SIAM J. Control Optim.
Vol. 52, No. 3, pp. 1403–1422

AN OPTIMAL CONTROL PROBLEM IN POLYCONVEX HYPERELASTICITY[*]

LARS LUBKOLL[†], ANTON SCHIELA[‡], and MARTIN WEISER[†]

Abstract. We consider an implant shape design problem arising in the context of facial surgery. The aim is to find the shape of an implant that deforms the soft tissue of the skin in a desired way. Assuming sufficient regularity, we introduce a reformulation as an optimal control problem where the control acts as a boundary force. The solution of that problem can be used to recover the implant shape from the optimal state. For a simplified problem, in the case where the state can be modeled as a minimizer of a polyconvex hyperelastic energy functional, we show existence of optimal solutions and derive—on a formal level—first order optimality conditions. Finally, preliminary numerical results are presented for the original optimal control formulation.

Key words. polyconvex elasticity, implant design, optimal control

AMS subject classifications. 49J20, 74B20, 65N21, 65N30

DOI. 10.1137/120876629

1. Introduction.

Facial bones of patients with severe trauma or congenital malformations are often partially replaced or augmented by implants. In contrast to load-bearing implants such as knee or hip joint prostheses, facial bone implants have a direct impact on the visual appearance of the patient and hence on social acceptance. An important criterion for the shape design of facial bone implants is therefore—besides a restoration of functionality—the resulting shape of the face.

For load-bearing and dental implants, mathematical topology and shape optimization is quite common; see, e.g., [43, 19, 41]. In contrast, shape optimization for facial bone implants has barely been considered [21]. Current medical practice is a manual selection of the implant's shape, which renders the postoperative outcome very much dependent on the experience of the surgeon. An algorithmic inverse approach deriving the implant shape from a desired result could improve this situation.

In the present work we consider this implant shape design problem. The facial appearance is determined by the soft tissue deformation due to the implant. This induces two major challenges: First, the deformation is, in principle, an obstacle problem such that the optimization leads to a complex mathematical program with equilibrium constraints (MPEC). Second, the relatively large deformations that the soft tissue undergoes incur nonlinear elastomechanics, which is notoriously known for its resistance to analytical treatment.

A simplified approach was proposed in [21], where a linearly elastic soft tissue problem, with the facial surface fixed at the desired shape and natural boundary conditions at the tissue-implant boundary, was solved. The implant shape is then determined retrospectively from the soft tissue displacement at the tissue-implant boundary. While easy and practical, this approach neglects material and geometric nonlinearities as well as the fact that, in the real-life situation, very different bound-

[*]Received by the editors May 10, 2012; accepted for publication (in revised form) November 12, 2013; published electronically May 1, 2014. This research was supported by the DFG Research Center Matheon "Mathematics for key technologies".
http://www.siam.org/journals/sicon/52-3/87662.html

[†]Zuse Institute Berlin, Berlin 14195, Germany (lubkoll@zib.de, weiser@zib.de).
[‡]Technische Universitat Berlin, Berlin 10623, Germany (schiela@math.tu-berlin.de).

ary conditions hold. In particular, the normal stress vanishes at the tissue-implant boundary but not on the skin surface, whereas in reality it is just the opposite.

In the current work, we consider a more realistic approach. In section 2 we reformulate the original MPEC as a simpler but equivalent control constrained optimal control problem with the normal force exerted by the implant acting as a control variable. Here too, the implant shape will be recovered from the soft tissue deformation. Existence of an optimal solution is treated in section 3. Due to missing boundary regularity, existence can only be shown for a slightly simplified problem with dead loads. We focus the discussion on the widely accepted class of polyconvex hyperelastic materials, which allow us to transfer Ball's elegant existence proof for deformations [6] to optimal solutions of the simplified shape design problem. To the best of our knowledge, this is the first rigorous existence result in optimal control of polyconvex elasticity.

Sections 4 and 5 are devoted to a formal derivation of first order optimality conditions, illustrated for a compressible Mooney–Rivlin material. In general, hyperelastic theory it is not even clear whether a local minimizer of the elastic energy functional satisfies the weak form of the corresponding Euler–Lagrange equation (see [8, Problems 5 and 6] and [14, 31]). In section 4 we discuss conditions under which this is the case, and the energy functional is well behaved, locally. However, the rigorous derivation of first order optimality conditions to our optimal control problem currently appears to be out of reach and can only be performed in a formal way. The main reason is, as always, the lack of rigorous regularity results for practically relevant configurations. Despite being related to the regularity of minimizers [7, 27, 28], polyconvexity and coercivity are not sufficient for the determination of such results. Finally, a very preliminary numerical example for the original shape design problem is presented in section 6.

Notation. By $\mathbb{M}^n$ we denote a set of $n \times n$ matrices equipped with the scalar product $F : G = \sum_{i,j} F_{ij}G_{ij}$ inducing the Euclidean norm $\|\cdot\|_M$. For invertible matrices $F \in \mathbb{M}^n$ the adjugate matrix is defined as $\operatorname{adj}(F) = \det(F)F^{-T}$. By $\mathbb{M}^n_+ := \{A \in \mathbb{M}^n \mid \det(A) > 0\}$ we denote the cone of orientation-preserving deformation gradients. Derivatives will be indicated by subscripts throughout, i.e., $\mathcal{E}_u = \frac{\partial}{\partial u}\mathcal{E}$.

2. Modeling.

In this section, we will derive from medical requirements a precise mathematical formulation of the implant shape design problem. First, we give a short summary of the properties that must be reflected by material-specific stress-strain relationships in biomechanics. Then, in section 2.2, we formulate the forward problem of finding the facial shape induced by a given implant shape as an obstacle problem. In section 2.3 we will see that the direct transcription of the forward problem into the inverse problem of finding an optimal implant shape so as to approximate a desired facial shape leads to a quite difficult optimization problem. Surprisingly, a simple reformulation turns out to be a standard optimal control problem.

2.1. Soft tissue biomechanics.

Stable configurations of biological soft tissues are, in general, characterized as minimizers of an energy functional

$$(1) \qquad \mathcal{E}^S(u) = \int_\Omega W(x, (I + \nabla u)(x))\,\mathrm{d}x = \int_\Omega W(x, \nabla\varphi(x))\,\mathrm{d}x,$$

where W is the stored energy function, u is the displacement, and $\varphi = I + u$ is the deformation. As soft tissues undergo comparatively large deformations, the energy functional $\mathcal{E}^S$ is nonquadratic for several reasons.

First, there is the geometric nonlinearity that describes the relation between deformation and the resulting strains. Its neglection may lead to overestimation of displacements (for an illustration, see Figure 4 in [45]). Another nonlinearity that is related only to geometry results from pressure boundary conditions imposed on the deformed domain. Using the Piola-transform in order to express this type of boundary condition on the undeformed domain implies a nonlinear transformation of the surface normal vectors which enters the problem formulation [6, 14] (see section 2.3).

Then there are the material-dependent constitutive nonlinearities that possibly must be taken into account. This is, to a great extent, a consequence of the distribution of collagen in most types of human soft tissue. As collagen is the main load-bearing element and the most common protein in human soft tissue, with a particularly high concentration in the skin and, in contrast to other muscles, the facial muscle tissue, it strongly determines the material behavior [20, 24, 25]. On the one hand, the collagen distribution leads to a nonlinear stress-strain relationship, which is mainly dependent on the collagen fiber morphology corresponding to the current stress state. This observation is outlined in [24] and reflected by Fung-type material laws [20]. On the other hand, the distribution of the collagen fibers endows the material with directional properties; i.e., while the stiffness increases with muscle contraction in direction of the collagen fibers, it remains constant in orthogonal directions [13, 24], thus leading to a strongly anisotropic behavior. This is complemented by the observation that these fiber directions may change during the deformation. Thus the accurate modeling of anisotropic effects is not trivial and requires the knowledge of collagen fiber orientations and distributions in the considered tissues.

There also is a constitutive nonlinear inequality that is associated with limited compressibility and takes the form

$$\det(\nabla\varphi(x)) = \det((I + \nabla u)(x)) > 0. \tag{2}$$

In the case of $\varphi \in C^1$, this inequality serves as a local "orientation-preserving" condition that locally prevents self-penetration of the considered material (see [14] and references therein).

Currently the most general class of stationary material laws that can incorporate the mentioned nonlinearities and is accessible to mathematical analysis are hyperelastic constitutive laws given by polyconvex stored energy functions [6]. This class, which will be considered in this paper, includes popular material laws for large strains such as neo-Hookean, Mooney–Rivlin [33, 38], Ogden-type [36], and Arruda–Boyce [3, 23] as well as carefully designed, possibly anisotropic, Fung-type material laws as in [20, 18, 42, 24, 10, 9]. As polyconvexity is closely related to the Legendre–Hadamard condition [16, 34], and thus to the stability of linearizations of the nonlinear problem, and does not impose nonphysical restrictions on models for biological soft tissues, it is widely accepted in both mathematical [14, 31, 16, 34, 32] and biomechanics [42, 23, 46, 18, 10, 9, 26] communities. Modern constitutive relations are directly constructed such that polyconvexity is guaranteed a priori.

