Technische Universität München
Fakultät für Mathematik

Second Order Shape Optimization with Geometric Constraints

Moritz Maximilian Keuthen

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		Univ.–Prof. Dr. Boris Vexler
Prüfer der Dissertation:	1.	Univ.–Prof. Dr. Michael Ulbrich
	2.	Ao. Univ.–Prof. Dr. Wolfgang Ring Karl–Franzens–Universität Graz, Österreich
	3.	Univ.–Prof. Dr. Roland Herzog Technische Universität Chemnitz

Die Dissertation wurde am 31.08.2015 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 07.12.2015 angenommen.

Abstract

This thesis is devoted to the analysis of shape optimization problems which are constrained by partial differential equations and subject to point-wise geometric constraints. We rigorously develop and analyze suitable optimization methods, in particular Newton-type strategies. Furthermore, we establish a close relation to optimization on manifolds. For a class of elliptic model problems we investigate an approximation of the Hessian via its operator symbol, and use it for preconditioning. The second focus of this thesis are point-wise geometric constraints, and their algorithmic treatment in function space. For a special class of constraints we extend the theory of Moreau-Yosida regularization in the field of optimal control, and show its applicability in shape optimization. A more general situation is addressed with a specialized projected descent method. The proposed methods are applied to several model problems and substantiated with numerical tests.

Zusammenfassung

Diese Arbeit beschäftigt sich mit der Analyse von Shape Optimierungsproblemen, die durch partielle Differentialgleichungen sowie punktweise geometrische Bedingungen restringiert sind. Wir entwickeln und analysieren geeignete Optimierungsmethoden, insbesondere Newton-artige Verfahren. Eine enge Verbindung zur Optimierung auf Mannigfaltigkeiten wird hergestellt. Des Weiteren untersuchen wir eine Approximation der Hesse anhand ihres Operatorsymbols für eine Klasse von elliptischen Modellproblemen, und verwenden sie als Vorkonditionierer.
Der zweite Fokus dieser Arbeit liegt auf punktweisen geometrischen Nebenbedingungen und ihrer algorithmischen Behandlung im Funktionenraum. Wir erweitern die Theorie der Moreau-Yosida Regularisierung im Bereich der Optimalsteuerung für eine bestimmte Klasse von Nebenbedingungen, und zeigen ihre Anwendbarkeit in der Shape Optimierung. Für eine allgemeinere Problemstellung verwenden wir ein spezialisiertes projiziertes Abstiegsverfahren. Die vorgestellten Methoden werden auf diverse Modellprobleme angewandt und durch numerische Tests untermauert.

Contents

1. Introduction

The broad field of shape optimization treats optimization problems which involve an objective functional depending on the 'shape' of some object. The extent to which the shape may be modified ranges from problems depending only on a few shape parameters, to problems where even the topology of the object is not specified a priori. While there is a rich literature discussing abstract and academic topics in shape optimization, this field of research has also been greatly influenced by many practical applications of shape optimization problems. For example, in engineering one is often interested in finding the optimal shape of some device or component. Usually, the function measuring optimality, i.e., the objective of the optimization problem, does not only depend directly on the shape of the considered objects, but also on the state of some shape-dependent physical quantities. These quantities might comprise the temperature distribution in the object, a flow through, or around the object, a wave field interacting with the object, its mechanical properties, or other shape-dependent variables. Two generic examples of shape optimization problems, which have received a lot of attention, are the drag minimization of some body traveling in a fluid, and the maximization of the mechanical stiffness of some elastic device. Typically, mathematical models of such physical quantities are based on partial differential equations. Another important aspect of shape optimization are restrictions on the shape of the objects posed by additional constraints. These may specify smoothness requirements, or global properties like the topology and the volume. We are specifically interested in point-wise geometric constraints enforcing, for example, forbidden regions. There are many more aspects of shape optimization one could highlight here. However, already now it should become clear that giving a full account of this rich field of research is beyond the scope of this thesis. Instead, we will focus on the two topics alluded to above. For a more comprehensive exposition of the broad field of shape optimization we mention the monographs [All02, All07, Ben95, DZ11, HM03, HP05, MP01, Pir84, SZ92].

This work is devoted to the study of shape optimization problems governed by partial differential equations (PDEs). In particular, we are motivated by situations which model a flow or waves. In this context, we expect that a physical optimal solution will have some regularity, such that the modeling assumptions are satisfied. For this reason, the shape optimization framework presented in this thesis does not allow for degenerate solutions, micro structures, or similar generalized solution concepts. We suppose furthermore, that the topology of the shapes under consideration is given a-priori. Following the standard approach of shape calculus we consider families of domains given as images of an initial domain with respect to at least Lipschitz continuous transformations. At the beginning of Chapter 2 we summarize some aspects of shape optimization and specify our setting in detail.

The high computational costs of PDE-constrained optimization motivate us to study in detail the convergence properties of shape optimization algorithms. In particular, second order methods, which offer the potential for fast local convergence, are of interest. However, shapes

represent a particularly intricate field of optimization. This is due to the fact that, in general, shape spaces are *nonlinear*. The situation rather resembles the setting of optimization on manifolds. While this has been known for years, only recently one has begun to transfer well established ideas from optimization on manifolds to shape optimization, cf., e.g., [RW12, Fre12, Sch14]. However, so far these approaches have either been focussed on the Riemannian manifold point of view, requiring a lot of smoothness, or they have been restricted to rather formal considerations. We explore in Section 2.7 the connection between these two worlds in the usual setting of shape calculus, i.e., using transformations which might only be Lipschitz continuous. Still, we observe that many important concepts from optimization on manifolds have natural counterparts in shape calculus. Some of these findings seem to be original. We emphasize that there is a natural expression for the second covariant derivative. This concept is used in optimization on manifolds in lieu of the second Eulerian derivative, which is usually studied in shape optimization. Another important aspect in optimization on manifolds are retractions, i.e., mappings from the tangent space in some point back to the manifold. In shape optimization the tangent space consists of vector fields in a suitable Banach space. There exist the competing concepts of transformations determined as flows of a vector field, or as perturbations of the identity. We discuss the merits and drawbacks of these concepts from the point of view of optimization on manifolds. In our opinion it is, at least in the context of second order methods, preferable to work with perturbations of the identity in the role of a retraction.

Based on those observations, we develop and analyze suitable algorithms. One possibility to cope with the special challenges of shape optimization is to assume that the initial domain is already close to a solution. In many practical applications this is a reasonable assumption, since one is tasked to improve a previous, or expert design, which is already good. In that case, the shape optimization problem may be reformulated on a fixed reference domain, and it suffices to operate only in the unit ball of the corresponding tangent space. This well established approach yields a nonlinear optimization problem in a Banach space setting, where standard optimization techniques can be applied, cf. Section 2.12. In most of this thesis, we consider such a setting, more precisely, we describe the admissible family of domains via perturbations of a reference boundary. These are then related to transformations of the whole domain with the help of suitable extension operators, e.g., via linear elasticity. If one wants to explore a larger family of admissible domains, one has to address the manifold-like nature of shape optimization. Inspired by linesearch methods along retractions, we develop a framework for a globally convergent linesearch descent method in Section 2.8. To the best of our knowledge such a rigorous and general analysis has not been presented before. It is related to the previous approach by considering a sequence of functionals defined in the unit ball of the current tangent space. It needs to be emphasized that, although our approach was motivated by optimization methods on Riemannian manifolds, we do not require results from that area. In particular, we do *not* require C^∞-smoothness, as it is usually done for Riemannian manifolds. Indeed, our results could be presented without any reference to optimization on manifolds. We feel however, that this perspective greatly helps to understand the special challenges of shape optimization. We further extend the available theory by discussing generalized Newton methods in this framework, cf. Section 2.9. These new algorithms are demonstrated on a simple showcase problem, cf. Section 2.10. Since they were developed only recently, and are still the subject of ongoing research, we do not present more involved applications in this thesis.

Interestingly, the analysis of shape optimization methods in the continuous setting has not received much attention so far. In contrast, there is a vast amount of literature concerned with existence of solutions of shape optimization problems, cf., e.g., [AH01, DZ11, HM03] and the references therein, as well as first and second order optimality conditions. For instance, Eppler and Harbrecht discuss in a series of papers second order necessary and sufficient conditions for shape optimization problems with smooth star-shaped domains, cf., e.g., [Epp00, EHS07, EH12]. In [EHS07] the second order sufficient conditions are exploited to obtain convergence of discrete solutions to a solution of the continuous problem. However, an analysis of the proposed Newton method is not carried out. Regarding the topic of global convergence of a general descent method we are aware of the paper [Hin05] by Hintermüller. He analyzes a linesearch descent method where the admissible transformations are given as the flow maps of 'sufficiently smooth' vector fields in an appropriate Hilbert space. However, several details are left unspecified. Usually, convergence of descent methods for shape optimization is, if at all, treated on the discrete level, cf., e.g., [ABV13, HM03, HLA08]. There are also contributions which discuss convergence of Newton-type methods for particular applications, cf., e.g., [Bur04, HR04, Hin07, Lau00]. However, conditions for fast local convergence are not discussed. Solvability of the Newton equation is either assured by the addition of a regularization term to the objective, or by a suitable modification of the unregularized Hessian. In a recent work of Schulz [Sch14] a connection between shape optimization and optimization on Riemannian manifolds is drawn. The author also studies the convergence of a Riemannian Newton method. Unfortunately, his analysis borrows heavily from the theory of C^∞-smooth infinite dimensional manifolds, and is not directly applicable to less restrictive situations. In [SSW14] an extension to a Lagrange-Newton approach for shape optimization with PDEs via Riemmanian vector space bundles is described. Finally, we would like to mention the thesis of Frey [Fre12]. It presents an interesting connection between state constrained optimal control with PDEs and shape optimization which is motivated by the study of state constrained optimal control of ordinary differential equations. Furthermore, very similar to our approach, the author draws a connection between shape optimization and optimization on Riemannian manifolds. However, most of his analysis is carried out on a formal level. In particular, he does not discuss convergence properties of his developed algorithms.

Obtaining shape derivatives of functionals which depend on the solution of a shape dependent PDE is a delicate issue. It usually plays a central role in publications from that field. Most strategies fall either in the category of function space embedding methods or in the category of function space parametrization methods, cf., e.g., [SZ92, DZ11]. We focus on the latter, and show how shape derivatives can be obtained in a general, systematic way in Section 2.14. Furthermore, we provide a convenient link to the proposed shape optimization methods. The approach is exemplified for several model problems. In particular, we describe it in detail for the example of potential flow pressure matching, see Chapter 3.

In PDE-constrained optimization it is prohibitive to assemble the full Hessian. Hence, matrix-free solution strategies like the method of conjugate gradients have to be employed to solve the Newton equation. The efficiency of these methods is highly dependent on the availability of good preconditioners, i.e., approximations of the Hessian. In the context of shape optimization, the Hessian is often characterized via its operator symbol, cf., e.g., [AT96, AV99, ESSI09, Sch10]. In Chapter 4 we derive exemplarily such an approximation using the symbol of the Hessian for an application in potential flow pressure matching. The approximation can either be used

instead of the Hessian in a Newton-type method, or as a preconditioner for the true Hessian. We verify numerically the accuracy of the approximation. Our numerical experiments indicate that this approach has the potential for significant savings in computational costs.

Besides the analysis of second order methods, the second focus of this thesis are point-wise geometric constraints which restrict the admissible shapes to be located inside or outside some given regions. We consider two quite different approaches to handle such situations.

In the special case where the free part of the boundary is required to be located inside some convex set, the situation resembles an optimal control problem with control constraints. However, due to the smoothness of the control, i.e., the transformations, similar problems as in state constrained optimal control arise. Basically three approaches have been proposed in the literature on state constraints to deal with the associated difficulties. Inexact primal-dual path following techniques based on Moreau-Yosida regularization were first investigated in [IK03, HK06a, HK06b], Lavrentiev regularization methods were proposed in [Trö05, MRT06, PTW08], and barrier methods were studied in [Sch09, SG11, Kru14]. The Lavrentiev regularization concept relaxes the state constraints to mixed control and state constraints, which feature Lagrange multipliers with higher regularity. However, in our setting the smoothness of the control causes the problems. The theory of barrier methods is only available for convex optimal control problems. Since our optimization problem contains a highly nonlinear state equation, we decide to follow the approach taken in [HK06a]. We published our findings in the context of shape optimization in [KU15]. However, our analysis is applicable in the more general framework of a nonlinear optimal control problem satisfying certain conditions. We present our results in Chapter 5 in this general framework since it may be useful also in other settings, cf., e.g., [BU15] for an application in seismic tomography. We introduce a Moreau-Yosida type penalty term and study the properties of the solutions to the associated subproblems. Facing a nonlinear problem we assume a strong second order condition to hold. We show local Lipschitz continuity of the regularized solutions, and prove convergence rate estimates similar to [HSW14]. The subproblems can be solved efficiently by a semismooth Newton method [Ulb11]. We demonstrate the applicability of the developed theory to shape optimization problems on the example of potential flow pressure matching.

