft tissues are formulated we now turn to specifications for different soft tissue types. In view of the splitting (5.1.1) we begin with the description of constitutive relations concerning tensile forces. A prominent role is played by proteins, in particular elastin and collagen, which are shortly introduced in Sec. 5.2.1. Largely based on descriptions for the influence of these proteins on the mechanical behavior, recent complex models for adipose, skin and muscle tissue are described in Sec. 5.2.2.

5.2.1. Proteins

Proteins are a subclass of polymers, macromolecules which are built from monomeric amino acids containing carbon, oxygen, hydrogen, nitrogen and possibly also sulfur. Their polymeric structure explains the close relationship between biomechanics and rubber elasticity [123, 147], which is reflected in state-of-the-art models for tissues consisting essentially of one of these proteins as well as the success of models derived from statistical mechanics such as the neo-Hookean [216] and Arruda-Boyce model [13]. They play an essential role in most processes in the human body [255]. In the context of this thesis their structural properties are of main interest and will be discussed for the most relevant proteins, elastin and collagen. More precisely we consider tissues that mainly consist of one of the proteins instead of the proteins themselves.

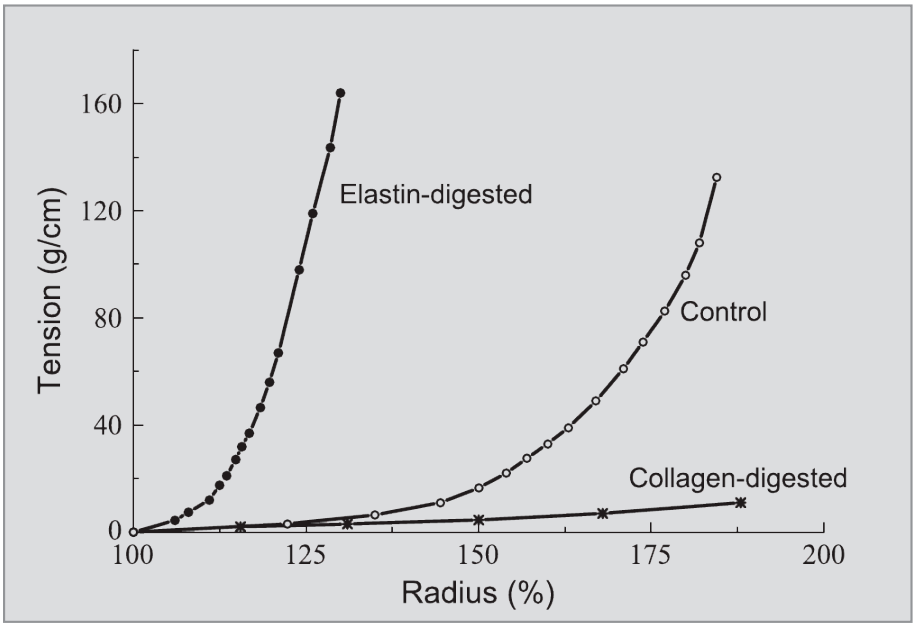


Figure 5.2.1.: Tension-radius responses of human iliac arteries.
control: untreated
collagen-digested: soft tissue behavior dominated by elastin
elastin-digested: soft tissue behavior dominated by collagen

Exemplarily for human soft tissue, a tension-radius curve for human iliac arteries is depicted in Fig. 5.2.1. This curve, here called *control* exhibits a *J-curve* shape, a shape that is observed for most biological soft tissues [137]. Removing the collagen content with formic acid, the *collagen-digested* curve is measured. We observe that in the absence of collagen a roughly linear tension-radius curve with small Young moduli occurs. In contrast, when removing elastin with the help of trypsin, a steeper J-curve is observed. The mechanical behavior of this *elastin-digested* tissue is dominated by the collagen content.

Mathematical descriptions for both cases of *collagen-* and *elastin-digested tissues* will be derived using statistical mechanics and educated guesses based on histological insights. Certainly the development of adequate mathematical homogenization methods [251] would both be challenging and highly desirable for further refinement and justification of biomechanical models. A similar direction is followed in Böl and Reese [38], where the authors examine the derivation of macroscopic descriptions based on finite element studies of the underlying microscopic structure.

5.2.1.1. Elastin

Elastin is the most "linearly" elastic protein occurring in the ECM, particularly in high concentrations in arteries and veins, which must exhibit strong elastic properties in order to smoothen the pulsatory blood flow [112]. Its fundamental building block is the protein tropoelastin, the most elastic of all known proteins, which can be stretched to eight times its resting molecular length and recoils without damage [278]. In human soft tissues it occurs as elastic fibers and is responsible for the elastic behavior, i.e. for the returning of soft tissue to its prior position if deformed and then released. In this function it often undergoes millions of load cycles without showing noteworthy fatigue [112, 118]. From a practical point of view it is responsible for the smoothness of the skin and the elasticity of various parts of the human body such as arteries and lung parenchyma [99]. Compared to other proteins it exhibits the smallest visco-elastic and plastic behavior and merely behaves like a rubbery material. Nonetheless, at smaller strain levels than in collagen, visco-elastic "strain softening", also known as "Mullins effect" [82], must be considered for time dependent soft tissue models and in the interpretation of experimental data. The production of elastin is turned off at puberty. In [112] more effects such as dependency on hydration, the time dependent strain rate (i.e. less elastic behavior at frequencies $f \gg 1$ Hz) and thermal agitation are addressed. In the same work the long lifetime of elastin, whose half-life is around the order of lifespan of the organism [49, 278], is confirmed in fatigue experiments.

The rubbery behavior of elastin fibers motivated the examination of a Mooney-Rivlin [193, 215] and a neo-Hookean [216] material law

$$W_{\text{neoHooke}} = c_0(\iota_1 - 3) = c_0\left(\|\nabla\varphi\|^2 - 3\right) \quad \text{with } c_0 > 0, \tag{5.2.1}$$

by Gundiah et al. [118]. For a successful application of the neo-Hookean material law in the description of the elastin contribution in arteries see also [137, 275]. Motivated by further experiments and histological observations on the fiber structure the orthotropic model

$$W_{\text{orthotropic}} = c_0(\iota_1 - 3) + c_1(\iota_4 - 1)^2 + c_2(\iota_6 - 1)^2 \quad \text{with } c_i > 0 \text{ for } i = 0, 1, 2,$$

was suggested in Gundiah et al. [119]. Recall that for structural tensors associated with fiber directions the invariants ι_4 and ι_6 coincide. However, for consistency with

the cited literature we use both invariants. In the above model these invariants are assumed to refer to different fiber directions. This admits to model complex two-dimensional patterns of elastic fibers.

A special case of this material law,

$$W_{\text{transverse}} = \frac{\mu_0}{2}(\iota_1 - 3) + c_1(\iota_6 - 1)^2 \quad \text{with } \mu_0 > 0 \text{ and } c_1 > 0,$$

was used by Weisbecker et al. [276] in experiments on the human thoracic aorta, confirming that the accurate description of the non-collageneous matrix requires anisotropic stress-strain relationships. In the case of the aorta, the elastin fibers are assumed to be aligned in circumferential direction thus leading to a transversely isotropic stress-strain relationships.

5.2.1.2. Collagen

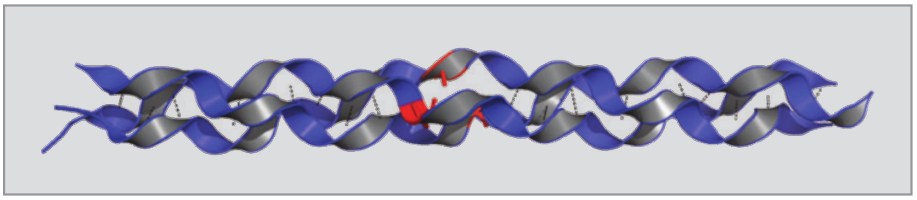


Figure 5.2.2.: Collagen triple helix.

Applied loads in biological soft tissues are mainly carried by collagen. It is the protein with the biggest weight contribution in mammalian soft tissues and the main load carrying element in most tissues. Therefore collagen and its distribution in the ECM play a prominent role in biomechanics. Collagen molecules are organized in triple-helical conformations (see Fig. 5.2.2), the tropocollagen. Most types of collagen assemble their molecules in cross-striated fibrils with diameters in the range $20 - 40nm$. Bundles of these fibrils form collagen fibers with a thickness between $0.2\mu m$ and $12\mu m$. In contrast to elastin, these fibers contain crystalline regions, leading to an increase in the fibers ability to sustain tensile forces and more pronounced anisotropic behavior. For sake of shortness, we do not distinguish the different types of collagen and other attained structures. Instead we refer the interested reader to Fung [99], Holzapfel [137] as well as Shoulders and Raines [237].

Mostly collagen fibers are arranged in wavy-like structures of different complexity. Tissues that take up unidirectional tensile forces, such as tendons or joint ligaments, exhibit roughly parallely arranged fibers, which are less regular in joint ligaments. The deviation from perfectly regular structures has an important influence on the materials stress-strain relationship [137, 267]. More complex fiber structures are observed in the skin, with a complex three-dimensional network with predominant fiber directions in a rhombic structure parallel to the surface, similar to the structure in blood vessels [99, 137, 138, 235].

Occurring in almost pure form in rat tail tendons, constitutive relations for pure collagen have been proposed in [129, 241]. However, the rigorous derivation of a macroscopic material law from the behavior of the collagen fibers is difficult and could, to date, not be realized. Instead, recognizing that the collageneous fibers are much stiffer than elastic fibers and the ground substance, the macroscopic behavior of collageneous tissues is explained by the partial recruitment of collagen fibers. This is illustrated in Fig. 5.2.3, distinguishing between three different phases. After the third phase fibers break. This is sometimes referred to as fourth phase. In most cases it is excluded from the modeling process and post-processing or state constraints must guarantee that computed states do not enter this phase.

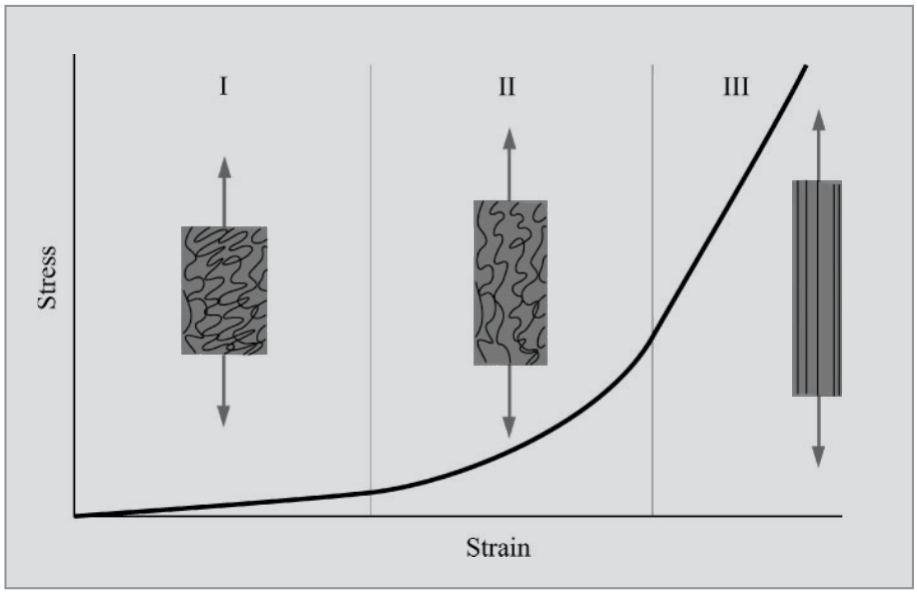


Figure 5.2.3.: Schematic drawing of a tensile stress-strain curve and the associated collagen fiber morphology.

- In phase I the collagen fibers are wavy and crimped. At this stage the mechanical behavior with respect to tensile stresses is dominated by the ground substance and possibly stretching of elastic fibers. The tissue behaves like an almost isotropic rubbery material with low Young moduli. Thus in this phase small applied loads induce relatively large deformations. The stress-strain relationship in this phase is roughly linear.
- In phase II collagen fibers start to contribute to the stress-strain relationship. With increasing load the fibers align along the directions of applied loads. They elongate, begin to uncrimp and take up forces.
- Eventually in phase III the collagen fibers straighten, now dominating the mechanical response of the soft tissue. In this phase the stress-strain relationship is, similar to models for tendon, roughly linear, but with significantly higher Young moduli as in phase I.

Today's state-of-the-art models for tissues that exhibit a "J-curve" stress-strain relationship, as in Fig. 5.2.3, are mostly based on Fung's model, first described in [97]. Starting from the observation that in rabbit mesentery, exposed to tensile stresses, Young's moduli at different stress levels do not deviate much from a linear relationship, Fung proposed the approximations

$$\frac{\partial \sigma}{\partial C} = a\sigma \quad \text{with } a > 0$$

and

$$\frac{\partial \sigma}{\partial C} = a\sigma(1 - b\sigma) \quad \text{with } a > 0 \text{ and } b \in \mathbb{R}.$$

Both lead to a stored energy function involving the exponential function. Extensions of the one dimensional considerations in Fung [97] to fully three dimensional models and different structures of the underlying fiber networks can, amongst others, be found in [22, 23, 99, 137, 138, 140, 141, 276]. These laws, which are called **Fung-elastic**, take the form

$$W_{\text{Fung}}(F) = c(\iota_4, \iota_5, \iota_6) \exp(q(\iota_4, \iota_5, \iota_6)), \tag{5.2.2}$$

for some, often quadratic, function q and structural tensor $M = v \otimes v$ with $|v| = 1$, that describes projections on the fiber orientation [137].

Models of this type are suitable if the collagen fibers are essentially aligned. If significant fiber dispersion is encountered this effect must also be included in the model, mostly in terms of a fiber density ρ_{fiber}, which provides a weighing between a principal and mixed invariant. This strategy has been pursued in Gasser et al. [104], where the formal dependencies are expressed as

$$W_{\text{Fung}}(F) = c(\rho_{\text{fiber}}, \iota_1, \dots, \iota_6) \exp(q(\rho_{\text{fiber}}, \iota_1, \dots, \iota_6)). \tag{5.2.3}$$

As in discussed in Sec. 5.1 we may replace the principal and mixed invariants by its modified counterparts.

Remark 5.12.

- The patient-specific determination of the fiber density and other material parameters is challenging. First attempts for the determination of the parameters for the model proposed in [104] from in vitro image data have recently been investigated by Annaidh et al. [9].
- As they do not satisfy a polynomial growth condition, Fung-elastic material laws seem not to fit to classical operator theory in Sobolev spaces. Similar material laws are one of the reasons for the investigation of operator equations and Galerkin schemes in Sobolev-Orlicz spaces [113, 165]. The latter admit non-polynomial growth, at the expense of a more involved convergence theory and weak limits that may no more be contained in the search space. However, in practice reasonable Fung-elastic material laws are polyconvex and satisfy the polyconvex coercivity condition of Ball [15]. Thus existence results still follows from Thm. 1.42 and Thm. 1.44.

5.2.2. Human soft tissues

We now turn to the description of the tensile elastic response for the most important soft tissue types in the human face. These are adipose, skin and muscle tissue. Bones are considered as solids and attached tendons as well as ligaments are neglected. We also neglect small irregularities in the tissues such as hair follicles, small hair muscles, nerves or blood vessels. Since the presented proteins play a prominent role in the determination of biological soft tissue's mechanical properties the descriptions presented in Sec. 5.2.1 will often appear as building blocks.

5.2.2.1. Adipose tissue

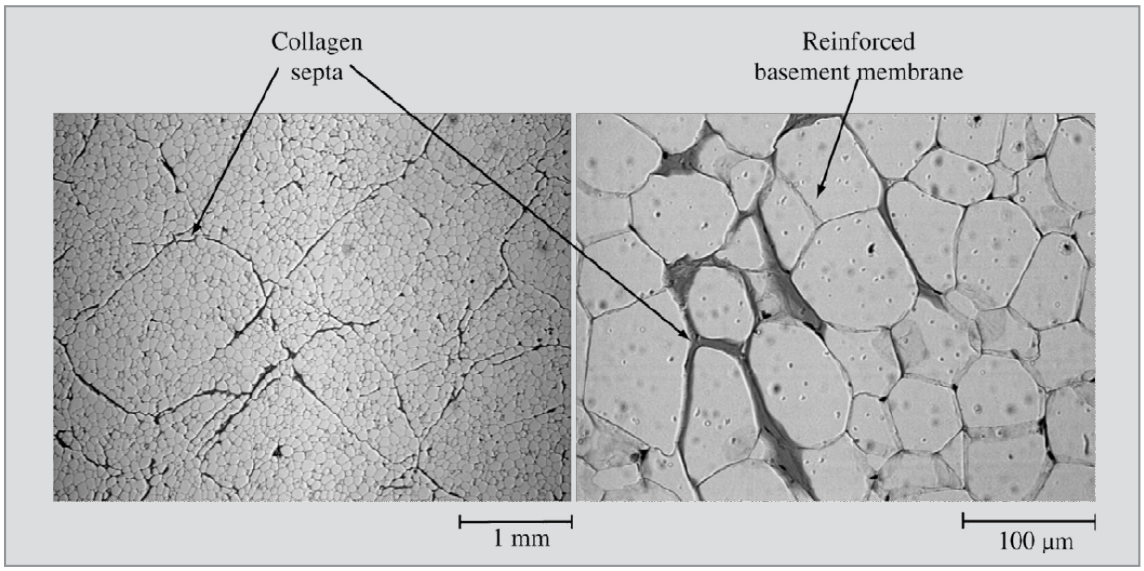


Figure 5.2.4.: Electron micrograph scan of porcine adipose tissue.

Adipose tissue is a loose connective tissue mainly consisting of adipocytes, the fat cells, which are embedded in the extracellular matrix. The ECM contains two collageneous structures, the reinforced basement membrane, a collagen mesh surrounding each adipocyte, and the interlobular septa, a network of long fibrous collagen bundles (see Fig. 5.2.4 and Fig. 5.2.5).

Despite its relevance, not only in plastic and reconstructive surgery [94, 167, 240], but also in forensics [145] and the understanding of the consequences of obesity [134, 263], the development of biomechanical models for (white) adipose tissue only started recently [6, 62, 63, 106, 240][3]. Bi- and triaxial experiments in Sommer et al. [240]

[3]Different types of adipose tissue are distinguished. Heat producing brown adipose tissue is mainly encountered in hibernating animals and mammal foetuses [255] and will not be considered here. Human adults mainly possess white adipose tissue with different structures depending on its desired functionality. As omental adipose tissue, which encloses the organs, it seems mainly

revealed the anisotropic behavior of fat tissue. Thus, strongly exploiting previous insights from the study of other soft tissues, mainly the arterial wall, a non-trivial anisotropic model has been proposed in [240].

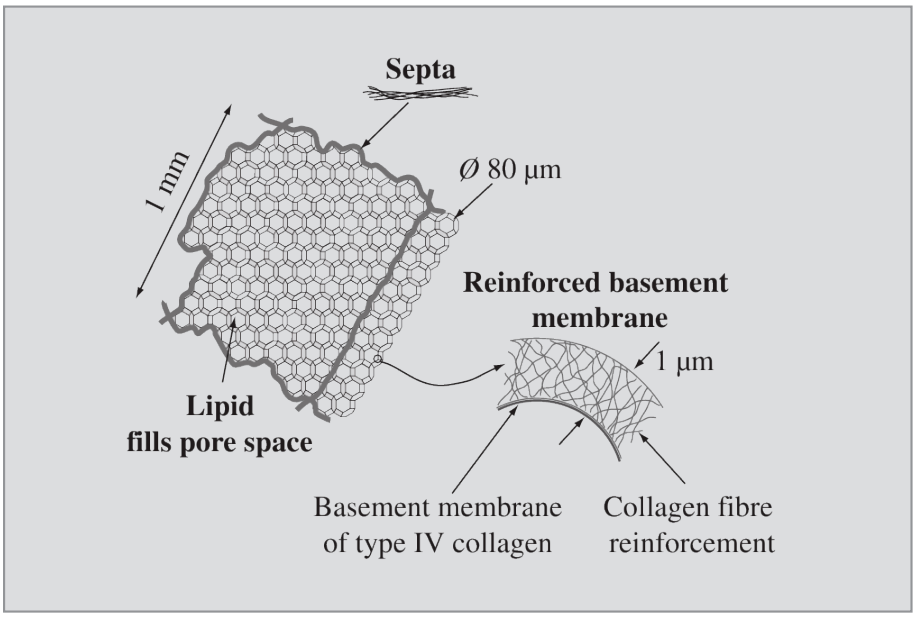


Figure 5.2.5.: Sketch of a lobule of adipose tissue.

Considering the assembly of adipocytes and its enclosing reinforced basement membrane as closed cell foam [63, 167, 240] a reasonable model for its contribution is given by a neo-Hookean material law

$$W_{\text{cells}} = \frac{c}{2}(\iota_1 - 3) \quad \text{with } c > 0.$$

The interlobular septa with its oriented collagen bundles can be described in the form of a Fung-elastic material via

$$W_{\text{septa}} = \frac{k_1}{k_2}\Big(\exp(k_2[\kappa\iota_1 + (1-3\kappa)\iota_4 - 1]^2) - 1\Big),$$

where k_1 is a stress-like parameter, k_2 is a dimensionless parameter and $0 \leq \kappa \leq \frac{1}{3}$ describes the fiber dispersion, cf. [240]. The mechanical response of adipose tissue with respect to tensile forces thus is given by

$$W_{\text{adipose}} = W_{\text{cells}} + W_{\text{septa}}.$$

As most of the volume is occupied by the adipocytes, see Fig. 5.2.4, water inside fat tissue can only move very slowly. Therefore it exhibits only very small compressibility and often is modeled using incompressible material laws [240].

to be responsible for the protection of the inner organs [6]. In contrast, subcutaneous adipose tissue is located under the skin and additionally takes the task of fat storage and acts as thermal insulator. Eventually, white adipose tissue plays a "wide-ranging role in metabolic regulation and physiological homeostasis, far beyond the simple paradigm of fat storage" [256] (see also [160, 257, 263, 272]).

5.2.2.2. Skin tissue

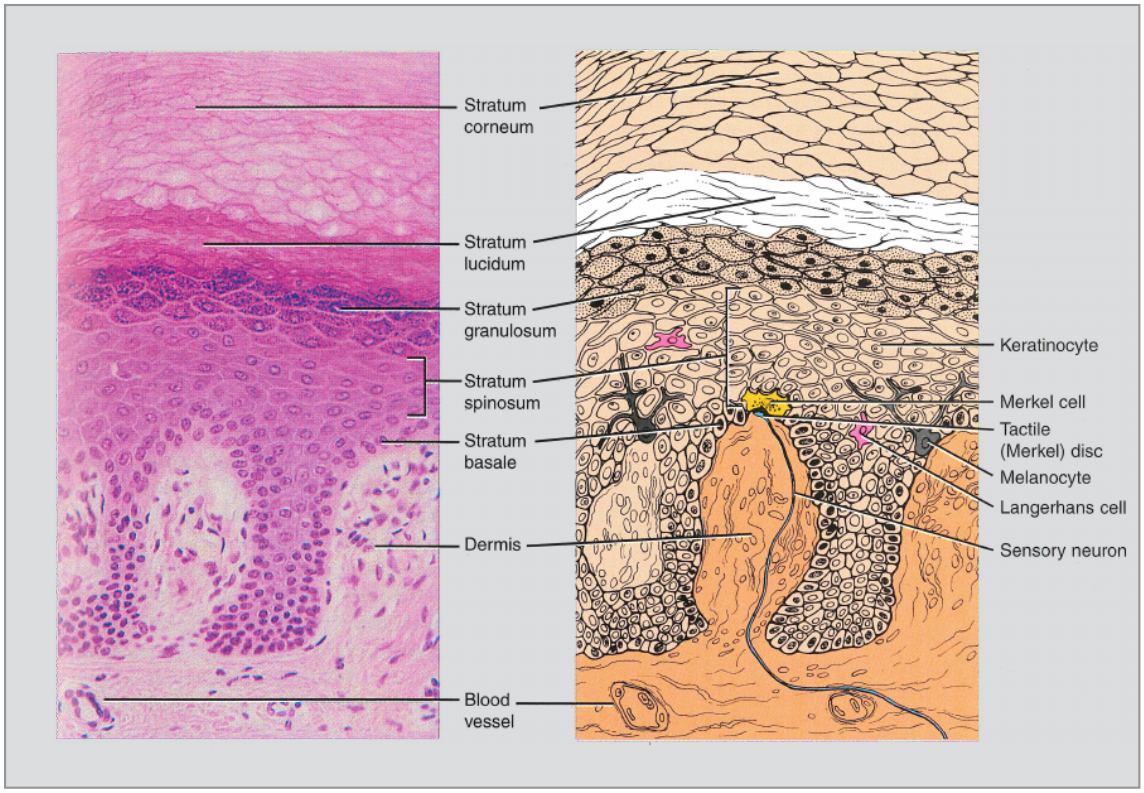


Figure 5.2.6.: Photomicrograph and diagram of a portion of the skin.

Human skin is a complex multi-layered soft tissue, cf. Fig. 5.2.6. The outer layers are comprised in the epidermis. Its main constituent are hydrophobic cells that contain an abundant amount of keratin, the keratinocytes. In direct contact with the environment is the stratum corneum. It protects the body from environmental influences such as heat, physical injuries, microbes or ultraviolet light [174]. Largely consisting of dead keratinocytes, it is relatively stiff and admits only small extension [105]. It occurs wrinkled, unfolds under tensile forces and only plays a minor role in the mechanical properties of skin tissue [105, 238]. Below, we may find the stratum lucidum. It only exists in the thick skin at the soles of our feet, the fingertips and the palm of hands and consists of some layers of flat, dead keratinocytes. The transition between living and dead keratinocytes happens in the following stratum granulosum. Expansibility of epidermis is mainly attributed to the stratum spinosum. Eventually the deepest layer of the epidermis is the stratum basale. Its most popular task is the production of new keratinocytes for the other four layers [255].

The second large part of the skin is the dermis. Forces exerted on the epidermis are transferred via the dermal-epidermal junction to the papillary dermis, the outer layer of the dermis. The adherence between the layers is enforced by cones reaching into

the papillary dermis, cf. Fig. 5.2.6. The latter contains a small and loose distribution of collagen and elastin fibers which are mainly oriented perpendicular to the dermal-epidermal junction. It connects the latter with the reticular dermis, the main load carrying constituent of the skin. It can be extended by roughly 25% under tensile forces [105], and due to its capacity to displace its ground substance it can be strongly compressed. Observed anisotropic properties of mammal skin tissue [10, 89, 238, 243] are attributed to a complex three dimensional network in this layer, mainly consisting of collagen [32] and partly of elastin fibers [205].

To the knowledge of the author a fiber reinforced model for skin tissue has not yet been published. Several others have been proposed, mainly based on phenomenological (membrane-)models. For a review of suggested models the interested reader is referred to [281].