2.2. Forward problem: Implant obstacle. The facial shape is determined by the elastic deformation of the soft tissue. In contrast, bone and implant are considered as rigid such that only the soft tissue domain Ω is considered.

In order to describe an implant's influence on the soft tissue, we restrict the discussion to implants of limited geometric complexity; i.e., we assume its manifold shape to be parametrized over Γ_c as continuous normal displacement

$$y \mapsto y + s(y)n(y) \quad \text{for } y \in \Gamma_c, \tag{3}$$

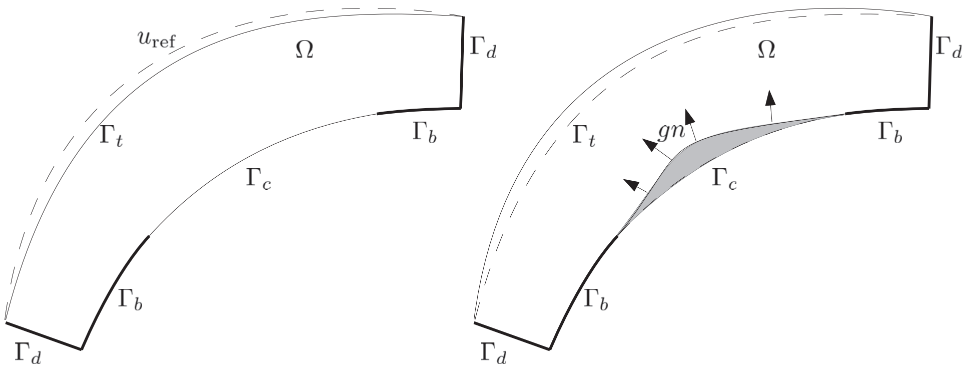


FIG. 1. *Cross-section of the reference configuration (left) and the deformed state due to the normal force gn defining the implant volume in gray (right).*

where $n(y)$ is the unit outer normal of Ω at $y \in \Gamma_c$ and Γ_c is the part of the interior soft tissue boundary where it normally is in contact with bone; see Figure 1. The implant displaces the soft tissue, which can freely glide over the implant surface but may not penetrate it. Hence, an obstacle condition has to be imposed on Γ_c. In a ring Γ_b around the implant region Γ_c we assume the soft tissue to be attached to the bone. Due to the quickly vanishing Green's function of elastomechanics, the soft tissue domain may be restricted to a bounded region in the vicinity of the implant by introducing an artificial boundary Γ_d cutting the soft tissue. Here, transparent boundary conditions [29] might be imposed. For simplicity, we just assume the tissue to be fixed on Γ_d. On the skin surface Γ_t, natural boundary conditions hold.

Thus with (1) the forward problem leads to an optimization problem subject to constraints given by the boundary conditions

$$\min_u \mathcal{E}^S(u) \tag{4a}$$

subject to

$$u = 0 \qquad \text{on } \Gamma_d \cup \Gamma_b, \tag{4b}$$

$$n(y)^T[x + u(x) - y] \geq s(y) \qquad \text{for all } x, y \in \Gamma_c \text{ with } x + u(x) \in y + \mathbb{R}\, n(y). \tag{4c}$$

In particular, the global nonpenetration condition (4c) is difficult to address algorithmically, as a direct mapping from y to x in Γ_c depends on the solution, is potentially multivalued, and is usually not readily available.

This problem can also be written in strong form if we introduce the first Piola–Kirchhoff tensor $\sigma(u) = \tilde{\sigma}(\nabla u)$, using the definition of hyperelasticity, i.e., the pointwise relation

$$\tilde{\sigma}(F) = \frac{\partial W}{\partial F}(x, F), \quad x \in \Omega,\ F \in \mathbb{M}^3_+\,. \tag{5}$$

Then we obtain as usual by formal partial integration,

$$-\operatorname{div}(\sigma(u)) = 0 \qquad \text{in } \Omega, \tag{6a}$$

$$u = 0 \qquad \text{on } \Gamma_b \cup \Gamma_d, \tag{6b}$$

$$n(y)^T[x + u(x) - y] \geq s(y) \qquad \text{for all } x, y \in \Gamma_c \text{ with } x + u(x) \in y + \mathbb{R}\, n(y). \tag{6c}$$

2.3. Inverse problem: Choice of design variable. Now the optimization problem consists of finding an implant shape given by the normal displacement $s(y)$ such that a desired facial shape is well approximated. Again for simplicity, we will consider the mismatch

$$J_0(u) = \frac{1}{2}\|u - u_{\mathrm{ref}}\|^2_{L^2(\Gamma_t)}$$

of displacement u and a desired displacement u_{ref} on the facial surface Γ_t, which is to be minimized subject to the obstacle problem (6). This formulation of the optimization problem as an MPEC has two mathematical drawbacks: it is algorithmically challenging, and the solutions are in general not unique (not even locally).

Theorem 5.3-1 in [14], as stated below in a simplified version, allows us to reformulate the MPEC as an optimal control problem with unilateral control constraints.

THEOREM 2.1. *Let Ω be a domain in $\mathbb{R}^3$, and let Γ_d, Γ_c be disjoint relatively open subsets of $\Gamma = \partial\Omega$ such that $\mathrm{vol}(\{\Gamma - (\Gamma_d \cup \Gamma_c)\}) = 0$ and $\mathrm{vol}(\Gamma_c) > 0$. Let the set of admissible solutions be of the form*

$$\Phi = \{\psi : \bar{\Omega} \to \mathbb{R}^3 |\ \det(\nabla\psi) > 0 \text{ in } \bar{\Omega}, \qquad \psi = \varphi_0 \text{ on } \Gamma_d, \psi(\Gamma_c) \subset C\},$$

where C is a given closed subset of $\mathbb{R}^3$, and let the total energy be defined by

$$I(\psi) = \int_\Omega W(\nabla\psi)\ \mathrm{d}x\,.$$

A smooth enough solution φ of the minimization problem

$$I(\varphi) = \inf_{\psi\in\Phi}\ I(\psi)$$

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

$$\begin{aligned}
-\operatorname{div}(\sigma(\nabla\varphi)) &= 0 \text{ in } \Omega,\\
\varphi &= \varphi_0 \text{ on } \Gamma_d,\\
\varphi(\Gamma_c) &\subset C,\\
\sigma(\nabla\varphi(x))n(x) &= 0 \text{ if } x \in \Gamma_c \text{ and } \varphi(x) \in \mathring{C},\\
\sigma(\nabla\varphi(x))n(x) &= g(x)\operatorname{adj}(x + \nabla u(x))^T n(x) \text{ if } x \in \Gamma_c \text{ and } \varphi(x) \in \partial C
\end{aligned}$$

with $g(x) \le 0$.

Note that the inequality $g \le 0$ corresponds to the implant not being able to pull the soft tissue. The lack of rigorous results on the regularity of solutions makes the statement in this theorem a formal one.

Remark 2.1. The boundary condition on Γ_c can be expressed as a boundary condition on the deformed boundary $\varphi(\Gamma_c)$ [14, p. 214]:

$$\sigma^\varphi(\varphi(x))n^\varphi(\varphi(x)) = 0 \text{ if } x \in \mathring{C}, \tag{7}$$

$$\sigma^\varphi(\varphi(x))n^\varphi(\varphi(x)) = g^\varphi(\varphi(x))n^\varphi(\varphi(x)) \text{ if } x \in \partial C \tag{8}$$

with $g^\varphi(x) \le 0$. Thus "the unilateral boundary condition of place on Γ_c constitutes a model of contact without friction with the obstacle ∂C. In this respect, the function $g^\varphi :\ \varphi(\Gamma_c) \to \mathbb{R}$, which measures the intensity of the contact load, is nothing but the Kuhn–Tucker multiplier associated with the constraint $\varphi(\Gamma_c) \subset C$" [14, p. 214].

Using the normal force g as the control variable instead of the obstacle shape, we obtain a control constrained optimization problem subject to the following equations of elasticity in strong form:

$$\min_{u,g} J(u,g) \tag{9a}$$

$$\text{subject to} \quad -\operatorname{div}(\sigma(u)) = 0 \quad \text{in } \Omega, \tag{9b}$$

$$u = 0 \quad \text{on } \Gamma_b \cup \Gamma_d, \tag{9c}$$

$$\sigma(u)n = -g \operatorname{adj}(I + \nabla u)^T n, \;\; g \geq 0 \quad \text{on } \Gamma_c. \tag{9d}$$

Due to the change of control variable from normal displacement s to normal force g, an explicit mapping between different points in Γ_c as required in (4c) is no longer needed. Eventually, from an optimal soft tissue displacement u solving (9), an implant shape can be reconstructed by filling the gap between the reference and deformed inner soft tissue boundary. Again it is parametrized over Γ_c, but now in the form

$$x \mapsto x + u(x) \quad \text{for } x \in \Gamma_c. \tag{10}$$

In this formulation, it is easy to satisfy additional medical requirements. For instance, no gaps should occur between the soft tissue and the implants since voids tend to be a source of infection. The retrospective construction of the implant shape by (10) obviously meets this requirement. In contrast, imposing it in the obstacle formulation (6) is quite involved.