We also consider more general geometric constraints in the form of some regions which should be contained, or which should not be contained in the optimal domain. In contrast to the Moreau-Yosida approach, we enforce the geometric constraints strictly, and propose a special variant of a projected descent method in Section 2.13. Since projecting shapes is a delicate issue, we instead project the search directions, i.e., the vector fields determining the transformations, onto a suitable admissible subspace. This can be implemented efficiently, and ensures feasibility of the generated iterates. However, in general, we can only expect to obtain a minimum with respect to a subset of the admissible family of domains.

The proposed approaches of handling point-wise geometric constraints in shape optimization are demonstrated further with the help of two model problems.

In Chapter 6 we consider the problem of minimizing the drag of a body traveling in a Stokes fluid. We derive first and second order derivatives of the reduced objective. The drag minimization problem is usually subject to constraints regarding the volume and the center of mass of the immersed body. We treat those in an Augmented Lagrangian framework. For the arising

subproblems we apply a trust-region globalized Newton method. The search directions are determined with the truncated conjugate gradients method. Subsequently, we add geometric constraints to the problem setting, and employ the Moreau-Yosida path following strategy in combination with a trust-region globalized semismooth Newton method.

Finally, in Chapter 7, we minimize the resonance of the harbor basin with respect to long range ocean waves. For this we may modify the shape of the breakwaters which protect the harbor. The simplified physical model is described by the Helmholtz equation. In this application several geometric constraints appear naturally. We describe the varying domains with the level set approach, and apply the proposed projected descent method. Furthermore, we experiment with some heuristic extensions of our method. Note that this application and the employed optimization strategy were already described in our paper [KK15].

A short conclusion and outlook is given in Chapter 8.

2. Aspects of shape optimization

In this chapter we lay the theoretical foundation of this thesis. Let us give a brief outline.

We introduce some notations and definitions which will be used throughout this thesis in the first section. Sections 2.2–2.6 mainly summarize selected, well established concepts and results in shape optimization. For readers which are familiar with the material it should suffice to browse through these sections. As described briefly in Section 2.2, there are several possibilities to construct a *metric* on certain families of sets. We focus on families of sets obtained as transformations of an initial set $\Omega_0 \subset \mathbb{R}^d$. The admissible transformations are obtained as *perturbations of the identity*. This is the standard setting of shape calculus. In Section 2.3 we draw the connection to the *velocity method*, where the transformations are obtained as the *flow* of some vector field. We proceed by treating *continuity of shape functionals* in Section 2.4 before deriving appropriate *first and second order derivative concepts* for shape functionals in Section 2.5 and 2.6. Our presentation mainly follows the excellent treatment of general shape optimization problems by Delfour and Zolésio in [DZ11]. For a more encompassing view of shape optimization we refer to the monographs [All02, All07, Ben95, HM03, HP05, MP01, Pir84, SZ92].

Sections 2.7–2.14 contain some of the core concepts of this thesis. In particular, we discuss different shape optimization algorithms and their convergence properties. In Section 2.7 we introduce some notions from the theory of *Riemannian manifolds*, and show that these have their natural counterparts in shape calculus. Inspired by this analogy, we translate the concept of a linesearch method along *retractions* into a *globally convergent linesearch descent shape optimization algorithm* in Section 2.8. We proceed by discussing a related class of *second order methods* in Section 2.9. The developed algorithms work directly with *transformations of the whole domain*, and are demonstrated on a simple showcase problem in Section 2.10. Some alternative *characterizations of shapes* are summarized in Section 2.11. In particular, we may describe shapes as *transformations of a reference boundary* or via the *level set method*. In Section 2.12 we present the established framework of shape optimization in terms of transformations of a reference boundary. Assuming that we start close to a solution, we obtain an *optimization problem in a Banach space framework*, and can apply standard techniques. We start our discussion of *point-wise geometric constraints* in Section 2.13, and propose a special version of a *projected descent method*. Finally, Section 2.14 is devoted to the special challenges of *PDE-constrained shape optimization*. Assuming the existence of a *design-to-state operator*, we draw the connection to the developed methodology using the *function space parametrization approach*.

2.1. Some notations and definitions

Most of our notation should be standard in the field of shape optimization problems with PDEs. Nevertheless, in this section we introduce some notations and recall some basic definitions.

- We denote the *interior* of a subset $A \subset \mathbb{R}^d$ by $\operatorname{int} A$ and the *closure* by $\overline{A}$.
- The *complement* of A in $\mathbb{R}^d$ is $A^c := \{x \in \mathbb{R}^d \mid x \notin A\}$ and the *boundary* of A is defined by $\partial A := \overline{A} \cap \overline{A^c}$.
- We denote the *power set* of a nonempty set $\mathcal{D}$ by $\mathcal{P}(\mathcal{D}) := \{A \mid A \subset \mathcal{D}\}$.
- Ω will usually be a bounded open subset of $\mathbb{R}^d$.
- For some normed space X we denote the *unit ball in* X around $0 \in X$ with $B^X(0,1)$.
- $\mathcal{L}(X,Y)$ denotes the space of *bounded linear functionals* from X to Y.
- The *dual* of X is denoted by $X^* := \mathcal{L}(X,\mathbb{R})$, and the *dual pairing* by $\langle\cdot,\cdot\rangle_{X^*,X}$.
- *Partial derivatives* of an operator $E\colon X \times Y \to Z$ are denoted by $E_x(x,y), E_y(x,y)$, etc.
- Df denotes the (weak) *Jacobian* matrix of a function $f : \mathbb{R}^d \to \mathbb{R}^m$.
- If $f\colon \mathbb{R} \to \mathbb{R}^m$ is a (pseudo-) time dependent function we usually write $\partial_t f$.
- Similarly if $f\colon \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^m$ the *partial (pseudo-) time derivative* is denoted by $\partial_t f(t,x) := \lim_{s \searrow 0} s^{-1}\left(f(t+s,x) - f(t,x)\right)$ and $Df(t,x) := Df(t)(x)$ is the *Jacobian with respect to the spacial variable*. Here and in the following we sometimes write, with a slight abuse of notation, $f(t) := f(t,\cdot)\colon \mathbb{R}^d \to \mathbb{R}^m$ for $t \in \mathbb{R}$.
- The letters U, V, W will be reserved for vector fields mapping a *domain* $\Omega \subset \mathbb{R}^d$, or the *whole* $\mathbb{R}^d$ to $\mathbb{R}^d$. In contrast, u, v, w are vector fields mapping from some *boundary* to $\mathbb{R}^d$.

We adopt the usual definitions and terminology of spaces of *continuous* and *continuously differentiable* functions, $C(\Omega)$, $C^k(\Omega)$, $k \in \mathbb{N}$, as well as the *Lebesgue spaces* $L^p(\Omega)$, $1 \le p \le \infty$, and the *Sobolev spaces* $W^{k,p}(\Omega)$, $1 \le k, p \le \infty$. The space of all k-times continuously differentiable functions with *compact support* in the open set Ω is denoted by $C_c^k(\Omega)$. A function which is bounded and uniformly continuous on Ω can be extended uniquely and continuously to the closure $\overline{\Omega}$. We write $f \in C^k(\overline{\Omega})$ if all partial derivatives of a function f up to order k are bounded and uniformly continuous on Ω. If $C^k(\overline{\Omega})$ is endowed with the norm

$$\|f\|_{C^k(\Omega)} := \max_{0 \le |\alpha| \le k} \sup_{x \in \Omega} |\partial^\alpha f(x)|$$

we obtain a Banach space. In particular, this holds for $C^k(\overline{\mathbb{R}^d}) \subsetneq C^k(\mathbb{R}^d)$. A function is (k,l)-*Hölder continuous* in Ω if $k \in \mathbb{N}$, $l \in (0,1]$, and

$$\forall \alpha, 0 \le |\alpha| \le k, \exists c_\alpha > 0\colon\ \forall x, y \in \Omega, \quad |\partial^\alpha f(x) - \partial^\alpha f(y)| \le c_\alpha |x-y|^l. \tag{2.1}$$

The space of (k,l)-Hölder continuous functions is denoted by $C^{k,l}(\Omega)$ and the corresponding subset of the space $C^k(\overline{\Omega})$ by $C^{k,l}(\overline{\Omega})$. Note that also

$$C^{k,l}(\overline{\mathbb{R}^d}) \subsetneq C^{k,l}(\mathbb{R}^d). \tag{2.2}$$

We obtain a Banach space if we endow $C^{k,l}(\overline{\Omega})$ with the norm

$$\|f\|_{C^{k,l}(\Omega)} := \max\Big(\|f\|_{C^k(\Omega)}, \max_{0\leq|\alpha|\leq k} c_\alpha \Big),$$

where c_α is the smallest constant provided by (2.1). Spaces of vector valued functions will be denoted by $L^p(\Omega,\mathbb{R}^m)$, $C^k(\Omega,\mathbb{R}^m)$, etc. Finally $(0,1)$-Hölder continuity is known as *Lipschitz continuity*. Let us particularly emphasize the following identity.

Lemma 2.1. *Let Ω be a bounded, open, path-connected, and Lipschitzian subset of $\mathbb{R}^d$. Then there holds*

$$W^{k+1,\infty}(\Omega) = C^{k,1}(\overline{\Omega})$$

both algebraically and topologically for all integers $k \geq 0$.

Proof. We refer to Theorem 2.2.6 and the following Corollary in [DZ11]. □

Remark 2.2. In view of this result we denote by Df the *weak derivative* of $f \in C^{0,1}(\overline{\Omega})$ for suitable Ω.

We call a set $\Omega \subset \mathbb{R}^d$ with $\partial\Omega \neq \emptyset$ *Lipschitzian* if it is a $C^{0,1}$ *epigraph* and *equi-Lipschitzian* if it is an *equi-$C^{0,1}$ epigraph*, see [DZ11, Definition 2.5.2] for a precise definition. If $\partial\Omega$ is compact the notions coincide, cf. [DZ11, Theorem 2.5.3]. Furthermore a set Ω is equi-Lipschitzian if and only if Ω satisfies the *uniform cone condition*. We refer to [DZ11, Section 2.6.4.1] for a precise definition and the equivalence result.

Before we can start our discussion of shape optimization problems we need to answer the question how one can quantify the *difference* of two sets. In other words, given two sets we want to have a measure of the *distance* between those sets. This is the topic of the next section, where we formalize the distance between two admissible sets in the notion of a metric.

2.2. Metrics on families of sets

In this section we give an overview of the construction of some metrics on certain families of sets. We begin with the intuitive notion of deforming a fixed reference subset by a family of transformations. This is the basis of the classical shape calculus.

2.2.1. Images of a set

Apparently Micheletti [Mic72] was one of the first to introduce a complete metric topology on a family of domains of class C^k which she obtained as the images of C^k-diffeomorphisms of a fixed open domain. In particular, she studied the quotient group with respect to reparametrizations of the initial domain. She coined the term *Courant metric* for the quotient metric. The following exposition is based on [DZ11, Chapter 3].

For a normed vector space Θ of maps from $\mathbb{R}^d$ to $\mathbb{R}^d$ we consider the family of transformations

$$\mathcal{F}(\Theta) := \{\tau \colon \mathbb{R}^d \to \mathbb{R}^d \mid \tau = \mathrm{Id} + U,\ U \in \Theta, \tau \text{ bijective and } \tau^{-1} - \mathrm{Id} \in \Theta\}. \tag{2.3}$$

Under suitable assumption on Θ one can show that the family $\mathcal{F}(\Theta)$ together with the composition $(F \circ G)(x) := F(G(x))$ for $F, G \in \mathcal{F}(\Theta)$ is a *group*. There exist several equivalent right-invariant metrics on $\mathcal{F}(\Theta)$. A metric d is right-invariant if $d(F, G) = d(F \circ H, G \circ H)$ for all H. One of these metrics is defined by

$$d_{\mathcal{F}}(\mathrm{Id}, F) := \inf_{\substack{F = F_1 \circ \cdots \circ F_n \\ F_i \in \mathcal{F}(\Theta)}} \sum_{i=1}^{n} \|F_i - \mathrm{Id}\|_\Theta + \|F_i^{-1} - \mathrm{Id}\|_\Theta, \tag{2.4}$$

and extended to all $F, G \in \mathcal{F}(\Theta)$ via $d_{\mathcal{F}}(F, G) := d_{\mathcal{F}}(\mathrm{Id}, G \circ F^{-1})$. Completeness follows from some additional requirements on Θ. The family of images of a set $\Omega_0 \subset \mathbb{R}^d$ associated with $\mathcal{F}(\Theta)$ is given by

$$\mathcal{O}_\Theta(\Omega_0) := \{F(\Omega_0) \mid F \in \mathcal{F}(\Theta)\}. \tag{2.5}$$

Note however, that two different transformations in $\mathcal{F}(\Theta)$ might generate the same image set. To obtain an isomorphism between the group of transformations $\mathcal{F}(\Theta)$ and the family of images $\mathcal{O}_\Theta(\Omega_0)$ the quotient group $\mathcal{F}(\Theta)/\mathcal{G}(\Omega_0)$ may be studied, where $\mathcal{G}(\Omega_0)$ denotes the subgroup of transformations which retain Ω_0, i.e.,

$$\mathcal{G}(\Omega_0) := \{F \in \mathcal{F}(\Theta) \mid F(\Omega_0) = \Omega_0\}.$$

Given a subgroup $\mathcal{G}$ the equivalence class of F in the quotient group $\mathcal{F}/\mathcal{G}$ is given by $[F] := F \circ \mathcal{G}$. The quotient metric is defined as the infimum of the distance between all members of two equivalence classes

$$d_{\mathcal{G}}([F], [H]) := \inf_{G, \tilde{G} \in \mathcal{G}} d_{\mathcal{F}}(F \circ G, H \circ \tilde{G}) = \inf_{G \in \mathcal{G}} d_{\mathcal{F}}(F, H \circ G). \tag{2.6}$$

For the choice $\mathcal{G} = \mathcal{G}(\Omega_0)$ the quotient metric is called the *Courant metric*. We will not go into the details here, but we point out that the following theorem applies also to other spaces not considered in this thesis. Recall from (2.2) that $C^k(\overline{\mathbb{R}^d}) \subsetneq C^k(\mathbb{R}^d)$ and $C^{k,l}(\overline{\mathbb{R}^d}) \subsetneq C^{k,l}(\mathbb{R}^d)$.