One of the more recent models, suggested in Bischoff et al. [36], is based on statistical mechanics. It is similar to the neo-Hookean model which we use for isotropic contributions of elastin and adipocytes. Aiming at the incorporation of material stiffening at large strains, the Gaussian statistics for the neo-Hookean model are replaced by an 8-chain model using Langevin chain statistics. This yields the isotropic Arruda-Boyce model [13]

$$W_{\text{Arruda-Boyce}} = Nk\Theta\sqrt{n}\Big(\beta_{\text{chain}}\lambda_{\text{chain}} + \sqrt{n}\ln\Big(\frac{\beta_{\text{chain}}}{\sinh(\beta_{\text{chain}})}\Big)\Big), \tag{5.2.4}$$

$$\beta_{\text{chain}} = l^{-1}\Big(\frac{\lambda_{\text{chain}}}{\sqrt{n}}\Big),\ \lambda_{\text{chain}} = \frac{1}{\sqrt{3}}\sqrt{\iota_1}, \tag{5.2.5}$$

where, according to the reinterpretation in [36], N is the collagen's free fiber length, n the collagen fiber density, k is Boltzmann's constant, $\Theta = 298\,K$ is the absolute environmental temperature and $l(x) = \coth(x) - \frac{1}{x}$ is the Langevin function. An advantage of this model is the implicit incorporation of exponential stiffening of collagen fibers within the inverse of the Langevin function l^{-1}. With this approach displacement-stress curves for various experimental results were reproduced, including in vitro experimental data on rat tissue in [32] as well as scar tissue and data from in vivo experiments [36]. A similar approach is used in Tepole et al. [253] for the investigation of models for skin expansion and strain-triggered tissue growth.

The investigations in the PhD-thesis of Hendriks [132] suggest that the in vivo treatment of the skin as one layer is not sufficient to adequately reproduce experimental data. Instead a model with two layers is suggested, one for the reticular dermis and one comprising the papillary dermis, the dermal-epidermal junction and the epidermis. For small strains the author determines Young's moduli for both layers, which is reported to be roughly of a factor of 10^3 bigger in the reticular dermis than in the comprised remaining layers. However, two-layer models require fine spatial resolutions. Since already single-layer volumetric models of the skin lead to a very large problems [26, 183], these more accurate two-layer models are challenging to incorporate efficiently in finite element computations.

5.2.2.3. Muscle tissue

Muscle tissue differs significantly from other biological soft tissues due to its ability to generate stresses by contraction and the induced changes in the tissues geometry and mechanical properties [249]. Three different muscle types are distinguished.

- *Smooth muscles* occur in various parts of the body, such as the skin or blood vessels, and can not be controlled voluntarily [99, Chap. 11]. In mechanical models for skin tissue these are in general neglected [36, 105].
- In contrast, *striated skeletal muscles* can be controlled voluntarily but also may contract unvoluntarily. Occupying a major part of animal bodies, they are "the prime mover of animal locomotion" [99, p. 392]. They can be stimulated at high frequencies in which case they retain maximal tension. This state is referred to as *tetanized*. The understanding of their passive and active behavior is of great interest, amongst others in sports and medical sciences [39, 75, 101, 103, 107, 236].
- A special case of striated muscles are *heart muscles* [99, Chap. 10]. Due to their specific task they can neither tetanize nor be controlled voluntarily.

Here, we focus on striated skeletal muscles. For details about cardiac and smooth muscles we refer the interested reader to Fung [99]. Skeletal muscles, with focus on medical relevance, are also nicely introduced by Oatis [198].

As depicted in Fig. 5.2.7, skeletal muscle is made of aligned bundles of muscle fibers, the muscle fasciculi. Each of these fibers consists of several hundred up to several thousand myofibrils. These are organized in repeated structures, the sarcomeres. Adjacent sarcomeres as well as myofibrils are connected by the zwischenscheibe (Z-disc). The contraction of muscles is realized by the proteins actin and myosin which are located in the sarcomere. Actin filaments are directly attached to the Z-disc, whereas myosin filaments are located in the center of the sarcomere. The myosin filaments are held in place by titin, one of the largest proteins in the human body [125]. During muscle contraction cross-bridges between myosin and actin filaments displace the actin filaments toward the center of the sarcomere [111, 143, 258, 284, 285]. This is illustrated in Fig. 5.2.8. In this thesis, we focus on static models. Regarding the active behavior of muscles the interested reader is referred to the recent publications [121, 130, 217, 218].

The contributors to the passive elastic behavior of skeletal muscles are not yet fully revealed. It is assumed that stretching of cross-links, stretching of proteins and the deformation of the connective tissue [101] as well as active contractibility of intramuscular connective tissue [227] play important roles in the transmission of forces, cf. [146, 192]. The presence of muscle fibers suggests that a transversely isotropic description with fiber direction perpendicular to the plane of symmetry is promising. This assumption is confirmed in [195]. In [40] a relatively simple incompressible and viscoelastic Ogden-type material law has been applied to describe the passive mechanical properties of skeletal muscle under compression transversely

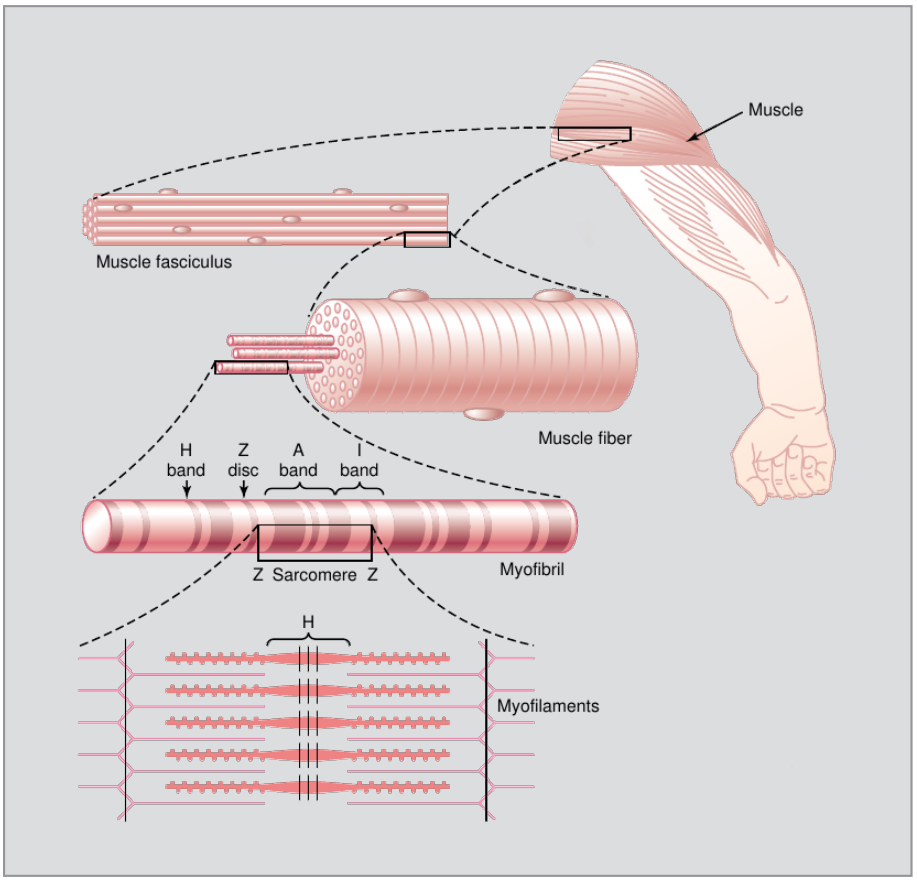


Figure 5.2.7.: Structure of skeletal muscles.

to the muscle fiber directions. Full transversely isotropic models for muscle tissue have already been used for the modeling of the tibialis anterior muscle of a rat in [107] and for the modeling of passive cardiac tissue in Humphrey and Yin [148, 149]. Their model has been extended to the description of active and passive components of skeletal muscles in Martins et al. [182]. Using the common decomposition into isotropic and fiber-related contributions

$$W_{\text{muscle}} = W_{\text{iso}} + W_{\text{fiber}},$$

the tensile response of their model is specified through

$$\begin{aligned} W_{\text{iso}} &= c\Big(\exp[b(\bar{\iota}_1 - 3] - 1\Big), \\ W_{\text{fiber}} &= A\Big(\exp[a(\bar{\iota}_5 - 1)^2] - 1\Big), \end{aligned}$$

with modified strain invariants as defined in Sec. 5.1. In our numerical experiments we will use a similar model. Only the volumetric model $W_{\text{vol}} = \frac{1}{D}(\sqrt{\iota_3} - 1)^2$ will be replaced by a more reasonable one. Eventually, we mention the successful application of models incorporating the muscle fibers using chain statistics in Böl [37].

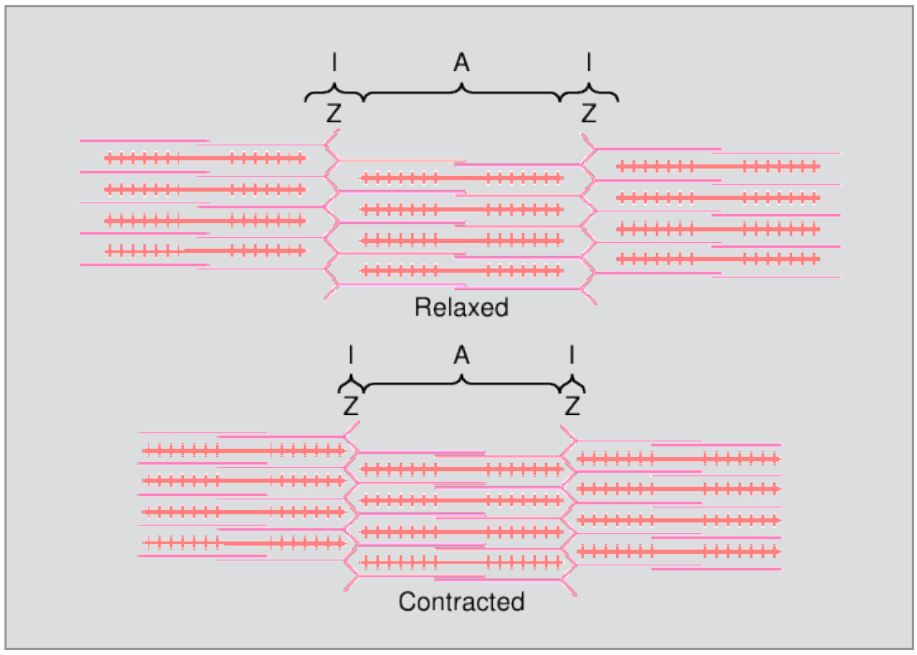


Figure 5.2.8.: Sketch of relaxed and contracted sarcomere.

5.3. Elastic response with respect to volumetric deformations

The models which were presented above for the description of the mechanical reaction with respect to applied tensile forces exploit significant insights into the tissue's micro-structure. In contrast, the description of the reaction to compressive forces is solely based on elementary physical requirements, such as the limit behavior

$$W_{\mathrm{c}}(F) \to \infty \quad \text{for } \det(F) \searrow 0, \text{ resp. } \det(F) \to \infty.$$

Various different models have been proposed for the case that W_{c} only depends on the third principal strain invariant, for an overview cf. Doll and Schweizerhof [84], Hartmann and Neff [128]. Focusing on the forward problem, sometimes even violations of the few requirements on W_{c} are tolerated. For slightly compressible materials the simple quadratic model

$$W_{\mathrm{c}}(\nabla\varphi) = c_0(\det(\nabla\varphi) - 1)^2, \ c_0 > 0 \tag{5.3.1}$$

has been successfully applied [84, 128, 144, 182], despite the fact that it does not satisfy

$$\lim_{\det(F)\searrow 0} W_{\mathrm{c}}(F) \to \infty.$$

This is justified due to relatively big constants c_0 implying $\iota_3 \approx 1$ for reasonable applied forces. In the context of inverse problems we do not know a priori which of the admissible forces will be chosen. Therefore W_{c} must satisfy (1.1.14).

Unfortunately it is not possible to reproduce this limit experimentally to verify the validity of a chosen model[4]. Therefore, from a modeling point of view, beyond the range of experimental verification any derived stored energy function lacks justification. For small values of ι_3, we better interpret W_{c} as barrier function, similar as in interior point methods [197, 224, 225], guaranteeing that iterates stay in the admissible set Φ.

Models that are reasonable in both cases, $\det(\nabla\varphi) \to 0$ and $\det(\nabla\varphi) \to \infty$, are mostly constructed as sum of two models $W_{\mathrm{c},1}, W_{\mathrm{c},2}$ satisfying

$$\lim_{\det(\nabla\varphi)\to 0} W_{\mathrm{c},1}(\nabla\varphi) < \infty \quad \text{and} \quad \lim_{\det(\nabla\varphi)\to\infty} W_{\mathrm{c},1}(\nabla\varphi) = \infty,$$

resp.

$$\lim_{\det(\nabla\varphi)\to 0} W_{\mathrm{c},2}(\nabla\varphi) = \infty \quad \text{and} \quad \lim_{\det(\nabla\varphi)\to\infty} W_{\mathrm{c},2}(\nabla\varphi) < \infty.$$

Again the interaction between both summands is not well understood and motivated rather heuristically.

Nonetheless, there exist some results which help in the choice of models. For the compressive part $W_{\mathrm{c},2}$ a widely used choice is

$$\begin{aligned} W_{\mathrm{c},2}(\nabla\varphi) &= -2c_1 \log(\det(\nabla\varphi)) \\ &= -2c_1 \log\left(\sqrt{\det(\nabla\varphi^T\nabla\varphi)}\right) \\ &= -c_1 \log(\det(C)) = -c_1 \log(\iota_3), \end{aligned}$$

where $C = C(\nabla\varphi) = \nabla\varphi^T\nabla\varphi$ is the Cauchy-Green strain tensor, ι_3 the third principal strain invariant and $c_1 > 0$. As has been shown in [56, 58] this allows for

$$W_{\mathrm{c},1}(\nabla\varphi) = c_0(\det(\nabla\varphi)^2 - 1) = c_0(\iota_3 - 1)$$

to choose constants $c_0 > 0$ and $c_1 > 0$ such that for $\varphi \to \mathrm{id}$ an Ogden-type material law approaches the description of the (constitutively) linearized theory of elasticity. Moreover the function

$$\mathbb{S}^n_+ \ni C \mapsto -\log(\det(C)) \in \mathbb{R}$$

is convex [248]. Thus the above choice of $W_{\mathrm{c},2}$ guarantees monotonicity of the volumetric stress-strain relationship, if considered independently of other models. Recently this has been generalized.

Theorem 5.13. *Let $f \in \mathrm{C}^2(\mathbb{R}_+)$. Then the function*

$$f \circ \det : \mathbb{S}^n_+ \to \mathbb{R},\ A \mapsto f(\det(A))$$

is convex if and only if

$$f''(s) + \frac{n-1}{ns} f'(s) \geq 0 \quad \textit{and} \quad f'(s) \leq 0 \quad \textit{for all } s \in \mathbb{R}_+. \tag{5.3.2}$$

[4]The lack of deeper insights in the development of models for the compressive part is emphasized by the fact that all proposed models, regardless whether they satisfy the elementary requirements from Sec. 1.1.4 or not, can be nicely fitted to experimental data.

Proof. See [172]. □

Besides the logarithmic example given above, this is satisfied for all functions $f(s) = s^{-k},\ k > 0$. Thus, also for the second summand of the model proposed in [128],

$$W_{\mathrm{c}}(\nabla\varphi) = \frac{c_0}{50}(\det(\nabla\varphi)^5 + \det(\nabla\varphi)^{-5} - 2) \quad \text{with } c_0 > 0.$$

Remark 5.14. Note that, as a direct consequence of the reasonable property $W_{\mathrm{c}}'(\mathrm{id}) = 0$, the combination of this model with an Ogden-type material law does not approach the descriptions of linearized elasticity for $\varphi \to \mathrm{id}$, cf. the computations in [56, pp. 186].

5.4. In vivo material parameters

In the first sections of this chapter models for different soft tissue types were introduced. As nice illustrated by the reported measurements in [240], the material parameters for these models vary strongly between individuals. Some of the influencing factors are rather general such as dependencies on the age, gender and the gender-specific differences in the natural aging process [4, 32, 150, 142]. Others are more individual such as nutrition, short- and long-term environmental circumstances as well as changes in metabolism [99, 139]. As a consequence of these effects, which accumulate during life time, the use of "averaged" material parameters severely degrades the quality of computational results.

Instead patient-specific parameters are mandatory, but difficult to attain. The complicated, and at most superficially understood, chemical and mechanical interaction between the constituents of biological soft tissue induces significant differences between *in vivo* and *in vitro* experiments. In [105, 212] skin samples are reported to admit in vitro twice the maximal stretch observed in vivo. Therefore results from in vitro experiments must be handled with care.

In the recent years the development of new experiments for measuring the relevant quantities, such as stresses and strains, in vivo received increased interest. Many soft tissues types can be distinguished by their different optical scattering properties. Thus, measuring the interference patterns of scattered light (speckle tracking), time-dependent strain rates can be measured in vivo up to some millimeters depths with a resolution of 1-10μm [176]. The corresponding imaging technique for the in vivo measurement of stress-strain relationships is optical coherence elastography (OCE) [159, 176, 250]. On the one hand this admits non-invasive detection of tissue irregularities such as tumors or, using intravascular (catheter-based) OCE, aneurysms [158, 250]. On the other hand, assuming that reasonable constitutive laws are available, this technique offers the potential to access the corresponding patient-specific in vivo material parameters. Regarding skin tissue this approach has recently been investigated in [175] and, using the material model from [104] for

arterial walls, in [9]. In the latter publication focus was on the determination of fiber directions.

Often in OCE the required stress fields are applied on the skin. Better control of the stress field, and thus the quality of the measured stress-strain relationship, is expected for volume forces. In this regard greatest potential is attributed to acoustomotive OCE which uses acoustic three dimensional radiation forces, cf. [154]. Also promising with respect to experimental setups is the magnetomotive OCE, where magnetic nanoparticles are introduced into the tissue.

Besides the difficulties in measuring in vivo strains or stresses, the reliable determination of these quantities is further complicated by various properties of biological soft tissues which we ignored so far. Any attempt to give a complete list of influencing factors certainly will fail. Instead we only comment on some effects which seem to play major roles and refer the interested reader to [99].

As mentioned earlier, the fluid content of biological soft tissues leads to viscoelastic behavior [99, 106, 220, 238, 252]. One of the phenomena attributed to the viscoelastic nature of soft tissues is the *Mullins effect*, i.e. strain dependent softening of the tissue, and its induced anisotropy [213]. As a consequence of the Mullins effect biological soft tissues exhibit varying stress-strain curves during cyclic loading. On one side there is a hysteresis loop, following different stress-strain curves during loading and unloading. Therefore different material parameters have to be determined for loading and unloading phases. In engineering literature biological soft tissues are also called *pseudo-elastic* to emphasize the point that elastic material laws do not describe a tissue's mechanical behavior. Only under certain fixed conditions these laws are meaningful.

Moreover each loading, resp. unloading, curve varies in cyclic loading. This variation is strongest in the first loading loop and becomes negligible after several loops. In the biomechanics community this issue is addressed by tissue preconditioning, i.e. by monitoring the response to cyclic loads and postponing the measurement of stresses and stretches until the curves repeat.

The presence of residual stresses and prestrain should be taken into account in the development of accurate models [87]. These are believed to be generated by elastin fibers which are stretched during development, whereas collagen and smooth muscle cells, due to their shorter life time, do not contribute to this matter [49, 270]. In arteries the residual stresses seem to be responsible to homogenize the stress distribution in the arterial walls [98], similar to prestrain and residual stresses in pavilion-roof constructions. Considering the influence of prestrain and residual stresses in mitral leaflets, inconsistencies between experimental data, measured ex vivo, and in vivo simulations could be explained in [212]. As aging entails degradation of elastin residual stresses are in general smaller in comparable tissues of older people.

5.5. Summary

Having an idea of the modeling of biological soft tissues we turn to the choice of constitutive relations for our numerical experiments. In this regard we have to keep in mind that chosen models should admit efficient implementations of the stored energy function and its first three derivatives. This is challenging for models based on chain statistics, such as the skin model of Bischoff et al. [36]. Even if the inverse of the Langevin function is replaced by a cheaper approximation, such as proposed in [60, 152], the complex derivatives and possibly non-negligible costs for the computation of the logarithm and the hyperbolic sine do not easily admit efficient implementations, even with the advanced function generation framework used in the numerical experiments in the next chapter. For this reason, we do not consider these models in our experiments and focus on fiber-reinforced models, which provide the basis for most state-of-the-art models.

For the description of muscle tissue a modified version of the model of Martins et al. [182] will be employed. It combines isotropic and anisotropic exponential stiffening. The unphysical quadratic volumetric penalty term is replaced by the model proposed in [128]. Setting

$$j = \det(\nabla\varphi) = \sqrt{\det(C)},$$

we get

$$W_{\text{muscle}} = c\Big(\exp[b(\bar{\iota}_1 - 3] - 1\Big) + d\Big(\exp[a(\bar{\iota}_5 - 1)^2] - 1\Big) + \frac{c_0}{50}(j^5 + j^{-5} - 2).$$

A similar model, proposed by Sommer et al. [240], is the basis for the description of adipose tissue. It consists of a polynomial part for the isotropic contribution of the adipocytes. The collagen septa is modeled in terms of an exponential Fung-type model that incorporates a simple model for fiber dispersion. In contrast to the assumed incompressibility in [240], we again use a compressible model with the same volumetric penalty as in the muscle model:

$$W_{\text{adipose}} = \frac{c}{2}(\iota_1 - 3) + \frac{k_1}{k_2}\Big(\exp(k_2[\kappa\iota_1 + (1 - 3\kappa)\iota_4 - 1]^2) - 1\Big) + \frac{c_0}{50}(j^5 + j^{-5} - 2).$$

Regarding skin tissue, to the knowledge of the author, no fiber-reinforced model has yet been proposed. Having ruled out models based on chain statistics, in our numerical examples we use a simpler model, proposed by Hendriks [132]. Using the same volumetric penalty as for muscle and adipose tissue, it takes the form of a compressible extended Mooney-Rivlin law:

$$W_{\text{skin}} = c_{10}\,(\iota_1 - 3) + c_{01}\,(\iota_1 - 3)\,(\iota_2 - 3) + \frac{c_0}{50}\left(j^5 + j^{-5} - 2\right).$$

The models for muscle and adipose tissue were chosen as they represent state-of-the-art understanding of the mechanical properties of the underlying tissues. Regarding skin tissue, up to date, less involved models were proposed.

Regarding the specification of material parameters we rely on reported data from literature for our numerical experiments. For real patient-specific computations the individual and spatially localized determination of in vivo material parameters is necessary.

6. Numerical Results

We now turn to experiments that demonstrate the behavior of the proposed affine covariant composite step method and the derived model for implant shape design. First we consider a two-dimensional example of control of nonlinear heat transfer on the unit square. This example is used to explain the behavior of our composite step method and to test its robustness.

Then we turn to problems in implant shape design. Obtaining the required patient-specific material parameters and fiber orientations for fiber-reinforced models on complex geometries is difficult. Therefore, on real world geometries, we have to rely on simpler, isotropic models. In this regard, we split the numerical examples for elastic materials into two groups, each containing two examples. First we use state-of-the-art material laws on simple geometries, where we can define initial fiber orientations manually. Then we present two examples on real patient geometries using an isotropic, compressible Mooney-Rivlin model.

In all examples the algorithmic contraction parameters from Sec. 3.3 are

$$\Theta_{\text{acc}} = 0.75,\ \Theta_{\text{aim,x}} = 0.5 \text{ and } \Theta_{\text{aim,n}} = 0.25.$$

Regarding the minimal required accuracy in the computation of tangential step we set $\delta_0 = 0.25$ to capture at least the two most significant binary digits. If not specified otherwise we use the hybrid conjugate gradient method (HCG) for the computation of tangential steps. Truncation is accepted if the relative energy error is smaller than δ_0. All appearing function spaces are discretized using linear finite elements.

There will be a slight deviation from the presented algorithmic setting. Instead of a fixed norm $\|\cdot\|_M$, local norms $\|\cdot\|_{M(x_k)}$ will be employed in each iteration.

The computations were performed with the help of the finite-element toolbox Kaskade 7 [114], which is based on the Dune-framework [27, 28]. For direct solvers we use the factorizations provided by Umfpack [73, 74] and, for multigrid solvers, Mumps [7, 8]. For visualization ParaView [131] was used in the first examples. In the last two examples, which require some more involved post-processing for the visualization of implant shapes, Zibamira [242] was employed.

6.1. Nonlinear heat transfer

In our first example we consider an optimal control problem in two dimensions with distributed control and observation on the unit square $\Omega = [0,1]^2$. We denote the unknown heat distribution by $y \in \mathrm{Y} := \mathrm{W}^{1,2}(\Omega)$ and the control by $u \in \mathrm{U} := \mathrm{L}^2(\Omega)$. The cost functional is a Tikhonov-regularized tracking type functional

$$J(y,u) = \frac{1}{2}\|y - y_{\mathrm{ref}}\|^2_{\mathrm{L}^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{\mathrm{L}^2(\Omega)}. \tag{6.1.1}$$

The constraints are given by a simple nonlinear model of heat transfer, which we consider in its weak formulation

$$c(y,u)v := \int_\Omega \kappa(y)(\nabla y, \nabla v)\,\mathrm{d}\mu - \langle u, v\rangle_{\mathrm{L}^2(\Omega)},$$

for some test function $v \in \mathrm{W}^{1,2}(\Omega)$ and

$$\kappa(y) = c\|y\|^2_{\mathrm{L}^2(\Omega)} + d$$

where $(\cdot,\cdot)$ denotes the Euclidean scalar product. With the nonnegative parameters c and d we can modify the influence of the nonlinear part and the distance to a singular problem. Thus, we can use these parameters to adjust the difficulty of our problem and test the robustness of our algorithm.

The corresponding optimal control problem was analyzed in [55]. There it was shown that $y \in \mathrm{C}(\overline{\Omega})$ for all $u \in \mathrm{L}^2(\Omega)$ which implies boundedness of $\kappa(y)$. Thus our choice of $\mathrm{W}^{1,2}(\Omega)$ as search space for the heat distribution is admissible.