Moreover, stresses large enough to damage the soft tissue or hurt the patient are undesirable. As, at least heuristically, the occurring stresses are essentially determined by the forces exerted on the soft tissue, we add a simple control penalization term to the cost functional:

$$J(u,g) := \frac{1}{2}\|u - u_{\text{ref}}\|^2_{L^2(\Gamma_t)} + \frac{\alpha}{2}\|g\|^2_{L^2(\Gamma_c)}. \tag{11}$$

Note that the applied penalization coincides with the well-known Tikhonov regularization for inverse problems.

In the following, for the sake of clarity, we will denote the parts of the boundary where homogeneous Dirichlet boundary conditions are imposed by $\Gamma_d = \Gamma_d \cup \Gamma_b$.

2.4. Simplification via dead loads. Although the problem (9) is already much more comprehensible than the original model, it is still untractable from an analytic point of view. The main difficulty lies in the boundary condition (9d). If we consider g as fixed, then it is not clear whether the corresponding boundary value problem has a solution u. For this, one would need a corresponding hyperelastic energy functional for which u is a minimizer.

Thus, of necessity, there must be an energy functional that corresponds to the equilibrium of forces, imposed at the boundary, which reads in our case as

$$\sigma n = -g \operatorname{adj}(I + \nabla u)^T n. \tag{12}$$

Unfortunately, except for the case of spatially constant control $g(x) = \mathit{const}$ on Γ_c (so-called pressure boundary conditions), a conservative formulation of these boundary conditions is in general not available [6, 12], leaving it as an open issue to model these conditions correctly.

For this reason, we will switch to a simplified setting; namely, we will replace (12) by one of the following two dead load boundary conditions:

$$\begin{aligned} &\sigma n = -gn, \quad g : \Gamma_c \to \mathbb{R}, \\ &\sigma n = -g, \quad g : \Gamma_c \to \mathbb{R}^3. \end{aligned} \tag{13}$$

Both conditions naturally enter linearly into the energy functional (see (17)) and can be augmented by a positivity constraint such as $g \geq 0$ in the first case or $g^T n \geq 0$ in the second case.

A comparison of (12) and (13) shows that our simplification is reasonable if $\mathrm{adj}(I + \nabla u) \approx I$. In this case, from a practical point of view, one can expect that a solution of the simplified problem will yield an implant form that is suboptimal with respect to the original problem but still reasonable. Of course, if dead loads are assumed, and the new implant form is reconstructed by the computed displacements, as above, then the computed soft tissue and the computed implant will not be in equilibrium physically because (13) and not the required (12) holds.

3. Existence of solutions. Our first step in the analysis of problem (9) is the study of existence of optimal controls g and corresponding deformations u for our simplified problem with dead loads.

In the context of nonlinear elasticity this is already a delicate issue, since there is hardly more analytical structure available than polyconvexity of the stored energy function and thus weak lower semicontinuity of the energy functional [6, 14, 31, 37]. To render the discussion precise, we state a list of standard assumptions in nonlinear elasticity that will be used throughout this paper.

Assumption 3.1.

1. Ω is a bounded Lipschitz-domain, and $\partial\Omega = \overline{\Gamma_d \cup \Gamma_t \cup \Gamma_c}$, $|\Gamma_c| > 0, |\Gamma_t| > 0$, is a measurable partition of its boundary.
2. The space of admissible deformations is contained in
$$U := \{u \in \mathbf{W}^{1,p}(\Omega) :\ \mathrm{adj}(I + \nabla u) \in \mathbf{L}^q(\Omega),\ \det(I + \nabla u) \in L^r(\Omega)\},$$
where $p \geq 2$, $q \geq p/(p-1)$, and $r > 1$.
3. On Γ_d, Dirichlet boundary conditions are imposed:
$$u|_{\Gamma_d} = u_0 \in \mathbf{W}^{\frac{p-1}{p},p}(\Gamma_d).$$
4. The stored energy function $W : \Omega \times \mathbb{M}^3 \to \mathbb{R} \cup \{+\infty\}$ exhibits the following properties.

   Polyconvexity. For almost all $x \in \Omega$ there exists a convex lower semicontinuous function
$$\mathbb{W}(x, \cdot, \cdot, \cdot) : \mathbb{M}^3 \times \mathbb{M}^3 \times \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$$
such that
$$\mathbb{W}\left(x, F, \mathrm{adj}(F), \det(F)\right) = W\left(x, F\right) \quad \text{for all } F \in \mathbb{M}^3_+,$$
and
$$\mathbb{W}(\cdot, F, H, \delta) :\ \Omega \to \mathbb{R} \cup \{+\infty\}$$
is measurable for all $(F, H, \delta) \in \mathbb{M}^3 \times \mathbb{M}^3 \times]0, \infty[$.

Nonself penetration. For almost all $x \in \Omega$ it holds that

$$\lim_{\det(F)\to 0^+} W(x,F) = +\infty \tag{14}$$

and

$$W(x,F) = +\infty \quad \text{for all } F \in \mathbb{M}^3 \setminus \mathbb{M}^3_+ \,. \tag{15}$$

Coercivity. There exist constants $\alpha > 0,\ \beta \in \mathbb{R}^3$ such that

$$W(x,F) \geq \alpha\left(\|F\|_M^p + \|\operatorname{adj}(F)\|_M^q + |\det(F)|^r\right) + \beta \tag{16}$$

for all $F \in \mathbb{M}^3_+$ and almost all $x \in \Omega$.

The elastic strain energy is given by

$$\mathcal{E}^S(u) = \int_\Omega W\left(x, I + \nabla u(x)\right) \mathrm{d}x,$$

and there exists at least one admissible deformation $\overline{u}$ such that $\mathcal{E}^S(\overline{u}) < \infty$.

Remark 3.1. While the usual definition of polyconvexity considers $\operatorname{dom}(W) = \mathbb{M}^3_+$ and $\operatorname{ran}(W) = \mathbb{R}_+$ we chose an equivalent formulation that allows the use of the extended real numbers in the image space. This reduces the length of the following proofs, as we do not have to examine the orientation-preserving condition $\det(I + \nabla u) > 0$ a.e. explicitly. Instead, this property is a direct consequence of the assumption $\mathcal{E}(u,g) < \infty$.

The above assumptions do not impose nonphysical restrictions (as long as we choose meaningful material parameters). Thus classical models for rubber-type materials such as neo-Hookean, Mooney–Rivlin [33, 38], Ogden-type [36], and Arruda–Boyce [3, 23] models satisfy these assumptions. Modern constitutive relations for biological soft tissues are constructed such that violation of these assumptions can be ruled out a priori [9, 10, 18, 23, 32, 42].

In view of section 2 we will impose the following assumptions on the control and the objective functional.

Assumption 3.2.

1. The control g is taken to be an element of $G = \mathbf{L}^2(\Gamma_c)$ and enters the total elastic energy functional via

$$\mathcal{E}(u,g) = \mathcal{E}^S(u) - \mathcal{E}^{\Gamma_c}(u,g) \tag{17}$$

with $\mathcal{E}^{\Gamma_c}(u,g) = \int_{\Gamma_c} g(s)u(s)\,\mathrm{d}s$.

2. The cost functional $J(u,g) : U \times G \to \mathbb{R}$ is weakly lower semicontinuous, and there exist a constant $\alpha_J > 0$ such that

$$J(u,g) \geq \alpha_J \|g\|_G^2. \tag{18}$$

Remark 3.2. Note that the above assumptions include mixed *displacement-traction* as well as *pure traction problems.* With respect to the latter, adequate choices of the cost functional may remove the "indeterminacy up to rigid translations" [14, 15].

Theorem 3.1. *Suppose that Assumptions* 3.1 *and* 3.2 *hold. Then the optimal control problem*

$$\min_{(u,g)\in U\times G} J(u,g) \quad \textit{subject to} \quad u \in \operatorname*{argmin}_{v\in U} \mathcal{E}(v,g) \tag{19}$$

has at least one solution.

Before turning to the proof of this theorem, we will first state two important lemmas that will be required therein. We start with a result on compensated compactness, which has been stated in [6, section 6] and, in a clearer version, in [14, Chapter 7]. It can be viewed as the main step in the proof of existence of energy minimizers in nonlinear elasticity.

LEMMA 3.2. *Let* $\Phi \in \mathbf{W}^{1,p}(\Omega)$, $p \geq 2$, *and* $r, q > 0$ *such that* $r^{-1} = p^{-1} + q^{-1} \leq 1$. *Then the following implication holds:*

$$\left.\begin{aligned} \varphi^k \rightharpoonup \varphi \ &\text{in } \mathbf{W}^{1,p}(\Omega),\ p \geq 2 \\ \operatorname{adj}(\nabla \varphi^k) \rightharpoonup \rho \ &\text{in } \mathbf{L}^q(\Omega),\ \frac{1}{p} + \frac{1}{q} \leq 1 \\ \det(\nabla \varphi^k) \rightharpoonup \delta \ &\text{in } L^r(\Omega), r \geq 1 \end{aligned}\right\} \Rightarrow \begin{cases} \rho = \operatorname{adj}(\nabla \varphi), \\ \delta = \det(\nabla \varphi). \end{cases}$$

Proof. See [14, Thm. 7.6-1]. □

Using the above result and the theorem of Mazur, one can prove the sequential weak lower semicontinuity of $\mathcal{E}^S$ with respect to sequences u_k for which $\mathcal{E}^S$ remains bounded (see [14, Proof of Thm. 7.7-1]). This result can be extended in the following way.