Theorem 2.3. *[DZ11, Theorem 3.2.9] Let $k \geq 0$ and Θ be equal to $C^k(\overline{\mathbb{R}^d}, \mathbb{R}^d)$ or $C^{k,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$.*

(i) The group $(\mathcal{F}(\Theta), d_{\mathcal{F}})$ is a complete right-invariant metric space. For Θ equal to $C^k(\overline{\mathbb{R}^d}, \mathbb{R}^d)$ it is also a topological group.

(ii) For any closed subgroup $\mathcal{G}$ of $\mathcal{F}(\Theta)$, the function $d_{\mathcal{G}}\colon \mathcal{F}(\Theta)\times\mathcal{F}(\Theta)\to\mathbb{R}$ defined in (2.6) is a right-invariant metric on $\mathcal{F}(\Theta)/\mathcal{G}$, and the space $(\mathcal{F}(\Theta)/\mathcal{G}, d_{\mathcal{G}})$ is complete. The topology induced by $d_{\mathcal{G}}$ coincides with the quotient topology of $\mathcal{F}(\Theta)/\mathcal{G}$.

(iii) If $\Omega_0\subset\mathbb{R}^d$ is nonempty and either closed or satisfies $\Omega_0 = \operatorname{int}\overline{\Omega_0}$, then $\mathcal{G}(\Omega_0)$ is a closed subgroup of $\mathcal{F}(\Theta)$.

Remark 2.4. Murat and Simon [MS76] followed a similar approach in 1976. They studied spaces which are equivalent to either $C^{k+1}(\overline{\mathbb{R}^d},\mathbb{R}^d)$ or $C^{k,1}(\overline{\mathbb{R}^d},\mathbb{R}^d)$, and constructed a metric via the semimetric $d_0(I,F) := \|F-\mathrm{Id}\|_\Theta + \|F^{-1}-\mathrm{Id}\|_\Theta$.

Obviously we have $\mathcal{F}(\Theta)\subsetneq \mathrm{Id}+\Theta$. For sufficiently small *perturbations $U\in\Theta$ of the identity* there holds also $(\mathrm{Id}+U)\in\mathcal{F}(\Theta)$, and the map $t\mapsto \mathrm{Id}+tU$ defines a C^1-path in $\mathcal{F}(\Theta)$.

Theorem 2.5. *[DZ11, Theorem 3.2.14 and 3.2.17] Let $k\geq 0$ and Θ be equal to $C^{k+1}(\overline{\mathbb{R}^d},\mathbb{R}^d)$ or $C^{k,1}(\overline{\mathbb{R}^d},\mathbb{R}^d)$.*

(i) The map $U\mapsto \mathrm{Id}+U\colon B^\Theta(0,1)\subset\Theta\to\mathcal{F}(\Theta)$ is well defined and continuous.

(ii) For all $F\in\mathcal{F}(\Theta)$ the tangent space to $\mathcal{F}(\Theta)$ in F is given by Θ.

Sketch of the proof. (i) The invertibility of $\mathrm{Id}+U$ is obtained by showing that for every $y\in\mathbb{R}^d$ the map $x\mapsto y-U(x)$ has a fixed-point. The implicit function theorem is employed to show that $(\mathrm{Id}+U)^{-1}-\mathrm{Id}\in\Theta$. Via the estimate $\|(\mathrm{Id}+U)^{-1}-\mathrm{Id}\|_\Theta \leq c\,\|(\mathrm{Id}+U)-\mathrm{Id}\|_\Theta = c\,\|U\|_\Theta$ the continuity property is established.
(ii) Every tangent element to $\mathcal{F}(\Theta)\subset\mathrm{Id}+\Theta$ will be contained in Θ. On the other hand, for any $U\in\Theta$ and $F\in\mathcal{F}(\Theta)$, the map $t\mapsto(\mathrm{Id}+tU)\circ F$ defines a continuous path in $\mathcal{F}(\Theta)$ for small $t\geq 0$. In particular the tangent vector to this path is U. □

Remark 2.6. Let us stress the importance of the second item in the above theorem. Knowledge of the appropriate tangent space will be the key to defining derivatives of functionals which depend on a shape. This sets the group $\mathcal{F}(\Theta)$ apart from its quotient groups and other groups of shapes. Although these may have desirable attributes, to the best of our knowledge there is so far *no rigorous* characterization of their tangent spaces available. For the quotient group $\mathcal{F}(\Theta)/\mathcal{G}(\Omega_0)$ the natural conjecture is that the tangent space consists of all vector fields which are *normal* to the boundary. This guess is based on the *structure theorem* of shape optimization, see Theorem 2.78 or [DZ92, Theorem 3.2]. This conjecture is true in the case of the Riemannian manifold given by the quotient group of all C^∞-embeddings of the unit circle in the plane, where the equivalence classes are specified by reparametrizations of the circle, cf. [MM06]. See also the discussion in [Fre12, Section 2.6.2].

Theorem 2.5 states in particular that $(U_n)\to 0$ in Θ implies $d_{\mathcal{F}}(\mathrm{Id},\mathrm{Id}+U_n)\to 0$. We also have the opposite implication.

Lemma 2.7. *Let $k\geq 0$ and Θ be equal to $C^{k+1}(\overline{\mathbb{R}^d},\mathbb{R}^d)$ or $C^{k,1}(\overline{\mathbb{R}^d},\mathbb{R}^d)$. For every sequence $(F_n)\subset\mathcal{F}(\Theta)$ which satisfies $d_{\mathcal{F}}(\mathrm{Id},F_n)\to 0$ for $n\to\infty$, it holds*

$$\|F_n-\mathrm{Id}\|_\Theta\to 0.$$

Proof. It suffices to show that there exists a constant $c(\Theta) > 0$ such that, for any sequence $(F_n) \subset \mathcal{F}(\Theta)$ with $d_{\mathcal{F}}(\mathrm{Id}, F_n) < \varepsilon_n < 1$ and $\varepsilon_n \to 0$, we have the bound $\|F_n - \mathrm{Id}\|_\Theta < \varepsilon_n c(\Theta)$. Let such a sequence be given. For every n there exists a finite factorization $F_n = F_1^n \circ \cdots \circ F_\nu^n$ such that

$$\sum_{i=1}^{\nu} \|F_i^n - \mathrm{Id}\|_\Theta + \|(F_i^n)^{-1} - \mathrm{Id}\|_\Theta \leq d(\mathrm{Id}, F_n) + \varepsilon_n \leq 2\varepsilon_n.$$

The existence of a constant $c(\Theta) > 0$ such that $\|F_n - \mathrm{Id}\|_\Theta < \varepsilon_n c(\Theta)$ is now provided by [DZ11, Assumption 3.2.2]. It is satisfied for the considered spaces, see [DZ11, Sections 2.5, 2.6]. □

Whereas transformations in $\mathcal{F}(C^1(\overline{\mathbb{R}^d}, \mathbb{R}^d))$ conserve Lipschitz domains, this is in general *not* true for $\mathcal{F}(C^{0,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d))$, see [DZ11, Example 2.5.1]. Nevertheless, the next result shows that the regularity is preserved for *small* Lipschitz transformations.

Lemma 2.8. *[BFCLS97, Lemma 3] Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then there exists a constant $0 < c(\Omega) < 1$ such that*

$$\begin{aligned} &\tau(\Omega) \text{ is a bounded Lipschitz domain for all } \tau = \mathrm{Id} + U, \\ &\text{satisfying } U \in \Theta := C^{0,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d) \text{ and } \|U\|_\Theta \leq c(\Omega). \end{aligned}$$

Finally, we recall the following classical result of Nečas, see also [MS76, Lemma 4.1].

Lemma 2.9. *[Neč12, Section 2.3.1] Let $\Omega \subset \mathbb{R}^d$ be a bounded open domain, $1 \leq p \leq \infty$, and $F \in \mathcal{F}(C^{0,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d))$. Then*

(i) $f \in L^p(F(\Omega)) \Leftrightarrow f \circ F \in L^p(\Omega)$.

(ii) $f \in W^{1,p}(F(\Omega)) \Leftrightarrow f \circ F \in W^{1,p}(\Omega)$.

(iii) $f \in W_0^{1,p}(F(\Omega)) \Leftrightarrow f \circ F \in W_0^{1,p}(\Omega)$, *if* $1 < p < \infty$.

2.2.2. Other groups of sets and associated metrics

Considering only images of a fixed set Ω_0 is convenient, but also quite restrictive. We give a brief overview of alternative constructions. It is out of the scope of this thesis to cover these in more detail. As already mentioned, the reason why we focus on $\mathcal{O}_\Theta(\Omega_0)$ is that we know the tangent space of $\mathcal{F}(\Theta)$.

Metrics via characteristic functions

A much larger class of sets than the one considered in section Section 2.2.1 is the class of Lebesgue measurable sets. They are identified with their *characteristic function*

$$\chi_\Omega(x) := \begin{cases} 1, & \text{if } x \in \Omega, \\ 0, & \text{if } x \notin \Omega. \end{cases}$$

Generalizing the notion of the *perimeter* of a set provides us with a *compactness* result. We summarize some results from [DZ11, Chapter 5], where this approach is treated in more detail.

In this paragraph we consider a nonempty holdall $\mathcal{D} \subset \mathbb{R}^d$ which is measurable and bounded. The space of characteristic functions on $\mathcal{D}$ is denoted by

$$\mathcal{X}(\mathcal{D}) := \{\chi_\Omega \mid \Omega \subset \mathcal{D} \text{ Lebesgue measurable}\} \subset L^\infty(\mathcal{D}).$$

Define

$$A \Delta B := (A \cap B^c) \cup (B \cap A^c), \text{ and } \chi_A \Delta \chi_B := |\chi_A - \chi_B| = \chi_{A\Delta B}.$$

If $\mathcal{X}(\mathcal{D})$ is endowed with the symmetric set difference Δ as multiplication, and the neutral multiplicative element $\chi_\emptyset$, then $\mathcal{X}(\mathcal{D})$ is an Abelian group. One may introduce equivalence classes of Lebesgue measurable sets, by identifying them with the L^p-equivalence class of their characteristic functions

$$[A] \leftrightarrow \chi_A \in L^p(\mathcal{D}).$$

The function $d_{\mathcal{X},p}([A_2],[A_1]) := \|\chi_{A_2} - \chi_{A_1}\|_{L^p(\mathcal{D})}$ defines a complete metric structure on $\mathcal{X}(\mathcal{D})$, that makes it a topological Abelian group for $1 \le p < \infty$, cf. [DZ11, Theorem 5.2.2]. The topologies induced by $L^p(\mathcal{D})$ on $\mathcal{X}(\mathcal{D})$ are all equivalent for $1 \le p < \infty$, cf. [DZ11, Theorem 5.2.3]. A sequence is said to be strongly convergent if it is strongly convergent in $L^p(\mathcal{D})$ for some $p \in [1, \infty)$. One can approximate an arbitrary Lebesgue measurable subset of $\mathbb{R}^d$ by a strongly convergent sequence of open C^∞-domains, see [DZ11, Theorem 5.3.1].

We extend now the notion of the perimeter of a set to $\mathcal{X}(\mathcal{D})$ and state the announced compactness property. Recall the *space of (vectorial) bounded measures* $M^1(\mathcal{D}, \mathbb{R}^d) = C_c(\mathcal{D}, \mathbb{R}^d)^*$, and the space of *functions of bounded variation* $BV(\mathcal{D}) := \{f \in L^1(\mathcal{D}) \mid \nabla f \in M^1(\mathcal{D}, \mathbb{R}^d)\}$ with the norm $\|f\|_{BV(\mathcal{D})} = \|f\|_{L^1(\mathcal{D})} + \|\nabla f\|_{M^1(\mathcal{D},\mathbb{R}^d)}$.

Definition 2.10. *Let Ω be a Lebesgue measurable subset of $\mathbb{R}^d$. The* perimeter of Ω with respect to an open subset $\mathcal{D}$ *of $\mathbb{R}^d$ is given by*

$$p_{\mathcal{D}}(\Omega) := \|\nabla \chi_\Omega\|_{M^1(\mathcal{D},\mathbb{R}^d)}.$$

We set $B\mathcal{X}(\mathcal{D}) := \{\, \chi_\Omega \in \mathcal{X}(\mathcal{D}) \mid \chi_\Omega \in BV(\mathcal{D})\}$. We say that a set Ω has finite perimeter *if $\chi_\Omega \in B\mathcal{X}(\mathbb{R}^d)$. Sets of finite perimeter are also called* Caccioppoli sets.