The desired solution, which is illustrated in Fig. 6.1.1, is given by

$$y_{\mathrm{ref}}(x) = 8(1-x_0)x_0(1-x_1)x_1 \quad \text{for } x = (x_0,x_1)^T \in \Omega.$$

The corresponding Lagrangian then is

$$\mathcal{L}(y,u,p) = J(y,u) + c(y,u)p.$$

Recall, that the choice of the scalar product on $\mathrm{Y} \times \mathrm{U}$ influences the direction of the normal step and thus affects the number of required outer iterations of our algorithm. In order to at least partially capture some part of the structure of the Lagrangian, we use local scalar products of the form

$$\langle (z_0,v_0),(z_1,v_1)\rangle_{M(y_k)} = \langle z_0, z_1\rangle_{M_y(y_k)} + \langle v_0, v_1\rangle_{M_u}$$

with

$$\langle z_0, z_1\rangle_{M_y(y_k)} = \int_\Omega \kappa(y_k)(\nabla z_0, \nabla z_1)\,\mathrm{d}\mu + \langle z_0, z_1\rangle_{\mathrm{L}^2(\Omega)},$$

and

$$\langle v_0, v_1\rangle_{M_u} = \alpha\langle v_0, v_1\rangle_{\mathrm{L}^2(\Omega)}.$$

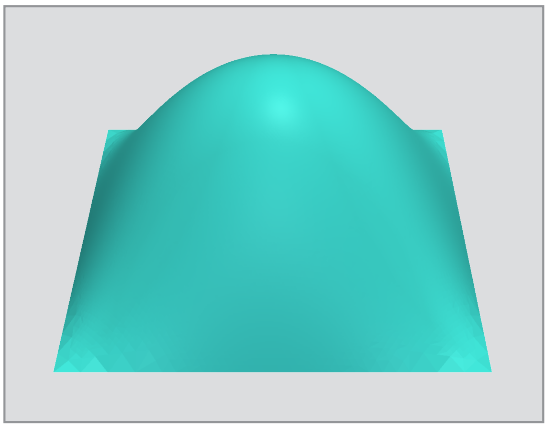


Figure 6.1.1.: Reference solution.

Since $\mathcal{L}_{uu}(y,u,p)v_0v_1 = \langle v_0, v_1\rangle_{M_u}$, application of the this inner product as preconditioner renders the PPCG-method independent of the Tikhonov regularization parameter α. Using the same discretization for state, control and adjoint state, with n degrees of freedoms for each, we get from Thm. 4.2 that at least $2n$ of the $3n$ eigenvalues of the preconditioned matrix for the tangential step cluster at 1.

For this example, and the following ones on error estimation, a direct factorization was used for the computation of (simplified) normal steps, adjoint corrections and in preconditioning the system that determines the tangential step.

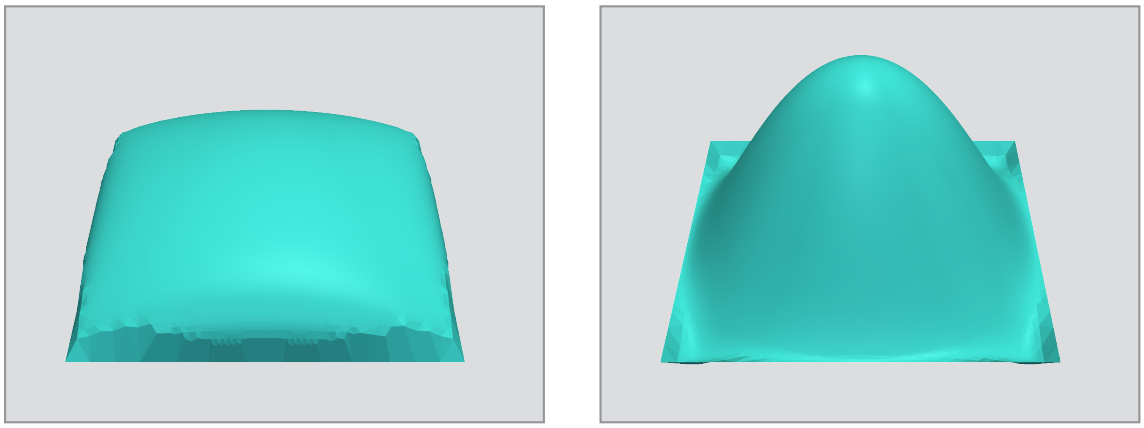
Alg.	HCG						RCG					
d \ c	1	10^1	10^2	10^3	10^4	10^5	1	10^1	10^2	10^3	10^4	10^5
10^{-5}	5	6	17	20	14	16	5	9	22	19	28	24
10^{-4}	5	6	13	28	22	12	5	8	22	57	17	14
10^{-3}	4	6	17	23	17	16	5	8	15	24	15	15
10^{-2}	4	6	13	15	17	19	6	8	14	20	17	18
10^{-1}	4	6	10	19	21	19	6	8	13	24	20	20
1	5	6	9	14	23	18	7	8	11	14	16	17

Table 6.1.: Required iterations for different parameters c and d on a fixed uniform grid with $h_{\max} = 2^{-7}$, and $\alpha = 10^{-6}$.

In Tab. 6.1 iteration numbers for various choices of the model parameters are given for $\alpha = 10^{-6}$ and both the regularized and the hybrid conjugate gradient method. Iteration numbers differ slightly from the ones of Tab. 4.1. This is mainly due to the different minimal required accuracy in the computations of the tangential steps, which is here given by $\delta_0 = 0.25$, whereas, in order to highlight the behavior of the

different conjugate gradient methods in the presence of nonconvexities, $\delta_0 = 10^{-3}$ was chosen in the computations for Tab. 4.1.

We observe that for the simplest problems ($c = 1,\ d \leq 10^{-4}$) both methods do not differ. In this case our algorithm does not encounter nonconvexities and all conjugate gradient methods for nonconvex problems coincide. For slightly more complicated problems ($c \leq 10$) truncation reduces the number of required iterations of the outer loop. In average this stays true if we further increase the parameter c that controls the nonlinear part. However, there also exist cases where RCG performs a bit better than HCG, such as for $c = 10^4$ and $d = 10^{-4}$ or $d = 1$.



(a) Computed state. **(b)** Computed control.

Figure 6.1.2.: State and control for an example of control of nonlinear heat transfer.

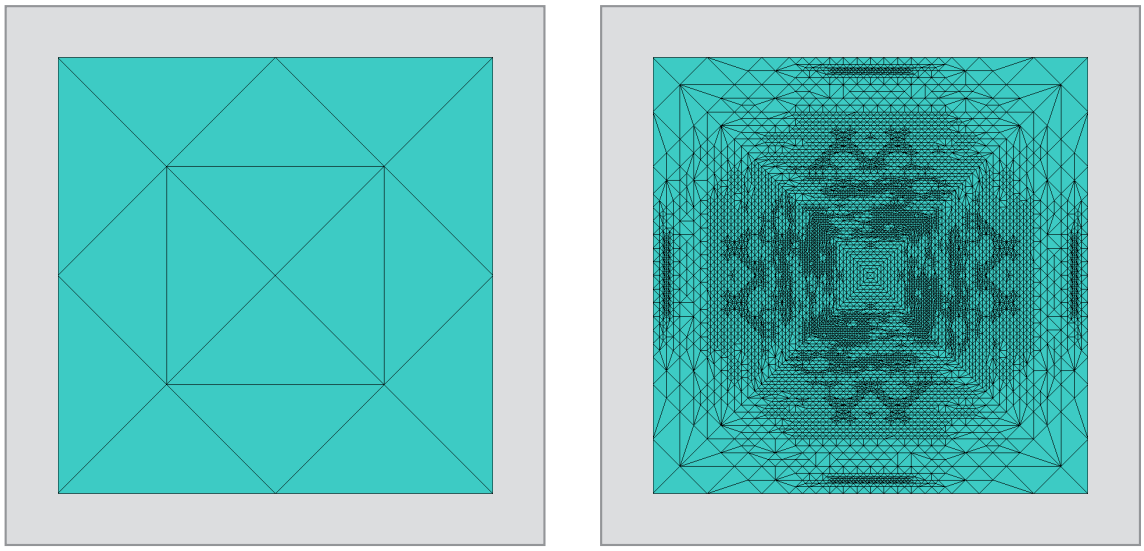
In Fig. 6.1.2 computed control and solution are exemplarily given for

$$c = 10^3,\ d = 10^{-1},\ \alpha = 10^{-6} \text{ and } \varepsilon_{\text{tol}} = 10^{-6}.$$

In Fig. 6.1.3(b) we see an adaptively refined grid, that was generated from the coarse grid in Fig. 6.1.3(a). Since the spatial resolution is far below the printer resolution the fully refined grid is not displayed. As explained in Sec. 4.4, the estimator aims at an accurate resolution of the control variable. This can be observed from Fig. 6.1.2, where we see that the grid refinement leads to discretizations that nicely captures local features of the control variable, whereas the boundary layer of the state variable plays a subordinate role and is not fully resolved.

Algorithmic parameters are plotted in Fig. 6.1.4. Grid refinement, marked by orange triangles, occurs in the iterations 4-6, 8-14, 16-21. Then the maximal number of vertices $n_{\max} = 50\,000$, allowed in the fine grid, is exceeded and no further refinement is allowed. The mesh is refined from 5 vertices (15 degrees of freedom) to 66 623 vertices (199 869 degrees of freedom).

The chosen example is highly nonlinear and several adaptive refinement steps are required to capture the problem structure. On the coarse grid the strong decrease



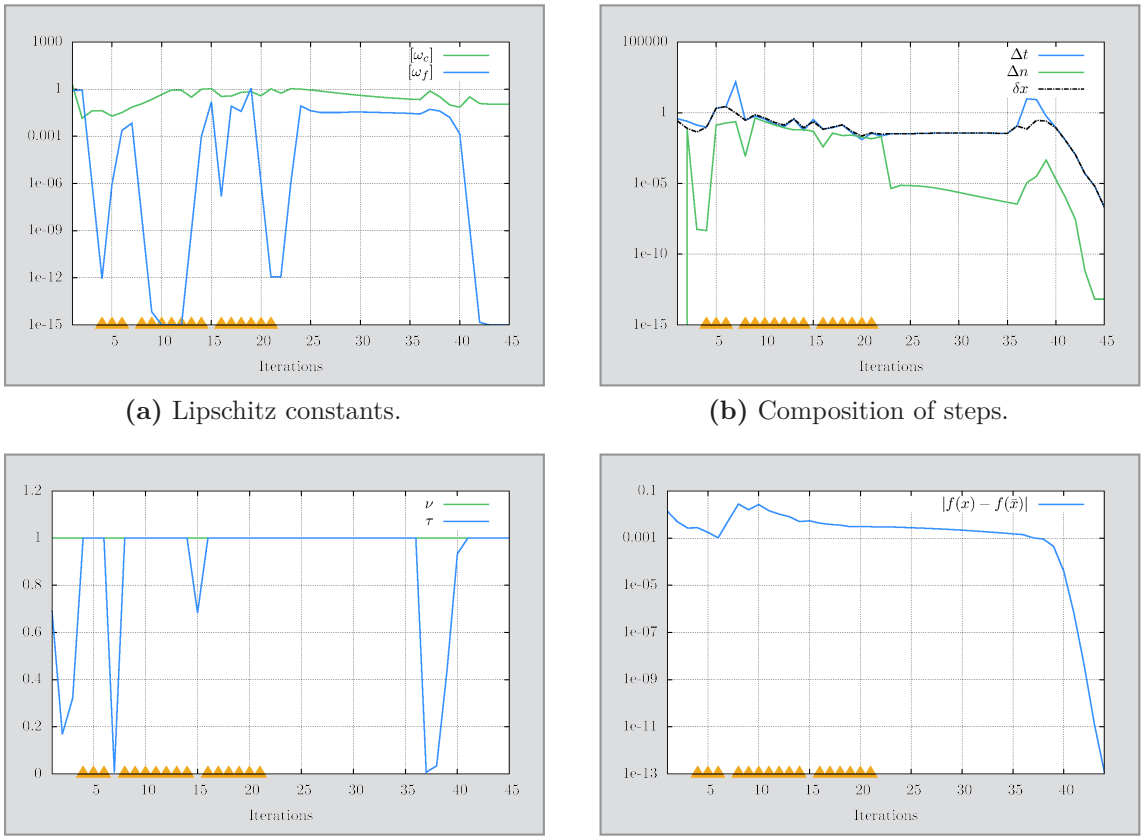
(a) Initial grid. **(b)** Adaptively refined grid (7621 vertices).

Figure 6.1.3.: Initial and refined grid.

of the $[\omega_f]$ indicates that the algorithm makes fast progress to an optimal solution. Following the mesh refinements in the iterations four to six a large tangential direction is computed in the seventh iteration. This is related to both, the newly discovered problem structure on the fine mesh, and unreliability of the computed tangential direction since we are too far from the feasible region. For this reason in this step the tangential step is strongly damped such that feasibility can be restored first and the value of the cost functional $f(x_k) = J(y_k, u_k)$ increases. Also in the following steps both mesh refinement and the normal steps lead to further increase in f. After the mesh refinement ends we observe on the one hand a strong increase in $[\omega_f]$ indicating that we are far from having attained optimality on the fine grid. On the other hand, since feasibility is now violated only due to the presence of tangential steps but no more as a consequence of mesh refinement, we see that there is a sudden drop in the size of the normal steps. The chosen example problem is already relatively complicated and about 20 further steps are required to get sufficiently close to an optimal solution on the fine grid. These last iterations are costly. In contrast the computational costs for the previous steps, that were computed on significantly coarser grids, are negligible.

How challenging this problem is for our algorithm can be estimated from the shape of the curve in Fig. 6.1.4(d)[1]. On the fine mesh the semi-logarithmic plot is concave. Only slow progress towards the optimal function value is made until the region of local convergence is reached in iteration 41.

[1]The plot in Fig. 6.1.4(d) illustrates the absolute difference between the cost function value at the iterates and the final value of the cost functional, the latter serving as an approximation of the optimal value. Therefore, this plot stops one iteration before the last.



(a) Lipschitz constants.

(b) Composition of steps.

(c) Damping factors.

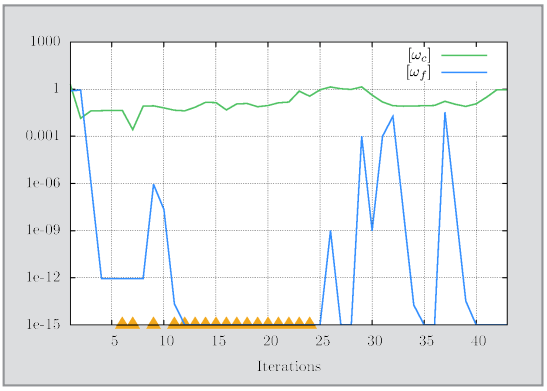
(d) Cost functional.

Figure 6.1.4.: Algorithmic behavior for $c = 1\,000$, $d = 0.1$, $\alpha = 10^{-6}$, $\varepsilon_{\text{tol}} = 10^{-6}$ and computation of the normal step with a direct factorization. Iterations in which refinement occurs are marked by orange triangles.

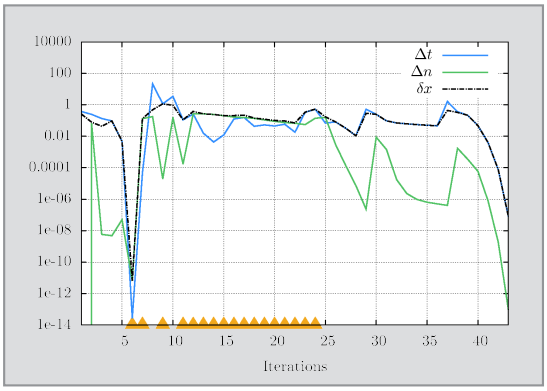
The tangential steps have to be damped at several points, in the first iterations, after mesh refinement and shortly before reaching the region of fast local convergence. In contrast, damping of the normal step is not required. This is in accordance with the reasoning in Chap. 3, where we designed our algorithm such that the computed iterates stay in the Kantorovich region of local contraction of the constraint.

Observe that in the last two iterations the normal steps stagnate at a length of approximately 10^{-13}. This indicates that the maximal attainable accuracy in the computation of the normal step is of the same size and we should not expect to be able to achieve a higher relative accuracy (except for the case of vanishing right hand sides, such as in the first iteration).

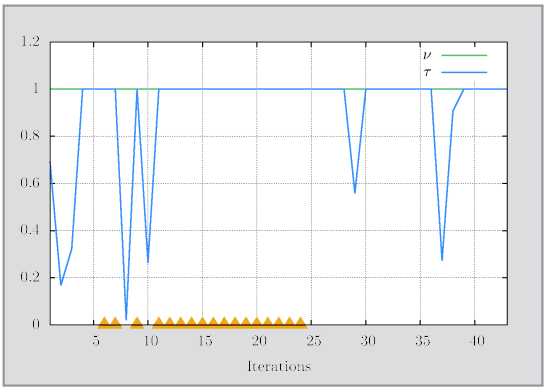
In Fig. 6.1.5 algorithmic parameters are plotted for the same example, but replacing the direct factorization in the computation of the normal step with a PPCG method, as described in Sec. 4.2, where we require a relative accuracy of 10^{-6}. This effects the computation of all parts of the possibly second order corrected composite step. In particular we will apply a block triangular constraint preconditioner, as discussed in



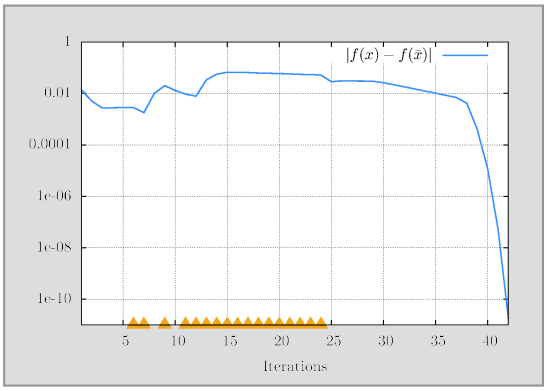
(a) Lipschitz constants.



(b) Composition of steps.



(c) Damping factors.



(d) Cost functional.

Figure 6.1.5.: Algorithmic behavior for $c = 1\,000$, $d = 0.1$, $\alpha = 10^{-6}$, $\varepsilon_{\text{tol}} = 10^{-6}$ and computation of the normal steps with a PPCG method. Iterations in which refinement occurs are marked by orange triangles.

Sec. 4.1 and Sec. 4.6, in the computation of (simplified) normal step, adjoint variable *and* tangential step.

The algorithmic behavior is similar to the case where we use a direct factorization. Most sensitive is the estimate $[\omega_f]$, which starts to increase significantly earlier than in the previous case.

Error estimation

For testing the proposed error estimator simple linear problems are used. We begin with the above introduced example on a square geometry $\Omega = [0, 1]^2$ with distributed observation and control. Let $c = 0$ and $d = 1$, which corresponds to the linear model

of heat transfer. We assume homogeneous Dirichlet boundary conditions on $\partial\Omega$ and consider the problem

$$\min J(y,u)$$
$$\text{subject to} \qquad \int_\Omega (\nabla y, \nabla v)\,\mathrm{d}\mu = \int_\Omega uv\,\mathrm{d}\mu \quad \text{for all } v \in \mathrm{W}^{1,2}(\Omega),$$

with $\alpha = 10^{-6}$ and J as defined above in (6.1.1). We begin with a coarse criss-cross triangulation consisting of four triangles and five vertices, at the corners and the center of Ω.

In Fig. 6.1.6 the estimated error is compared with the real error. Since the latter is not easily available analytically we use the solution on an adequate finer grid. For this we use the adaptively refined grid, refine it three times uniformly and compute the solution on the thus generated grid. In Sec. 4.4.2 it was explained that error estimates are in general not suitable for balancing error contributions. This is confirmed in Fig. 6.1.6. Quantitatively the error estimates differ significantly from the real error. For this example it underestimates the real error by roughly a factor of 10. However, qualitatively it largely captures the discretization error which is efficiently reduced on the refined meshes.

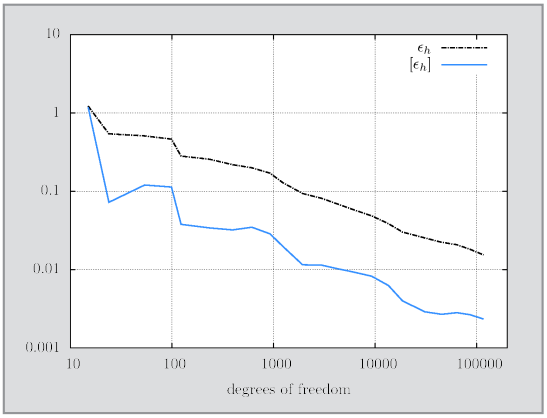


Figure 6.1.6.: Comparison of estimated and true error for a quadratic constrained optimization problem.

To illustrate the influence of corner singularities on the local mesh resolution we consider two examples of boundary control on a T-shaped domain

$$\Omega = \left\{x \in \mathbb{R}^2 :\ |x_0| \le 1 \text{ and } 0 \le x_1 \le 2\right\} \cup \left\{x \in \mathbb{R}^2 :\ |x_0| \le 2 \text{ and } 2 \le x_1 \le 3\right\}.$$

Both examples consider the problem

$$\min J(y,u) = \frac{1}{2}\|y - y_{\mathrm{ref}}\|^2_{\mathrm{L}^2(\Gamma_{\mathrm{o}})} + \frac{\alpha}{2}\|u\|^2_{\mathrm{L}^2(\Gamma_{\mathrm{c}})}$$
$$\text{subject to} \qquad \int_\Omega (\nabla y, \nabla v) + yv\,\mathrm{d}\mu = \int_{\Gamma_{\mathrm{c}}} uv\,\mathrm{ds} \quad \text{for all } v \in \mathrm{W}^{1,2}(\Omega),$$

with $\alpha = 0.01$, but with different observation boundaries. The control boundary is given through

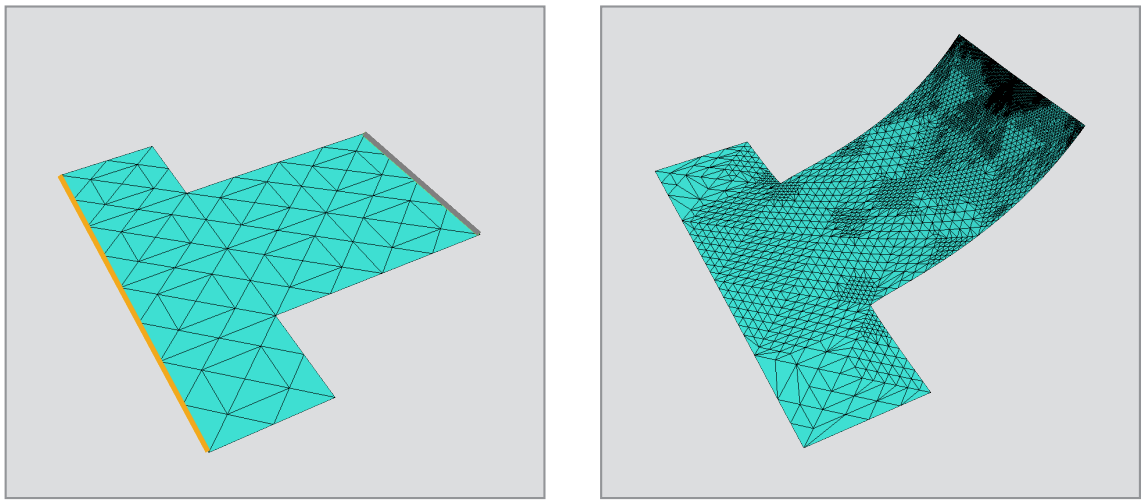
$$\Gamma_c = \left\{ x \in \mathbb{R}^2 : \ x_1 = 0 \right\}.$$

In the first example the observation is located on the opposite side of Ω on

$$\Gamma_o = \left\{ x \in \mathbb{R}^2 : \ x_1 = 3 \right\}.$$

The reference solution is given by

$$y_{\text{ref}}(x) = 1 - |x|^2 \quad \text{for } x \in \Gamma_o.$$



(a) Initial grid (orange: observation boundary, dark gray: control boundary).

(b) Computed state on the adaptively refined grid.

Figure 6.1.7.: Initial grid and computed solution for a problem of boundary control.

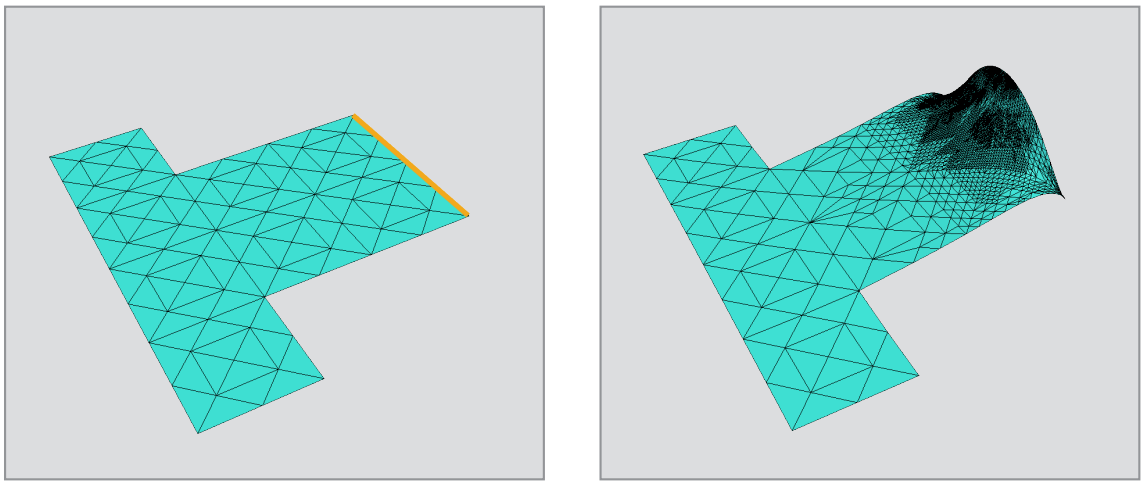
The coarse grid for this problem is given in Fig. 6.1.7(a) and the corresponding state on the adaptively refined grid in Fig. 6.1.7(b). Since the control is located on a part of the boundary, marked in dark gray in Fig. 6.1.7(a), the grid is strongly refined near this control boundary. Increased mesh refinement also occurs near the intruding corners, where lower regularity than in convex domains is expected [116]. Moreover, we observe stronger mesh refinement in triangles whose longest edge is in y-direction. This is a grid-dependent effect that occurs since state and control are almost constant in x-direction.

In the second example the observation boundary coincides with the control boundary, i.e. $\Gamma_o = \Gamma_c$. The reference solution is again chosen to be

$$y_{\text{ref}}(x) = 1 - |x|^2 \quad \text{for } x \in \Gamma_o.$$

The coarse grid for this problem is given in Fig. 6.1.8(a) and the corresponding state on the adaptively refined grid in Fig. 6.1.8(b). Since control and observation are

located on the same part of the boundary, the grid is mainly refined in its vicinity. On the control boundary refinement is strongest at the corners and in the center, the regions where the curvature of control and observation are largest. For the same reasons as in the last example grid dependent effects can be observed. In this example the state is almost vanishing near the intruding corners. Thus no corner singularities occur and there is no need to increase the spatial resolution in their vicinity.



(a) Initial grid (orange: control and observation boundary).

(b) Computed state on the adaptively refined grid.

Figure 6.1.8.: Initial grid and computed solution.