LEMMA 3.3. *Let Assumptions* 3.1 *and* 3.2 *hold. Consider a weakly converging sequence* $(u_k, g_k) \rightharpoonup (\tilde{u}, \tilde{g})$ *in* $U \times G$ *such that*

$$u_k \in \operatorname*{argmin}_{v \in U} \mathcal{E}(v, g_k),$$

and $\mathcal{E}(u_k, g_k)$ *is bounded from above. Then*

$$\lim_{k \to \infty} \mathcal{E}(u_k, g_k) = \mathcal{E}(\tilde{u}, \tilde{g}) = \min_{v \in U} \mathcal{E}(v, \tilde{g}). \tag{20}$$

Proof. First, we show the weak lower semicontinuity of $\mathcal{E}$ for sequences that leave the energy bounded from above.

Weak lower semicontinuity of the first part $\mathcal{E}^S$ with respect to u_k follows as in [14, Proof of Thm. 7.7-1] from Lemma 3.2 and convexity of the functional $\mathbb{W}$ with respect to its arguments. The second part

$$\mathcal{E}^{\Gamma_c}(u_k, g_k) = \int_{\Gamma_c} u_k g_k \, ds$$

is even *weakly continuous.* This follows via compactness of the trace mapping $\mathbf{W}^{1,p}(\Omega) \hookrightarrow \mathbf{L}^2(\Gamma_c)$ by strong convergence $u_k|_{\Gamma_c} \to \tilde{u}|_{\Gamma_c}$ in $\mathbf{L}^2(\Gamma_c)$ and weak convergence $g_k \rightharpoonup \tilde{g}$ in $\mathbf{L}^2(\Gamma_c)$. In summary, we can conclude weak lower semicontinuity of $\mathcal{E}$:

$$\mathcal{E}(\tilde{u}, \tilde{g}) \leq \liminf_{k \to \infty} \mathcal{E}(u_k, g_k),$$

and, if u is fixed,

$$\lim_{k \to \infty} \mathcal{E}(u, g_k) = \mathcal{E}(u, \tilde{g}).$$

Next, by the minimizing property of u_k, we obtain $\mathcal{E}(u_k, g_k) \leq \mathcal{E}(\tilde{u}, g_k)$ and

$$\limsup_{k \to \infty} \mathcal{E}(u_k, g_k) \leq \limsup_{k \to \infty} \mathcal{E}(\tilde{u}, g_k) = \lim_{k \to \infty} \mathcal{E}(\tilde{u}, g_k) = \mathcal{E}(\tilde{u}, \tilde{g}),$$

implying

$$\limsup_{k\to\infty} \mathcal{E}(u_k, g_k) \le \mathcal{E}(\tilde{u}, \tilde{g}) \le \liminf_{k\to\infty} \mathcal{E}(u_k, g_k),$$

and thus

$$\lim_{k\to\infty} \mathcal{E}(u_k, g_k) = \mathcal{E}(\tilde{u}, \tilde{g}).$$

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

$$\mathcal{E}(\bar{u}, \tilde{g}) \le \mathcal{E}(\tilde{u}, \tilde{g}) = \lim_{k\to\infty} \mathcal{E}(u_k, g_k) \le \lim_{k\to\infty} \mathcal{E}(\bar{u}, g_k) = \mathcal{E}(\bar{u}, \tilde{g}). \qquad \square$$

Observe the two structural properties that make this proof work: First, linearity of $\mathcal{E}^{\Gamma_c}$ with respect to g, and second, compactness of the trace mapping $\mathbf{W}^{1,p}(\Omega) \hookrightarrow \mathbf{L}^2(\Gamma_c)$. Our proof extends to any $\mathcal{E}^{\Gamma_c}$ with the same abstract properties.

Proof of Theorem 3.1. First, we show that we can apply Lemma 3.3. Then, using the weak lower semicontinuity of J, we will show that there exists an admissible minimizing sequence $(u_k, g_k)_{k\in\mathbb{N}}$ converging weakly in $U \times G$ to a minimizer $(\tilde{u}, \tilde{g})$ of the optimal control problem. Eventually exploiting the coerciveness of $\mathcal{E}$ will lead to the admissibility of the weak limit $(\tilde{u}, \tilde{g})$, i.e.,

$$\operatorname{adj}(\nabla\tilde{u}) \in \mathbf{L}^q(\Omega) \quad \text{and} \quad \det(I + \nabla\tilde{u}) \in L^r(\Omega).$$

Existence of a weakly convergent subsequence. As has been shown in [6, Thms. 7.3 and 7.6], [14, Thm. 7.7-1], for every $g_k \in G$ there exists a displacement $u_k \in U$ such that $u_k \in \operatorname{argmin}_{v\in U} \mathcal{E}(v, g_k)$. Thus, as the energy functional $J(u, g)$ is bounded from below, there exists a minimizing sequence $(u_k, g_k)_{k\in\mathbb{N}}$ of J with $g_k \in G$, $u_k \in U$, and u_k being a minimizer of $\mathcal{E}(\cdot, g_k)$. From (18) we deduce that the sequence $\{g_k\}_{k\in\mathbb{N}}$ is bounded in G by some constant C_g, and by reflexivity of G there exists a weakly convergent subsequence, which will again be denoted as $\{g_k\}_{k\in\mathbb{N}}$ with weak limit $\tilde{g} \in G$.

First, we have to show that the sequence $\{\mathcal{E}(u_k, g_k)\}_{k\in\mathbb{N}}$ is bounded from above. Setting $\|\cdot\|_U := \|\cdot\|_{\mathbf{W}^{1,p}(\Omega)}$, $\|\cdot\|_G := \|\cdot\|_{\mathbf{L}^2(\Gamma_c)}$ and using Hölder's inequality and the continuity of the trace operator, we get an estimate for the sensitivity of the elastic energy functional with respect to changes in the Neumann boundary conditions:

$$\mathcal{E}(u, g_k) - \mathcal{E}(u, 0) = \int_{\Gamma_c} u(0 - g_k)\,\mathrm{d}s \le \|u\|_U \|g_k\|_G \le C_g \|u\|_U. \tag{21}$$

Thus as u_k minimizes $\mathcal{E}(\cdot, g_k)$, the boundedness of $\{\mathcal{E}(u_k, g_k)\}_{k\in\mathbb{N}}$ is a consequence of (21), inserting $u = \overline{u}$ as defined at the end of Assumption 3.1:

$$\mathcal{E}(u_k, g_k) \le \mathcal{E}(\overline{u}, g_k) \le C_g \|\overline{u}\|_U + \mathcal{E}(\overline{u}, 0) < \infty.$$

Now the boundedness of $\{u_k\}_{k\in\mathbb{N}}$ follows from the coercivity of $\mathcal{E}$; i.e., there exist constants $\tilde{\gamma} > 0$, $\tilde{\beta} \in \mathbb{R}$ such that

$$\tilde{\gamma} \|u_k\|_U^p \le \mathcal{E}(u_k, g_k) + \tilde{\beta} \le C_g \|\overline{u}\| + \mathcal{E}(\overline{u}, 0) + \tilde{\beta}.$$

Again reflexivity implies the existence of a subsequence $u_k \rightharpoonup \tilde{u}$ in U.

Admissibility of $(\tilde{u}, \tilde{g})$. Now we can apply Lemma 3.3 to get

$$\lim_{k\to\infty} \mathcal{E}(u_k, g_k) = \mathcal{E}(\tilde{u}, \tilde{g}) = \min_{v\in U} \mathcal{E}(v, \tilde{g}).$$

Thus the pair $(\tilde{u}, \tilde{g})$ is an admissible candidate for a minimizer of J and a weak limit of the minimizing sequence (u_k, g_k) of J. As J is weakly lower semicontinuous, $(\tilde{u}, \tilde{g})$ indeed minimizes J. Moreover, the coercivity inequality (16) in combination with Lemma 3.2 guarantees that

$$\operatorname{adj}(I + \nabla\tilde{u}) \in \mathbf{L}^q(\Omega) \quad \text{and} \quad \det(I + \nabla\tilde{u}) \in L^r(\Omega),$$

and condition (15) ensures that $\det(I + \nabla\tilde{u}) > 0$ a.e. in Ω. □

4. Weak formulation. In the following, we discuss weak formulations corresponding to the energy minimization problem $\min_{u\in U} \mathcal{E}(u, g)$. This means that we derive first order necessary optimality conditions for the constraint of the optimal control problem under consideration. For the sake of clarity from now on we will suppress the dependence on x; i.e., we will assume that the material under consideration is homogeneous. The derived results also hold for heterogeneous materials.