Theorem 2.11. *[DZ11, Theorem 5.6.3] Assume that $\mathcal{D}$ is a bounded open domain in $\mathbb{R}^d$ with a Lipschitzian boundary. Let $\{\Omega_n\}$ be a sequence of measurable domains in $\mathcal{D}$ with uniformly bounded perimeter $p_{\mathcal{D}}(\Omega_n) \leq c$ for some $c > 0$. Then there exists a measurable set $\Omega \subset \mathcal{D}$ and a subsequence $\{\Omega_{n_k}\}$, such that*

$$\chi_{\Omega_{n_k}} \to \chi_\Omega \text{ in } L^1(\mathcal{D}) \text{ as } k \to \infty, \text{ and } p_{\mathcal{D}}(\Omega) \leq \liminf_{k\to\infty} p_{\mathcal{D}}(\Omega_{n_k}) \leq c.$$

Moreover $\lim_{k\to\infty} \langle \nabla \chi_{\Omega_{n_k}}, \varphi \rangle_{M^1(\mathcal{D},\mathbb{R}^d),C_c(\mathcal{D},\mathbb{R}^d)} \to \langle \nabla \chi_\Omega, \varphi \rangle_{M^1(\mathcal{D},\mathbb{R}^d),C_c(\mathcal{D},\mathbb{R}^d)} \ \forall \varphi \in C_c(\mathcal{D}, \mathbb{R}^d)$.

In particular the following family of sets has uniformly bounded perimeter.

Theorem 2.12. *[DZ11, Theorem 5.6.11] Let $\mathcal{D}$ be a bounded open set in $\mathbb{R}^d$ with uniformly Lipschitzian boundary. The family of characteristic functions of all Lebesgue measurable subsets of $\mathcal{D}$ which satisfy the uniform cone property [DZ11, Section 2.6.4.1] with the* same constant parameters *is* compact *in $L^p(\mathcal{D})$ for all $p \in [1, \infty)$.*

Remark 2.13.

(i) It is a standard technique in optimization to show the existence of an optimal solution by combining compactness of the admissible set with lower semicontinuity of the objective. The above results make it possible to apply this line of reasoning also to shape optimization.

(ii) Let us stress, that in contrast to $\mathcal{F}(\Theta)$, it is not at all clear what the tangent space to the group $X(\mathcal{D})$ is, see also [DZ11, Remark 5.2.3].

(iii) If one wants to guarantee the conditions of Theorem 2.12 one needs to impose strong smoothness conditions on the admissible domain deformations. Instead, one often adds a *perimeter penalty* to the objective functional to obtain the existence of a solution.

(iv) Shape calculus, and hence shape optimization, is usually based on the group $\mathcal{F}(\Theta)$ and *not* the group $\mathcal{X}(\mathcal{D})$. This is due to (ii). Hence there is a deplorable gap between the standard analysis of existence of solutions, and the standard optimization framework which tries to find those solutions. In most of this thesis we will not concern ourselves with the question of existence of solutions. Instead, we suppose that solutions exist, and focus on the question how to find one.

Metrics via the distance or oriented distance functions

One can also construct a metric on the family of all *distance functions* of subsets of $\mathcal{D}$, where the distance function from a point x to the set $A \subset \mathbb{R}^d$ is given by

$$d_A(x) := \begin{cases} \inf_{y\in A} |y - x|, & A \neq \emptyset, \\ +\infty, & A = \emptyset. \end{cases} \tag{2.7}$$

This metric is equivalent to the *Pompéiu-Hausdorff metric*. We refer to [DZ11, Chapter 6] for more details regarding this subject as well as the *Hausdorff complementary metric*.

Another possibility is to study the *oriented distance function* from $x \in \mathbb{R}^d$ to $A \subset \mathbb{R}^d$

$$b_A(x) := d_A(x) - d_{A^c}(x).$$

The oriented distance functions are a fascinating subject to study. They are intrinsically linked to various geometric properties of their associated set, like smoothness of the boundary, convexity of the set, or the unit outward normal on the boundary of the set. Furthermore, they can be used to construct a uniform metric topology on suitable equivalence classes of sets. We will revisit the oriented distance function in Section 2.11.3. A thorough exposition covering these and many more aspects can be found in [DZ11, Chapter 7].

2.3. Transformations generated by velocities

Recall that the tangent space of the group of transformations $\mathcal{F}(\Theta)$ is given by Θ, cf. Theorem 2.5. We will now discuss a special class of transformations, which describe the *flow* associated with some vector field over an artificial time interval $[0,1]$ with values in Θ. This approach is termed *velocity* or *speed method*, and goes back to Zolésio [Zol73, Zol79]. Besides shape optimization, this approach finds a lot of applications in the wide field of imaging and motion capturing. We refer to [You10] for an introduction to this point of view. The current section is mainly based on [DZ11, Chapter 4], and there the reader may find also a more detailed overview and further references.

2.3.1. The subgroup of flow maps

In this paragraph we assume that the tangent space is given by $\Theta = C^{0,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$. We consider velocity fields $\mathscr{V}\colon [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, and writing $\mathscr{V}(t) := \mathscr{V}(t,\cdot)$ we define

$$\begin{aligned} \|\mathscr{V}\|_{L^1([0,1],\Theta)} &:= \int_0^1 \|\mathscr{V}(t)\|_\Theta \, \mathrm{d}t, \text{ and} \\ L^1([0,1],\Theta) &:= \{\mathscr{V}\colon [0,1] \times \mathbb{R}^d \to \mathbb{R}^d \mid \|\mathscr{V}\|_{L^1([0,1],\Theta)} < \infty\}. \end{aligned}$$

For a velocity field $\mathscr{V} \in L^1([0,1],\Theta)$ and every $x_0 \in \mathbb{R}^d$ the differential equation

$$\partial_t x(t) = \mathscr{V}(t, x(t)), \quad x(0) = x_0, \tag{2.8}$$

has a unique solution denoted by $x_\mathscr{V}(\cdot; x_0)$ in $W^{1,1}((0,1),\mathbb{R}^d) \subset C([0,1],\mathbb{R}^d)$, cf. [DZ11, Section 4.2.1] and [You10, Appendix C].

Definition 2.14. *For $\vartheta > 0$ and $\mathscr{V} \in L^1([0,\vartheta],\Theta)$ the* flow *or* flow map *$T_\mathscr{V}$ associated with $\mathscr{V}$ is given by*

$$T_\mathscr{V}\colon [0,\vartheta] \times \mathbb{R}^d \to \mathbb{R}^d, \quad (t, x_0) \mapsto T_\mathscr{V}(t, x_0) := x_\mathscr{V}(t; x_0). \tag{2.9}$$

Furthermore, we abbreviate $T_\mathscr{V}(t) := T_\mathscr{V}(t,\cdot)$ and define $U_\mathscr{V}(t) := T_\mathscr{V}(t) - \mathrm{Id}$. For an autonomous vector field $\mathscr{V}(t,x) \equiv V(x)$ we write T_V.

Theorem 2.15. *[DZ11, Theorem 4.2.1] The set*

$$\mathcal{G}_\Theta := \{T_\mathscr{V}(1) \mid \mathscr{V} \in L^1([0,1],\Theta)\}$$

is a subgroup of $\mathcal{F}(\Theta)$, and for every $\mathscr{V} \in L^1([0,1],\Theta)$ it holds

$$\sup_{0\le t\le 1} \|U_\mathscr{V}(t)\|_\Theta \le \|\mathscr{V}\|_{L^1([0,1],\Theta)} \exp\left(2 + 2\|\mathscr{V}\|_{L^1([0,1],\Theta)}\right). \tag{2.10}$$

In particular, for all $t \in [0,1]$ and every $\mathscr{V} \in L^1([0,1],\Theta)$, it holds $T_\mathscr{V}(t) \in \mathcal{F}(\Theta)$, and the map $t \mapsto T_\mathscr{V}(t)\colon [0,1] \to \mathcal{F}(\Theta)$ defines a continuous path in $\mathcal{F}(\Theta)$.

Sketch of the proof. To show that $\mathcal{G}_\Theta$ is closed under composition, i.e., $T_{\mathscr{V}_1}(1) \circ T_{\mathscr{V}_2}(1) \in \mathcal{G}_\Theta$ one constructs a suitable concatenation of $\mathscr{V}_1$ and $\mathscr{V}_2$ which is in $L^1([0,1],\Theta)$. To realize that each $T_\mathscr{V}(1)$ has an inverse in $\mathcal{G}_\Theta$ one verifies that for $\mathscr{V}^-(t,x) := -\mathscr{V}(1-t,x)$ it holds

$$T_{\mathscr{V}^-}(1) = (T_\mathscr{V}(1))^{-1}. \tag{2.11}$$

Finally, one needs to check $\mathcal{G}_\Theta \subset \mathcal{F}(\Theta)$. This is done via (2.10), which is obtained after some estimations and exploiting the choice $\Theta = C^{0,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$. □

Remark 2.16. (i) Note that $T_V(\varepsilon t) = T_{\varepsilon V}(t)$ for $\varepsilon > 0$ and all $V \in \Theta$. In particular, for any autonomous velocity field $V \in \Theta$ and any $T > 0$ the associated flow satisfies $T_V(T) \in \mathcal{F}(\Theta)$.

(ii) An attractive property of the flow map is the characterization (2.11), i.e., its inverse is easily obtained by inverting the time and the direction of the associated velocity field.

One can construct a metric on $\mathcal{G}_\Theta$ that is of geodesic type.

Theorem 2.17. *[DZ11, Theorem 4.2.2] Let $k \ge 0$, and Θ be equal to $C^{k,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$ or $C^{k+1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$. Then the function $d_{\mathcal{G}_\Theta}\colon \mathcal{G}_\Theta \times \mathcal{G}_\Theta \to \mathbb{R}$, given by*

$$d_{\mathcal{G}_\Theta}(\mathrm{Id}, T_\mathscr{V}(1)) := \inf\left\{ \int_0^1 \|\mathscr{W}(t)\|_\Theta \, \mathrm{d}t \;\middle|\; \mathscr{W} \in L^1([0,1],\Theta),\ T_\mathscr{W}(1) = T_\mathscr{V}(1) \right\}$$
$$d_{\mathcal{G}_\Theta}(T_\mathscr{V}(1), T_\mathscr{W}(1)) := d_G(\mathrm{Id}, T_\mathscr{V}(1) \circ T_\mathscr{W}(1)^{-1}),$$

defines a right-invariant metric on $\mathcal{G}_\Theta$.

Sketch of the proof. By construction $d_{\mathcal{G}_\Theta}$ is non-negative, symmetric, and right-invariant. The triangle inequality is verified by considering a suitable concatenation of velocity fields. Finally, the identity of indiscernibles is checked with the help of Theorem 2.15. □

Unfortunately, the completeness of $(\mathcal{G}_\Theta, d_{\mathcal{G}_\Theta})$ and related questions are still open, we refer to the discussion in [DZ11, Section 4.4.2].

2.3.2. Equivalence between transformations and velocities

There is a close connection between a velocity and its associated transformation. We summarize here the results in the case of a region of interest $\emptyset \neq \mathcal{D} \subset \mathbb{R}^d$. Of course this includes the choice $\mathcal{D} = \mathbb{R}^d$. We recall the following characterization of the closed and convex *Clarke tangent cone* $C_{\mathcal{D}}(x)$ to $\overline{\mathcal{D}}$ at $x \in \overline{\mathcal{D}}$.

Definition 2.18. *[Cla90, Theorem 2.4.5] A vector $v \in \mathbb{R}^d$ belongs to the Clarke tangent cone $C_{\mathcal{D}}(x)$ at $x \in \mathcal{D}$ if and only if, for every sequence $(x_n) \subset \mathcal{D}$ converging to x and every sequence $(t_n) \subset (0, \infty)$ decreasing to 0, there exists a sequence $(v_n) \subset \mathbb{R}^d$ converging to v such that $x_n + t_n v_n \in \mathcal{D}$ for all n.*

We require the following for a vector field $\mathscr{V}\colon [0, \vartheta] \times \overline{\mathcal{D}} \to \mathbb{R}^d$, cf. [DZ11, Section 4.5.1].

Assumption 2.1. *There exists a $\vartheta > 0$ and a vector field $\mathscr{V}\colon [0, \vartheta] \times \overline{\mathcal{D}} \to \mathbb{R}^d$ such that*

$$\begin{aligned} \forall x \in \overline{\mathcal{D}}, &\quad \mathscr{V}(\cdot, x) \in C([0, \vartheta], \mathbb{R}^d), \\ \exists c > 0, \forall x, y \in \overline{\mathcal{D}}, &\quad \|\mathscr{V}(\cdot, x) - \mathscr{V}(\cdot, y)\|_{C([0,\vartheta], \mathbb{R}^d)} \leq c|x - y|, \\ \forall t \in [0, \vartheta], \forall x \in \overline{\mathcal{D}}, &\quad \mathscr{V}(t, x) \in \{-C_{\mathcal{D}}(x) \cap C_{\mathcal{D}}(x)\}. \end{aligned}$$

Recall the short notation $\mathscr{V}(t) := \mathscr{V}(t, \cdot)$. In particular, the above conditions imply that $\mathscr{V}(\cdot) \in C([0, \vartheta], C(\overline{\mathcal{D}}, \mathbb{R}^d))$ for an open, bounded holdall $\mathcal{D} \subset \mathbb{R}^d$. On the other hand we may require the following for a transformation $T\colon [0, \vartheta] \times \overline{\mathcal{D}} \to \mathbb{R}^d$.