6.2. Examples from biomechanics

Now we present some examples related to the problem of implant shape design. The coarse segmentations, that were provided for real-world geometries, neither contain information on different tissue types nor on directional properties such as fiber directions which are necessary to define anisotropic constitutive relations. Therefore, we provide numerical examples involving state-of-the-art anisotropic material laws only on simple geometries where we can analytically define fiber directions. For geometries attained from medical image data we have to content ourselves with a single isotropic material law.

In all computations we consider a Tikhonov-regularized tracking type cost functional

$$J(\varphi, g) = \frac{1}{2}\|\varphi - \varphi_{\text{ref}}\|^2_{\mathrm{L}^2(\Gamma_{\text{o}})} + \frac{\alpha}{2}\|g\|^2_{\mathrm{L}^2(\Gamma_{\text{c}})}$$

as cost functional and a pressure-type boundary control (see Sec. 2.2), i.e. nonlinear boundary conditions of the form

$$\sigma(\nabla\varphi)n = g\text{cof}(\nabla\varphi)n \quad \text{on } \Gamma_{\text{c}}.$$

The observation is located on a part of the domain's boundary $\Gamma_{\text{o}} \subset \partial\Omega$. Since the corresponding control constraint was never violated in any of the experiments, performed during the work on this thesis, the additional requirement $g \leq 0$ was neglected in the computations.

In particular for complex Fung-type material laws a maximal attainable accuracy of $\varepsilon_{\max} \gg 10^{-16}$ was observed in the solution of the linear saddle point systems. Thus, the maximal attainable accuracy in all computations was set to $\varepsilon_{\max} = 10^{-9}$. Recall that the material laws are expressed in terms of the corresponding strain invariants, i.e. the principal invariants of the Cauchy-Green strain tensor $C(\nabla\varphi) = \nabla\varphi^T\nabla\varphi$, while the unknown in our computations will be the deformation φ.

6.2.1. State-of-the-art material laws on simple geometries

In this subsection we work on the simple geometry

$$\Omega = [-1, 1] \times [-1, 1] \times [-0.25, 0.25].$$

We consider two different combinations of material laws presented in Chap. 5. Requiring not only the stored energy function, but also its first, second and third derivative, the manual implementation of these material laws is highly error prone. For this reason a template-based automatic function generation toolbox has been added to KASKADE 7. With its help we can easily generate material laws of arbitrary complexity as well as their first three derivatives. Using template programming

techniques[2] instead of virtual inheritance or tree structures [262] allows the C++-compiler to generate highly optimized code, thus admitting efficient evaluations in the assembly process.

In the following two examples we use the scalar products

$$\langle v, w \rangle_{M_u} = \int_{\Gamma_c} vw \, \mathrm{ds} \quad \text{in U}$$

and

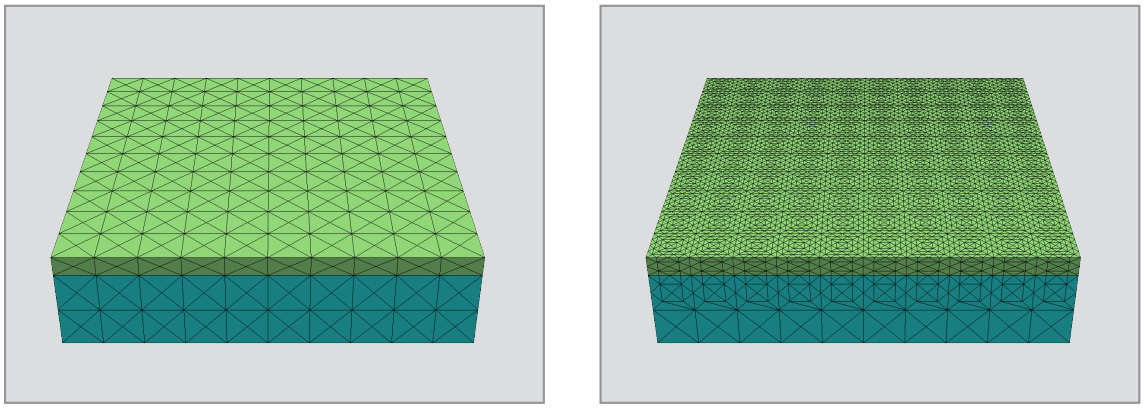
$$\langle v, w \rangle_{M_y} = \int_{\Omega} \nabla v : \nabla w + vw \, \mathrm{dx} \quad \text{in } \Phi.$$

In both examples we require a relative accuracy of $\epsilon_{\text{tol}} = 10^{-3}$. On

$$\Gamma_{\text{d}} = \left\{ x \in \mathbb{R}^3 : \max_{i \in \{0,1\}} |x_i| = 1 \right\}$$

we assume homogeneous Dirichlet boundary conditions.

6.2.1.1. Skin and adipose tissue



(a) Initial grid. **(b)** Adaptively refined grid.

Figure 6.2.1.: Initial and final grid.
Green: skin tissue. Cyan: adipose tissue.

In our first example we combine a model for adipose tissue (Sec. 5.2.2.1) with one for skin tissue (Sec. 5.2.2.2). As illustrated in Fig. 6.2.1, there is a thicker layer of adipose tissue (cyan)

$$\Omega_{\text{adipose}} := \{ x \in \Omega : \ -0.25 \leq x_2 \leq 0.15 \}$$

[2] In particular SFINAE-based techniques (**s**ubstitution **f**ailure **i**s **n**ot **a**n **e**rror) admit to remove computational overhead that can not be eliminated by the compiler, such as addition or multiplication with zeros that are known at compile-time.

and a thinner layer of skin tissue (green)

$$\Omega_{\text{skin}} := \{x \in \Omega : \ 0.15 < x_2 \leq 0.25\}\,.$$

For adipose tissue we employ the material law proposed in Sommer et al. [240], augmented by the penalty function of Hartmann and Neff [128] to enforce small compressibility. Then, with

$$j = \det(\nabla\varphi) = \sqrt{\iota_3},$$

the stored energy function is given by

$$W_{\text{adipose}} := \frac{c}{2}(\iota_1 - 3) + \frac{k_1}{k_2}\Big(\exp(k_2[\kappa\iota_1 + (1-3\kappa)\iota_4 - 1]^2) - 1\Big) + \frac{d}{50}\left(j^5 + j^{-5} - 2\right).$$

From [240, Table 2] we take the mean material parameters of the considered specimens,

$$c = 0.3\,\text{kPa}, \quad \kappa = 0.09, \quad k_1 = 0.8\,\text{kPa}, \quad k_2 = 47.3,$$

and for the penalty function we heuristically set $d = 10\,\text{kPa}$. We assume that the fiber directions of the interlobular septa, which are incorporated in the model via the first mixed invariant ι_4, are initially aligned along the z-axes.

As discussed in Sec. 5.5, we will use the extended Mooney-Rivlin law proposed in Hendriks [132] for the description of the skin. Again it is augmented by the same penalty function to account for compressibility. This yields the stored energy function

$$W_{\text{skin}} := c_{10}\,(\iota_1 - 3) + c_{01}\,(\iota_1 - 3)\,(\iota_2 - 3) + e\left(j^5 + j^{-5} - 2\right).$$

The material parameters of the corresponding incompressible model for human dermis in vivo are taken from [281],

$$c_{10} = 9.4\,\text{kPa} \quad \text{and} \quad c_{01} = 82\,\text{kPa}.$$

For the volumetric part we again use the choice $e = 10\,\text{kPa}$. We let a pressure-type control act on the control boundary

$$\Gamma_{\text{c}} = \left\{x \in \mathbb{R}^3 : \ x_2 = -0.25\right\}.$$

On the observation boundary

$$\Gamma_{\text{o}} = \left\{x \in \mathbb{R}^3 : \ x_2 = 0.25\right\}$$

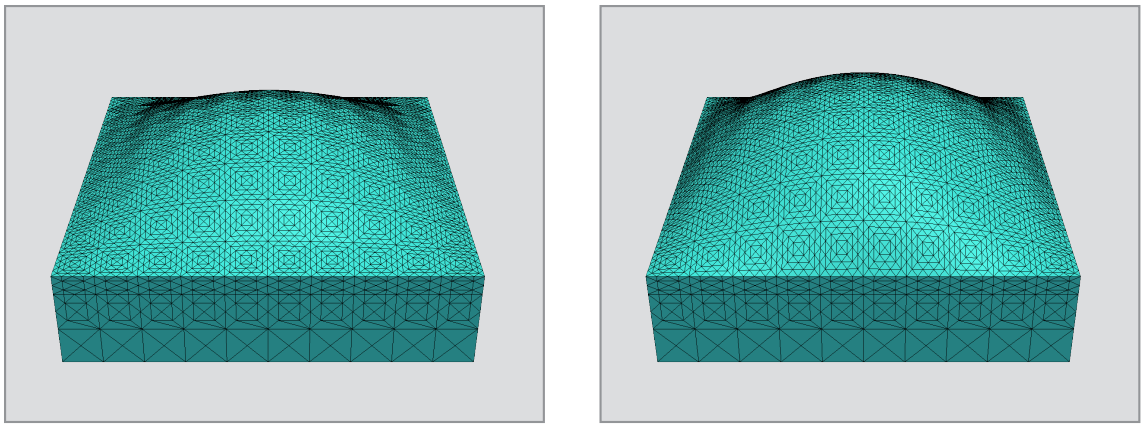
we define a reference displacement in z-direction, see Fig. 6.2.2(b), via

$$u_{\text{ref,z}}(x) = \frac{1}{2}(1 - x_0^2)(1 - x_1^2).$$

This induces a reference deformation

$$\varphi_{\text{ref}}(x) = \begin{pmatrix} x_0 \\ x_1 \\ x_2 + u_{\text{ref,z}}(x) \end{pmatrix}.$$

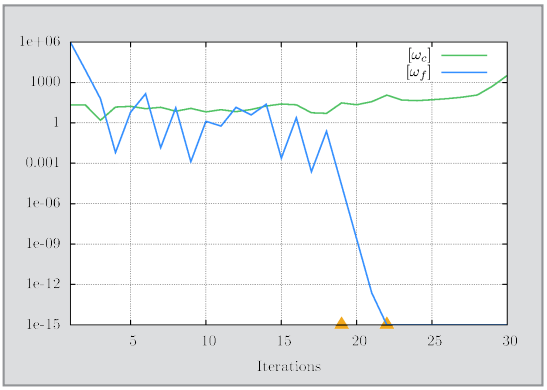
In this setting the control essentially compresses the fibers that are aligned along the z-axes. Thus these buckle and their contribution to the mechanical response is negligible. Since the model for adipose tissue incorporates fiber dispersion still a part of the fibers is expected to yield relevant contributions. The Tikhonov regularization parameter is set to $\alpha = 10^{-2}$.



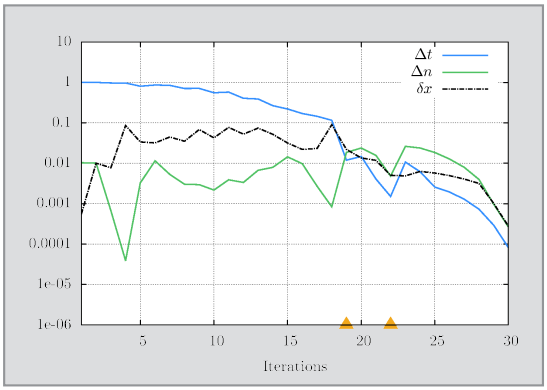
(a) Computed deformation. **(b)** Reference deformation.

Figure 6.2.2.: Deformations of the reference configuration for an example for adipose and skin tissue.

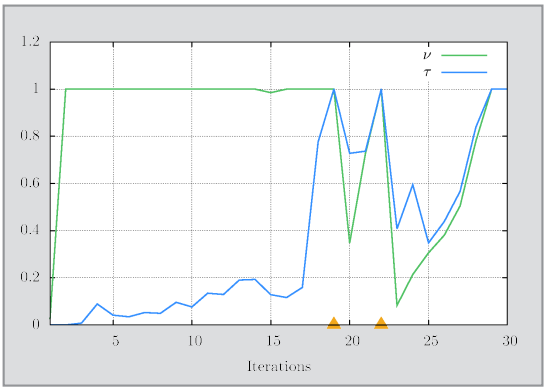
The corresponding computed solution is given in Fig. 6.2.2(a). This example is relatively complex. This can be observed from the algorithmic quantities that are plotted in Fig. 6.2.3 as well as the large number of required steps. The strong nonlinearity of the constraint and the Lagrangian become manifest in large estimates for the Lipschitz constants (Fig. 6.2.3(a)). Consequently only small steps in tangential direction are allowed. This can be observed from Fig. 6.2.3(c) where we see the damping factors ν and τ for the normal resp. tangential step. In the first 17 iterations the tangential steps are strongly damped, such that feasibility can be restored by the normal steps. We observe from Fig. 6.2.3(b) that in this phase the tangential step, which the algorithm would like to take, is significantly bigger than the normal step and is reduced only slowly. In contrast to the previous example of control of nonlinear heat transfer we have to dampen the normal step. In the first iteration this is a consequence of the fact that the employed models do not yield an equilibrium of forces for $\varphi = \text{id}$. For such models our algorithm needs its first iteration(s) to get a more reasonable "initial" iterate. Secondly the normal steps are dampened after mesh refinement. In this case it can not be guaranteed that the



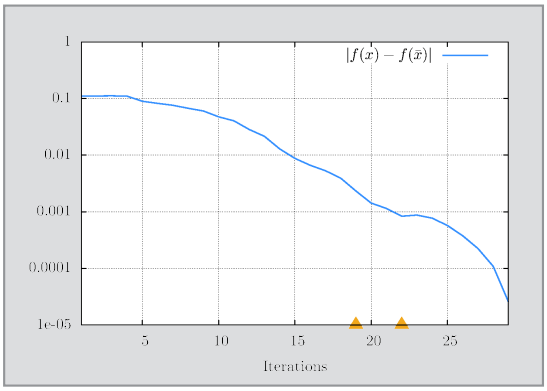
(a) Lipschitz constants.



(b) Composition of steps.



(c) Damping factors.



(d) Cost functional.

Figure 6.2.3.: Algorithmic behavior for a two-phase model composed of adipose and skin tissue. Orange triangles on the horizontal axes indicate mesh refinement.

normal steps stay in the Kantorovich region of the constraint since a priori we do not know how far our iterates are from this region on the refined mesh.

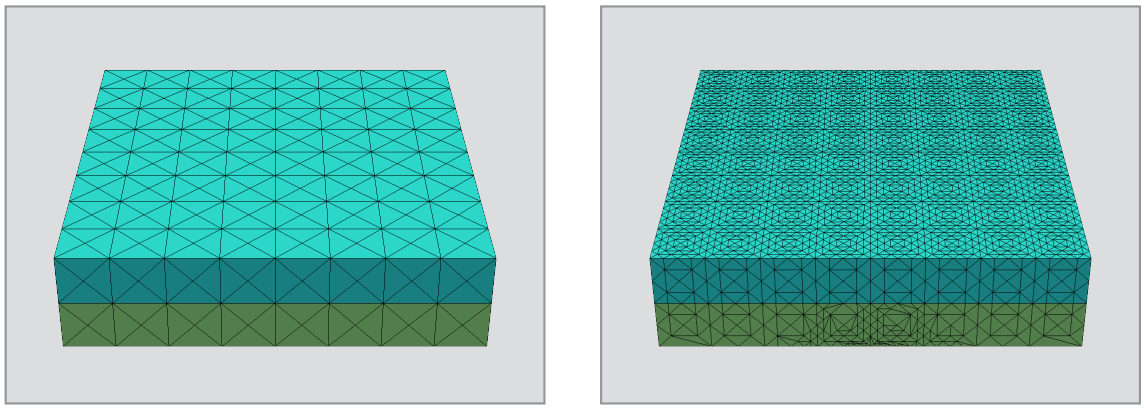
In iteration 19 the algorithm seems to reach the region of local convergence. The grid is refined in this iteration. After further three iterations, during which $[\omega_f]$ is rapidly reduced to its lower bound 10^{-15}, again undamped steps are accepted and the grid is refined again. As this problem contains complex nonlinearities our algorithm needs again some iterations to recover and eventually reaches the region of fast local convergence in iteration 29.

Same as in the given example of nonlinear heat transfer we observe that the logarithmic plot of the deviation of the cost functional $f(x) = J(\varphi, g)$ from the optimal value $f(\bar{x}) = J(\bar{\varphi}, \bar{g})$ is rather concave, except for the iterations after grid refinement[3].

[3]The plot in Fig. 6.2.13(d) illustrates the absolute difference between the cost function value at the iterates and the final value of the cost functional, the latter serving as an approximation of the optimal value. Therefore, this plot stops one iteration before the last.

In this example we set the upper bound of allowed vertices to $n_{\mathrm{v,max}} = 25\,000$. The grid was refined from 1 844 vertices (11 285 degrees of freedom) to 40 194 vertices (245 705 degrees of freedom). We observe that the skin tissue requires a significantly higher resolution than the adipose tissue, see Fig. 6.2.1(b).

6.2.1.2. Adipose and muscle tissue



(a) Initial grid. **(b)** Adaptively refined grid.

Figure 6.2.4.: Initial and final grid.
Green: skin tissue. Cyan: adipose tissue.

In the second example we again consider the model of Sommer et al. [240] for the description of the adipose tissue (Sec. 5.2.2.1) and retain the initial fiber direction of the interlobular septa. We employ two layers of same thickness. As illustrated in Fig. 6.2.4, the adipose tissue (cyan) is located in

$$\Omega_{\text{adipose}} = \{x \in \Omega : \ 0 \leq x_2 \leq 0.25\}$$

and the muscle tissue (green) in

$$\Omega_{\text{skin}} = \{x \in \Omega : \ -0.25 \leq x_2 < 0\}\,.$$

For the description of muscular tissue (Sec. 5.2.2.3) we use the model of Martins et al. [182]. We replace their model for the volumetric part by the physically more reasonable model of Hartmann and Neff [128],

$$W_{\text{vol}}(\iota_3) = \frac{c_{\text{vol}}}{50}\left(j^5 + j^{-5} - 2\right),$$

where again

$$j = \det(\nabla\varphi) = \sqrt{\iota_3}.$$

This yields the model

$$W_{\text{muscle}} = c\left(\exp\left(b(\bar{\iota}_1 - 3)\right) - 1\right) + A\left(\exp\left(a(\bar{\iota}_6 - 1)^2\right) - 1\right) + \frac{c_{\text{vol}}}{50}\left(j^5 + j^{-5} - 2\right),$$

with $c_{\text{vol}} = 10\,\text{kPa}$ and the remaining material constants chosen as in [182, Ex. 3.1],

$$c = 0.387\,\text{kPa}, \quad b = 23.46, \quad A = 0.584\,\text{kPa} \quad \text{and} \quad a = 12.43.$$

The muscle fibers are assumed to be initially aligned along the x-axes. We let a pressure-type control act on

$$\Gamma_{\mathrm{c}} = \left\{ x \in \mathbb{R}^3 : \ x_2 = -0.25 \right\}.$$

On the observation boundary

$$\Gamma_{\mathrm{o}} = \left\{ x \in \mathbb{R}^3 : \ x_2 = 0.25 \right\}$$

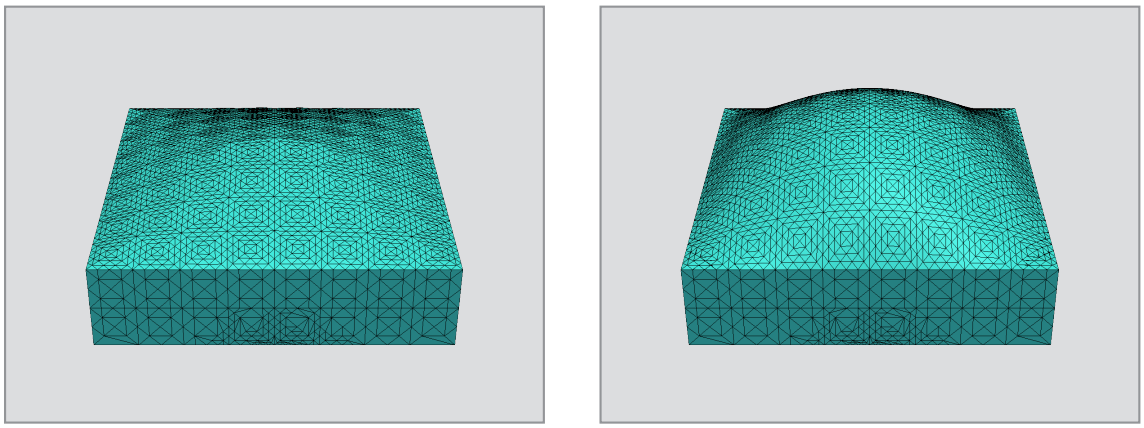
we define the reference displacement, see Fig. 6.2.5(b), via

$$u_{\mathrm{ref,z}}(x) = \frac{1}{2}(1 - x_1^2)(1 - x_2^2),$$

resp.

$$\varphi_{\mathrm{ref}}(x) = \begin{pmatrix} x_0 \\ x_1 \\ x_2 + u_{\mathrm{ref,z}}(x) \end{pmatrix}.$$

The Tikhonov regularization parameter is set to $\alpha = 10^{-4}$.



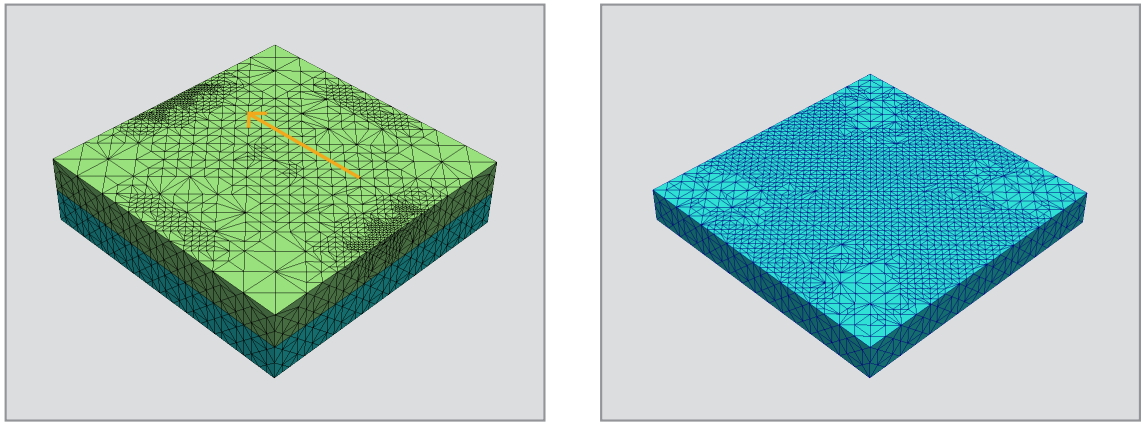
(a) Computed deformation. **(b)** Reference deformation.

Figure 6.2.5.: Deformations of the reference configuration for an example for adipose and muscle tissue.

Againt the control essentially compresses the fibers in the adipose tissue that are aligned along the z-axes. Thus these buckle and their contribution to the mechanical response is negligible. Since the model for adipose tissue incorporates fiber dispersion still a part of the fibers is expected to yield relevant contributions.

The model of the muscle tissue requires larger forces to deform, compared to the models of skin or adipose tissue. This is a consequence of both, the exponential isotropic contribution and the exponential contribution for the fibers. As these are initially aligned along the x-axes, they take up a considerable amount of the exerted forces.

The computed solution for this example is given in Fig. 6.2.5(a). Here, the grid is refinement at several parts. At the observation boundary increased uniform refinement occurs, see Fig. 6.2.5(a). On the control boundary we observe anisotropic mesh refinement, which is strongest near the Dirichlet boundary, where the muscle fibers are assumed to be attached (see Fig. 6.2.6(a)). Eventually, in Fig. 6.2.6(b) we also observe increased mesh refinement at the interior contact surface between muscle and adipose tissue. There discontinuous gradients are expected due to the different models for the tissue types.

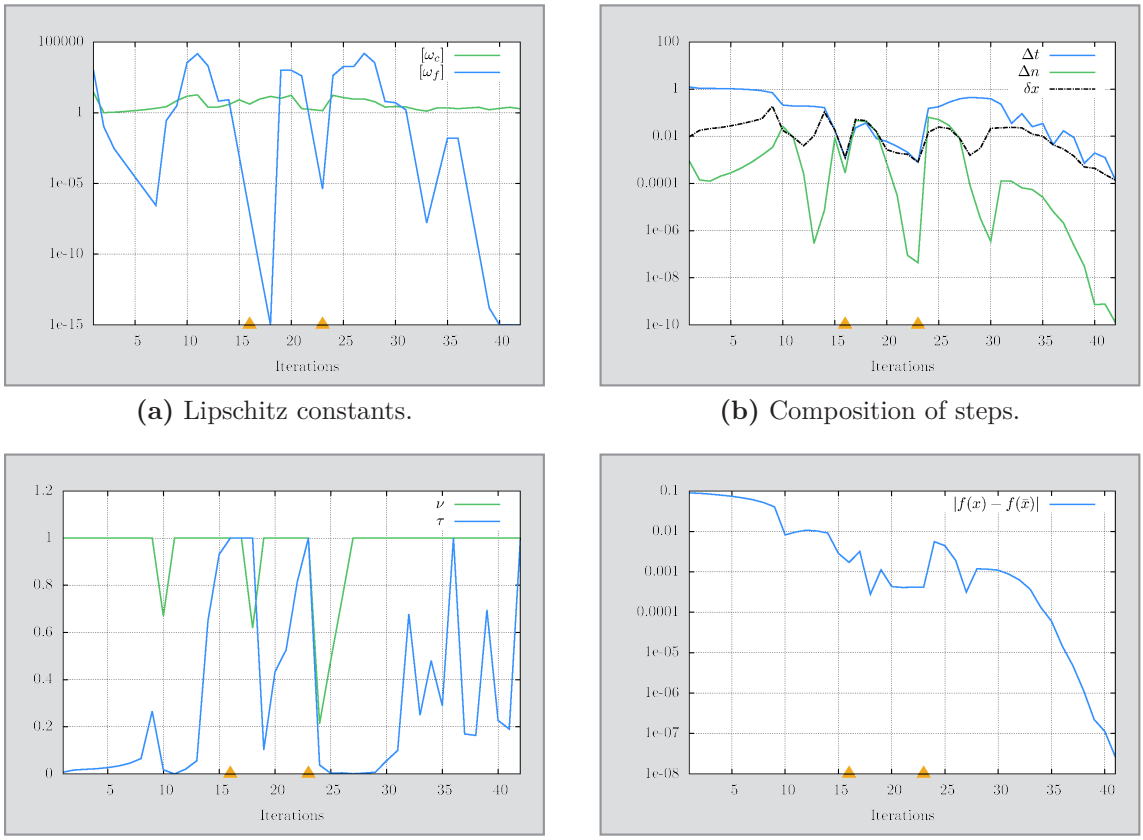


(a) Control boundary, muscle fiber direction indicated by orange arrow.