As noted in the introduction, it is, in general, not clear whether a local minimizer of the elastic energy functional satisfies the weak formulation (see [8, Problems 5 and 6])

$$\mathcal{E}'(u, g)h = 0 \quad \text{for all } h \in C^\infty(\Omega).$$

In the context of compressible material laws, the main difficulties are caused by condition (14). While being necessary in order to avoid local self-penetration and to model the observed material behavior in a qualitatively correct way, the introduced singularity leads to severe analytical difficulties.

In particular, it implies for the strain energy, that

$$\mathcal{E}^S(u) = \int_\Omega W(\nabla\varphi)\,\mathrm{d}x = \infty$$

on a dense subset of $\mathbf{W}^{1,p}(\Omega)$ for any $p < \infty$ and thus also on a dense subset of U; i.e., for every $u \in U$ with $\mathcal{E}(u, g) < \infty$ one can construct a sequence $u_k \to u$ in U such that

$$\mathcal{E}(u_k, g) = \infty \quad \text{for all } k \in \mathbb{N}.$$

Thus, we cannot expect differentiability in spaces weaker than $\mathbf{W}^{1,\infty}(\Omega)$.

To make this discussion concrete, in the following we consider a compressible Mooney–Rivlin material law. This widely used constitutive relation is a special case of a compressible Ogden-type material. It is polyconvex and isotropic and may be written in terms of the (right) Green–St.Venant strain tensor

$$E(u) = \frac{1}{2}\left(\nabla u^T + \nabla u + \nabla u^T \nabla u\right)$$

and the deformation gradient $\nabla\varphi = I + \nabla u$:

$$W(u) = a\operatorname{tr}(E) + b\left(\operatorname{tr}(E)\right)^2 + c\operatorname{tr}(E^2) + \Gamma\left(\det(\nabla\varphi)\right) + f$$

and $\lim_{s\to 0^+}\Gamma(s)=\infty$. Setting $\alpha = a-2b,\ \beta=-c$, W can be represented in the following way:

$$W(\nabla\varphi)=\frac{\alpha}{2}\|\nabla\varphi\|^2+\frac{\beta}{2}\|\operatorname{adj}\nabla\varphi\|^2+\Gamma\left(\det(\nabla\varphi)\right)+e.$$

Popular choices for Γ take the form (see [35, 36])

$$\Gamma(t)=\frac{1}{2}dt^2-e\ln(t)\quad\text{or}\quad\Gamma(t)=\frac{1}{2}dt^2+\frac{e}{k}t^{-k},\ k>0. \tag{22}$$

Remark 4.1. In both cases the first summand t^2 guarantees, with $\alpha>0$, $\beta>0$, $d>0$, the validity of the coerciveness inequality (16) with $p=q=r=2$. Moreover, for small strain, the material behaves like a St.Venant–Kirchhoff material. Thus, near $E=0$ the stored energy function W should be a second order approximation of the stored energy function of a St.Venant–Kirchhoff material. In the case of $\Gamma(t)=dt^2-e\ln(t)$ it is always possible to determine $\alpha>0$, $\beta>0$, $d>0$, $e>0$ such that this is the case [14, Thm. 4.10-2]. This property comes at the expense of the model's quality, restricting its validity to rather academic questions. Thus we will focus on a nonlogarithmic form as proposed in [35]. In this case the choice of parameters is dependent on the Poisson ratio $\nu=\frac{\lambda}{2(\lambda+\mu)}$. More precisely, the inequality

$$k<-1+\frac{1}{1-2\nu}\ \Leftrightarrow\ \nu>\frac{k}{2(k+1)} \tag{23}$$

restricts the possible range for k for given ν and vice versa; i.e., $k\geq 9$ requires $\nu>0.45$, thus possibly implying the risk of constitutive locking (Poisson locking [11]). While being independent of Young's modulus, this inequality becomes less restrictive with growing ν.

With respect to the weak formulation, we first focus on the energy minimization problem

$$\min_{u\in U}\ \mathcal{E}(u,g)\quad\text{for given fixed } g\in G. \tag{24}$$

In the following we study the derivatives of $\mathcal{E}$ with respect to u, starting with a pointwise computation of the derivatives of W at nonsingular F in direction δF:

$$W'(F)\delta F=\alpha F:\delta F+\beta\operatorname{adj}F:\operatorname{adj}'(F)\delta F+\Gamma'(\det F)\left(\operatorname{adj}F:\delta F\right). \tag{25}$$

Here we used the differentiation rule $\det'(F)\delta F=\operatorname{adj}F:\delta F$. Further, we may also compute the second derivative:

$$\begin{aligned}(26)\quad & W''(F)(\delta F_1,\delta F_2)\\ &=\alpha\delta F_1:\delta F_2+\beta\operatorname{adj}'(F)\delta F_1:\operatorname{adj}'(F)\delta F_2+\beta\operatorname{adj}F:\operatorname{adj}''(F)(\delta F_1,\delta F_2)\\ &\quad+\Gamma'(\det F)\left(\operatorname{adj}'(F)\delta F_1:\delta F_2\right)+\Gamma''(\det F)\left(\operatorname{adj}F:\delta F_1\right)(\operatorname{adj}F:\delta F_2).\end{aligned}$$

The validity of the above pointwise formulae follows, for $F\in\mathbb{M}^3_+$, $\delta F_1,\delta F_2\in\mathbb{M}^3$, directly from the definitions of det, adj, and Γ. Having stated differentiability properties of W as a nonlinear function of the matrix $F\in\mathbb{M}^3$, we now turn to their study as superposition operators.

To this end, we consider the space $\mathbf{L}^p(\Omega)$ of p-integrable matrix valued functions $F:\Omega\to\mathbb{M}^3$, insert the matrix valued function $F\in\mathbf{L}^p(\Omega)$ pointwise into W, and consider the result in another L^p-space. For this purpose we first need some properties of $\operatorname{adj}, \Gamma$ and an additional assumption on local minimizers of the energy functional $\mathcal{E}$.

Lemma 4.1. *Let $F \in \mathbf{L}^p(\Omega)$. Then the mapping*

$$\operatorname{adj}'(F): \ \mathbf{L}^{p'}(\Omega) \to \mathbf{L}^1(\Omega) \tag{27}$$

is linear and continuous for $p^{-1} + (p')^{-1} \leq 1$. Moreover, the mapping

$$\operatorname{adj}''(F): \ \mathbf{L}^{s_1}(\Omega) \times \mathbf{L}^{s_2}(\Omega) \to \mathbf{L}^1(\Omega) \tag{28}$$

is independent of F and bilinear and continuous for $s_1^{-1} + s_2^{-1} \leq 1$. For $N > 2$ we have $\operatorname{adj}^{(N)} = 0$.

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

Definition 4.1. *Let $\varphi \in \mathbf{W}^{1,p}(\Omega)$ with $p \geq 1$. We call φ* nondegenerate *if there exists a constant $\epsilon > 0$ such that*

$$\det(\nabla\varphi) \geq \epsilon \quad a.e.\ in\ \Omega. \tag{29}$$

In the context of elasticity theory we also call the displacement $u \in U$ nondegenerate if $\varphi = \mathrm{Id} + u$ is nondegenerate.

Remark 4.2. In the similar framework of barrier regularizations, examples can be given where the violation of an analogue to nondegeneracy in the above sense yields minimizers that do not satisfy the formal optimality conditions [39].

Suppose there exists a local minimizer $u \in U$ of $\mathcal{E}_g$ that is degenerate; i.e., there exists a sequence

$$(x_k)_{k\in\mathbb{N}} \subset \Omega, \ x_k \to x \in \Omega \quad \text{such that} \quad \det(I + \nabla u(x_k)) \to 0.$$

Physically this corresponds to a deformation that becomes singular at $x \in \Omega$, and thus is reasonable only in the modeling of cutting or piercing processes. In these cases other effects, such as plasticity, become dominant. In the context of applications like implant shape design, the elastic behavior is predominant, thus justifying the nondegeneracy assumption on minimizers of $\mathcal{E}$.

Lemma 4.2. *Assume that $F \in \mathbf{L}^p(\Omega)$ is nondegenerate, $\operatorname{adj} F \in \mathbf{L}^q(\Omega)$, and $\det F \in L^r(\Omega)$. Assume that the integrability indices $s_i \in [1, \infty], i = 1, \dots, N$, satisfy*

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

which is possible only if the expressions on the corresponding right-hand sides are nonnegative.

Then, for the choice

$$\delta F_i \in \mathbf{L}^{s_i}(\Omega), \quad s_i \in [1, \infty], \ i = 1, \dots, N,$$

we obtain

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

Proof. Differentiating Γ from (22), we get

$$\Gamma'(t) = t - t^{-(k+1)}, \quad \Gamma''(t) = 1 + (k+1)t^{-(k+2)},$$
$$\Gamma'''(t) = -(k+1)(k+2)t^{-(k+3)}.$$

Thus under our assumption of *nondegeneracy* it follows that $\frac{d}{dF}\Gamma(\det F)$ grows linearly in $\det F$ and that $\frac{d^2}{dF^2}\Gamma(\det F)$ is bounded independently of $\det F$. Then, using Hölder's inequality, inspection of the relevant terms in (25) and (26) yields our results for $N = 1$ and $N = 2$. For $N = 3$, we compute

$$\begin{aligned}\frac{d^3}{dF^3}\Gamma(\det F)(\delta F_1, \delta F_2, \delta F_3) = \Gamma'(\det F)&(\operatorname{adj}''(F)(\delta F_1, \delta F_2) : \delta F_3)\\ &+ 3\Gamma''(\det F)(\operatorname{adj}'(F)\delta F_1 : \delta F_2)(\operatorname{adj} F : \delta F_3)\\ &+ \Gamma'''(\det F)(\operatorname{adj} F : \delta F_1)(\operatorname{adj} F : \delta F_2)(\operatorname{adj} F : \delta F_3)\end{aligned}$$

and use again Hölder's inequality. □

Now we can turn to the study of the derivatives of W.