Assumption 2.2. *There exists a $\vartheta > 0$, and a map $T\colon [0, \vartheta] \times \overline{\mathcal{D}} \to \mathbb{R}^d$ such that*

$$\begin{aligned} \forall x \in \overline{\mathcal{D}}, &\quad T(\cdot, x) \in C^1([0, \vartheta], \mathbb{R}^d) \\ \exists c_1 > 0, \forall x, y \in \overline{\mathcal{D}}, &\quad \|T(\cdot, x) - T(\cdot, y)\|_{C^1([0,\vartheta], \mathbb{R}^d)} \leq c_1|x - y|, \end{aligned} \tag{2.12}$$

$$\forall t \in [0, \vartheta] \text{ the map } T_t\colon \overline{\mathcal{D}} \to \overline{\mathcal{D}} \text{ given by } T_t(x) := T(t, x) \text{ is bijective,} \tag{2.13}$$

$$\begin{aligned} \forall \tilde{x} \in \overline{\mathcal{D}}, &\quad T^{-1}(\cdot, \tilde{x}) \in C([0, \vartheta], \mathbb{R}^d) \\ \exists c_2 > 0, \forall \tilde{x}, \tilde{y} \in \overline{\mathcal{D}}, &\quad \left\|T^{-1}(\cdot, \tilde{x}) - T^{-1}(\cdot, \tilde{y})\right\|_{C([0,\vartheta], \mathbb{R}^d)} \leq c_2|\tilde{x} - \tilde{y}|, \end{aligned} \tag{2.14}$$

where we define $T^{-1}(t, \tilde{x}) := T_t^{-1}(\tilde{x})$.

We have the following equivalence result available.

Theorem 2.19. *[DZ11, Theorems 4.5.1 and 4.5.2]*

(i) If Assumption 2.1 is satisfied for $\mathscr{V}$, then Assumption 2.2 holds for the flow map $T_{\mathscr{V}}$ from Definition 2.14.

(ii) If Assumption 2.2 is satisfied for a map T, then the map

$$\mathscr{V}\colon [0,\vartheta]\times\overline{\mathcal{D}}\to\mathbb{R}^d,\ (t,x)\mapsto\mathscr{V}(t,x):=\partial_t T\left(t,T_t^{-1}(x)\right)$$

satisfies Assumption 2.1. If additionally $T(0,\cdot)=\mathrm{Id}$, then the solution of (2.8) for that $\mathscr{V}$ is given by $T(\cdot,x_0)$.

(iii) Given a $\vartheta>0$ and a map $T\colon[0,\vartheta]\times\overline{\mathcal{D}}\to\overline{\mathcal{D}}$, with $T(0,\cdot)=\mathrm{Id}$, satisfying the conditions (2.12) and (2.13), there exists a $\tilde{\vartheta}>0$ such that the conclusions of part (ii) hold on the intervall $[0,\tilde{\vartheta}]$.

Sketch of the proof. (i) Existence and uniqueness of viable solutions of (2.8) is shown via a special version of *Nagumo's theorem*. Bijectivity is obtained by considering the reverse flowmap. The condition (2.14) is then easily checked.
(ii) The continuity properties of $\mathscr{V}$ are verified directly using Assumption 2.2. The condition $\mathscr{V}(t,x)\in\{-C_{\mathcal{D}}(x)\cap C_{\mathcal{D}}(x)\}$ is proven in two steps with an elemental ε - δ argument.
(iii) Choosing $\tilde{\vartheta}=\min\{\vartheta,1/(2c_1^2)\}$ one can verify (2.14) and hence apply (ii). □

Remark 2.20. (i) In particular, the flow map $T_{\mathscr{V}}(t)\colon\overline{\mathcal{D}}\to\overline{\mathcal{D}}$ is a homeomorphism which maps interior points onto interior points and boundary points onto boundary points if Assumption 2.1 is satisfied, cf. [DZ11, Remark 4.5.1].

(ii) The theorem can be extended to smoother mappings. In [DZ11, Section 4.4.3] this is done for the choice $\mathcal{D}=\mathbb{R}^d$.

2.4. Continuity of shape functionals

In this section we introduce a concept of continuity of a functional with regard to shape changes. More precisely we introduce continuity with respect to the metric $d_{\mathcal{F}}$ from (2.4). It can be shown that notion is equivalent to continuity along velocity fields. Recall the group of transformations $\mathcal{F}(\Theta)$ from (2.3), and the corresponding images $\mathcal{O}_\Theta(\Omega_0)$ of a set Ω_0 from (2.5).

Definition 2.21. *[DZ11, Definition 4.3.1] Let $\emptyset\neq\mathcal{D}\subset\mathbb{R}^d$ be given with $\mathcal{O}\subset\mathcal{P}(\mathcal{D})$ and let $\mathcal{B}$ be a topological space. It is common to call $\mathcal{O}$ the* admissible family of sets. *A* shape functional *is a map*

$$\mathfrak{j}\colon\mathcal{O}\to\mathcal{B}.$$

Note that a shape functional can not discern between to different members of the group of transformations $\mathcal{F}(\Theta)$ which describe the same set. To be more precise let $\Omega_0\subset\mathbb{R}^d$, $\Omega\in\mathcal{O}_\Theta(\Omega_0)$, and $\mathfrak{j}\colon\mathcal{O}_\Theta(\Omega_0)\to\mathcal{B}$ be a shape functional into some topological space $\mathcal{B}$. Then for all $F,H\in\mathcal{F}(\Theta)$ with $\Omega=F(\Omega_0)=H(\Omega_0)$ it holds

$$\mathfrak{j}(\Omega)=\mathfrak{j}(F(\Omega_0))=\mathfrak{j}(H(\Omega_0)).$$

Remark 2.22. (i) This constancy of a shape functional with respect to the members of an equivalence class of $\mathcal{F}(\Theta)/\mathcal{G}(\Omega_0)$ motivates the investigation of the quotient group. It would be a very interesting and worthwhile endeavor to extend the analysis of this chapter also to $\mathcal{F}(\Theta)/\mathcal{G}(\Omega_0)$, but this is beyond the scope of this thesis.

(ii) In practice one usually has no problems when using $\mathcal{F}(\Theta)$ in a derivative based optimization algorithm since a linesearch along a descent direction will leave the current equivalence class. The situation is more delicate when considering Newton's method, as the Hessian will always have a nontrivial kernel consisting of directions which do not change the shape of the domain. We discuss this issue in more detail in Section 2.9.

One may define the continuity of a shape functional in quite general terms if one has a suitable metric structure associated with $\mathcal{O}$. However, for the sake of brevity, we present only the setting for $\mathcal{F}(\Theta)$. This suffices for the purposes of this thesis.

Definition 2.23. *Consider a nonempty set $\Omega_0 \subset \mathbb{R}^d$, and let Θ be equal to $C^{k+1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$ or $C^{k,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$ for $k \geq 0$. A shape functional $\mathfrak{j}\colon \mathcal{O}_\Theta(\Omega_0) \to \mathcal{B}$, mapping into some Banach space $\mathcal{B}$, is* continuous at $F \in \mathcal{F}(\Theta)$ with respect to $d_{\mathcal{F}}$, *if for all $\varepsilon > 0$ there exists a $\delta > 0$, such that*

$$\forall G \in \mathcal{F}(\Theta),\ \textit{with } d_{\mathcal{F}}(F,G) < \delta,\ \textit{it holds } \ \|\mathfrak{j}(G(\Omega_0)) - \mathfrak{j}(F(\Omega_0))\|_{\mathcal{B}} < \varepsilon.$$

It may be easier to check the continuity only along velocity fields. The next result shows that this is enough to guarantee continuity with regard to $d_{\mathcal{F}}$. Let α be a *multiindex* and $\partial^\alpha G$ be the corresponding partial derivative of a function $G\colon \mathbb{R}^d \to \mathbb{R}^d$. We define

$$\operatorname{Lip}(G) := \sup_{y \neq x} \frac{|G(y) - G(x)|}{|y - x|} \text{ and } \forall k \geq 0 :\ \operatorname{Lip}_k(G) := \sum_{|\alpha| = k} \operatorname{Lip}(\partial^\alpha G).$$

Theorem 2.24. *Let Ω_0 be a nonempty, open subset of $\mathbb{R}^d$ satisfying $\partial\Omega_0 \backslash \partial\overline{\Omega_0} = \emptyset$, $\mathcal{B}$ be a Banach space, and $\Theta = C^{k,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$ for some $k \geq 0$. Consider a shape functional $\mathfrak{j}\colon \mathcal{O}_\Theta(\Omega_0) \to \mathcal{B}$. Then $\mathfrak{j}$ is continuous at* Id *for the metric $d_{\mathcal{F}}$ if and only if*

$$\lim_{t \searrow 0} \mathfrak{j}(T_{\mathscr{V}}(t, \Omega_0)) = \mathfrak{j}(\Omega_0),$$

for all $\mathscr{V} \in C([0,\vartheta], C^k(\overline{\mathbb{R}^d}, \mathbb{R}^d))$ satisfying the uniform Lipschitz condition $\operatorname{Lip}_k(\mathscr{V}(t)) \leq L$ for some constant $L > 0$ independent of $t \in [0,\vartheta]$.

Sketch of the proof. One can directly apply the proof of [DZ11, Theorem 4.6.3]. It is sufficient to show the theorem for a real-valued shape functional. Otherwise one can work with the auxiliary functional $j(F) := \|\mathfrak{j}(F(\Omega)) - \mathfrak{j}(\Omega)\|_{\mathcal{B}}$. If $\mathfrak{j}$ is $d_{\mathcal{F}}$-continuous we can combine Theorem 2.19 and the inequality

$$d_{\mathcal{F}}(T_{\mathscr{V}}(t), \mathrm{Id}) \leq \|T_{\mathscr{V}}(t)^{-1} - \mathrm{Id}\|_\Theta + \|T_{\mathscr{V}}(t) - \mathrm{Id}\|_\Theta$$

to obtain the claim. Conversely, given a sequence $(F_n) \subset \mathcal{F}(\Theta$ with $d_{\mathcal{F}}(F_n, \mathrm{Id}) \to 0$, one constructs a velocity $\mathscr{V}$ satisfying Assumption 2.1 and again employs Theorem 2.19 to show that $\mathfrak{j}$ is $d_{\mathcal{F}}$-continuous. The velocity $\mathscr{V}$ is obtained by considering a suitable interpolation of the sequence F_n. □

Remark 2.25. In [DZ11, Theorem 4.6.3] this result is shown for the Courant metric on $\mathcal{F}(C^{k,1}(\overline{\mathbb{R}^d},\mathbb{R}^d))/\mathcal{G}(\Omega)$. A similar result applies for $C^{k+1}(\overline{\mathbb{R}^d},\mathbb{R}^d)$, cf. [DZ11, Theorem 4.6.2]. A set Ω satisfying the condition $\partial\Omega\backslash\partial\overline{\Omega}=\emptyset$ is called *crackfree*.

2.5. First order derivatives

Following the presentation in [DZ11] we develop now the notions of shape semiderivatives and derivatives. Compared to optimization problems posed in a linear vector spaces this is a more delicate issue. One has to work with difference quotients along paths instead of the usual directional derivatives. We consider here an admissible set of domains $\mathcal{O}\subset\mathcal{P}(\mathbb{R}^d)$. Most of the concepts developed in this and the next section can be carried over to the setting $\mathcal{O}\subset\mathcal{P}(\mathcal{D})$ for some $\mathcal{D}\subset\mathbb{R}^d$, cf., e.g., [DZ92].

Definition 2.26. *[DZ11, Definition 9.3.2] Consider a shape functional* $\mathfrak{j}\colon\mathcal{O}\to\mathbb{R}$ *on some admissible set* $\mathcal{O}\subset\mathcal{P}(\mathbb{R}^d)$.

(i) Let $\mathscr{V}$ *be a velocity field satisfying Assumption 2.1 with* $T_{\mathscr{V}}(t,\Omega)\in\mathcal{O}$ *for all* $t\in[0,\vartheta]$. *The shape functional* $\mathfrak{j}$ *has an* Eulerian semiderivative *at* $\Omega\in\mathcal{O}$ *in the direction* $\mathscr{V}$ *if the limit*

$$d\mathfrak{j}(\Omega;\mathscr{V}):=\lim_{t\searrow 0}\frac{1}{t}\left(\mathfrak{j}(T_{\mathscr{V}}(t,\Omega))-\mathfrak{j}(\Omega)\right)$$

exists. If $\mathscr{V}(t,x)=V(x)$ *is an autonomous velocity field we will also write* $d\mathfrak{j}(\Omega;V)$.