(b) Inner boundary between muscle and adipose tissue.

Figure 6.2.6.: Mesh refinement at parts of the inner and outer boundary.

The material laws in this example are more challenging than in the previous one. In the algorithmic parameters this is reflected by big values of $[\omega_f]$ and correspondingly strong damping of the tangential steps, in particular in the first iterations and after mesh refinement. Observe that in the tangential step in the ninth iteration is too optimistic and the following normal step is damped. Thus, to guarantee that we stay in the Kantorovich region of the constraints, at least in the case that no refinement occurs, we would have to choose stricter contraction parameters Θ_{acc}, $\Theta_{\text{aim,x}}$ and $\Theta_{\text{aim,n}}$. Following this too optimistic tangential step priority is attributed to restore feasibility. Thus the tangential steps are again strongly damped for some steps and the cost functional value increases slightly. Then the mesh is refined for the first time. As to be expected after mesh refinement feasibility must be restored again and larger normal steps occur as well as damping of the normal step in iteration 18. Observe that the estimate $[\omega_f]$ still decreases until the same iteration 18 and directly after increases rapidly. In Sec. 3.3.2 and Sec. 3.4 we described that our estimates for the decrease in the cost functional may not be reliable in the case of large normal steps, since we evaluate $f'(x)$ instead of the required quantity $f'(x + \delta n)$. For this reason in cases where our algorithm takes large normal steps to restore feasibility, such as in iterations 17 and 18, the estimates for $[\omega_f]$ may be misleading. However, as soon as the normal steps become smaller we get again reasonable estimates. After the second refinement step this effect does not occur. There $[\omega_f]$ directly increases



(a) Lipschitz constants.

(b) Composition of steps.

(c) Damping factors.

(d) Cost functional.

Figure 6.2.7.: Algorithmic behavior for a two-phase model for adipose and muscle tissue. Orange triangles on the horizontal axes indicate mesh refinement.

strongly and the tangential steps are strongly damped. While feasibility is restored the cost functional strongly increases in the next iterations[4]. Again we observe, except after mesh refinement, a rather concave shape of the absolute deviation of the function value from the optimal value. In the remaining iterations the algorithm reaches its region local convergence and terminates after 42 iterations. The strongly oscillating damping factors for the tangential step in these last steps indicate that the tangential damping factors are chosen too optimistic. A more regular behavior, can be attained by stricter choices for the algorithmic contraction parameters.

In this example the mesh is refined from initially 851 vertices (5 251 degrees of freedom) to 29 134 vertices (175 822 degrees of freedom), where mesh refinement was switched off after exceedance of $n_{\mathrm{v,max}} = 25\,000$.

[4]The plot in Fig. 6.2.7(d) illustrates the absolute difference between the cost function value at the iterates and the final value of the cost functional, the latter serving as an approximation of the optimal value. Therefore, this plot stops one iteration before the last.

Here Fig. 6.2.7(d) is a bit misleading since before mesh refinement the cost functional was underestimated. Thus the decrease in iterations 25-28 is actually an increase in the function value.

6.2.2. Isotropic models on real-world geometries

Our last examples are concerned with the computation of implants in the context of real patient data. We consider two cases of augmentation implants on the zygomatic bone resp. the mandible.

For both examples we will model the soft tissue using a compressible Mooney-Rivlin material law of the form

$$W(\nabla\varphi) = a_0\|\nabla\varphi\|^2 + a_1\|\mathrm{cof}(\nabla\varphi)\|^2 + a_2\det(\nabla\varphi)^2 - a_3\log(\det(\nabla\varphi)),$$

where we determine the parameters $a_0, \dots, a_3$ such that near $\varphi = \mathrm{id}$ our model approximates the descriptions of linearized elasticity with material parameters

$$E_L = 1,\ \nu_L = 0.45,$$

i.e.

$$a_0 = 0.08625, \quad a_1 = 0.08625, \quad a_2 = 0.68875, \quad a_3 = 1.895.$$

The parameter E_L is Young's modulus and ν_L the Poisson ratio. They are related to the Lamé constants via

$$\lambda_L = \frac{\nu_L E_L}{(1+\nu_L)(1-2\nu_L)} \quad \text{and} \quad \mu_L = \frac{E_L}{2(1+\nu_L)}.$$

The choice $E_L = 1$ is no restriction since it results from the simple rescaling

$$g \to E^{-1}g \text{ and } \alpha \to E^2\alpha.$$

Again we employ the Tikhonov regularized tracking-type cost functional

$$J(\varphi, u) = \frac{1}{2}\|\varphi - \varphi_{\mathrm{ref}}\|^2_{\mathrm{L}^2(\Gamma_o)} + \frac{\alpha}{2}\|g\|^2_{\mathrm{L}^2(\Gamma_c)},$$

with $\alpha = 0.05$.

For the definition of the norm we use the symmetric and positive definite part of the second derivative of a St.Venant-Kirchhoff material law, i.e. we define local inner products in $\mathrm{W}^{1,2}(\Omega)$ via

$$\begin{aligned}\langle v, w\rangle_{M(\varphi_k)} = \int_\Omega &(\lambda_L \mathrm{tr}\,(E'(\nabla u_k)\nabla v)\,\mathrm{tr}\,(E'(\nabla u_k)\nabla w)\\ &+2\mu_L\,(E'(\nabla u_k)\nabla v : E'(\nabla u_k)\nabla w))\,\mathrm{d}\mu,\end{aligned}$$

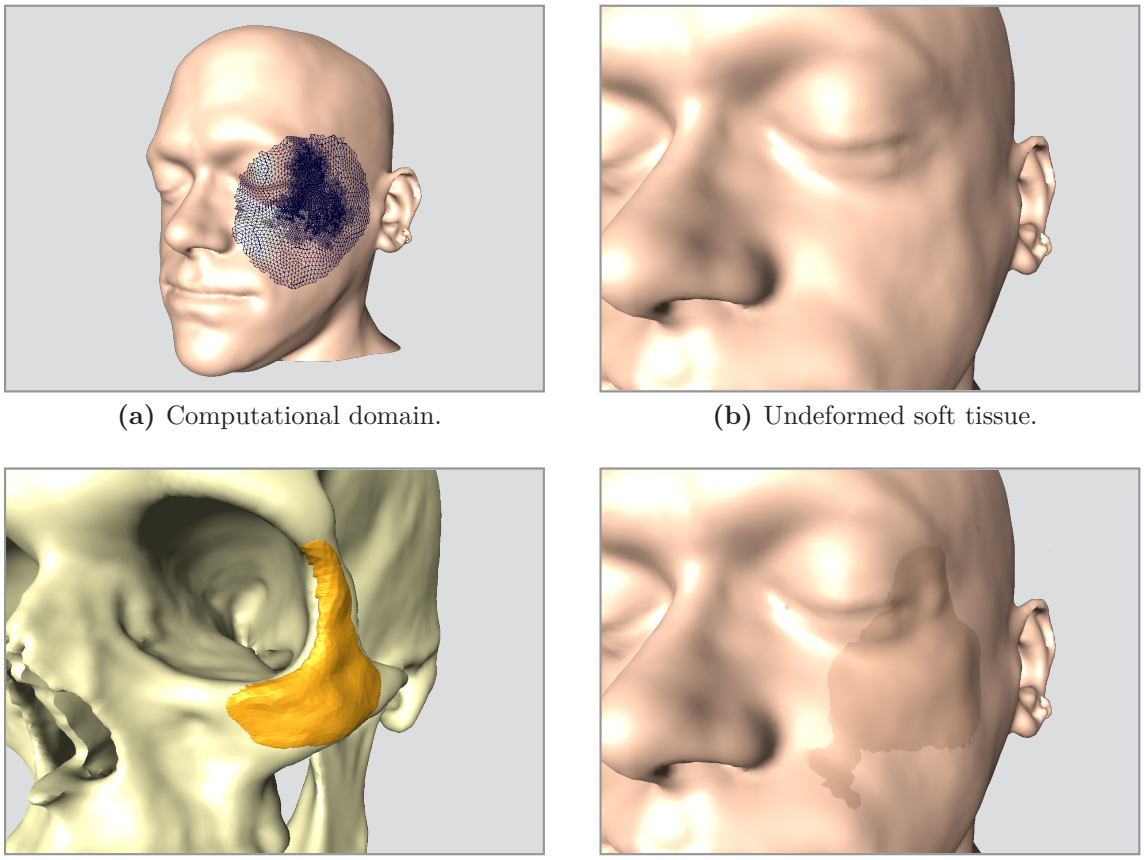
where

$$E(\nabla v) = \frac{1}{2}\left(\nabla v + \nabla v^* + \nabla v^*\nabla v\right)$$

is the strain tensor and $u_k = \varphi_k - \mathrm{id}$.

In the following all figures are generated with ZIBAMIRA [242].

6.2.2.1. An implant at the zygomatic bone



(a) Computational domain.

(b) Undeformed soft tissue.

(c) Reference.

(d) Reference surface. Differences to the undeformed surface are shaded.

Figure 6.2.8.: An implant on the zygomatic bone.

We begin with the task of computing an augmentation implant at the zygomatic bone (Fig. 6.2.8(c)) from a desired shape φ_{ref} as illustrated in Fig. 6.2.8(d). Thus the skin surface determines the observation boundary Γ_{o}, whereas the control boundary Γ_{c} is given by the contact surface between soft tissues and bones. The computational domain and the undeformed skin surface are illustrated in the first row of Fig. 6.2.8. On the artificial soft tissue boundary Γ_{d}, where it was virtually cut from its surrounding tissue, we impose homogeneous Dirichlet boundary conditions. Here, transparent boundary conditions would also be reasonable.

For the attainment of a prescribed deformation φ_{ref} on Γ_{o} an implant shape Ω_{rigid} was estimated (Fig. 6.2.8(c)). To compute the corresponding deformation of the forward problem with our model we need again a formulation in terms of the pressure exerted by the implant. For this we consider both observation and control on Γ_{c}. Then we

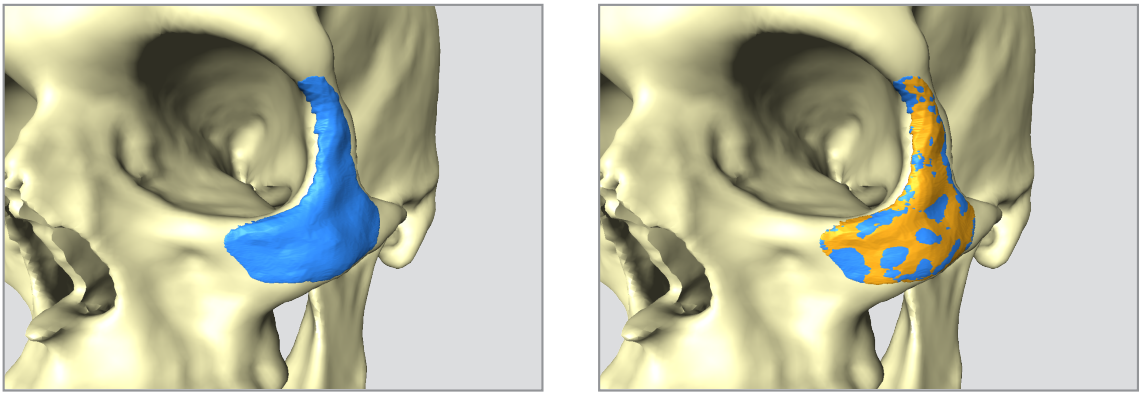
let a pressure-type boundary condition act on Γ_c such that the normal displacement on Γ_c lies on the implant boundary $\partial\Omega_{\text{rigid}}$, i.e. we solve the problem

$$\min_{(\varphi,g)\in\Phi\times L^2(\Gamma_c)} \frac{1}{2}\|\varphi-\varphi_{\text{ref},0}\|^2_{L^2(\Gamma_c)} + \frac{\alpha}{2}\|g\|^2_{L^2(\Gamma_c)} \tag{6.2.1a}$$

$$\text{subject to} \quad \mathcal{E}^{\text{str}}_{\varphi}(\varphi,g)v - \int_{\Gamma_c} g\,\text{cof}(\nabla\varphi)nv\,\text{ds} = 0 \tag{6.2.1b}$$

$$\varphi_{\text{ref},0} = zn \quad \text{for } z\in \text{argmax}_t\, x+tn \quad \text{a.e. on } \Gamma_c. \tag{6.2.1c}$$

Denoting with φ_0 a solution of this problem, we let $\varphi_{\text{ref}} = \varphi_0|_{\Gamma_c}$. This is illustrated in the second row of Fig. 6.2.8.



(a) Computed implant.

(b) Computed (blue) and reference (orange) implant.

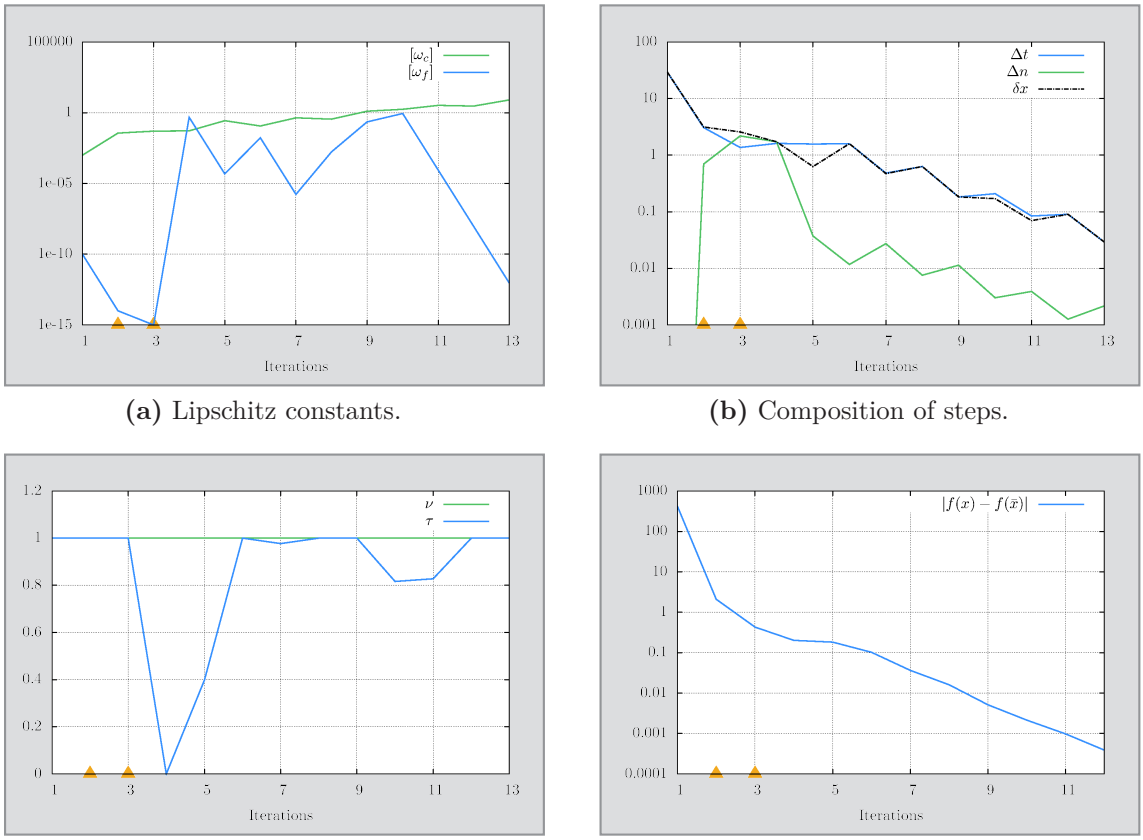
Figure 6.2.9.: Comparison of implants on the zygomatic bone.

In Fig. 6.2.9(a) the computed implant is shown and a comparison of both reference and computed implant in Fig. 6.2.9(b). Differences between both implant shapes are not visible.

The post-processing step for the visualization of implant shapes with ZIBAMIRA requires the projection of the unstructured onto Cartesian grids. This leads to a less regular boundary of the contact surface between implant and bone than on the unstructured grid. This can best be observed at the upper part of the implant.

Our choice of material parameters guarantees that the undeformed state is in equilibrium[5]. Thus, starting with $\varphi = \text{id}$, we begin feasible and the first normal step vanishes. Since the soft tissue between bone and skin surface is relatively thin this example is not too hard to solve. Computations on the coarse grid, as presented in [180], do not require globalization. Incorporating mesh refinement globalization

[5] As seen in the last examples, this is often not the case for more involved constitutive laws. These are typically considered useful if an equilibrium state not to far from the undeformed state exists. This is not related to pre- or residual stresses, where this behavior is to be expected. Instead, this is a deficiency of these models.



(a) Lipschitz constants.

(b) Composition of steps.

(c) Damping factors.

(d) Cost functional.

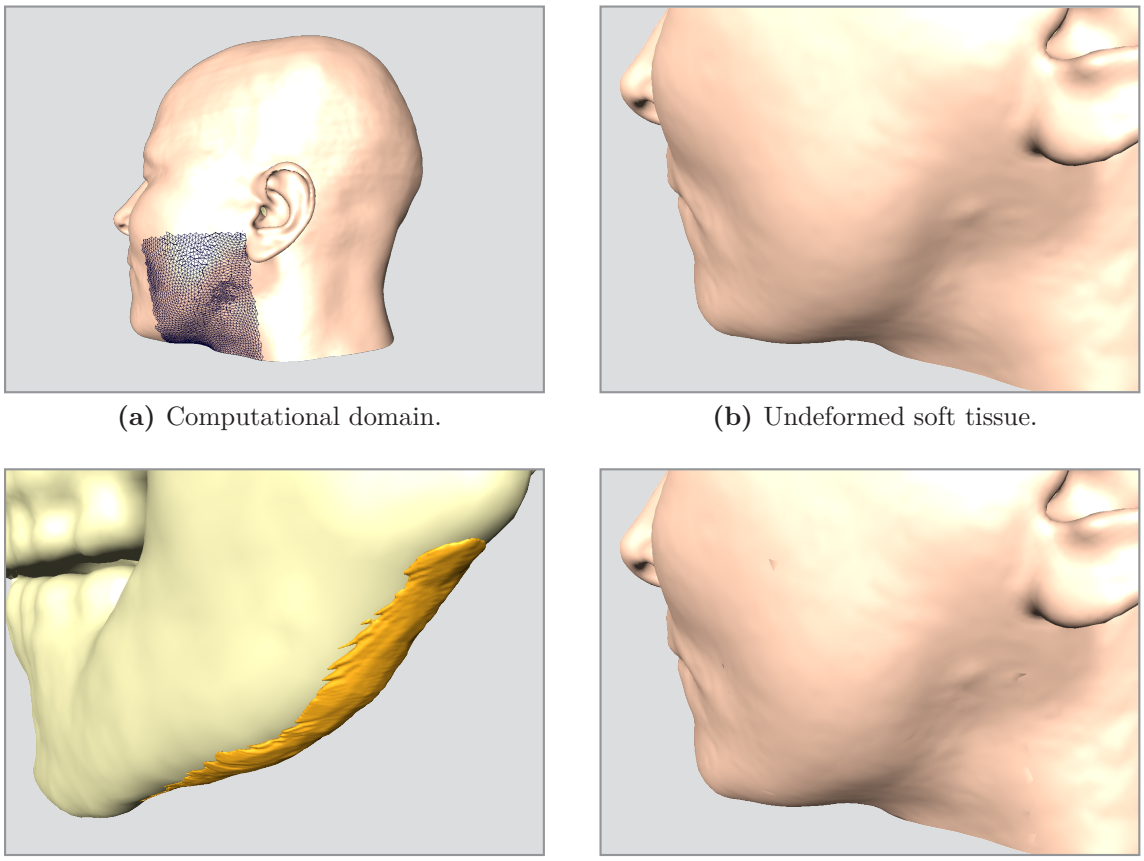
Figure 6.2.10.: Algorithmic behavior for the computation of an implant at the zygomatic bone.

is only required for the tangential steps, mainly after the mesh refinement in iterations two and three. In these two iterations the mesh is refined from initially 10 924 vertices (70 674 degrees of freedom) to 104 638 vertices (640 695 degrees of freedom). Then the upper bound on the number of vertices $n_{\mathrm{v,max}} = 100\,000$ is exceeded and no further refinement is allowed. From Fig. 6.2.10(b) we see that only directly after mesh refinement the composite step is dominated by the normal step. In all other iterations only small corrections are required to restore feasibility and the tangential step essentially determines the composite step. The semi-logarithmic plot of the deviation of the values of the cost functional from the optimal value on the refined grid in Fig. 6.2.10(d)[6] is, except for the first iterations after mesh refinement, of convex shape. Our algorithm is able to make fast progress towards optimality in the first iterations. When mesh refinement occurs, feasibility must be restored first, and the cost functional values stagnate in iterations four and five. In the following iterations

[6]The plot in Fig. 6.2.10(d) illustrates the absolute difference between the cost function value at the iterates and the final value of the cost functional, the latter serving as an approximation of the optimal value. Therefore, this plot stops one iteration before the last.

a roughly linear decrease is observed. This change from strict convexity to linearity indicates that the resolution of the coarse mesh was not fine enough to capture the complexity of the considered problem.

6.2.2.2. An implant at the mandible



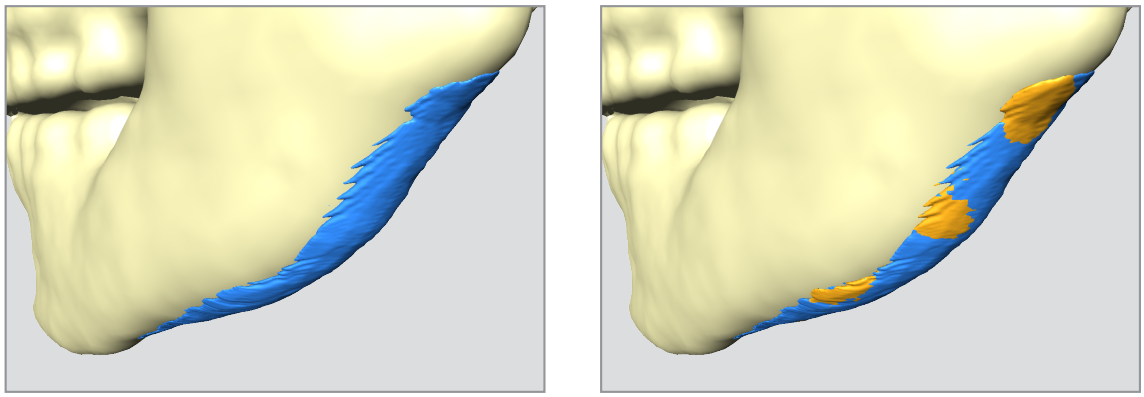
(a) Computational domain. **(b)** Undeformed soft tissue.

(c) Reference. **(d)** Reference surface.

Figure 6.2.11.: An implant on the mandible.

Finally, our last example is concerned with an augmentation implant at the mandible. Again, the soft tissue is described by a compressible Mooney-Rivlin material law, with the same material parameters as in the last example. The computational domain and the undeformed skin surface are given in the first row of Fig. 6.2.11. An estimated implant is inserted and the corresponding deformation is computed by solving (6.2.1). Reference implant and the corresponding skin surface are illustrated in the second row of Fig. 6.2.11. Differences in the position of points on the skin surface are difficult to observe. Rather differences in the surface normals and correspondingly, differences in shading, are visible. This might be an interesting point regarding "esthetic" cost functionals.

On the left side of Fig. 6.2.12 the computed implant is given. On the right both, reference and computed implant, are shown. Differences between both implant shapes are not visible. We observe a jagged implant boundary in both reference and computed implant. These spikes are not part of the solution but artifacts of the post-processing with ZIBAMIRA, which projects the unstructured grid to a uniform



(a) Computed implant.

(b) Computed (blue) and reference (orange) implant.

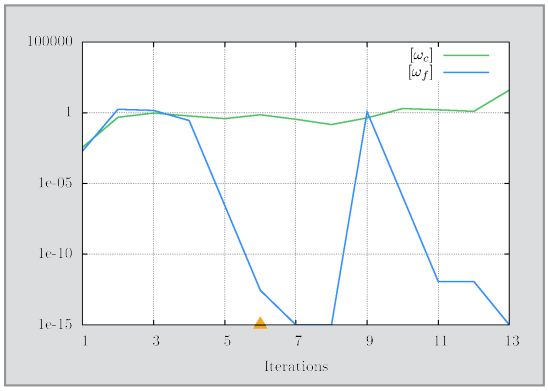
Figure 6.2.12.: Comparison of implants on the mandible.

grid of moderate accuracy. Since in this example the implant is thinner than the one in the previous example, this effect is more pronounced here.

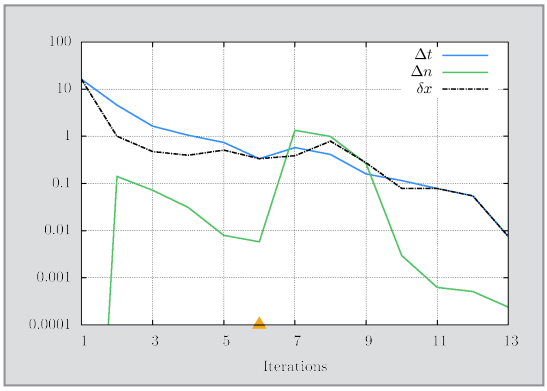
Due to the thicker soft tissue surrounding the implant this problem is slightly more difficult than the previously presented one. This can be first observed from Fig. 6.2.13(b) and (d). In Fig. 6.2.13(b) we observe that the composite step δx is almost of the same length as the tangential step δt. Moreover, the logarithmic plot of the cost functional still is roughly convex (Fig. 6.2.13(d)).

Same as in the last example our choice of material parameters guarantees that the undeformed state is in equilibrium. Thus, starting with $\varphi = \mathrm{id}$, we begin feasible and the first normal step vanishes. In the sixth iteration undamped steps are accepted and the mesh is refined from 14 825 vertices (92 655 degrees of freedom) to 37 294 vertices (229 373 degrees of freedom). Due to the thickness of the soft tissue only a small number of faces on the skin surface is refined. On the refined mesh the iterates are no more in the region of local convergence. In Fig. 6.2.13(b) we see that a large normal step is required to retrieve feasibility. With increasing damping factor τ for the tangential step the newly encountered problem structure on the refined mesh also affects the estimate $[\omega_f]$. In the ninth iteration the damped normal step is relatively large and significantly larger than the damped tangential step (Fig. 6.2.13(c)). Thus we observe an increase in the cost functional value in this iteration (Fig. 6.2.13(d))[7]. Same as in the previous example in most iterations the composite step is dominated by the tangential direction, only in the three iterations after mesh refinement the normal step has a larger influence.