PROPOSITION 4.3. *Assume that* $F \in \mathbf{L}^p(\Omega)$ *is nondegenerate,* $\operatorname{adj} F \in \mathbf{L}^q(\Omega)$*, and* $\det F \in L^r(\Omega)$*. (In the following, we take* $s_i \in [1, \infty]$ *and assume that the inequalities for the* s_i *are nonvoid.)*

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

$$W'(F)\delta F \in L^1(\Omega) \quad \textit{for all } \delta F \in \mathbf{L}^{s_1}(\Omega),$$

and $W'(F)$ *is linear and continuous in* δF*.*

If $0 \le s_1^{-1} + s_2^{-1} \le 1 - \max(2p^{-1}, r^{-1} + p^{-1}, 2q^{-1})$*, then*

$$W''(F)(\delta F_1, \delta F_2) \in L^1(\Omega) \quad \textit{for all } \delta F_i \in \mathbf{L}^{s_i}(\Omega),\ i = 1, 2,$$

and $W''(F)$ *is bilinear and continuous in* $(\delta F_1, \delta F_2)$*.*

If $0 \le s_1^{-1} + s_2^{-1} + s_3^{-1} = 1 - \max(r^{-1}, p^{-1} + q^{-1}, 3q^{-1})$*, then*

$$W'''(F)(\delta F_1, \delta F_2, \delta F_3) \in L^1(\Omega) \quad \textit{for all } \delta F_i \in \mathbf{L}^{s_i}(\Omega),\ i = 1, 2, 3,$$

and $W'''(F)$ *is trilinear and continuous in* $(\delta F_1, \delta F_2, \delta F_3)$*.*

Proof. The assertion follows from inspection of the particular terms in (25) for W' and in (26) for W''. For W''' a similar term can be computed. Well-definedness of the derivatives of Γ in suitable L_p-spaces has been shown in Lemma 4.2; the remaining terms are second and fourth order polynomials in the coefficients of F. With this information, our result follows from repeated application of the Hölder inequality. □

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

$$\mathcal{E}_u^S(u)v_1 := \int_\Omega W'(I + \nabla u)\nabla v_1 \,\mathrm{d}x, \tag{30}$$

$$\mathcal{E}_{uu}^S(u)(v_1, v_2) := \int_\Omega W''(I + \nabla u)(\nabla v_1, \nabla v_2) \,\mathrm{d}x, \tag{31}$$

$$\mathcal{E}_{uuu}^S(u)(v_1, v_2, v_3) := \int_\Omega W'''(I + \nabla u)(\nabla v_1, \nabla v_2, \nabla v_3) \,\mathrm{d}x \tag{32}$$

are well defined. Moreover, we have to verify whether the remainder terms vanish, i.e., under which conditions the defined functionals really are the directional derivatives of the strain energy. For given $u \in \mathbf{W}^{1,2}(\Omega)$ this is a delicate issue. Fortunately, the coerciveness inequality (16) retains that $\operatorname{adj}(I+\nabla u) \in \mathbf{L}^2(\Omega)$ and $\det(I+\nabla u) \in L^2(\Omega)$ if u is a minimizer of $\mathcal{E}$. Therefore we have the following.

COROLLARY 4.4. *Assume that $u \in U$ is nondegenerate and $\mathcal{E}^S(u)$ is finite. Then $\mathcal{E}^S_u(u)$ and $\mathcal{E}^S_{uu}(u)$ are well defined in $\mathbf{W}^{1,\infty}(\Omega)$, resp., $\mathbf{W}^{1,\infty}(\Omega) \times \mathbf{W}^{1,\infty}(\Omega)$. If, further,*

- *$\operatorname{adj}(I + \nabla u) \in \mathbf{L}^\infty(\Omega)$, then (30) is well defined for $v_1 \in \mathbf{W}^{1,2}(\Omega)$ and (32) for $v_i \in \mathbf{W}^{1,\infty}(\Omega)$, $i = 1, 2, 3$;*
- *$u \in \mathbf{W}^{1,\infty}(\Omega)$, then (30), (31), and (32) are well defined for $v_i \in \mathbf{W}^{1,s_i}(\Omega)$ with $\sum s_i^{-1} = 1$, respectively.*

Proof. By coercivity of $\mathcal{E}^S$ we conclude that $p = 2$, $q = 2$, $r = 2$. Thus, we can apply Proposition 4.3 for $s_i = \infty$ to obtain our first result for (30), (31).

If $q = \infty$, then $s = 2$ is admitted for (30), and Proposition 4.3 yields $\sum s_i \geq 0$ for (32).

Since adj and det are polynomials, it follows from $u \in \mathbf{W}^{1,\infty}(\Omega)$ and nondegeneracy that $p = q = r = \infty$ such that $\sum s_i^{-1} = 1$ can be chosen. □

PROPOSITION 4.5. *If $u \in U$ is nondegenerate and $u \in \mathbf{W}^{1,\infty}(\Omega)$, then $\mathcal{E}^S$ is directionally differentiable for each $\delta u \in \mathbf{W}^{1,\infty}(\Omega)$ with derivative given by (30). The corresponding remainder term is uniform in δu.*

Moreover, $\mathcal{E}^S$ is twice directionally differentiable with second derivative given by (31). For sufficiently small $\|\delta u\|_{\mathbf{W}^{1,\infty}(\Omega)}$ the corresponding remainder term can be estimated by

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

Proof. In order to prove the statement we consider for $\delta u \in \mathbf{W}^{1,\infty}(\Omega)$ the remainder term

$$\begin{aligned} |\mathcal{E}(u+\delta u, g) - \mathcal{E}(u, g) - \mathcal{E}_u(u, g)\delta u| &= \frac{1}{2}\left|\mathcal{E}_{uu}(u + \xi\delta u, g)(\delta u)^2\right| \\ &\leq \frac{1}{2}\|\,\mathcal{E}_{uu}(u + \xi\delta u, g)\|\|\delta u\|^2_{\mathbf{W}^{1,\infty}(\Omega)}. \end{aligned}$$

By Corollary 4.4 we know that $\mathcal{E}_{uu}(u, g)(\delta u)^2$ is finite, and since $\mathcal{E}_{uu}$ is continuous at u in $\mathbf{W}^{1,\infty}(\Omega)$, $\mathcal{E}_{uu}(u + \xi\delta u, g)(\delta u)^2$ is bounded.

The proof for the second derivative runs analogously, using the properties of W'''. □

Remark 4.3. Let us point out the following subtlety.[1] The question arises of whether for $u \notin \mathbf{W}^{1,\infty}(\Omega)$ a similar result can be achieved. To this end, continuity of $\mathcal{E}_{uu}$ at (u, g) with respect to perturbations $\delta u \in \mathbf{W}^{1,\infty}(\Omega)$ would be required. Since $\mathcal{E}_{uu}$ has a singularity in $\det(I + \nabla u)$, we need for its continuity the uniform continuity of the mapping $F \to \det F$, even at a nondegenerate point and even for $\delta u \in \mathbf{W}^{1,\infty}(\Omega)$. Since $\det F$ in $\mathbb{R}^{3\times 3}$ is a third order polynomial in the entries of F, this property holds only on bounded subsets of $\mathbb{R}^{3\times 3}$ so that $u \in \mathbf{W}^{1,\infty}(\Omega)$ is really needed.

The combination of these results allows us to prove the main theorem of this section.

[1] The authors are indebted to Dipl. Math. Simon Rösel, who communicated this issue to us and found a flaw in a preliminary version of this paper.

THEOREM 4.6. *Let $u \in U$ be a nondegenerate local minimizer of $\mathcal{E}$ with $\mathcal{E}(u) < \infty$ and $u \in \mathbf{W}^{1,\infty}(\Omega)$. Then it satisfies the following weak formulation:*

$$\mathcal{E}_u(u,g)\delta u = 0 \quad \textit{for all } \delta u \in \mathbf{W}^{1,\infty}(\Omega). \tag{33}$$

If, in turn, $u \in \mathbf{W}^{1,\infty}(\Omega)$ satisfies (33), and $\mathcal{E}_{uu}(u,g)v^2 \geq \delta\|v\|^2_{\mathbf{W}^{1,2}(\Omega)}$ for all $v \in \mathbf{W}^{1,\infty}(\Omega)$, then for sufficiently small $\delta u \in \mathbf{W}^{1,\infty}(\Omega)$ and some $\varepsilon > 0$ we have the growth condition

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

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

Proof. The proof is standard. To show that $\mathcal{E}_u(u,g)\delta u = 0$, we compute

$$\mathcal{E}_u(u,g)(\pm\delta u) = \lim_{t\to 0}\frac{\mathcal{E}(u \pm t\delta u, g) - \mathcal{E}(u,g)}{t} \geq 0,$$

since u is a local minimizer of $\mathcal{E}$.