(ii) Let $\Theta\subset C^{0,1}(\overline{\mathbb{R}^d},\mathbb{R}^d)$ *be a Banach space and* $\mathcal{O}=\mathcal{O}_\Theta(\Omega)$ *for some* $\Omega\subset\mathbb{R}^d$. *The shape functional* $\mathfrak{j}$ *has a* Hadamard semiderivative *at* Ω *in the direction* $V\in\Theta$ *if, for some* $\vartheta>0$ *and all* $\mathscr{V}\in C^0([0,\vartheta],\Theta)$ *satisfying* $\mathscr{V}(0)=V$ *the limit*

$$d_H\mathfrak{j}(\Omega;V):=\lim_{t\searrow 0}\frac{1}{t}\left(\mathfrak{j}(T_{\mathscr{V}}(t,\Omega))-\mathfrak{j}(\Omega)\right)$$

exists, depends only on V, *and is independent of the choice of* $\mathscr{V}$ *satisfying Assumption 2.1. If the Hadamard semiderivative exists there holds obviously* $d_H\mathfrak{j}(\Omega;V)=d\mathfrak{j}(\Omega;V)$.

(iii) Let $\Theta\subset C^{0,1}(\overline{\mathbb{R}^d},\mathbb{R}^d)$ *be a Banach space and* $\mathcal{O}=\mathcal{O}_\Theta(\Omega)$ *for some* $\Omega\subset\mathbb{R}^d$. *The shape functional* $\mathfrak{j}$ *has a* Hadamard derivative *at* Ω *with respect to* Θ *if it has a Hadamard semiderivative in every directions* $V\in\Theta$, *and if the map*

$$d\mathfrak{j}(\Omega;\cdot)\colon\Theta\to\mathbb{R},\quad V\mapsto d\mathfrak{j}(\Omega;V)$$

is linear and continuous. The Hadamard derivative will be denoted by $\mathfrak{j}'(\Omega)\in\Theta^*$.

Lemma 2.27. *[DZ11, Theorem 9.3.1] Let* Θ *be a Banach subspace of* $C^{0,1}(\overline{\mathbb{R}^d},\mathbb{R}^d)$, $\mathfrak{j}\colon\mathcal{O}\to\mathbb{R}$ *be a shape functional,* $\Omega\subset\mathcal{O}$ *and* $\vartheta>0$. *Suppose that the Eulerian semiderivative* $d\mathfrak{j}(\Omega;\mathscr{V})$ *exists for all* $\mathscr{V}\in C([0,\vartheta],\Theta)$, *and the map*

$$C([0,\vartheta],\Theta)\to\mathbb{R},\quad \mathscr{V}\mapsto d\mathfrak{j}(\Omega;\mathscr{V}),$$

is continuous. Then the Hadamard semiderivative with respect to Θ *exists at* Ω *for all* $\mathscr{V}\in C([0,\vartheta],\Theta)$ *in the direction* $\mathscr{V}(0)$. *Furthermore, it holds* $d\mathfrak{j}(\Omega;\mathscr{V})=d_H\mathfrak{j}(\Omega;\mathscr{V}(0))$.

Due to Theorem 2.24, Hadamard semidifferentiability is enough to obtain continuity with respect to $d_{\mathcal{F}}$:

Corollary 2.28. *Let $k \geq 0$, Θ be equal to $C^{k,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$ or $C^{k+1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$, and let Ω be nonempty, open, and satisfy $\partial\Omega \backslash \partial\overline{\Omega} = \emptyset$. If a shape functional $\mathrm{j}\colon \mathcal{O}_\Theta(\Omega) \to \mathbb{R}$ is Hadamard semidifferentiable at Ω for all $V \in \Theta$, then it is continuous in* Id *with respect to $d_{\mathcal{F}}$.*

Proof. The claim follows directly from the definition of Hadamard semidifferentiability and Theorem 2.24. □

In [DZ11, Theorem 9.3.3] the assertion of Corollary 2.28 is formulated for the Courant metric. Recall from Theorem 2.5 that the map $U \to \mathrm{Id} + U :\ B^\Theta(0,1) \to \mathcal{F}(\Theta)$ is well defined and continuous. Hence we can relate a shape functional $\mathrm{j}\colon \mathcal{O}_\Theta(\Omega_0) \to \mathbb{R}$ locally to a functional defined on the unit ball in the tangent space Θ.

Definition 2.29. *Let $k \geq 0$, Θ be equal to $C^{k+1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$ or $C^{k,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$, and consider a shape functional $\mathrm{j}\colon \mathcal{O}_\Theta(\Omega) \to \mathbb{R}$ for some nonempty set $\Omega \subset \mathbb{R}^d$. We abbreviate*

$$\tau_U := \mathrm{Id} + U \text{ for } U \in \Theta,$$

and define

$$\mathrm{j}_\Omega\colon B^\Theta(0,1) \subset \Theta \to \mathbb{R}, \quad \mathrm{j}_\Omega(U) := \mathrm{j}(\tau_U(\Omega)).$$

If j is continuous in Ω for the metric $d_{\mathcal{F}}$ on $\mathcal{F}(\Theta)$, then j_Ω is continuous in $U = 0$. For the localized functional j_Ω we have the usual notions of Gâteaux and Fréchet derivatives available

Definition 2.30. *[DZ11, Definition 9.3.3] Let the conditions of Definition 2.29 be satisfied.*

(i) The functional j_Ω is said to have a Gâteaux semiderivative *at $U \in B^\Theta(0,1)$ in the direction $V \in \Theta$ if the following limit exists and is finite:*

$$d\mathrm{j}_\Omega(U;V) := \lim_{t \searrow 0} \frac{1}{t} \left(\mathrm{j}_\Omega(U + tV) - \mathrm{j}_\Omega(U)\right).$$

(ii) The functional j_Ω is said to be Gâteaux differentiable *at U if it has a Gâteaux semiderivative in all directions $V \in \Theta$, and the map*

$$d\mathrm{j}_\Omega(U;\cdot)\colon \Theta \to \mathbb{R}, \quad V \mapsto d\mathrm{j}_\Omega(U;V)$$

is linear and continuous. This map will be denoted by $\mathrm{j}'_\Omega(U) \in \Theta^*$.

(iii) If the functional j_Ω is Gâteaux differentiable at U and

$$\lim_{\|V\|_\Theta \to 0} \frac{|\mathrm{j}_\Omega(U+V) - \mathrm{j}(U) - \langle \mathrm{j}'_\Omega(U), V\rangle_{\Theta^*,\Theta}|}{\|V\|_\Theta} = 0,$$

then we speak of Fréchet differentiability *of j_Ω at U.*

As usual, if $\mathfrak{j}_\Omega$ is Gâteaux differentiable and the map $U \mapsto \mathfrak{j}'_\Omega(U)$ is continuous, then $\mathfrak{j}_\Omega$ is Fréchet differentiable. Let us emphasize the following important connection between the differentiability concepts of $\mathfrak{j}$ and $\mathfrak{j}_\Omega$.

Theorem 2.31. *[DZ11, Theorem 9.3.4] Let the conditions of Definition 2.29 be satisfied and consider some $U \in B^\Theta(0,1)$.*

(i) If $\mathfrak{j}$ has a Hadamard semiderivative at $\tau_U(\Omega)$ in the direction $V \circ \tau_U^{-1}$, then $\mathfrak{j}_\Omega$ has a Gâteaux semiderivative at U in the direction V and it holds

$$d\mathfrak{j}_\Omega(U;V) = d_H\mathfrak{j}(\tau_U(\Omega); V \circ \tau_U^{-1}).$$

Conversely, if $\mathfrak{j}_\Omega$ has a Gâteaux semiderivative at U in the direction $V \circ \tau_U$, then $\mathfrak{j}$ has a Hadamard semiderivative at $\tau_U(\Omega)$ in the direction V.

(ii) If either $d\mathfrak{j}_\Omega(U;\cdot)$ or $d\mathfrak{j}(\Omega;\cdot)$ is linear and continuous with respect to all $V \in \Theta$ so is the other, and for all $V \in \Theta$ it holds

$$\langle \mathfrak{j}'_\Omega(U), V\rangle_{\Theta^*,\Theta} = \langle \mathfrak{j}'(\tau_U(\Omega)), V \circ \tau_U^{-1}\rangle_{\Theta^*,\Theta},$$
$$\langle \mathfrak{j}'(\tau_U(\Omega)), V\rangle_{\Theta^*,\Theta} = \langle \mathfrak{j}'_\Omega(U), V \circ \tau_U\rangle_{\Theta^*,\Theta}.$$

Sketch of the proof. Realizing that the respective differential quotients of $d\mathfrak{j}_\Omega(U;V)$ and $d_H\mathfrak{j}(\tau_U(\Omega); V \circ \tau_U^{-1})$ are equal the claim follows. □

Remark 2.32. The element $V \circ \tau_U^{-1} \in \Theta$ corresponds to the *parallel transport* of $V \in \Theta$ from the tangent space at $\Omega \in \mathcal{O}$ to the tangent space at $\tau_U(\Omega) \in \mathcal{O}$. We will discuss this interpretation and its connection to the theory of manifolds in Section 2.7.

It is common to define the *shape derivative* as a *vector distribution* [DZ11, Definition 9.3.4].

Definition 2.33. *Let $\mathfrak{j}\colon \mathcal{O} \subset \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ be a shape functional and $\Omega \in \mathcal{O}$.*

(i) The functional $\mathfrak{j}$ is said to be shape differentiable *at Ω if it is Hadamard differentiable with respect to $\Theta = C_c^\infty(\mathbb{R}^d, \mathbb{R}^d)$.*

(ii) The map $\mathfrak{j}'(\Omega) \in C_c^\infty(\mathbb{R}^d, \mathbb{R}^d)^*$ is called the shape derivative *of $\mathfrak{j}$ at Ω.*

(iii) If there exists some finite $k \geq 0$, such that $\mathfrak{j}'(\Omega)$ is continuous for the $C_c^k(\mathbb{R}^d, \mathbb{R}^d)$-topology, then we say that the shape derivative is of order k.

Remark 2.34. (i) The Hadamard-Zolésio structure theorem (cf. [DZ11, Theorem 9.3.6]) shows that the support of the vector distribution $\mathfrak{j}'(\Omega) \in C_c^\infty(\mathbb{R}^d, \mathbb{R}^d)^*$ is contained in $\partial\Omega$. Furthermore, the distribution is *normal* to $\partial\Omega \cap \mathcal{D}$ in an appropriate sense. In particular, assuming enough smoothness one may obtain a representation of the form

$$\langle \mathfrak{j}'(\Omega), V\rangle_{\Theta^*,\Theta} = \int_{\partial\Omega\cap\mathcal{D}} g\, V^T n \,\mathrm{d}S,$$

where n denotes the unit exterior normal. In Section 2.9.2 we discuss the structure theorem and its implications in more detail, in particular with regard to solvability of the Newton equation.

(ii) As we will see in Section 2.14 a shape functional which depends on a shape dependent solution of a PDE can be conveniently related to $\mathfrak{j}_\Omega$. Hence, one can exploit Theorem 2.31 to obtain the shape derivative. Often the corresponding directional derivative has a natural representation as a volume integral. If the boundary $\partial\Omega$ is smooth enough, this expression can be related via the Gauß divergence theorem to a boundary representation in accordance with the structure predicted by the Hadamard-Zolésio theorem.

(iii) If the boundary $\partial\Omega$ is compact then there exist $k, s \geq 0$ such that $\mathfrak{j}'(\Omega)$ is continuous for the $C_c^k(\mathbb{R}^d, \mathbb{R}^d)$-topology and $\mathfrak{j}'(\Omega) \in H^{-s}(\mathbb{R}^d, \mathbb{R}^d)$, cf. [DZ11, Remark 9.3.1].

(iv) The derivative $\mathfrak{j}'(\Omega) \in C_c^\infty(\mathbb{R}^d, \mathbb{R}^d)^*$ is often called the *shape gradient*. We think that the terms derivative and gradient should be clearly separated. The *derivative* is an element of the dual space of some vector space. In our terminology the *gradient* is the Riesz representative of the derivative with respect to some scalar product. So if $\mathfrak{j}'(\Omega) \in H^*$ for some Hilbert space H the gradient $\nabla\mathfrak{j}(\Omega)$ with respect to the H-scalar product is given by

$$(\nabla\mathfrak{j}(\Omega), V)_H = \langle \mathfrak{j}'(\Omega), V\rangle_{\Theta^*,\Theta}, \quad \forall V \in H.$$

2.6. Second order derivatives

We begin this section by defining the Eulerian semiderivative of the shape derivative. This is the traditional way of introducing second order derivatives of shape functionals. As was shown by Delfour and Zolésio, under certain conditions, the second order Eulerian semiderivative can be decomposed into a canonical symmetric part plus the shape derivative in some specific direction. We call this symmetric part the *shape Hessian*. In Section 2.7 we will show that this term corresponds to the second covariant derivative as it is known in the theory of manifolds. Recall the notion of a flowmap $T_{\mathscr{V}}$ from Definition 2.14.