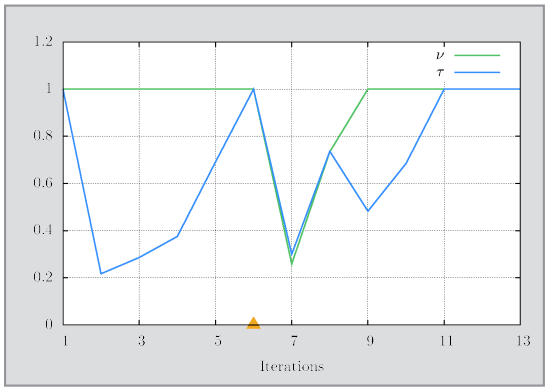
[7]The plot in Fig. 6.2.13(d) illustrates the absolute difference between the cost function value at the iterates and the final value of the cost functional, the latter serving as an approximation of the optimal value. Therefore, this plot stops one iteration before the last.



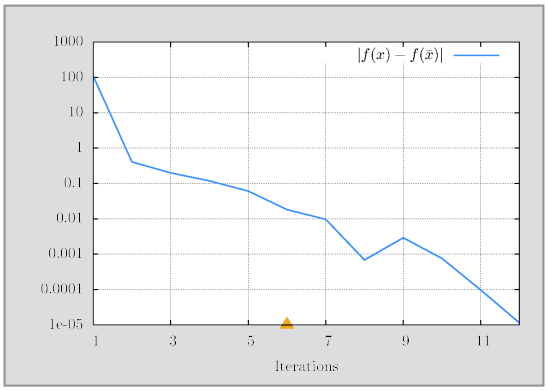
(a) Lipschitz constants.



(b) Composition of steps.



(c) Damping factors.



(d) Cost functional.

Figure 6.2.13.: Algorithmic behavior for the computation of an implant at the mandible.

Conclusion

In the course of this thesis a model for the implant shape design problem has been developed. The influence of the implant on the surrounding soft tissue naturally yields an obstacle problem in polyconvex hyperelasticity. Solving such a problem numerically involves the treatment of a challenging contact problem. In order to avoid this, the geometric model of the implant as obstacle was replaced by a mechanical model, wherein its influence is incorporated indirectly in terms of the force exerted on the surrounding soft tissue. This leads to a formulation with a nonlinear pressure-type boundary condition and the implant shape design problem can be described as a nonconvex equality constrained optimization problem

$$\begin{aligned} &\min_{(\varphi,g)\in\Phi\times \mathrm{L}^2(\Gamma_c)} J(\varphi,g) = J_0(\varphi) + \frac{\alpha}{2}\|g\|^2_{\mathrm{L}^2(\Gamma_c)} \\ \text{subject to} \qquad &\mathcal{E}^{\mathrm{str}}_{\varphi}(\varphi,g)v - \int_{\Gamma_c} g\,\mathrm{cof}(\nabla\varphi)nv\,\mathrm{ds} = 0 \quad \text{for all } v \in \mathrm{W}^{1,p}_0(\Omega;\mathbb{R}^3). \end{aligned}$$

with $g \leq 0$. Analytical results are difficult to attain in this setting. The nonlinear pressure-type boundary condition does not admit a potential and thus the incorporation into the hyperelastic setting is intricate. Slightly simplifying the pressure-type to Neumann boundary conditions, existence of solutions of the corresponding bi-level optimization problem was established in Thm. 2.3.

For computations based on this model an affine covariant composite step method was proposed. Its way to cope with the double aim of feasibility and optimality is to split the full Lagrange-Newton step δx into a *normal step* δn and a *tangential step* δt. Inspired by the affine covariant Newton method for underdetermined systems of Deuflhard [76], the normal step is computed as a minimum norm correction with respect to a suitable norm. The tangential step aims at the minimization of the cost functional in the kernel of the linearized constraints. In this regard, the cubic regularization method, as employed in Weiser et al. [277] and Schiela [226], is extended to equality constrained optimization problems. A simplified Newton step takes the role of a second order correction and helps us to avoid the Maratos effect. Eventually, the globalization mechanism was adapted to the particular requirements of elasticity theory, namely validity of the orientation preservation condition

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

For a practical realization of the composite step method, several strategies have been proposed. The computation of the (simplified) normal step and the adjoint correction is equivalent to finding the solution to strictly convex constrained optimization

problems. Thus, these quantities are uniquely determined and – depending on the problem structure – direct factorizations or projected preconditioned conjugate gradient methods (PPCG) are adequate.

The computation of the tangential step involves the solution of a nonconvex optimization problem. In this context, a regularized conjugate gradient method (RCG) was found to be more robust, but also more expensive, than the truncated conjugate gradient method (TCG). In order to profit from both, the increase robustness of RCG and the performance of TCG both approaches were combined in a hybrid method (HCG). This hybrid method can be interpreted as a safeguarding mechanism for the detection of fast local convergence by algorithmic quantities.

We stress that there is no structure for nonconvex problems and each of the discussed conjugate gradient methods for nonconvex problems may outperform the others for suitable problems or even lead to failure of the outer iteration. Thus, it would be of interest to identify classes of nonconvex problems for which regularization or truncation works particularly well. A candidate may be given through strictly polyconvex problems. These contain a strictly convex part for the deformation gradient, that can be used as preconditioner, and may admit meaningful interpretations of the regularized problem.

Error estimation for adaptive mesh refinement was considered in the affine covariant framework. Based on a hierarchical extension of the ansatz space, the error in the primal variables of the Lagrange-Newton step is estimated in terms of the employed norm. For efficient estimation of the discretization error the underlying KKT-system is simplified to (perturbed) block-triangular form. The choice which blocks to eliminate was motivated by the implant shape design problem, more precisely by the question of determining the influence of the resolution of the implant shape on the computed deformation. The generated meshes for our examples seem to be reasonable. If this is indeed the case is a question that cannot be answered easily. The problem is that we would need a detailed understanding of the relevance of the discretization accuracy of the constraint, the cost functional and their interplay.

Moreover, we shortly discussed state-of-the-art models for the most relevant tissues, skin, adipose and muscle tissue. These are non-convex and often exhibit anisotropy and exponential growth of the elastic energy. The incorporation of these models in numerical experiments yields further challenges for numerical solvers.

In the presented numerical examples it was demonstrated that the proposed algorithm is able to cope with both state-of-the-art fiber-reinforced models and complex geometries. It extends previous attempts to simulate therapeutic outcomes [163, 282], wherein only simple material models have been employed. To the knowledge of this author the question of how to predict an implant shape, based on a desired therapeutic outcome, has not been considered before.

Open questions

The applicability of the developed model and algorithm requires further validation. This involves the need to answer several open questions in different fields of research. In the following, the most relevant for the task of implant shape design are shortly addressed.

Selection of open theoretical questions. The mathematical understanding of elastic problems is limited. For a non-exhaustive but more detailed discussion of open problems, the interested reader is referred to Ball [17, 19].

1. The orientation preservation condition

   $$\det(\nabla\varphi) > 0$$

   and the associated limit behavior of the stored energy function

   $$\lim_{\det(F)\searrow 0} W(F) = \infty$$

   lead to analytical difficulties. The set

   $$\left\{\varphi \in \mathrm{W}^{1,\mathrm{p}}(\Omega) :\ \mathcal{E}^{\mathrm{str}}(\varphi) = \int_\Omega W(\nabla\varphi)\,\mathrm{d}\mu = \infty\right\}$$

   is dense in $\mathrm{W}^{1,\mathrm{p}}(\Omega)$, for $p < \infty$, which implies that the energy functional $\mathcal{E}$ is not differentiable in $\mathrm{W}^{1,\mathrm{p}}(\Omega)$. Thus, well-posedness of the formal first order conditions cannot be proven. However, as discussed exemplarily for a Mooney-Rivlin material law in Sec. 1.3, well-posedness can be expected in finite element spaces, at least if our algorithm guarantees nondegeneracy of the iterates. A better understanding of this issue is desirable.

2. The incorporation of the implant, either directly as obstacle, or indirectly via pressure-type boundary conditions, is not trivial. Considering the implant as an obstacle requires the treatment of a contact problem. This is not only challenging numerically, but also theoretical results are hard to attain and a better understanding of these problems is desirable.
   Other challenges arise in the analytical treatment of the pressure-type boundary conditions. Besides the fact that these are only formally related to the obstacle problem, these boundary conditions change when the boundary changes. Thus, they are non-conservative. An existence result including pressure-type boundary conditions certainly is of interest.

Selection of open algorithmic questions. The affine covariant composite step method leaves space for improvement.

1. The inexact computation of the different algorithmic quantities requires a thorough investigation, as it influences our algorithm at various points.
   a) First we have to make sure that the estimates of the Lipschitz constants are at most slightly affected. Close to the solution the relative accuracies in the computation of the different quantities should admit at least linear convergence.
   b) Due to the inexact step computation orthogonality relations, such as Galerkin orthogonality and the orthogonality between normal and tangential step, do not hold any more. This should be taken into account in the derivation of error estimators to properly separate discretization and algebraic errors, cf. [12, 127].
   c) Currently major parts of the computation time of the inexact solvers are spent in the preconditioners, more specifically in the application of the multigrid solvers. Therefore, one could replace the multigrid solvers by suitable multigrid preconditioners, that asymptotically approach multigrid solvers. If this can be realized on a sound theoretical basis, we expect a significant increase in the performance of the proposed composite step method. However, in the chosen affine covariant setting this is challenging, since in multigrid methods the termination criterion is based on suitable decrease of the residuals.
2. The complexity of realistic geometries imposes challenges regarding efficiency. In order to accurately resolve the complex geometries, occurring in medical applications, large number of degrees of freedom are necessary. To efficiently cope with these geometries, new mesh refinement and coarsening strategies must be developed. Standard mesh refinement strategies, such as the red-green refinement for simplicial meshes [25], do not alter the covered domain. Requiring an adequate representation of the computational domain already on the coarse grid, the efficiency of adaptive algorithms can only be partially exploited. We are so far lacking a balanced mesh coarsening and refinement strategy that admits to recover the geometry obtained from the segmentation of medical image data during refinement. However, convergence of such a refinement scheme in three dimensions is a delicate matter, even more if grid regularity should be retained. In addition this requires adjustments of error estimators to take into account the additional geometric information.
3. The proposed error estimator nicely fits into the affine covariant setting. However, error estimation for optimal control problems is largely motivated by heuristic arguments and a better theoretical and practical understanding is needed.
4. The usual problem with affine covariance is of interest here. We would like to have a rigorous proof of convergence. Due to the well-known difficulties with cycling [76] in affine covariant Newton methods for PDEs such an existence proof will need some adjustments of the proposed algorithm.

Selection of open questions in biomechanics. Despite the impressive progress since the emergence of the scientific field of biomechanics in the middle of the last century, the understanding of biological processes in the human body still is largely shaded in mystery.

1. One of the big challenges is the determination of patient-specific local material parameters. It is well known that the mechanical properties of biological soft tissues changes due to "extrinsic" aging, related to hostile environmental conditions, and natural "intrinsic" aging [4, 99, 150]. This is only one of the reason for the fact that material parameters have to be determined individually and spatially localized. Thus we need measurement techniques that are capable of attaining in vivo localized information for different tissue types.
   Previous attempts to estimate material parameters typically tried to compute them from the gray-scale values of the medical image data [163]. This procedure is limited in its accuracy and not able to determine the multiple different parameters of complex material laws.
   For superficial structures, such as the skin or artery walls accessed via catheters, optical coherence elastography (OCE) [158, 159, 175, 176, 250], which distinguishes different tissues by their optical scattering properties, has recently been successfully applied [158, 175]. However this technique can not be used for deeper tissues such as adipose or skeletal muscle tissue.
2. Regarding the long term prediction of the outcome of medical interventions, we have to better understand tissue growth [169, 264, 288]. Due to altered loading conditions and transfer of forces the insertion of implants may trigger tissue growth [83, 212, 253, 288]. As a first step it is important to understand under which conditions implants lead to growth that has a significant influence on the long-term esthetic outcome. Secondly, further understanding of the growth process itself may help in the design of implants and the prediction of therapeutical results.

Selection of general open questions. Eventually, we need a unified framework for the incorporation of simulation and optimization algorithms in clinical applications. In this regard, the realization of a model for TIMMS is a crucial step.

1. Validation of the computed implants and stress distributions with data from real surgical interventions is necessary. Besides the obvious need of verifying the outcome of any significant therapeutical validation, this might allow to optimize the complicated descriptions for biological soft tissues.
2. Standardization must be in the center of such an attempt, in particular, standardization of interfaces is mandatory. It helps to avoid waste of resources in siloed solutions. Moreover, unified interfaces strongly simplify validation processes and interdisciplinary cooperation.

For most of the remaining open questions the ways to their solution are well-known to the experts in the respective fields. These can be solved with reasonable effort

in few years. The highly accurate determination of anisotropic in vivo material parameters is expected to require more time, as adequate technical equipment must be developed and produced.

In the last decades significant progress was made in many directions. The achievements in the calculus of variations in the second half of the 20^{th} century did yield a reasonable setting for the description of elastic material through the property of *polyconvexity*. In the same time span both progress in mathematics and, more important, the development of new imaging and measurement techniques did admit the development of increasingly accurate descriptions for biological soft tissues by the biomechanics community. Today highly accurate, mathematically tractable models for many biological soft tissue types exist. As demonstrated in this thesis, modern algorithmic techniques admit solutions to the problems arising in the application of these models.

Due to valuable contributions from many fields of research, instead of asking the question, whether the accurate computational design of implants is possible, we are at the point to ask when a suitable software will be available in clinical practice. With the concentrated effort of experts in the involved fields this can certainly be realized within few years.

A. Functional analysis and the calculus of variations

Definition A.1. Let T be a metric space.

- For compact T we denote by $\mathrm{C}(T) = \mathrm{C}(T, \|\cdot\|_\infty)$ the Banach space of continuous functions on T endowed with the supremum norm

$$\|v\|_\infty := \sup_{x \in T} |v(x)|.$$

- We denote by $\mathrm{M}(T)$ the Banach space of regular Borel measures endowed with the norm of total variation

$$\|\nu\|_{tv} := \sup_{Z \in \mathcal{Z}(T)} \sum_{E \in Z} |\nu(E)|,$$

  where $\mathcal{Z}(T)$ is the set of all finite, pairwise disjoint decompositions of T into measurable subsets.

Theorem A.2. *(Implicit function theorem)*

Let $\mathrm{X}, \mathrm{Y}, \mathrm{Z}$ *be Banach spaces,* $(x_0, y_0) \in \mathrm{X} \times \mathrm{Y}$ *and let* $F : U(x_0, y_0) \subseteq \mathrm{X} \times \mathrm{Y} \to \mathrm{Z}$ *be defined on an open neighborhood* $U(x_0, y_0)$ *and* $F(x_0, y_0) = 0$*. Further, assume that*

- F is Fréchet differentiable in its second argument,
- $F_y : \mathrm{Y} \to \mathrm{Z}$ is bijective and continuous in $U(x_0, y_0)$,
- and, at (x_0, y_0), the function F is Lipschitz-continuous in its first argument and continuous in its second.

Then,

1. for every $x \in B_{r_0}(x_0)$, there exist a positive numbers r_0 and r such that there is exactly one element $y(x) \in \mathrm{Y}$, satisfying $\|y(x) - y_0\| \leq r$ and $F(x, y(x)) = 0$,
2. and $y(\cdot)$ is Lipschitz continuous.

Proof. From the implicit function theorem of Hildebrandt and Graves [286, Thm. 4.B], we get that 1. holds and $y(\cdot)$ is continuous. Leaves to show that $y(\cdot)$ is Lipschitz

continuous. For sake of shortness not the whole proof of the implicit function theorem is repeated here. Instead we begin, for $x \in B_{r_0}(x_0)$, with the inequality, cf. [286, Prop. 1.2] with $k = \frac{1}{2}$,

$$\|y(x_1) - y(x_0)\| \leq 2 \|T(x_1)y(x_0) - T(x_0)y(x_0)\|,$$

where

$$T(x)y = y - F_y(x_0, y_0)^{-1} F(x, y).$$

Then,

$$\begin{aligned} \|y(x_1) - y(x_0)\| &\leq 2 \|T(x_1)y(x_0) - T(x_0)y(x_0)\| \\ &= 2 \left\| F_y(x_0, y_0)^{-1} \left(F(x_1, y(x_0) - F(x_0, y(x_0)) \right) \right\| \\ &\leq 2 \left\| F_y(x_0, y_0)^{-1} \right\| L_F \|x_1 - x_0\|. \end{aligned}$$

Thus, $y(\cdot)$ is Lipschitz continuous at x_0. □

Theorem A.3. *(Riesz representation theorem)*
Let K be a compact metric (or topological) linear space. Then $\mathrm{C}(K)^*$ is isometrically isomorphic to $\mathrm{M}(K)$ under the mapping

$$T : \mathrm{M}(K) \to \mathrm{C}(K)^*, \quad (T\nu)(f) = \int_K f \, \mathrm{d}\nu.$$

Proof. See [279, Thm. II.2.5]. □

Theorem A.4. *Let $\{u_j\}_j$ be a bounded sequence in $\mathrm{W}^{1,\mathrm{p}}(\Omega)$ with $1 \leq p \leq \infty$ and affine boundary values given by $u_F = Fx$, $F \in \mathbb{M}^{m,n}$. Denote the associated gradient parametrized measure by $\nu = \{\nu_x\}_{x \in \Omega}$. Then there exists another bounded sequence $\{v_j\}_j$ with the same boundary values such that the associated gradient parametrized measure $\bar{\nu}$ is homogeneous an*

$$\langle \bar{\nu}, \varphi \rangle = \frac{1}{|\Omega|} \int_\Omega \int_{\mathbb{R}^m} \varphi(\lambda) \, \mathrm{d}\nu_x(\lambda) \, \mathrm{dx},$$

for any $\varphi \in \mathrm{C}_0(\mathbb{R}^m)$.

Proof. See [207, Thm. 8.1]. □

Theorem A.5. *Consider a sequence* $\{u_j\}_j$*, bounded in* $\mathrm{W}^{1,p}(\Omega)$ *with* $1 \leq p \leq \infty$ *and its associated* $\mathrm{W}^{1,\mathrm{p}}$*-parametrized measure* $\nu = \{\nu_x\}_{x\in\Omega}$*. Let*

$$F_a = \int_{\mathbb{M}^{m,n}} A \,\mathrm{d}\nu_a(A)$$

and $u_a(x) = F_a x$ *for* $a \in \Omega$*. Then, there exists, for almost all* $a \in \Omega$*, a bounded sequence* $\left\{w_j^a\right\}_j$ *in* $\mathrm{W}^{1,\mathrm{p}}(\Omega)$ *such that* $w_j^a - u_a \in \mathrm{W}_0^{1,\mathrm{p}}(\Omega)$*, for all* j*, and the associated* $\mathrm{W}^{1,\mathrm{p}}$*-parametrized measure is* ν_a*.*

Proof. See [207, Thm. 8.4]. □

Lemma A.6. *Let* $\{v_j\}_j$ *be a bounded sequence in* $\mathrm{W}^{1,\mathrm{p}}(\Omega)$ *with* $1 \leq p < \infty$*. Then there always exists another sequence* $\{u_j\}_j$ *of Lipschitz functions such that* $\{|u_j|^p\}_j$ *is equiintegrable and the sequences of gradients possess the same underlying* $\mathrm{W}^{1,\mathrm{p}}$*-parametrized measure.*

Proof. See [207, Lem. 8.15]. □

Theorem A.7 (Tonelli)**.** *Let* Ω *be a domain in* $\mathbb{R}^n$ *and* $m \geq 1$*. For functions* $u : \Omega \to \mathbb{R}^m$ *and continuous* $f : \mathbb{R}^m \to \bar{\mathbb{R}}$ *define the nonlinear function*

$$\mathcal{F}(u) := \int_\Omega f\Big(u(x)\Big)\,\mathrm{dx}$$

Then the function $\mathcal{F}$ *is weakly lower semi-continuous on* $\mathrm{L}^{\mathrm{p}}(\Omega)$ *with* $1 < p < \infty$ *and weak-star lower semi-continuous on* $\mathrm{L}^\infty(\Omega)$ *if and only if* $u \mapsto f(u)$ *is convex.*

Proof. For a sketch of the proof see [214, Thm. 10.16]. For the complete proof see [92, Thm. 1] or [181, Box 4.1]. □

Theorem A.8. *Let* $\{z_j\}_j \subset \mathbb{M}^{m,n}$ *be sequence of measurable functions with associated parametrized measure* $\nu = \{\nu_x\}_{x\in\Omega}$*. Then*

$$\liminf_{j\to\infty} \int_E \Psi(x, z_j(x))\,\mathrm{dx} \geq \int_E \int_{\mathbb{M}^{m,n}} \Psi(x,\lambda)\,\mathrm{d}\nu_x(\lambda)\,\mathrm{dx},$$

for every non-negative Carathéodory function Ψ *and every measurable subset* $E \subset \Omega$*.*

Proof. Analogous to [207, Thm. 6.11]. □

Lemma A.9 (Lemma of Mazur)**.**

Let X *be a metrizable locally convex space and* $\{x_j\}_j$ *a weakly convergent sequence* $x_j \rightharpoonup x \in \mathrm{X}$*. Then there is a sequence* $\{y_j\}_j$ *with corresponding index sets*

$$K_j \subset \mathbb{Z},\ |K_j| < \infty,\ \operatorname{argmin}_{k\in K_j} \geq j,$$

such that for all $j \in \mathbb{N}$

$$y_j = \sum_{k \in K_j} \alpha_k^j x_k, \quad \sum_{k \in K_j} \alpha_k^j = 1,\ \alpha_k^j \geq 0,\ K_j \subset \mathbb{Z},\ |K_j| < \infty,$$

i.e. y_j *is a finite convex combination of elements of* $\{x_j\}_j$*, and*

$$y_j \to x.$$

Proof. See [221, Thm. 3.13]. □

Acknowledgments

First of all my gratitude goes to my supervisor Prof. Dr. Anton Schiela. For sharing my interest in both analytical as well as algorithmical questions, for always having time to discuss occurring problems with me and for a remarkable talent in keeping me motivated when my own motivation found a descent direction. I thank you for the last years. It was a pleasure to learn from and work with you.

I thank Dr. Stefan Zachow for providing the patient data used in this thesis.

I also want to take the opportunity to express my gratitude to my colleagues and friends, Dr. Michael Baldus, Dr. Sunayana Ghosh, Dr. Sebastian Götschel, Marian Moldenhauer and Moritz Nagel, who spent their free time with proof-reading, sorting out typos, notational inconsistencies as well as grammatical mistakes and pointing out weaknesses in my argumentation.

Eventually I thank my family and my friends. Since a nice family and good friends make everything easier you all played an important role during the formation of this work.

List of Figures

List of Tables

List of Algorithms

Bibliography

[1] Digital imaging and communications in medicine (DICOM) standard.

[2] E. Acerbi and N. Fusco. Semicontinuity problems in the calculus of variations. *Arch. Ration. Mech. Anal.*, 86:125–145, 1984.

[3] R. A. Adams and J. J. F. Fournier. *Sobolev spaces*. Elsevier, 2nd edition, 2003.

[4] P. G. Agache, C. Monneur, J. L. Leveque, and J. De Rigal. Mechanical properties and Young's modulus of human skin in vivo. *Arch. Dermatol. Res.*, 269: 221–232, 1980.

[5] M. Ainsworth and J. T. Oden. *A Posteriori Error Estimation in Finite Element Analysis*. Wiley, 2000.

[6] N. Alkhouli, J. Mansfield, E. Green, J. Bell, B. Knight, N. Liversedge, J. C. Tham, R. Welbourn, A. C. Shore, K. Kos, and C. P. Winlove. The mechanical properties of human adipose tissues and their relationships to the structure and composition of the extracellular matrix. *Am. J. Physiol. Endocrinol. Metab.*, 305:E1427–E1435, 2013.

[7] P. R. Amestoy, I. S. Duff, J. Koster, and J.-Y. L'Excellent. A fully asynchronous multifrontal solver using distributed dynamic scheduling. *SIAM J. Mat. Anal. and Appl.*, 23(1):15–41, 2001.

[8] P. R. Amestoy, A. Guermouche, J.-Y. L'Excellent, and S. Pralet. Hybrid scheduling for the parallel solution of linear systems. *Par. Comp.*, 32(2):136–156, 2006.

[9] A. N. Annaidh, K. Bruyère, M. Destrade, M. D. Gilchrist, C. Maurini, M. Otténio, and G. Saccomandi. Automated estimation of collagen fibre dispersion in the dermis and its dontribution to the anisotropic behaviour of skin. *Ann. Biomed. Eng.*, 40(8):1666–1678, 2012.

[10] A. N. Annaidh, K. Bruyère, M. Destrade, M. D. Gilchrist, and M. Otténio. Characterising the anisotropic mechanical properties of excised human skin. *J. Mech. Behav. Biomed.*, 5:139–148, 2012. doi: 10.1016/j.jmbbm.2011.08.016.

[11] M. Arioli. A stopping criterion for the conjugate gradient algorithm in a finite element method framework. *Numer. Math.*, 97:1–24, 2004. doi: 10.1007/s211-03-0500-y.

[12] M. Arioli, J. Liesen, A. Miedlar, and Z. Strakoš. Interplay between discretization and algebraic computation in adaptive numerical solution of elliptic PDE problems. *GAMM. Mitt.*, 36(1):102–129, 2013.

[13] E. M. Arruda and M. C. Boyce. A three-dimensional constitutive model for the large stretch behavior of rubber elastic materials. *J. Mech. Phys. Solids*, 41(2):389–412, 1993.

[14] I. Babuška, J. R. Whiteman, and T. Strouboulis. *Finite Elements: An Introduction to the Method and Error Estimation*. Oxford University Press, 2011.

[15] J. M. Ball. Convexity conditions and existence theorems in nonlinear elasticity. *Arch. Rational Mesh. Anal.*, 63:337–403, 1977.

[16] J. M. Ball. Discontinuous equilibrium solutions and cavitation in nonlinear elasticity. *Phil. Trans. Roy. Soc. London*, 306(1496):557–611, 1982.

[17] J. M. Ball. *Geometry, Mechanics and Dynamics*, chapter "Some open problems in elasticity", pages 3–59. Springer, 2002.

[18] J. M. Ball. *Mathematical models of martensitic microstructure. Materials Science and Engineering A*, volume 378, chapter Mathematical models of martensitic microstructure, pages 61–69. Elsevier, 2004.

[19] J. M. Ball. Progress and puzzles in nonlinear elasticity. In J. Schröder and P. Neff, editors, *Poly-, Quasi- and Rank-One Convexity in Applied Mechanics*, volume 516 of *CISM International Centre for Mechanical Sciences*, pages 1–15. Springer Vienna, 2010. ISBN 978-3-7091-0173-5. doi: 10.1007/978-3-7091-0174-2_1.

[20] J. M. Ball and J. C. Currie. Null lagrangians, weak continuity, and variational problems of arbitrary order. *J. Funct. Anal.*, 41:135–174, 1981.