For our second assertion, we note that

$$\begin{aligned}\mathcal{E}(u+\delta u) - \mathcal{E}(u) &= \frac{1}{2}\,\mathcal{E}_{uu}(u,g)\delta u^2 + r(\delta u)\\ &\geq \frac{\delta}{2}\|\delta u\|^2_{\mathbf{W}^{1,2}(\Omega)} + r(\delta u).\end{aligned}$$

Due to Proposition 4.5,

$$r(u,\delta u) \leq c\|\delta u\|_{\mathbf{W}^{1,\infty}(\Omega)}\|\delta u\|^2_{\mathbf{W}^{1,2}(\Omega)}$$

so that, for $\|\delta u\|_{\mathbf{W}^{1,\infty}(\Omega)} \to 0$, we obtain

$$\mathcal{E}(u+\delta u) - \mathcal{E}(u) \geq \left(\frac{\delta}{2} - c\|\delta u\|_{\mathbf{W}^{1,\infty}(\Omega)}\right)\|\delta u\|^2_{\mathbf{W}^{1,2}(\Omega)} \geq \varepsilon\|\delta u\|^2_{\mathbf{W}^{1,2}(\Omega)}. \qquad \square$$

5. Formal first order optimality conditions. Next, we discuss first order optimality conditions of our optimal control problem. As we have seen above, differentiability of the equality constraints $\mathcal{E}_u(u,g) = 0$ requires the choice of $\mathbf{W}^{1,\infty}(\Omega)$ (or stronger) as a topological framework. Thus we have to restrict our discussion to a formal level, as on the one hand, we lack an existence result in this space, and on the other hand, existing regularity results do not admit the application of the implicit function theorem in order to show that the set $\mathcal{E}_u(u,g) = 0$ is a smooth manifold. Its application requires continuous invertibility of the linearized weak formulation in suitable spaces. One possible framework would be to consider

$$\mathcal{E}_{uu} : \mathbf{W}^{2,p}(\Omega) \to L^p(\Omega)$$

for $\mathbf{W}^{2,p}(\Omega) \hookrightarrow \mathbf{W}^{1,\infty}(\Omega)$ (cf., e.g., [14, Chapter 6]). However, the class of problems for which suitable regularity results hold is small.

Formally, the first order optimality conditions of (19) can be derived via the Lagrangian function

$$L(u,g,p) = J(u,g) + p(\mathcal{E}_u(u,g)).$$

Computing the formal derivatives of L with respect to u, g, and p yields the system

$$J_u(u,g) + \mathcal{E}_{uu}(u,g)^* p = 0 \quad \text{in } U^*, \tag{34a}$$

$$J_g(u,g) + \mathcal{E}_{ug}(u,g)^* p = 0 \quad \text{in } G^*, \tag{34b}$$

$$\mathcal{E}_u(u,g) = 0 \quad \text{in } U^*. \tag{34c}$$

If J is the sum of a measure of the error and a Tikhonov regularization term, i.e., if J is of the form $J(u,g) = J^{err}(u) + \frac{\alpha}{2}\|g\|^2_{L^2(\Gamma_c)}$, where α is the Tikhonov regularization parameter (see (9a)), then these conditions can be written explicitly:

$$J_u^{err}(u) + \mathcal{E}_{uu}(u,g)^* p = 0 \quad \text{in } U^*, \tag{35a}$$

$$\alpha g(x) + \left(\operatorname{adj}(I+\nabla u)^T n\right) p(x) = 0 \quad \text{a.e. on } \Gamma_c, \tag{35b}$$

$$\mathcal{E}_u(u,g) = 0 \quad \text{in } U^*. \tag{35c}$$

Elimination of g via (35b) reduces system (35) to

$$J_u^{err}(u) + \mathcal{E}_{uu}(u)^* p = 0, \tag{36a}$$

$$\mathcal{E}_u\left(u, -\frac{\left(\operatorname{adj}(I+\nabla u)^T n\right) p}{\alpha}\right) = 0. \tag{36b}$$

6. Numerical results. In order to perform first numerical experiments, we consider the cost functional

$$J(u,g) = \frac{\beta}{2}\|u - u_{\text{ref}}\|^2_{\mathbf{L}^2(\Gamma_t)} + \frac{\alpha}{2}\|g\|^2_{L^2(\Gamma_c)}, \tag{37}$$

where the additional parameter $\beta \in]0,1]$ is introduced in order to establish a numerical continuation scheme $\beta \to 1$. In a direct approach, the occurring nonlinearities would lead to too small Newton steps in the solution of the nonlinear problem for $\beta = 1$.

Then, setting $\tilde{n} = \operatorname{adj}(I+\nabla u)^T n$, the reduced optimality system reads

$$\int_\Omega W''(\nabla\varphi)\nabla p \nabla v \, \mathrm{d}x + \int_{\Gamma_t} \beta(u - u_{\text{ref}}) v \, \mathrm{d}s = 0 \quad \text{for all } v \in U, \tag{38a}$$

$$\int_\Omega W'(\nabla\varphi)\nabla w \, \mathrm{d}x + \int_{\Gamma_c} \frac{\tilde{n}p}{\alpha}\tilde{n} w \, \mathrm{d}s = 0 \quad \text{for all } w \in U. \tag{38b}$$

Further, in view of possibly large values for Young's modulus $\mathbb{E}$, we perform a rescaling of the problem via

$$W \mapsto \mathbb{E}^{-1} W \quad \text{and} \quad \alpha \mapsto \mathbb{E}^2 \alpha. \tag{39}$$

This is a problem formulation that is invariant with respect to Young's modulus. This is of advantage, because in the presence of large $\mathbb{E}$, appropriate Tikhonov parameters satisfy $\alpha \sim \mathbb{E}^{-2}$ (see [30]) and thus may become very small. This in turn affects the condition number of the Newton matrix and thus the numerical accuracy of the Newton steps. As the coefficients of W depend linearly on Young's modulus, the application of the transformations (39) is equivalent to setting $\mathbb{E} = 1$.

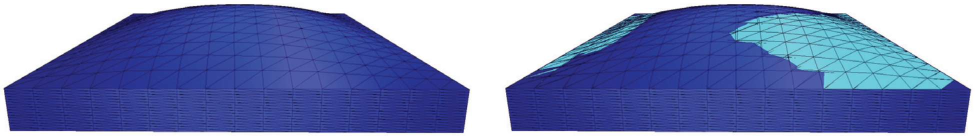
In summary, we solve a sequence of problems

$$(P_k) \quad \left\{ \begin{aligned} \int_\Omega W''(\nabla\varphi)\nabla p \nabla v \, \mathrm{d}x + \int_{\Gamma_t} \beta_k (u - u_{\text{ref}}) v \, \mathrm{d}x = 0, \\ \int_\Omega W'(\nabla\varphi) h \, \mathrm{d}x + \int_{\Gamma_c} \frac{\tilde{n}p}{\alpha}\tilde{n} h \, \mathrm{d}s = 0 \end{aligned} \right. \tag{40}$$

with $0 < \beta_0 < \cdots < \beta_N \leq 1,\ N > 0,\ \mathbb{E} = 1$. The second material parameter of linearized elasticity, the Poisson ratio ν, is close to $\frac{1}{2}$ in order to correspond to a quasi-incompressible material, as encountered in soft tissue models. As constitutive locking is a commonly observed phenomenon for $\nu \to \frac{1}{2}$ [5, 11] that should be excluded in order to monitor the influences of the nonlinearities, we set $\nu = 0.45$. This choice keeps the risk of constitutive locking small while staying reasonable from a modeling point of view. In general, in order to allow the Poisson ratio to attain all values in the admissible range $[0, 0.5[$, mixed formulations, and/or adjusted discretization schemes for the forward problem of elastostatics [4, 40, 44] must be used and adapted to the optimal control problem.

As noted in section 4, a logarithmic dependence of Γ on the volume change is not sufficient to accurately model the soft tissue's behavior. For the sake of numerical simplicity, we nevertheless choose the logarithmic penalty term $\Gamma(s) = s^2 + \log(s)$.

The systems (P_k) have been discretized on the cuboid $[-1, 1] \times [-1, 1] \times [-0.1, 0.1]$ with the finite element toolbox Kaskade7 [22] using linear elements. The resulting finite dimensional, nonlinear equations are solved with a covariant damped Newton method as presented in [17, Chapter 3]. For the solution of the arising linear systems of equations, we use the distributed multifrontal solver MUMPS [1, 2]. First numerical results are given in Figure 2.



(a) Computed deformation.

(b) Comparison between computed solution and desired deformation (lighter areas).

FIG. 2. *Numerical results for* $\alpha = 0.1$.