Definition 2.35. *[DZ11, Definition 9.6.1] Let* $\mathfrak{j}\colon \mathcal{O} \subset \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ *be a shape functional and* $\Omega \in \mathcal{O}$*. Consider velocity fields* $\mathscr{V}$ *and* $\mathscr{W}$ *satisfying Assumption 2.1 for* $\vartheta > 0$*. Suppose that* $d_H\mathfrak{j}(T_{\mathscr{W}}(t,\Omega); \mathscr{V}(t))$ *exists for all* $t \in [0,\vartheta]$*. We say that* $\mathfrak{j}$ *has a* second order Eulerian semiderivative *at* Ω *in the direction* $(\mathscr{V}, \mathscr{W})$ *if the following limit exists*

$$d^2\mathfrak{j}(\Omega; \mathscr{V}; \mathscr{W}) := \lim_{t \searrow 0} \frac{1}{t} \left(d\mathfrak{j}(T_{\mathscr{W}}(t,\Omega); \mathscr{V}(t)) - d\mathfrak{j}(\Omega; \mathscr{V}(0)) \right). \tag{2.15}$$

The definition is compatible with the second order expansion of the function $f(t) := \mathfrak{j}(T_{\mathscr{V}}(t,\Omega))$ at $t = 0$. If $\mathscr{V} = V$ and $\mathscr{W} = W$ are autonomous vector fields we write $d^2\mathfrak{j}(\Omega; V; W)$. Under suitable conditions the second order Eulerian semiderivative depends only on $\mathscr{V}(0), \mathscr{W}(0)$, and $\partial_t\mathscr{V}(0)$. For convenience we introduce the following spaces. For integers $m, k \geq 0$, and some $\vartheta > 0$, we choose either $\Theta = C^{k,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$ or $\Theta = C^{k+1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$ and define the spaces

$$\mathcal{V}^{m,k} := C^m\Big([0,\vartheta], \Theta\Big), \text{ and } \mathcal{V}^k := \Theta. \tag{2.16}$$

Clearly for any $m, k \geq 0$, every vector field $\mathscr{V} \in \mathcal{V}^{m,k}$ satisfies Assumption 2.1 for the choice $\mathcal{D} = \mathbb{R}^d$. The definition of $\mathcal{V}^k$ may seem superfluous, but allows us to conveniently express a higher smoothness assumption. Note that in [DZ11, Section 9.3] the notations $\mathcal{V}^{m,k}, \mathcal{V}^k$ are used in a slightly different context, since they allow for tangent spaces Θ which are not Banach spaces.

In analogy to Definition 2.26 we introduce the notion of twice Hadamard differentiability.

Definition 2.36. *Let $k \geq 0$, Θ be equal to $C^{k,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$ or $C^{k+1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$, and consider a shape functional $\mathfrak{j}\colon \mathcal{O}_\Theta(\Omega_0) \to \mathbb{R}$ for some nonempty set $\Omega_0 \subset \mathbb{R}^d$. Suppose furthermore that the functional $\mathfrak{j}$ is Hadamard differentiable with respect to Θ.*

(i) We say that the functional $\mathfrak{j}$ has a second order Hadamard semiderivative *at $\Omega \in \mathcal{O}_\Theta(\Omega_0)$ in the direction $(V, W) \in \Theta \times \Theta$, if there exists a $\vartheta > 0$ such that for any $\mathscr{V}, \mathscr{W} \in \mathcal{V}^{1,k}$ satisfying $\mathscr{V}(0) = V$, $\mathscr{W}(0) = W$, the second order Eulerian semiderivative $d^2\mathfrak{j}(\Omega; \mathscr{V}; \mathscr{W})$ exists and satisfies*

$$d^2\mathfrak{j}(\Omega; \mathscr{V}; \mathscr{W}) - d\mathfrak{j}(\Omega; \partial_t \mathscr{V}(0)) = d^2\mathfrak{j}(\Omega; V; W),$$

where we used again the notation $\partial_t \mathscr{V}(t)(x) = \partial_t \mathscr{V}(t, x)$. The second order Hadamard semiderivative is denoted by $d^2\mathfrak{j}(\Omega; V; W)$.

(ii) We say that the functional $\mathfrak{j}$ is twice Hadamard differentiable *at $\Omega \in \mathcal{O}_\Theta(\Omega_0)$ with respect to Θ, if it has a second order Hadamard semiderivative $d^2\mathfrak{j}(\Omega; V; W)$ for all $V, W \in \Theta$, and if the map*

$$d^2\mathfrak{j}(\Omega; \cdot; \cdot)\colon \Theta \times \Theta \to \mathbb{R}$$

is bilinear and continuous.

As for first order derivatives we have a simple continuity condition available which implies twice Hadamard differentiability.

Theorem 2.37. *[DZ11, Theorem 9.6.2] Let $k \geq 0$, Θ be equal to $C^{k,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$ or $C^{k+1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$, and consider a shape functional $\mathfrak{j}\colon \mathcal{O}_\Theta(\Omega_0) \to \mathbb{R}$ for some nonempty set $\Omega_0 \subset \mathbb{R}^d$. Suppose that for $\Omega \in \mathcal{O}_\Theta(\Omega_0)$ and $\vartheta > 0$*

(i) the Eulerian semiderivative $d^2\mathfrak{j}(\Omega; \mathscr{V}; \mathscr{W})$ exists for all $\mathscr{V} \in \mathcal{V}^{1,k}$ and $\mathscr{W} \in \mathcal{V}^{0,k}$,

(ii) $\forall \mathscr{W} \in \mathcal{V}^{0,k}$, $\forall t \in [0, \vartheta]$, the functional $\mathfrak{j}$ is Hadamard differentiable with respect to Θ at $T_{\mathscr{W}}(t, \Omega) \in \mathcal{O}_\Theta(\Omega_0)$,

(iii) $\forall V \in \mathcal{V}^k$, the map $\mathcal{V}^{0,k} \to \mathbb{R}$, $\mathscr{W} \mapsto d^2\mathfrak{j}(\Omega; V; \mathscr{W})$ is continuous.

Then we can decompose the second order Eulerian semiderivative for all $\mathscr{V} \in \mathcal{V}^{1,k}$ and $\mathscr{W} \in \mathcal{V}^{0,k}$ into

$$d^2\mathfrak{j}(\Omega; \mathscr{V}; \mathscr{W}) = d^2\mathfrak{j}(\Omega; \mathscr{V}(0); \mathscr{W}(0)) + d\mathfrak{j}(\Omega; \partial_t \mathscr{V}(0)).$$

In particular, $\mathfrak{j}$ has a second order Hadamard semiderivative $d^2\mathfrak{j}(\Omega; V; W)$ for all $V, W \in \Theta$.

Sketch of the proof. The differential quotient (2.15) can be split into

$$\frac{1}{t}\Big(d\mathrm{j}(T_{\mathscr{W}}(t,\Omega);\mathscr{V}(0)) - d\mathrm{j}(\Omega;\mathscr{V}(0))\Big) + \frac{1}{t}\Big(d\mathrm{j}(T_{\mathscr{W}}(t,\Omega);\mathscr{V}(t)) - d\mathrm{j}(T_{\mathscr{W}}(t,\Omega);\mathscr{V}(0))\Big).$$

Combining conditions (i) and (iii) with a suitable auxiliary sequence of velocities $(\mathscr{W}_n)$ satisfying $d^2\mathrm{j}(\Omega;\mathscr{V}(0);\mathscr{W}_n) = d^2\mathrm{j}(\Omega;\mathscr{V}(0);\mathscr{W})$ one shows that the first term converges to $d^2\mathrm{j}(\Omega;\mathscr{V}(0);\mathscr{W}) = d^2\mathrm{j}(\Omega;\mathscr{V}(0);\mathscr{W}(0))$. For the second term one considers the vector field $\tilde{\mathscr{V}}(t) = \frac{1}{t}(\mathscr{V}(t) - \mathscr{V}(0))$. Exploiting the linearity of the Hadamard derivative one may then show that the second term converges to $d\mathrm{j}(\Omega;\partial_t\mathscr{V}(0))$. The last assertion follows directly from Definition 2.36. □

Remark 2.38. The term $d^2\mathrm{j}(\Omega;\mathscr{V}(0);\mathscr{W}(0))$ is in general *not* symmetric. The following result tells us more about its structure.

Theorem 2.39. *[DZ11, Theorem 9.6.5] Let $k \geq 0$, Θ be equal to $C^{k,1}(\overline{\mathbb{R}^d},\mathbb{R}^d)$ or $C^{k+1}(\overline{\mathbb{R}^d},\mathbb{R}^d)$, and consider a shape functional $\mathrm{j}\colon \mathcal{O}_\Theta(\Omega_0) \to \mathbb{R}$ for some nonempty set $\Omega_0 \subset \mathbb{R}^d$. Furthermore let $\Omega \in \mathcal{O}_\Theta(\Omega_0)$, $U \in B^\Theta(0,1)$, and recall $\mathrm{j}_\Omega\colon B^\Theta(0,1) \to \mathbb{R}$ from Definition 2.29.*

(i) Given $V, W \in \Theta$, assume that there exists a $\vartheta > 0$ such that

$$\forall t \in [0,\vartheta] \text{ the derivative } d\mathrm{j}_\Omega(U+tW;V) \text{ exists.}$$

Then the second Gâteaux semiderivative at $U \in B^\Theta(0,1)$

$$d^2\mathrm{j}_\Omega(U;V;W) := \lim_{t\searrow 0}\frac{1}{t}\Big(d\mathrm{j}_\Omega(U+tW;V) - d\mathrm{j}_\Omega(U;V)\Big)$$

exists if and only if $d^2\mathrm{j}(\tau_U(\Omega);\mathscr{V}_V;\mathscr{W}_W)$ exists for the velocity fields

$$\mathscr{V}_V(t) := V\circ(\mathrm{Id}+U+tW)^{-1}, \text{ and } \mathscr{W}_W(t) := W\circ(\mathrm{Id}+U+tW)^{-1}.$$

In this case it holds $d^2\mathrm{j}_\Omega(U;V;W) = d^2\mathrm{j}(\tau_U(\Omega);\mathscr{V}_V;\mathscr{W}_W)$.

(ii) If $U \in \mathcal{V}^{k+1}$ and j is twice Hadamard differentiable at $\tau_U(\Omega)$ with respect to Θ, then the second derivatives of j and j_Ω are related by

$$\begin{aligned} d^2\mathrm{j}_\Omega(U;V;W) &= d^2\mathrm{j}(\tau_U(\Omega);V\circ\tau_U^{-1};W\circ\tau_U^{-1}) - d\mathrm{j}(\tau_U(\Omega);D(V\circ\tau_U^{-1})(W\circ\tau_U^{-1})),\\ d^2\mathrm{j}(\tau_U(\Omega);V;W) &= d^2\mathrm{j}_\Omega(U;V\circ\tau_U;W\circ\tau_U) + d\mathrm{j}(\tau_U(\Omega);DVW), \end{aligned}$$

for all $V \in \mathcal{V}^{k+1}$ and $W \in \Theta$. In particular, this implies that j_Ω has a second order Gâteaux derivative at U with respect to $\mathcal{V}^{k+1}$.

(iii) If $\mathscr{V} \in \mathcal{V}^{1,k+1}$, $\mathscr{W} \in \mathcal{V}^{0,k}$, and j has a second order Hadamard semiderivative in the direction $(\mathscr{V}(0),\mathscr{W}(0))$, then

$$d^2\mathrm{j}(\Omega;\mathscr{V};\mathscr{W}) = d^2\mathrm{j}_\Omega(0;\mathscr{V}(0);\mathscr{W}(0)) + d\mathrm{j}(\Omega;D\mathscr{V}(0)\mathscr{W}(0) + \partial_t\mathscr{V}(0)).$$

Sketch of the proof. (i) Due to Theorem 2.31 it holds $d\mathfrak{j}_\Omega(U+tW;V) = d\mathfrak{j}(\tau_{U+tW}(\Omega); V\circ\tau_{U+tW}^{-1})$. It can be easily verified, that τ_{U+tW} corresponds to the transformation given by the flowmap $T_{\mathscr{W}_W}(t)$ where $\mathscr{W}_W$ is defined above. Furthermore $\mathscr{V}_V(t) = V \circ \tau_{U+tW}^{-1}$. Hence it holds $d\mathfrak{j}_\Omega(U + tW; V) = d\mathfrak{j}(T_{\mathscr{W}_W}(t, \tau_U(\Omega)); \mathscr{V}_V(t))$ and $d\mathfrak{j}_\Omega(U; V) = d\mathfrak{j}(\tau_U(\Omega); \mathscr{V}_V(0))$. Thus the differential quotients of $d^2\mathfrak{j}_\Omega(U; V; W)$ and $d^2\mathfrak{j}(\tau_U(\Omega); \mathscr{V}_V; \mathscr{W}_W)$ coincide

$$\frac{1}{t}\Big(d\mathfrak{j}_\Omega(U + tW; V) - d\mathfrak{j}_\Omega(U; V)\Big) = \frac{1}{t}\left(d\mathfrak{j}(T_{\mathscr{W}_W}(t, \tau_U(\Omega)); \mathscr{V}_V(t)) - d\mathfrak{j}(\tau_U(\Omega); \mathscr{V}_V(0))\right).$$

In particular, if either $d^2\mathfrak{j}_\Omega(U; V; W)$ or $d^2\mathfrak{j}(\tau_U(\Omega); \mathscr{V}_V; \mathscr{W}_W)$ exists, so does the other and the terms are equal.
(ii) Noting that $\partial_t \mathscr{V}_V(0) = -D(V \circ \tau_U^{-1})(W \circ \tau_U^{-1})$ the assertion follows from (i) and the definition of twice Hadamard differentiability since

$$\begin{aligned} d^2\mathfrak{j}(\tau_U(\Omega); V \circ \tau_U^{-1}; W \circ \tau_U^{-1}) &= d^2\mathfrak{j}(\tau_U(\Omega); \mathscr{V}_V(0); \mathscr{W}_W(0)) \\ &= d^2\mathfrak{j}(\tau_U(\Omega); \mathscr{V}_V; \mathscr{W}_W) - d\mathfrak{j}(\tau_U(\Omega); \partial_t \mathscr{V}_V(0)) \\ &= d^2\mathfrak{j}_\Omega(U; V; W) + d\mathfrak{j}(\tau_U(\Omega); D(V \circ \tau_U^{-1})(W \circ \tau_U^{-1})). \end{aligned}$$

(iii) This is a direct consequence of (ii) and Definition 2.36. □

Analogously to Definition 2.33 a shape functional is called twice shape differentiable if the second order Eulerian semiderivative is a vector distribution. We give the definition from [DZ11] for completeness, but will usually work with Hadamard differentiability with respect to some suitable Banach space Θ.