[21] J. M. Ball and F. Murat. $W^{1,p}$-quasiconvexity and variational problems for multiple integrals. *J. Funct. Anal.*, 58:225–253, 1984.

[22] D. Balzani. *Polyconvex Anisotropic Energies and Modeling of Damage Applied to Arterial Walls*. PhD thesis, Universität Duisburg-Essen, 2006.

[23] D. Balzani, P. Neff, J. Schröder, and G. A. Holzapfel. A polyconvex framework for soft biological tissues. Adjustment to experimental data. *Int. J. Solids Struct.*, 43:6052–5070, 2006.

[24] W. Bangerth and R. Rannacher. *Adaptive Finite Element Methods for Differential Equations*. Birkhäuser, 2003.

[25] R. E. Bank, A. H. Sherman, and A. Weiser. *Scientific Computing (Applications of Mathematics and Computing to the Physical Sciences)*, chapter Some refinement algorithms and data structures for regular local mesh refinement, pages 3–17. North-Holland, 1983.

[26] G. G. Barbarino, M. Jabareen, J. Trzewik, A. Nkengne, G. Stamatas, and E. Mazza. Development and validation of a three-dimensional finite element model of the face. *J. Biomech. Eng.*, 131(4):041006–1 – 041006–11, 2009. doi: 10.1115/1.3049857.

[27] P. Bastian, M. Droske, C. Engwer, R. Klöfkorn, T. Neubauer, M. Ohlberger, and M. Rumpf. Towards a unified framework for scientific computing. In *Proc. of the 15th International Conference on Domain Decomposition Methods*, 2005.

[28] P. Bastian, M. Blatt, C. Engwer, A. Dedner, R. Klöfkorn, S. P. 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.

[29] R. Becker and R. Rannacher. An optimal control approach to error estimation and mesh adaptation in finite element methods. *Acta Numerica*, 10:1–102, 2001. doi: 10.1017/S0962492901000010.

[30] R. Becker, H. Kapp, and R. Rannacher. Adaptive finite element methods for optimal control of partial differential equations: basic concept. *SIAM J. Optim. Control*, 39:113–132, 2000.

[31] R. Becker, M. Braack, D. Meidner, R. Rannacher, and B. Vexler. Adaptive finite element methods for pde-constrained optimal control problems. In W. Jäger, R. Rannacher, and J. Warnatz, editors, *Reactive Flows, Diffusion and Transport*, pages 177–205. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-28379-9. doi: 10.1007/978-3-540-28396-6_8.

[32] S. M. Belkoff and R. C. Haut. A structural model used to evaluate the changing microstructure of maturing rat skin. *J. Biomech.*, 24(8):711–720, 1991.

[33] O. Benedix and B. Vexler. A posteriori error estimation and adaptivity for elliptic optimal control problems with state constraints. *Comp. Opt. Appl.*, 44 (1):3–25, 2009.

[34] M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point systems. *Acta Numerica*, 14:1–137, 2005.

[35] L. Berliner and H. U. Lemke. *Modellgestützte Therapie*, chapter From Model-Guided Therapy to Model-Based Evidence: Potential Impact on Medical Outcomes, Economics and Ethics, pages 253–270. Health Academy, 2008.

[36] J. E. Bischoff, E. M. Arruda, and K. Grosh. Finite element modeling of human skin using an isotropic, nonlinear elastic constitutive model. *J. Biomech.*, 33: 645–652, 2000.

[37] M. Böl. Micromechanical modelling of skeletal muscles: from the single fibre to the whole muscle. *Arch. Appl. Mech.*, 80:557–567, 2010. doi: 10.1007/s00419-009-0378-y.

[38] M. Böl and S. Reese. Finite element modelling of rubber-like polymers based on chain statistics. *Int. J. Solids Struct.*, 43:2–26, 2006.

[39] M. Böl, S. Reese, K. K. Parker, and E. Kuhl. Computational modeling of muscular thin films for cardiac repair. *Comp. Mech.*, 43:535–544, 2009. doi: 10.1007/s00466-008-0328-5.

[40] E. M. H. Bosboom, M. K. C. Hesselink, C. W. J. Oomens, C. V. C. Bouten, M. R. Drost, and F. P. T. Baaijens. Passive transverse mechanical properties of skeletal muscle under in vivo compression. *J. Biomech.*, 34:1365–1368, 2001.

[41] D. Braess. *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer, 4$^{\text{th}}$ edition, 1992.

[42] J. H. Bramble and J. E. Pasciak. A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems. *Math. Comp.*, 50(181), 1988. doi: 10.1090/S0025-5718-1988-0917816-8.

[43] J. H. Bramble, J. E. Pasciak, and J. Xu. Parallel multilevel preconditioners. *Math. Comp.*, 55:1–22, 1990.

[44] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer, 2008.

[45] W. L. Briggs, V. E. Henson, and S. F. McCormick. *A Multigrid Tutorial*. SIAM, 2000.

[46] H. Bufler. Pressure loaded structures under large deformations. *Z. Angew. Math. Mech.*, 64(7):287–295, 1984.

[47] R. Byrd, M. Hribar, and J. Nocedal. An interior point algorithm for large-scale nonlinear programming. *SIAM J. Opt.*, 9(4):877–900, 1999. doi: 10.1137/S1052623497325107.

[48] R. Byrd, J. Gilbert, and J. Nocedal. A trust region method based on interior point techniques for nonlinear programming. *Math. Prog.*, 89(1, Ser. A):149–185, 2000. doi: 10.1007/PL00011391.

[49] L. Cardamone, A. Valentin, J. F. Eberth, and J. D. Humphrey. Origin of axial prestretch and residual stress in arteries. *Biomech. Model. Mechanobiol.*, 8(6): 431–446, 2009. doi: 10.1007/s10237-008-0146-x.

[50] M. Carozza and A. Passarelli di Napoli. Model problems from nonlinear elasticity: partial regularity results. *ESAIM Contr. Optim. Ca.*, 13(1), 2007. doi: 10.1051/cocv:2007007.

[51] C. Carstensen. Some remarks on the history and future of averaging techniques in a posteriori finite element error analysis. *Z. Angew. Math. Mech.*, 84(1): 3–21, 2004. doi: 10.1002/zamm.200410101.

[52] C. Carstensen. Reliable and efficient averaging techniques as universal tool for a posteriori finite element error control on unstructured grids. *Int. J. Numer. Anal. Mod.*, 3(3):333–347, 2006.

[53] C. Cartis, N. I. M. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unsconstrained optimization. Part I: motivation, convergence and numerical results. *Math, Prog.*, 127(2):245–295, 2011. doi: 10.1007/s10107-009-0286-5.

[54] C. Cartis, N. I. M. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function- and derivative-evaluation complexity. *Math. Prog.*, 130(2):295–319, 2011. doi: 10.1007/s10107-009-0337-y.

[55] E. Casas and F. Tröltzsch. First- and second-order optimality conditions for a class of optimal control problems with quasilinear elliptic equations. *SIAM J. Control Optim.*, 48(2):688–718, 2009. ISSN 0363-0129. doi: 10.1137/080720048.

[56] P. G. Ciarlet. *Mathematical Elasticity Vol. I: Three-dimensional Elasticity*. North-Holland, 1988.

[57] P. G. Ciarlet. *An Introduction to Differential Geometry with Applications to Elasticity*. Springer, 2005.

[58] P. G. Ciarlet and G. Geymonat. On constitutive equations in compressible nonlinear elasticity. *C. R. Acad. Sci. Paris*, 295:423–426, 1982.

[59] J. A. Clark, J. C. Y. Cheng, and K. S. Leung. Mechanical properties of normal skin and hypertrophic scar. *Burns*, 22(6):443–446, 1996.

[60] A. Cohen. A Padé approximant to the inverse langevin function. *Rheol. Acta*, 30:270–273, 1991.

[61] T. F. Coleman and A. Pothen. The null space problem I. Complexity. *SIAM J. Alg. Disc. Meth.*, 7(4):527–537, 1986.

[62] K. Comley and N. A. Fleck. The mechanical response of porcine adipose tissue. *Preprint*, pages 1–30, 2009.

[63] K. Comley and N. A. Fleck. A micromechanical model for the Young's modulus of adipose tissue. *Int. J. Solids Struct.*, 47:2982–2990, 2010.

[64] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust-Region Methods*. MPS-SIAM Series on Optimization, 2000.

[65] D. T. Corr and D. A. Hart. Biomechanics of scar tissue and uninjured skin. *Adv. Wound Care*, 2(2):37–43, 2013.

[66] M. Costabel and M. Dauge. Construction of corner singularities for Agmon-Douglis-Nirenberg elliptic systems. *Math. Nachr.*, 162:209–237, 1993.

[67] C. L. Cox and G. J. Fix. On the accuracy of least squares methods in the presence of corner singularities. *Comp. & Maths. with Appls.*, 10(6):463–475, 1984.

[68] J. C. Criscione, J. D. Humphrey, A. S. Douglas, and W. C. Hunter. An invariant basis for natural strain which yields orthogonal stress response terms in isotropic hyperelasticity. *J. Mech. Phys. Solids*, 48:2445–2465, 2000.

[69] J. C. Criscione, A. D. McCulloch, and W. C. Hunter. Constitutive framework optimized for myocardium and other high-strain, laminar materials with one fiber family. *J. Mech. Phys. Solids*, 50:1681–1702, 2002.

[70] J.-P. Crouzeix and J. A. Ferland. Criteria for quasi-convexity and pseudo-convexity: relationships and comparisons. *Math. Prog.*, 23:193–205, 1982.

[71] A. L. F. da Silva, A. M. Borba, N. R. Simão, F. L. M. Pedro, A. H. Borges, and M. Miloro. Customized polymethyl methacrylate implant for the reconstruction of craniofacial osseous defects. *Case Reports in Surgery*, 2014(358569): 1–8, 2014. doi: 10.1155/2014/358569.

[72] B. Dacorogna. *Direct Methods in the Calculus of Variations*. Springer, 2nd edition, 2008.

[73] T. A. Davis and I. S. Duff. An unsymmetric-pattern multifrontal method for sparse LU factorization. *SIAM J. Mat. Anal. and Appl.*, 18(1):140–158, 1997.

[74] T. A. Davis and I. S. Duff. A combined unifrontal/multifrontal method for unsymmetric sparse matrices. *ACM Trans. Math. Softw.*, 25(1):1–19, 1999.

[75] P. G. De Deyne. Application of passive stretch and its implications for muscle fibers. *Phys. Ther.*, 81:819–827, 2001.

[76] P. Deuflhard. *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms*. Springer, 2004.

[77] P. Deuflhard and M. Weiser. *Computational science for the 21st century*, chapter Local inexact Newton multilevel FEM for nonlinear elliptic problems, pages 129–138. Wiley, 1997.

[78] P. Deuflhard and M. Weiser. *Multigrid Methods V, Lecture Notes in Computational Science and Engineering*, chapter Global inexact Newton multilevel FEM for nonlinear elliptic problems, pages 71–89. Springer, 1998.

[79] P. Deuflhard and M. Weiser. *Numerische Mathematik 3: adaptive Lösung partieller Differentialgleichungen*. De Gruyter, 2011.

[80] P. Deuflhard, P. Leinen, and H. Yserentant. Concepts of an adaptive hierarchical finite element code. *Impact Comp. Sci. Eng.*, 1:3–35, 1989.

[81] P. Deuflhard, A. Schiela, and M. Weiser. Mathematical cancer therapy planning in deep regional hyperthermia. *Acta Numerica*, 21:307–378, 2012. doi: 10.1017/S0962492912000049.

[82] J. Diani, B. Fayolle, and P. Gilormini. A review on the mullins' effect. *Eur. Polym. J.*, pages 601–612, 2009.

[83] D. E. Discher, P. Janmey, and Y. Wang. Tissue cells feel and respond to the stiffness of their substrate. *Science*, 310(5751):1139–1143, 2005.

[84] S. Doll and K. Schweizerhof. On the development of volumetric strain energy functions. *J. Appl. Mech.*, 67(1):17–21, 1999. doi: 10.1115/1.321146.

[85] H. S. Dollar, N. I. M. Gould, W. H. A. Schilders, and A. J. Wathen. Using constraint preconditioners with regularized saddle-point problems. *Comp Optim. Appl.*, 36(2-3):249–270, 2007. doi: 10.1007/s10589-006-9004-x.

[86] W. Dörfler and R. H. Nochetto. Small data oscillation implies the saturation assumption. *Numer. Math.*, 91:1–12, 2002. doi: 10.1007/s002110100321.

[87] M. G. Dunn, F. H. Silver, and D. A. Swann. Mechanical analysis of hypertrophic scar tissue: Structural basis for apparent increased rigidity. *J. Invest. Dermatol.*, 84:9–13, 1985.

[88] V. Ebbing. *Design of Polyconvex Energy Functions for All Anisotropy Classes*. PhD thesis, Universität Duisburg-Essen, 2010.

[89] C. Edwards and R. Marks. Evaluation of biomechanical properties of human skin. *Clin. Dermatol.*, 13:375–380, 1995.

[90] W. Ehlers and G. Eipper. The simple tension problem at large volumetric strain computed from finite hyperelastic material laws. *Acta Mech.*, 130:17–27, 1998.

[91] A. E. Ehret and M. Itskow. A polyconvex hyperelastic model for fiber-reinforced materials in application to soft tissues. *J. Mater. Sci.*, 42:8853–8863, 2007.

[92] L. C. Evans. *Partial Differential Equations*. Oxford University Press, 1998.

[93] R. C. C. A. Filho, R. M. Oliveira, N. L. Neto, C. Gurgel, and R. C. C. Abdo. Reconstruction of bony facial contour deficiencies with polymethylmethacrylate implant: case report. *J. Appl. Oral Sci.*, 9(4):426–430, 2011. doi: 10.1590/S1678-77572011000400021.

[94] C. Fischbach. *Development of a 3-D Model System of Adipogenesis*. PhD thesis, Universität Regensburg, 2003.

[95] B. Fischer and R. Freund. Chebyshev polynomials are not always optimal. *J. Approx. Th.*, 65(3):261–272, 1991.

[96] R. Fletcher and S. Leyffer. Nonlinear programming without a penalty function. *Math. Program.*, 91(2, Ser. A):239–269, 2002. ISSN 0025-5610.

[97] Y. C. Fung. Elasticity of soft tissues in simple elongation. *Am. J. Physiol.*, 213:1532–1544, 1967.

[98] Y. C. Fung. What are residual stresses doing in our blood vessels? *Ann. Biomed. Eng.*, 19:237–249, 1991.

[99] Y. C. Fung. *Biomechanics: Mechanical Properties of Living Tissues*. Springer, 2nd edition, 1993.

[100] N. Fusco and J. E. Hutchinson. Partial regularity and everywhere continuity for a model problem from non-linear elasticity. *J. Austral. Math. Soc.*, 57: 158–169, 1994.

[101] R. L. Gajdosik. Passive extensibility of skeletal muscle: review of the literature with clinical implications. *Clin. Biomech.*, 16:87–101, 2001.

[102] H. Gajewski, K. Gröger, and K. Zacharias. *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*. Akademie-Verlag Berlin, 1974.

[103] W. E. Garrett, P. K. Nikolaou, B. M. Ribbeck, R. R. Glisson, and A. V. Seaber. The effect of muscle architecture on the biomechanical failure properties of skeletal muscle under passive extension. *Am. J. Sports Med.*, 16(1):7–12, 1988.

[104] T. C. Gasser, R. W. Ogden, and G. A. Holzapfel. Hyperelastic modelling of arterial layers with distributed collagen fibre orientations. *J. R. Soc. Interface*, 3:15–35, 2006.

[105] M. Geerligs. *Skin layer mechanics*. PhD thesis, Technische Universiteit Eindhoven, 2010.

[106] M. Geerligs, W. M. P. Gerrit, P. A. J. Ackermans, C. W. J. Oomens, and F. P. T. Baaijens. Linear viscoelastic behavior of subcutaneous adipose tissue. *Biorheol.*, 45:677–688, 2008.

[107] A. W. Gielen, C. W. J. Oomens, P. H. M. Bovendeerd, T. Arts, and J. D. Janssen. A finite element approach for skeletal muscle using a distributed moment model of contraction. *Comput. Methods Biomech. Biomed. Eng.*, 3 (3):231–244, 2000. doi: 10.1080/10255840008915267.

[108] E. Gladilin. *Biomechanical Modeling of Soft Tissue and Facial Expressions for Craniofacial Surgery Planning*. PhD thesis, Freie Universität Berlin, 2002.

[109] E. Gladilin, S. Zachow, P. Deuflhard, and H.-C. Hege. Anatomy- and physics-based facial animation for craniofacial surgery simulations. *Med. Biol. Eng. Comput.*, 42:167–170, 2004.

[110] G. H. Golub and R. S. Varga. Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods. *Numer. Math.*, 3:147–156, 1961.

[111] R. S. Goody. The missing link in the muscle cross-bridge cycle. *Nat. Struct. Biol.*, 10(10):773–775, 2003.

[112] J. Gosline, M. Lillie, E. Carrington, P. Guerette, C. Ortlepp, and K. Savage. Elastic proteins: biological roles and mechanical properties. *Phil. Trans. Roy. Soc. London*, 357:121–132, 2002.

[113] J.-P. Gossez. Nonlinear elliptic boundary value problems for equations with rapidly (or slowly) increasing coefficients. *Trans. Amer. Mat. Soc.*, 190:163–205, 1974.

[114] S. Götschel, M. Weiser, and A. Schiela. *Advances in DUNE*, chapter Solving Optimal Control Problems with the Kaskade7 Finite Element Toolbox, pages 101–112. Springer, 2012.

[115] N. I. M. Gould, M. E. Hribar, and J. Nocedal. On the solution of equality constrained quadratic programming problems arising in optimization. *SIAM J. Sci. Comput.*, 23(4):1376–1395, 2001. doi: 10.1137/S1064827598345667.

[116] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. SIAM, 2011.

[117] M. Guidorzi. Partial regularity in non-linear elasticity. *Manuscripta Math.*, 107:25–41, 2002.

[118] N. Gundiah, M. B. Ratcliffe, and L. A. Pruitt. Determination of strain energy function for arterial elastin: Experiments using histology and mechanical tests. *J. Biomech.*, 40:586–594, 2007.

[119] N. Gundiah, M. B. Ratcliffe, and L. A. Pruitt. The biomechanics of arterial elastin. *J. Mech. Behav. Biomed. Mat.*, 2:288–296, 2009.

[120] A. Günnel. *Numerical Aspects in Optimal Control of Elasticity Models with Large Deformations*. PhD thesis, Technische Universität Chemnitz, 2014.

[121] M. Günther, O. Röhrle, D. F. B. Haeufle, and S. Schmitt. Spreading out muscle mass within a hill-type model: A computer simulation study. *Comp. Math. Method. M.*, 2012(ID 848630):1–13, 2012. doi: 10.1155/2012/848630.

[122] M. E. Gurtin. *An Introduction to Continuum Mechanics*. Academic Press, 1981.

[123] E. Guth. Muscular contraction and rubber elasticity. *Ann. N. Y. Acad. Sci.*, 47:715–766, 1947. doi: 10.1111/j.1749-6632.1947.tb31734.x.

[124] M. H. Gutknecht and S. Röllin. The Chebyshev iteration revisited. *Par. Comp.*, 28:263–283, 2002.

[125] A. C. Guyton and J. E. Hall. *Textbook of Medical Physiology*. Elsevier, 11$^{\text{th}}$ edition, 1956.

[126] C. Hamburger. Partial regularity of minimizers of polyconvex variational integrals. *Calc. Var.*, 18:221–241, 2003.

[127] J. Harbrecht and R. Schneider. On error estimation in finite element methods without having galerkin orthogonality. Berichtsreihe des SFB 611 Preprint 457, Universität Bonn, 2009.

[128] S. Hartmann and P. Neff. Polyconvexity of generalized polynomial-type hyperelastic strain energy functions for near-incompressibility. *Int. J. Solids Struct.*, 40:2767–2791, 2003. doi: 10.1016/S0020-7683(03)00086-6.

[129] R. C. Haut and R. W. Little. A constitutive equation for collagen fibers. *J. Biomech.*, 5:423–430, 1972.

[130] T. Heidlauf and O. Röhrle. Modeling the chemoelectromechanical behavior of skeletal muscle using the parallel open-source software library openCMISS. *Comp. Math. Method. Med.*, 2013(517287):1–14, 2013. doi: 10.1155/2013/517287.

[131] A. Henderson, J. Ahrens, and C. Law. *The ParaView Guide*. Kitware, 2004.

[132] F. M. Hendriks. *Mechanical behaviour of human epidermal and dermal layers in vivo*. PhD thesis, Technische Universiteit Eindhoven, 2005.

[133] M. R. Hestenes and E. Stiefel. Methods of conjugate of gradients for solving linear systems. *J. Res. Nat. Bur. Standards*, 49:409–436, 1952.

[134] A. P. Hills, E. M. Hennig, N. M. Byrne, and J. R. Steele. The biomechanics of adiposity – structural and functional limitations of obesity and implications for movement. *Obes. Rev.*, 3:35–43, 2002.

[135] G. A. Holzapfel. *Nonlinear solid mechanics*. Wiley, 2000.

[136] G. A. Holzapfel. *Handbook of Material Behavior: Nonlinear Models and Properties*, chapter "Biomechanics of Soft Tissue". Academic Press, 2001.

[137] G. A. Holzapfel. *Arterial Walls: Biomechanical Aspects*, chapter Collagen, pages 285–324. Springer, 2008.

[138] G. A. Holzapfel and R. W. Ogden. *Biomechanics of Soft Tissue in Cardiovascular Systems*. Springer, 2003.

[139] G. A. Holzapfel and R. W. Ogden. *Mechanics of biological tissue*. Springer, 2006.

[140] G. A. Holzapfel and H. W. Weizsäcker. Biomechanical behavior of the arterial wall and its numerical characterization. *Comp. Biol. Med.*, 1998.

[141] G. A. Holzapfel, T. C. Gasser, and R. W. Ogden. A new constitutive framework for arterial wall mechanics and a comparative study of material models. *J. Elas.*, 61:1–48, 2000.

[142] H. Holzmann, G. W. Korting, D. Kobelt, and H. G. Vogel. Prüfung der mechanischen Eigenschaften von menschlicher Haut in Abhängigkeit von Alter und Geschlecht. *Arch. Klin. exp. Derm.*, 239:355–367, 1971.

[143] F. C. Hoppensteadt and C. S. Peskin. *Mathematics in Medicine and Life Sciences*. Springer, 1992.

[144] C. O. Horgan and J. G. Murphy. On the volumetric part of strain energy functions used in the constitutive modeling of slightly compressible solid rubbers. *Int. J. Solids Struct.*, 46:3078–3085, 2009.

[145] L. Huang, N. Bakker, J. Kim, J. Marston, I. Grosse, J. Tis, and D. Cullinane. A multi-scale finite element model of bruising in soft connective tissues. *J. Forensic. Biomech.*, 3(235579):1–5, 2012. doi: 10.4303/jfb/235579.

[146] P. A. Huijing. Muscle as a collagen fiber reinforced composite: a review of force transmission in muscle and whole limb. *J. Biomech.*, 32:329–345, 1999.

[147] J. D. Humphrey. Continuum mechanics of soft biological tissues. *Proc. Roy. Soc.*, A(459):3–46, 2003. doi: 10.1098/rspa.2002.1060.

[148] J. D. Humphrey and F. C. Yin. Constitutive relations and finite deformations of passive cardiac tissue II: stress analysis in the left ventricle. *Circ. Res.*, 65: 805–817, 1989. doi: 10.1161/01.RES.65.3.805.

[149] J. D. Humphrey and F. D. Yin. On constitutive relations and finite deformations of passive cardiac tissue: I. pseudostrain-energy function. *ASME J. Biomech. Eng.*, 109:298–304, 1987.

[150] S. H. Hussain, B. Limthongkul, and T. R. Humphreys. The biomechanical properties of the skin. *Dermatol. Surg.*, 39:193–203, 2013.

[151] M. Itskow and N. Aksel. A class of orthotropic and transversely isotropic hyperelastic constitutive models based on a polyconvex strain energy function. *Int. J. Solids Struct.*, 41:3833–3848, 2004.

[152] M. Itskow, R. Dargazany, and K. Hörnes. Taylor expansion of the inverse function with application to the Langevin function. *Math. Mech. Solids*, pages 1–9, 2011. doi: 10.1177/1081286511429886.

[153] H. M. Jensen and J. Christoffersen. Kink band formation in fiber reinforced materials. *J. Mech. Phys. Solids*, 45(7):1121–1136, 1997.

[154] P. Jordàn. *Image-Based Mechanical Characterization of Soft Tissue using Three Dimensional Ultrasound*. PhD thesis, Harvad University, 2008.

[155] D. Kainmüller, T. Lange, and H. Lamecker. Shape constrained automatic segmentation of the liver based on a heuristic intensity model. In *Proc. MICCAI Workshop 3D Segmentation in the Clinic: A Grand Challenge*, pages 109–116, 2007.

[156] D. Kainmüller, H. Lamecker, and S. Zachow. Multi-object segmentation with coupled deformable models. *Annals of the BMVA*, (5):1–10, 2009.

[157] C. Keller, N. I. M. Gould, and A. J. Wathen. Constraint preconditioning for indefinite linear systems. *SIAM J. Matrix Anal. Appl.*, 21:1300–1317, 2000.

[158] B. F. Kennedy, X. Liang, S. G. Adie, D. K. Gerstmann, B. C. Quirk, S. A. Boppart, and D. D. Sampson. In vivo three-dimensional optical coherence elastography. *Opt. Express*, 19(7):6623–6634, 2011. doi: 10.1364/OE.19.006623.

[159] B. F. Kennedy, R. A. McLaughlin, K. M. Kennedy, L. Chin, A. Curatolo, A. Tien, B. Latham, C. M. Saunders, and D. D. Sampson. Optical coherence micro-elastography: mechanical-contrast imaging of tissue microstructure. *Biomed. Opt. Express*, 5(7):2113–2124, 2014. doi: 10.1364/BOE.5.002113.

[160] T. Khan, E. S. Muise, P. Iyengar, Z. V. Wang, M. Chandalia, N. Abate, B. B. Zhang, P. Bonaldo, S. Chua, and P. E. Scherer. Metabolic dysregulation and adipose tissue fibrosis: Role of collagen VI. *Mol. Cell. Biol.*, 29(6):1575–1591, 2009.

[161] J. K. Knowles and E. Sternberg. On the failure of ellipticity of the equations for finite elastostatic plane strain. *Archive for Rational Mechanics and Analysis*, 63:321–336, 1976. ISSN 0003-9527. doi: 10.1007/BF00279991. 10.1007/BF00279991.

[162] J. K. Knowles and E. Sternberg. On the failure of ellipticity and the emergence of discontinuous deformation gradients in plane finite elastostatics. *Journal of Elasticity*, 8:329–379, 1978. ISSN 0374-3535. doi: 10.1007/BF00049187. 10.1007/BF00049187.