7. Conclusion. In this work, basic analytical and numerical results for the mathematical treatment of an implant design problem have been established. Since the direct approach in section 2.2 appears very difficult to handle, a structurally much simpler optimal control reformulation, where the boundary forces act as a control is proposed in section 2.3. The shape of the implant is reconstructed in a second step from the computed deformation of the soft tissue. This can be motivated by a classical theorem [14, Thm. 7.9-1]: under the reasonable assumption of sufficient regularity of the soft tissue deformation, the solutions of the optimal control problem also solve the obstacle problem between soft tissue and implant.

Due to the well-known intricacies of nonlinear hyperelasticity, the analytical treatment of the optimal control problem is still quite challenging. Even for proving the existence of an optimal solution we had to resort to a simplified problem setting either using a dead load or allowing tangential forces. Currently, stronger results such as uniqueness or continuity of solutions with respect to parameters appear out of reach.

From a numerical point of view, the dead load simplification is not necessary, and the original optimal control problem can be addressed directly. However, our numerical experience indicates that the problem is quite challenging, and refined algorithmic ideas are necessary to treat the nonlinear implant shape problem to full satisfaction. This includes on the one hand globalization techniques for the nonlinear solver, and on the other hand adaptivity and iterative solution techniques for the linear systems.

REFERENCES

[1] P.R. Amestoy, I.S. Duff, J.-Y. L'Excellent, and J. Koster, *A fully asynchronous multifrontal solver using distributed dynamic scheduling*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 15–41.

[2] P.R. Amestoy, A. Guermouche, J.-Y. L'Excellent, and S. Pralet, *Hybrid scheduling for the parallel solution of linear systems*, Parallel Comput., 32 (2006), pp. 136–156.

[3] 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 (1993), pp. 389–412.

[4] O. Axelsson and A. Padiy, *On a robust and scalable linear elasticity solver based on a saddle point formulation*, Internat. J. Numer. Methods Engrg., 44 (1999), pp. 801–818.

[5] I. Babuška and M. Suri, *Locking effects in the finite element approximation of elasticity problems*, Numer. Math., 62 (1992), pp. 439–463.

[6] J.M. Ball, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Rational Mech. Anal., 63 (1977), pp. 337–403.

[7] J.M. Ball, *Strict convexity, strong ellipticity, and regularity in the calculus of variations*, Math. Proc. Cambridge Philos. Soc., 87 (1980), pp. 501–513.

[8] J.M. Ball, *Some open problems in elasticity*, in Geometry, Mechanics, and Dynamics, Springer, New York, 2002, pp. 3–59.

[9] D. Balzani, *Polyconvex Anisotropic Energies and Modeling of Damage Applied to Arterial Walls*, Ph.D. thesis, Department of Civil Engineering, Universität Duisburg-Essen, Essen, Germany, 2006.

[10] D. Balzani, P. Neff, J. Schröder, and G.A. Holzapfel, *A polyconvex framework for soft biological tissues. Adjustment to experimental data*, Internat. J. Solids Structures, 43 (2006), pp. 6052–5070.

[11] D. Braess, *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie,* 4th ed., Springer, Berlin, 1992.

[12] H. Bufler, *Pressure loaded structures under large deformations*, Z. Angew. Math. Mech., 64 (1984), pp. 287–295.

[13] M. Chabanas, V. Luboz, and Y. Payan, *Patient specific finite element model of the face soft tissues for computer-assisted maxillofacial surgery*, Medical Image Analysis, 7 (2003), pp. 131–151.

[14] P.G. Ciarlet, *Mathematical Elasticity Vol.* I: *Three-dimensional Elasticity*, North–Holland, Amsterdam, 1988.

[15] P.G. Ciarlet, *An Introduction to Differential Geometry with Applications to Elasticity*, Springer, Dordrecht, 2005.

[16] B. Dacorogna, *Direct Methods in the Calculus of Variations*, 2nd ed., Springer, New York, 2008.

[17] P. Deuflhard, *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms*, Springer-Verlag, Berlin, 2004.

[18] A.E. Ehret and M. Itskow, *A polyconvex hyperelastic model for fiber-reinforced materials in application to soft tissues*, J. Mater. Sci., 42 (2007), pp. 8853–8863.

[19] P.R. Fernandes, R.B. Ruben, and J. Folgado, *Bone implant design using optimization methods*, in Biomechanics of Hard Tissues: Modeling, Testing, and Materials, A. Öchsner and W. Ahmed, eds., John Wiley, New York, 2010, pp. 267–296.

[20] Y.C. Fung, *Biomechanics: Mechanical Properties of Living Tissues*, 2nd ed., Springer, New York, 1993.

[21] E. Gladilin, A. Ivanov, and V. Roginsky, *Implant shape optimization using reverse FEA*, in Medical Imaging 2005: Visualization, Image-Guided Procedures, and Display, R.L. Galloway and K.R. Cleary, eds., SPIE Proc. 5744, 2005, pp. 200–205.

[22] S. Götschel, A. Schiela, and M. Weiser, *Solving optimal control problems with the Kaskade 7 finite element toolbox*, in Advances in DUNE, Springer, Berlin, pp. 101–112.

[23] S. Hartmann and P. Neff, *Polyconvexity of generalized polynomial-type hyperelastic strain energy functions for near-incompressibility*, Int. J. Solids Struct., 40 (2003), pp. 2767–2791.

[24] G.A. Holzapfel, *Biomechanics of soft tissue*, in The Handbook of Materials Behavior Models, J. Lemaitre, ed., Academic Press, Boston, 2001, pp. 1049–1063.

[25] G.A. Holzapfel and R.W. Ogden, *Mechanics of Biological Tissue*, Springer, Berlin, New York, 2006.

[26] J.D. Humphrey, *Continuum mechanics of soft biologicial tissues*, Proc. R. Soc., Lond. A, 8 (2003), pp. 3–46.

[27] J.K. Knowles and E. Sternberg, *On the failure of ellipticity of the equations for finite elastostatic plane strain*, Arch. Rational Mech. Anal., 63 (1976), pp. 321–336.

[28] J.K. Knowles and E. Sternberg, *On the failure of ellipticity and the emergence of discontinuous deformation gradients in plane finite elastostatics*, J. Elasticity, 8 (1978), pp. 329–379.

[29] S. Lee, R.E. Caflisch, and Y.-J. Lee, *Exact artificial boundary conditions for continuum and discrete elasticity*, SIAM J. Appl. Math., 66 (2006), pp. 1749–1775.

[30] L. Lubkoll, *Optimal Control in Implant Shape Design*, Master's thesis, ZIB, TU Berlin, 2010.

[31] J.E. Marsden and J.R. Hughes, *Mathematical Foundations of Elasticity*, Prentice–Hall, New York, 1983.

[32] A. Mielke, *Necessary and sufficient conditions for polyconvexity of isotropic functions*, J. Convex Anal., 12 (2005), pp. 291–314.

[33] M. Mooney, *A theory of large elastic deformation*, J. Appl. Phys., 11 (1940), pp. 582–592.

[34] C.B. Morrey, *Multiple Integrals in the Calculus of Variations*, reprint of the 1966 edition, Springer-Verlag, Berlin, 2008.

[35] F.D. Murnaghan, *Finite Deformation of an Elastic Solid*, John Wiley, New York, 1951.

[36] R.W. Ogden, *Large deformation isotropic elasticity: On the correlation of theory and experiment for compressible rubber-like solids*, Proc. R. Soc. London A, 27 (1972), pp. 567–583.

[37] R.W. Ogden, *Non-linear Elastic Deformations*, Dover, New York, 1997.

[38] R.S. Rivlin, *Large elastic deformations of isotropic materials.* IV. *Further developments of the general theory*, Phil. Trans. Roy. Soc. London, 241 (1948), pp. 379–397.

[39] A. Schiela, *Barrier methods for optimal control problems with state constraints*, SIAM J. Optim., 20 (2009), pp. 1002–1031.

[40] K. Shavan, B.P. Lamichhane, and B. Wohlmuth, *Locking-free finite element methods for linear and nonlinear elasticity in* 2*d and* 3*d*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 4075–4086.

[41] L. Shi, H. Li, A.S.L. Fok, C. Uncer, H. Devlin, and K. Horner, *Shape optimization of dental implants*, Int. J. Oral Maxillofacial Implants, 22 (2007), pp. 911–920.

[42] D.J. Steigmann, *Frame-invariant polyconvex strain-energy functions for some anisotropic solids*, Math. Mech. Solids, 8 (2003), pp. 497–506.

[43] A. Tovar, S.E. Gano, J.J. Mason, and J.E. Renaud, *Optimum design of an interbody implant for lumbar spine fixation*, Advances in Engineering Software, 36 (2005), pp. 634–642.

[44] M. Vogelius, *An analysis of the p-version of the finite element method for nearly incompressible materials*, Numer. Math., 41 (1983), pp. 39–53.

[45] M. Weiser, P. Deuflhard, and B. Erdmann, *Affine conjugate adaptive Newton methods for nonlinear elastomechanics*, Optim. Methods Softw., 22 (2007), pp. 414–431.

[46] P.J. Wilber and J.R. Walton, *The convexity properties of a class of constitutive models for biological soft issues*, Math. Mech. Solids, 7 (2002), pp. 217–235.