[163] R. M. Koch. *Methods for Physics Based Facial Surgery Prediction*. PhD thesis, ETH Zürich, 2001.

[164] R. Kornhuber, R. Krause, O. Sander, P. Deuflhard, and S. Ertel. A monotone multigrid solver for two body contact problems in biomechanics. *Computing and Visualization in Science*, 11(1):3–15, 2008. ISSN 1432-9360. doi: 10.1007/s00791-006-0053-6.

[165] M. A. Krasnosel'skiĭ and Y. B. Rutickiĭ. *Convex Functions and Orlicz Spaces*. P. Noordhoff, 1961.

[166] J. Kristensen. On the non-locality of quasiconvexity. *Ann. de l'I.H.P.*, 16(1): 1–13, 1999.

[167] I. L. Kruglikov. General theory of body contouring: 2. modulation of mechanical properties of subcutaneous fat tissue. *J. Cosmet. Dermatol. Sci. Appl.*, 4: 117–127, 2014.

[168] K. Langer. On the anatomy and physiology of the skin. I. the cleavability of cutis. *Br. J. Plast. Surg.*, 31:3–8, 1978. Translated from Langer, K. (1861). Zur Anatomie und Physiologie der Haut. I. Über die Spaltbarkeit der Cutis, Sitzungsbericht der Mathematisch-naturwissenschaftlichen Classe der Kaiserlichen Academie der Wissenschaften, 44, 19.

[169] Y. Lanir. Mechanistic micro-structural theory of soft tissues growth and remodeling: tissues with unidirectional fibers. *Biomech. Model. Mechanobiol.*, pages 1–22, 2014. ISSN 1617-7959. doi: 10.1007/s10237-014-0600-x.

[170] J.-B. Le Cam and E. Toussaint. Cyclic volume changes in rubber. *Mech. Mater.*, 41:898–901, 2009.

[171] S. Lee, R. E. Caflisch, and Y.-J. Lee. Exact artificial boundary conditions for continuum and discrete elasticity. *SIAM J. Appl. Math.*, 66(5):1749–1775, 2006.

[172] S. Lehmich, P. Neff, and J. Lankeit. On the convexity of the function $c \mapsto f(\det(c))$ on positive-definite matrices. *Math. Mech. Solids*, 19(4):369–375, 2014. doi: 10.1177/1081286512466099.

[173] H. U. Lemke and L. Berliner. *Modellgestützte Therapie*, chapter Modellgestützte Therapie, patientenspezifisches Modell und modellbasiderte Evidenz, pages 13–24. Health Academy, 2008.

[174] K. Levi. *Biomechanics of human stratum corneum: dry skin conditions, tissue damage and alleviation*. PhD thesis, Stanford University, 2009.

[175] X. Liang and S. A. Boppart. Biomechanical properties of in vivo human skin from dynamic optical coherence elastography. *IEEE Trans. Biomed. Eng.*, 57 (4):953–959, 2010.

[176] X. Liang, V. Crecea, and S. A. Boppart. Dynamic optical coherence elastography: a review. *J. Innov. Opt. Health. Sci.*, 3(4):221–233, 2010. doi: 10.1142/S1793545810001180.

[177] J. Liesen and Z. Strakoš. *Krylov Subspace Methods: Principles and Analysis*. Oxford University Press, 2013.

[178] L. Lubkoll. Optimal control in implant shape design. Master's thesis, ZIB, TU Berlin, 2010.

[179] L. Lubkoll, A. Schiela, and M. Weiser. An optimal control problem in polyconvex hyperelasticity. *SIAM J. Control Optim.*, 52(3):1403–1422, 2014.

[180] L. Lubkoll, A. Schiela, and M. Weiser. An affine covariant composite step method for optimization with PDEs as equality constraints. *Preprint*, 2015.

[181] J. E. Marsden and J. R. Hughes. *Mathematical Foundations of Elasticity*. Prentice-Hall, 1983.

[182] J. A. C. Martins, E. B. Pires, R. Salvado, and P. B. Dinis. A numerical model of passive and active behavior of skeletal muscles. *Comp. Meth. Appl. Mech. Eng.*, 151:419–433, 1998.

[183] E. Mazza and G. G. Barbarino. 3d mechanical modeling of facial soft tissue for surgery simulation. *Facial Plast. Surg. Clin. North. Am.*, 19(4), 2011. doi: doi:10.1016/j.fsc.2011.07.006.

[184] J. Meixensberger. *Modellgestützte Therapie*, chapter Modellgestützte Therapie - Einfluss und Auswirkungen auf das Betätigungsfeld des Chirurgen, pages 271–277. Health Academy, 2008.

[185] J. Merodio. A note on tensile instabilities and loss of ellipticity for a fiber-reinforced nonlinearly elastic solid. *Arch. Mech.*, 58(3):293–303, 2006.

[186] J. Merodio and R. W. Ogden. Material instabilities in fiber-reinforced nonlinearly elastic solids under plane deformations. *Arch. Mech.*, 54(5-6):525–552, 2002.

[187] J. Merodio and T. J. Pence. Kink surfaces in a directional reinforced neo-Hookean material under plane deformation: II. Kink band stability and maximally dissipative broadening. *J. Elasticity*, 62:145–170, 2001.

[188] J. Merodio and T. J. Pence. Kink surfaces in a directional reinforced neo-hookean material under plane deformation: I. Mechanical equilibrium. *J. Elasticity*, 62:119–144, 2001.

[189] N. G. Meyers. Quasi-convexity and lower semi-continuity of multiple variational integrals of any order. *Trans. Amer. Mat. Soc.*, 119:125–149, 1965.

[190] A. Mielke. Necessary and sufficient conditions for polyconvexity of isotropic functions. *Journ. Conv. Anal.*, 12(2):291–314, 2005.

[191] G. Mingione. Regularity of minima: An invitation to the dark side of the calculus of variations. *Appl. Math.*, 51:355–425, 2006.

[192] R. J. Monti, R. R. Roy, J. A. Hodgson, and V. R. Edgerton. Transmission of forces within mammalian skeletal muscles. *J. Biomech.*, 32:371–380, 1999.

[193] M. Mooney. A theory of large elastic deformation. *Journ. App. Phys.*, 11(9): 582–592, 1940.

[194] C. B. Morrey. Quasi-convexity and the lower semicontinuity of multiple integrals. *Pacific J. Math.*, 2:25–53, 1952.

[195] D. A. Morrow, T. L. H. Donahue, G. M. Odegard, and K. R. Kaufman. Transversely isotropic tensile material properties of skeletal muscle tissue. *J. Mech. Behav. Biomed..*, 3(1):124–129, 2010. doi: 10.1016/j.jmbbm.2009.03.004.

[196] F. D. Murnaghan. *Finite deformation of an elastic solid*. Wiley, 1951.

[197] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 1999.

[198] C. A. Oatis. *Kinesiology: The Mechanics and Pathomechanics of Human Movement*, chapter Biomechanics of Skeletal Muscle, pages 45–68. Lippincott Williams & Wilkins, 2004.

[199] R. W. Ogden. Large deformation isotropic elasticity: on the correlation of theory and experiment for incompressible rubber-like solids. *Proc. Roy. Soc. London*, A(326):565–583, 1972.

[200] R. W. Ogden. Large deformation isotropic elasticity: on the correlation of theory and experiment for compressible rubber-like solids. *Proc. Roy. Soc. London*, A(328):567–583, 1972.

[201] R. W. Ogden. *Non-linear elastic deformations*. Dover, 1997.

[202] 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.

[203] J. S. Ovall. Two dangers to avoid when using gradient recovery methods for finite element error estimation and adaptivity. Technical Report 6, Max-Planck-Institut, Leipzig, 2006.

[204] J. S. Ovall. Function, gardient and hessian recovery using quadratic edge-bump functions. *SIAM J. Numer. Anal.*, 45(3):1064–1080, 2007.

[205] H. Oxlund, J. Manschot, and A. Viidik. The role of elastion in the mechanical properties of skin. *J. Biomech.*, 21(3):213–218, 1988.

[206] N. A. Papadakis and P. Y. Papalambros. Essentially quadratic penalty functions and the Maratos effect. Technical Report UM-MEAM-89-06, The University of Michigan, 1989.

[207] P. Pedregal. *Parametrized Measures and Variational Principles*. Birkhäuser, 1997.

[208] P. Pedregal. *Variational Methods in Nonlinear Elasticity*. SIAM, 2000.

[209] S. T. J. Peng and R. F. Landel. Stored energy function and compressibility of compressible rubberlike materials under large strain. *J. Appl. Phys.*, 46(6): 2599–2604, 1975.

[210] R. W. Penn. Volume changes accompanying the extension of rubber. *Trans. Soc. Rheol.*, 14(4):509–517, 1970.

[211] D. J. K. Rao, V. Malhotra, R. S. Batra, and A. Kukreja. Esthetic correction of depressed frontal bone fracture. *Natl. J. Maxillofac. Surg.*, 2(1):69–72, 2011.

[212] M. K. Rausch and E. Kuhl. On the effect of prestrain and residual stress in thin biological membranes. *J. Mech. Phys. Solids*, 61:1955–1969, 2013. doi: 10.1016/j.jmps.2013.04.005.

[213] M. Rebouah, G. Chagnon, and D. Favier. *Constitutive Models for Rubber VIII*, chapter Anisotropic modeling of the Mullins effect the residual strain of filled silicone rubber. CRC Press, 2013.

[214] M. Renardy and R. C. Rogers. *An Introduction to Partial Differential Equations*. Springer, 2nd edition, 1993.

[215] R. S. Rivlin. Large elastic deformations of isotropic materials. IV. further developments of the general theory. *Phil. Trans. Roy. Soc. London*, 241(835): 379–397, 1948.

[216] R. S. Rivlin. Large elastic deformations of isotropic materials. I. fundamental concepts. *Phil. Trans. Roy. Soc. London*, 240:459–490, 1948.

[217] O. Röhrle, J. B. Davidson, and A. J. Pullan. Bridging scales: A three-dimensional electromechanical finite element model of skeletal muscle. *SIAM J. Sci. Comput.*, 30(6):2882–2904, 2008. doi: 10.1137/070691504.

[218] O. Röhrle, J. B. Davidson, and A. J. Pullan. A physiologically based, multi-scale model of skeletal muscle structure and function. *Frontiers in Physiology*, 3(358):1–14, 2012.

[219] A. Rösch and D. Wachsmuth. A-posteriori error estimates for optimal control problems with state and control constraints. *Numer. Math.*, 120(4):733–762, 2010.

[220] C. S. Roy. The elastic properties of the arterial wall. *J. Physiol.*, 3(2):125–159, 1881.

[221] W. Rudin. *Functional Analysis*. McGraw-Hill, Inc, 2nd edition, 1973.

[222] Y. Saad. *Iterative Methods for Sparse Linear Systems*. PWS Publishing Company, 1996.

[223] O. Sander. A fast solver for finite deformation contact problems. Matheon Preprint 319, Freie Universität Berlin, 2006.

[224] A. Schiela. *The Control Reduced Interior Point Method*. PhD thesis, Freie Universität Berlin, 2006.

[225] A. Schiela. Barrier methods for optimal control problems with state constraints. *SIAM J. Optim.*, 20(2):1002–1031, 2009.

[226] A. Schiela. A flexible framework for regularization algorithms for non-convex optimization in function space. Technical report, Technische Universität Hamburg-Harburg, 2014.

[227] R. Schleip, I. L. Naylor, D. Ursu, W. Melzer, A. Zorn, H.-J. Wilke, F. Lehmann-Horn, and W. Klingler. Passive muscle stiffness may be influenced by active contractility of intramuscular connective tissue. *Med. Hypotheses*, 66:66–71, 2006. doi: 10.1016/j.mehy.2005.08.025.

[228] J. G. Schmidt, G. Berti, J. Fingberg, J. Cao, and G. Wollny. A finite element based tool chain for the planning and simulation of maxillo-facial surgery. In *ECCOMAS 2004*, 2004.

[229] J. Schöberl and W. Zulehner. Symmetric indefinite preconditioners for saddle point problems with applications to pde-constrained optimization problems. *SIAM J. Matrix Anal. Appl.*, 29(3):752–773, 2007. doi: 10.1137/060660977.

[230] J. Schröder and P. Neff. On the construction of polyconvex anisotropic free energy functions. In *Proceedings of the IUTAM Symposium on Computational Mechanics of Solid Materials at Large Strains*, pages 171–180, 2001.

[231] J. Schröder and P. Neff. Application of polyconvex anisotropic free energies to soft tissues. In *WCCM V Fifth World Congress on Computational Mechanics*, pages 1–8, 2002. Schröder, J.

[232] J. Schröder and P. Neff. Invariant formulation of hyperelastic transverse isotropy based on polyconvex free energy functions. *Int. J. Solids Struct.*, 40:401–445, 2003.

[233] J. Schröder, P. Neff, and D. Balzani. A variational approach for materially stable anisotropic hyperelasticity. *Int. J. Solids Struct.*, 42:4352–4371, 2005.

[234] J. Schröder, P. Neff, and V. Ebbing. Anisotropic polyconvex energies on the basis of crystallographic motivated structural tensors. *Journ. Mech. Phys. Sol.*, 56:3486–3506, 2008.

[235] R. E. Shadwick, A. P. Russell, and R. F. Lauff. The structure and mechanical design of rhinoceros dermal armour. *Phil. Trans. Roy. Soc. London*, 1282: 419–428, 1992.

[236] J. Shim, A. Grosberg, J. C. Nawroth, K. K. Parker, and K. Bertoldi. Modeling of cardiac thin films: Pre-stretch, passive and active behavior. *J. Biomech.*, 45:832–841, 2012. doi: 10.1016/j.jbiomech.2011.11.024.

[237] M. D. Shoulders and R. T. Raines. Collagen structure and stability. *Ann. Rev. Biochem.*, 78:929–958, 2009.

[238] H. S. Silver, J. W. Freeman, and D. DeVore. Viscoelastic properties of human skin and processed dermis. *Skin Res. Technol.*, 7:18–23, 2001.

[239] I. S. Sokolnikoff. *Mathematical Theory of Elasticity*. McGraw-Hill, 1946.

[240] G. Sommer, M. Eder, L. Kovacs, H. Pathak, L. Bonitz, C. Mueller, P. Regitnig, and G. A. Holzapfel. Multiaxial mechanical properties and constitutive modeling of human adipose tissue: A basis for preoperative simulations in plastic and reconstructive surgery. *Acta Biomater.*, 9:9036–9048, 2013.

[241] R. Sopakayang, R. De Vita, A. Kwansa, and J. W. Freeman. Elastic and viscoelastic properties of a type I collagen fiber. *J. Biomech.*, 293:197–205, 2012.

[242] D. Stalling, M. Westerhoff, and H.-C. Hege. *The Visualization Handbook*, chapter Amira: a highly interactive system for visual data analysis, pages 749–767. Elsevier, 2005.

[243] H. L. Stark. Directional variations in the extensibility of human skin. *Br. J. Plast. Surg.*, 30:105–114, 1977.

[244] D. J. Steigmann. Frame-invariant polyconvex strain-energy functions for some anisotropic solids. *Math. Mech. Solids*, 8:497–506, 2003.

[245] M. Stoll. *Solving Linear Systems using the Adjoint*. PhD thesis, University of Oxford, 2008.

[246] Z. Strakoš and J. Liesen. On numerical stability in large scale linear algebraic computations. *Z. Angew. Math. Mech.*, 85(5):307–325, 2005. doi: 10.1002/zamm.200410185.

[247] Z. Strakoš and P. Tichý. Error estimation in preconditioned conjugate gradients. *BIT Numer. Math.*, 45(4):789–817, 2005. ISSN 0006-3835. doi: 10.1007/s10543-005-0032-1.

[248] G. Strang. Inverse problems and derivatives of determinants. *Arch. Ration. Mech. Anal.*, 114:255–265, 1991.

[249] R. K. Strumpf, J. D. Humphrey, and F. C. P. Yin. Biaxial mechanical properties of passive and tetanized canine diaphragm. *Am. J. Physiol.*, 265:H469–475, 1993.

[250] C. Sun, B. Standish, and V. X. D. Yang. Optical coherence elastography: current status and future applications. *J. Biomed. Opt.*, 16(4):043001–1 – 043001–12, 2011. doi: 10.1117/1.3560294.

[251] L. Tartar. *The General Theory of Homogenization*. Springer, 2009.

[252] D. C. Taylor, J. D. Dalton, A. V. Seaber, and W. E. Garrett. Viscoelastic properties of muscle-tendon units: The biomechanical effects of stretching. *J. Sp. Med.*, 18(3):300–309, 1990.

[253] A. B. Tepole, A. K. Gosain, and E. Kuhl. Stretching skin: The physiological limit and beyond. *Int. J. Nonlinear Mech.*, 47:938–949, 2012. doi: 10.1016/j.ijnonlinmec.2011.07.006.

[254] G. J. Tortora and B. H. Derrickson. *Principles of Anatomy and Physiology*. Wiley, 6th edition, 1996.

[255] G. J. Tortora and B. H. Derrickson. *Anatomie und Physiologie*. Wiley-Blackwell, 2006.

[256] P. Trayhurn and J. H. Beattie. Physiological role of adipose tissue: white adipose tissue as an endocrine and secretory organ. *Proc. Nutr. Soc.*, 60: 329–339, 2001.

[257] P. Trayhurn, C. Bing, and I. S. Wood. Adipose tissue and adipokines - energy regulation from the human perspective. *J. Nutr.*, 136:1935S–1939S, 2006.

[258] R. T. Tregear and S. B. Marston. The crossbridge theory. *Ann. Rev. Physiol.*, 41:723–736, 1979.

[259] F. Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen: Theorie, Verfahren und Anwendungen*. Vieweg+Teubner, 2nd edition, 2005.

[260] U. Trottenberg, C. W. Oosterlee, and A. Schüller. *Multigrid*. Academic Press, 2001.

[261] F. Truesdell and W. Noll. *The Non-Linear Field Theories of Mechanics*. Springer, 2004.

[262] M. Ulbrich and S. Ulbrich. Automatic differentiation: A structure-exploiting forward mode with almost optimal complexity for kantorovič trees. Technical Report IAMS1996.1TUM, Technische Universität München, 1996.

[263] P. Valet, G. Tavernier, I. Castan-Laurell, J. S. Saulnier-Blache, and D. Langin. Understanding adipose tissue development from transgenic animal models. *J. Lipid. Res.*, 43:835–860, 2002.

[264] P. Vandewalle, F. Schutyser, J. Van Cleynenbreugel, and P. Suetens. Modelling of facial soft tissue growth for maxillofacial surgery planning environments. In N. Ayache and H. Delingette, editors, *Surgery Simulation and Soft Tissue Modeling*, volume 2673 of *Lecture Notes in Computer Science*, pages 27–37. Springer Berlin Heidelberg, 2003. ISBN 978-3-540-40439-2. doi: 10.1007/3-540-45015-7_3.

[265] A. Vardi. A trust region algorithm for equality constrained minimization: convergence properties and implementation. *SIAM J. Numer. Anal.*, 22(3): 575–591, 1985. ISSN 0036-1429. doi: 10.1137/0722035.

[266] R. S. Varga. A comparison of the successive overrelaxation method and semi-iterative methods using Chevyshev polynomials. *J. Soc. Indust. Appl. Math*, 5:39–46, 1957.

[267] M. Vasta, A. Pandolfi, and A. Gizzi. A fiber distributed model of biological tissues. In *Procedia IUTAM*, volume 6, pages 79–86, 2013.

[268] R. Verfürth. *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*. Wiley-Teubner, 1996.

[269] B. Vexler and W. Wollner. Adaptive finite elements for elliptic problems with control constraints. *SIAM J. Control Optim.*, 47(1):509–534, 2008.

[270] K. Y. Volokh. Stresses in growing soft tissues. *Acta Biomater.*, 2:493–504, 2006. doi: 10.1016/j.actbio.2006.04.002.

[271] V. Šverák. Rank-one convexity does not imply quasiconvexity. *Proc. Roy. Soc. Edinburgh*, A(120):185–189, 1992.

[272] B. L. Wajchenberg. Subcutaneous and visceral adipose tissue: Their relation to the metabolic system. *Endocr. Rev.*, 21(6):697–738, 2000.

[273] A. Wathen. Realistic eigenvalue bounds for the galerkin mass matrix. *IMA J. Numer. Anal.*, 7(4):449–457, 1987.

[274] A. Wathen and T. Rees. Chebyshev semi-iteration in preconditioning for problems including the mass matrix. *ETNA. Electr. Trans. Numer. Anal.*, 34: 125–135, 2008.

[275] P. N. Watton, Y. Ventikos, and G. A. Holzapfel. Modelling the mechanical response of elastin for arterial tissue. *J. Biomech.*, 42:1320–1325, 2009.

[276] H. Weisbecker, C. Viertler, D. M. Pierce, and G. A. Holzapfel. The role of elastin and collagen in the softening behavior of the human thoracic aortic media. *J. Biomech.*, 46:1859–1865, 2013.

[277] M. Weiser, P. Deuflhard, and B. Erdmann. Affine conjugate adaptive newton methods for nonlinear elastomechanics. *Opt. Meth. Softw.*, 22(3):414–431, 2007.

[278] A. S. Weiss. The science of elastin. Technical report, Elastagen Pty. Ltd., 2011. URL `http://www.elastagen.com/media/The_Science_of_Elastin.pdf`.

[279] D. Werner. *Funktionalanalysis*. Springer, 6th edition, 1995.

[280] M. G. Wertheim. Mémoire sur l'eastocote et la cohésion des principaux tissues du corp humain. *Ann. Chim. Phys.*, 21:385–414, 1847.

[281] F. Xu and T. Lu. *Introduction to Skin Biothermomechanics and Thermal Pain*, chapter Skin Biomechanics Modeling, pages 154–206. Springer and Science Press Beijing, 2011.

[282] S. Zachow. *Computergestützte 3D Osteotomieplanung in der Mund-Kiefer-Gesichtschirurgie unter Berücksichtigung der räumlichen Weichgewebeanordnung*. PhD thesis, Technische Universität Berlin, 2005.

[283] S. Zachow, M. Weiser, and P. Deuflhard. *Modellgestützte Therapie*, chapter Modellgestützte Operationsplanung in der Kopfchirurgie, pages 140–156. Health Academy, 2008.

[284] G. I. Zahalak. Non-axial muscle stress and stiffness. *J. theor. Biol.*, 182:59–84, 1996.

[285] G. I. Zahalak. The effects of cross-fiber deformation on axial fiber stress in myocardium. *J. Biomech. Eng.*, 121(4):376–385, 1999.

[286] E. Zeidler. *Nonlinear Functional Analysis and its Applications*, volume I: Fixed-Point Theorems. Springer New York, 1986.

[287] J. Ziems. *Adaptive Multilevel SQP-Methods for PDE-constrained Optimization*. PhD thesis, Technische Universität Darmstadt, 2010.

[288] A. M. Zöllner, A. B. Tepole, and E. Kuhl. On the biomechanics and mechanobiology of growing skin. *J. Theor. Biol.*, 297:166–175, 2012. doi: 10.1016/j.jtbi.2011.12.022.

[289] G. W. Zumbusch. Symmetric hierarchical polynomials and the adaptive h-p-version. In *ICOSAHOM'95*, 1996.

Nomenclature

Abbreviations

CG	Conjugate Gradient Method
CT	Computer Tomography
DICOM	Digital Imaging and Communications in Medicine
DLY	Deuflhard, Leinen, Yserentant (Error Estimator)
DWR	Dual Weighted Residual Method
ECM	Extracellular Matrix
FE	Finite Element
FMG	Full Multigrid Scheme
GMRES	Generalized Minimal Residual Method
HCG	Hybrid Conjugate Gradient Method
IGT	Image Guided Therapy
KKT	Karush-Kuhn-Tucker (Conditions)
MBT	Model Based Therapy
MINRES	Minimal Residual Method
MRT	Magnetic Resonance Tomography
OCE	Optical Coherence Elastography
PDE	Partial Differential Equation
PPCG	Projected Preconditioned Conjugate Gradient Method
PPCG	Projected Preconditioned Conjugate Gradient Method
RCG	Regularized Conjugate Gradient Method

SQP	Sequential Quadratic Programming
SSC	Second Order Sufficient Optimality Condition
TCG	Truncated Conjugate Gradient Method
TIMMS	Therapy Imaging and Model Management System

Notation

Φ	Set of Admissible Deformations
co	Convex Hull
φ	Deformation
u	Displacement (Elasticity) / Control (Optimal Control)
C	Left Cauchy-Green Strain Tensor
E	Strain Tensor
∇^s	Linearized Strain Tensor (Symmetric Gradient)
I	Unit Matrix
$\mathcal{E}^{\text{ext}}$	Energy Functional Associated with External Forces
$\mathcal{E}^{\text{str}}$	Strain Energy Functional
$\mathcal{E}$	(Full) Energy Functional
W	Stored Energy Function
id	Identity
χ	Indicator Function
$\mathcal{L}$	Lagrange Function
ds	Surface Measure
λ_L, μ_L	Lamé Constants
ν_L	Poisson Ratio
E_L	Young's Modulus
$\mathbb{M}^{m,n}$	Space of $m \times n$-Matrices
$\mathbb{M}^n$	Space of $n \times n$-Matrices
$\mathbb{M}^n_+$	Set of Orientation-Preserving Square Matrices
$\mathbb{S}^n$	Space of Symmetric $n \times n$-Matrices
$\mathbb{S}^n_+$	Set of Symmetric Orientation-Preserving $n \times n$-Matrices

$\mathbb{O}^n$	Set of Orthogonal $n \times n$-Matrices
$\mathbb{O}^n_+$	Set of Orthogonal Orientation Preserving $n \times n$-Matrices
μ	Lebesgue Measure
σ	First Piola-Kirchhoff Stress Tensor
Σ	Second Piola-Kirchhoff Stress Tensor
P_T	Piola Transform of a Tensor T
Q	Preconditioner
$\iota_1, \iota_2, \iota_3$	Principal Strain Invariants
$\iota_4, \iota_5, \iota_6$	Mixed Strain Invariants
$\bar{\iota}_1, \bar{\iota}_2, \bar{\iota}_3$	Modified Principal Strain Invariants
$\bar{\iota}_4, \bar{\iota}_5, \bar{\iota}_6$	Modified Mixed Strain Invariants
$\mathcal{R}(f)$	Range of f
$\mathbb{R}$	Space of Real Numbers
$\mathbb{R}_+$	Set of Nonnegative Real Numbers
Π_k	Space of Polynomials of Order k
$\mathrm{C}(\Omega)$	Space of Continuous Functions
$\mathrm{L}^p_\nu(\Omega)$	Lebesgue Space with Respect to Measure ν
$\mathrm{L}^p(\Omega)$	Lebesgue Space
$\mathrm{W}^{1,p}(\Omega)$	Sobolev Space
T_{def}	Cauchy stress tensor
tr	Trace of a Matrix
adj	Adjugate Matrix
cof	Cofactor Matrix
det	Determinant