Definition 2.40. *[DZ11, Definition 9.6.2 (i)] Let* $\mathfrak{j}\colon \mathcal{O} \subset \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ *be a shape functional. The functional* $\mathfrak{j}$ *is called* twice shape differentiable *at* $\Omega \in \mathcal{O}$*, if for all* $V, W \in C_c^\infty(\mathbb{R}^d, \mathbb{R}^d)$ *the second order Eulerian semiderivative* $d^2\mathfrak{j}(\Omega; V; W)$ *exists, and if the mapping*

$$C_c^\infty(\mathbb{R}^d, \mathbb{R}^d) \times C_c^\infty(\mathbb{R}^d, \mathbb{R}^d) \to \mathbb{R} : \ (V, W) \mapsto d^2\mathfrak{j}(\Omega; V; W)$$

is bilinear and continuous. If it is continuous for all $V, W \in C_c^k(\mathbb{R}^d, \mathbb{R}^d)$ *for some finite* $k \geq 0$*, then* $d^2\mathfrak{j}(\Omega; \cdot; \cdot)$ *is of order* k*.*

Remark 2.41. (i) The associated distribution in $(C_c^\infty(\mathbb{R}^d, \mathbb{R}^d) \times C_c^\infty(\mathbb{R}^d, \mathbb{R}^d))^*$ is usually termed the *shape Hessian*, cf. [DZ11, Definition 9.6.2 (ii)]. We will use this term in a slightly different context, since we know from Theorem 2.39, that $d^2\mathfrak{j}(\Omega; V; W)$ is in general *not* symmetrical. We are inspired by the theory of second derivatives on manifolds, cf. Section 2.7.

(ii) There is an analogue of the structure theorem for the second shape derivative. It states that the support of the associated distribution is a subset of $\partial\Omega \times \partial\Omega$ and the distribution is normal to the boundary in an appropriate sense. Compare Theorem 2.79 in Section 2.9.2.

(iii) If the boundary of Ω is smooth enough, one can again derive a boundary representation of the second shape derivative, cf. [DZ11, Theorem 9.6.4]. We will not pursue this further.

Definition 2.42. *Let the conditions of Definition 2.40 be satisfied. We define the* shape Hessian *of* $\mathfrak{j}$ *at* $\Omega \in \mathcal{O}$ *as*

$$\nabla^2\mathfrak{j}(\Omega)\colon C_c^\infty(\mathbb{R}^d,\mathbb{R}^d)\times C_c^\infty(\mathbb{R}^d,\mathbb{R}^d)\to\mathbb{R},\quad \nabla^2\mathfrak{j}(\Omega)[V,W] := d^2\mathfrak{j}(\Omega;V;W) - d\mathfrak{j}(\Omega;DVW).$$

Obviously this definition is motivated by Theorem 2.39. The next result shows that we can expect symmetry of the shape Hessian.

Corollary 2.43. *Let* $k \geq 0$, Θ *be equal to* $C^{k,1}(\overline{\mathbb{R}^d},\mathbb{R}^d)$ *or* $C^{k+1}(\overline{\mathbb{R}^d},\mathbb{R}^d)$*, and consider a shape functional* $\mathfrak{j}\colon \mathcal{O}_\Theta(\Omega_0)\to\mathbb{R}$ *for some nonempty set* $\Omega_0\subset\mathbb{R}^d$*. Suppose that* $\mathfrak{j}$ *is twice Hadamard differentiable with respect to* Θ *at* $\Omega\in\mathcal{O}_\Theta(\Omega_0)$*. Then*

$$\nabla^2\mathfrak{j}(\Omega)[V,W] = d^2\mathfrak{j}_\Omega(0;V;W)$$

for all $V\in\mathcal{V}^{k+1}$ *and* $W\in\Theta$*. In particular, the shape Hessian is* symmetric *if* $\mathfrak{j}_\Omega$ *is twice Fréchet differentiable at* 0 *with respect to* $\mathcal{V}^{k+1}$*.*

Proof. The identity follows directly from Theorem 2.39 and the definition of $\nabla^2\mathfrak{j}$. Symmetry of the second Fréchet derivative for real-valued functionals is well known. □

It is common in the theory of Riemannian manifolds to define the *Riemannian Hessian* as the linear mapping from the tangent space into itself which is obtained if the *Riemannian connection* is applied to the *Riemannian gradient*. Here the Riemannian gradient is the Riesz representative of the derivative with respect to the *Riemannian metric*. We explain these notions in Section 2.7. In [Sch14] such a construction was used in the context of shape optimization and called the *Riemannian shape Hessian*. He considered the manifold of all equivalence classes of C^∞-embeddings of the unit circle into the plane $\mathbb{R}^2$, where the equivalence relation is defined by the set of all C^∞-reparametrizations of the unit circle. However, it needs to be emphasized that this construction works only if a Riemannian metric and the associated Riemannian connection are available. On infinite dimensional manifolds the choice of a suitable Riemannian metric is a delicate issue, cf. [BBM14, Mic15]. A more general construction is the *second covariant derivative*. We will see in the next section that it corresponds to our choice of the shape Hessian. Although we use the theory of Riemannian manifolds to motivate this choice, it is not necessary for its definition, and this term has already been used for practical computations by many authors, see for example [NR95].

In practice, at least in the context of PDE-constrained optimization, it is usually prohibitive to compute the whole Hessian $\nabla^2\mathfrak{j}$. Rather one is interested in evaluating $V\mapsto\nabla^2\mathfrak{j}[V,\cdot]$, e.g., during an iterative solution strategy for Newton's equation. Thus, we also introduce the map

$$\mathfrak{j}''_\Omega(U)\in\mathcal{L}(\Theta,\Theta^*)\colon\quad \langle\mathfrak{j}''_\Omega(U)V,W\rangle_{\Theta^*,\Theta} := d^2\mathfrak{j}_\Omega(U;W;V)\ \text{ for all } V,W\in\Theta, \tag{2.17}$$

which is self-adjoint if $\mathfrak{j}_\Omega$ is twice Fréchet differentiable.

2.7. Relation to Riemannian manifolds

We will discuss now the connection between shape optimization on $\mathcal{F}(\Theta)$ and the theory of optimization on Riemannian manifolds. We refer to the monograph [AMS08] for a nice introduction to optimization on finite dimensional manifolds, and to [RW12, Sch14] for optimization on infinite dimensional manifolds and applications to shape spaces. The theory of infinite dimensional shape spaces as Riemannian manifolds is intricate, we refer to the surveys [BBM14, Mic15] and the references therein.

Despite the fact that $\mathcal{F}(\Theta)$ is *not* a Riemannian manifold in the traditional sense (in particular it is not C^∞-smooth), we will introduce now some notions from the theory of optimization on manifolds, and discuss their connection to the usual terminology of shape optimization. Of course, we are not the first to note the correlation between shape optimization and Riemannian manifolds, for example there are several such remarks in [DZ11], and the paper [Sch14] is based on this insight. However, to the best of our knowledge, the thesis of Frey [Fre12] is so far the only work which translates the setting of Riemannian manifolds directly into the theory of shape optimization. Unfortunately, many considerations are only carried out on a formal level. Furthermore, we believe that some notions should be interpreted from a slightly different perspective. Hence, while the general ideas presented in [Fre12] are very similar to our presentation here, we emphasize that several details differ. In particular, his notion of a retraction in shape optimization deviates from our interpretation.

For convenience we will first recall some traditional concepts of optimization on manifolds before drawing the connection to shape optimization. We will mainly follow [AMS08, RW12] and summarize only some important ideas. A more comprehensive introduction to infinite dimensional Riemannian geometry is given, for instance, in [Kli11].

Brief introduction to Riemannian manifolds

We denote with $\mathcal{M}$ a geodesically complete Riemannian manifold which is locally homeomorphic to some separable Hilbert space, cf. [Kli11, Section 1.1]. The vector space of all smooth real-valued functions on $\mathcal{M}$ is denoted by $C^\infty(\mathcal{M})$. A *curve* in $\mathcal{M}$ is a smooth mapping $\gamma\colon \mathbb{R} \ni t \mapsto \gamma(t) \in \mathcal{M}$.

Definition 2.44. *(i) A* tangent vector ξ_x *to the manifold* $\mathcal{M}$ *at a point* $x \in \mathcal{M}$ *is a linear operator* $\xi_x\colon C^\infty(\mathcal{M}) \to \mathbb{R}$ *such that there exists a curve* $\gamma\colon [0,1] \to \mathcal{M}$ *with* $\gamma(t_0) = x$, $t_0 \in (0,1)$, *satisfying*

$$\xi_x f = \dot\gamma(t_0) f := \partial_t \left(f \circ \gamma\right)(t_0) \quad \forall f \in C^\infty(\mathcal{M}).$$

For a tangent vector ξ_x *there exist infinitely many curves satisfying* $\xi_x = \dot\gamma(t_0)$.

(ii) The set of all tangent vectors ξ_x *is called the* tangent space $T_x\mathcal{M}$ *to* $\mathcal{M}$ *at* $x \in \mathcal{M}$.

(iii) The set of all tangent vectors to $\mathcal{M}$ *is called the* tangent bundle

$$T\mathcal{M} := \cup_{x\in\mathcal{M}} T_x\mathcal{M}$$

to $\mathcal{M}$. *It can be shown that* $T\mathcal{M}$ *is again a manifold.*

(iv) A vector field ξ *is a smooth mapping from* $\mathcal{M}$ *to the tangent bundle* $T\mathcal{M}$ *that assigns to each point* $x \in \mathcal{M}$ *a tangent vector* $\xi_x \in T_x\mathcal{M}$*. We introduce the vector field* $f\xi + g\zeta$ *as*

$$(f\xi + g\zeta)_x := f(x)\xi_x + g(x)\zeta_x \in T_x\mathcal{M}$$

for all $x \in \mathcal{M}$, $f, g \in C^\infty(\mathcal{M})$*, and vector fields* ξ, ζ*. The set of smooth vector fields will be denoted by* $\mathfrak{X}(\mathcal{M})$*.*

For every $x \in \mathcal{M}$ the tangent space $T_x\mathcal{M}$ is a vector space. Since we are dealing with a *Riemannian manifold*, it can be equipped with an inner product

$$g_x(\cdot,\cdot) : \ T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}.$$

This inner product is called the *Riemannian metric*, and varies smoothly with x. It induces a norm on $T_x\mathcal{M}$, which is denoted by $\|\cdot\|_x$. Given two vector fields $\xi, \zeta \in \mathfrak{X}(\mathcal{M})$ the map $g(\xi,\zeta) : x \mapsto g_x(\xi_x, \zeta_x)$ is an element of $C^\infty(\mathcal{M})$. One can now define the length of a curve and *geodesics*, i.e., shortest curves connecting two given points on $\mathcal{M}$, but since we will not use these concepts we omit them here for brevity. Let us only point out that in infinite dimensions the choice of a suitable Riemannian metric is delicate. There are examples where the geodesic distance vanishes for distinct points on $\mathcal{M}$, i.e., the geodesic distance is not point separating, cf. [BBM14, Mic15] and the references cited therein.

The dual space $T_x^*\mathcal{M}$ of $T_x\mathcal{M}$ is called the *cotangent space* of $\mathcal{M}$ at x, its elements are the *covectors*. A smooth map $x \mapsto \mu_x \in T_x^*\mathcal{M}$ is called a *covector field*, and for a vector field $\xi \in \mathfrak{X}(\mathcal{M})$ the function $x \mapsto (\langle \mu, \xi\rangle_{\mathfrak{X}(\mathcal{M})^*,\mathfrak{X}(\mathcal{M})})_x := \langle \mu_x, \xi_x\rangle_{T_x\mathcal{M}^*,T_x\mathcal{M}}$ is an element of $C^\infty(\mathcal{M})$. The *differential* of a function $f \in C^\infty(\mathcal{M})$ defined by

$$f'(x) \in \mathcal{L}(T_x\mathcal{M}, \mathbb{R}), \quad \langle f'(x), \xi_x\rangle_{T_x\mathcal{M}^*,T_x\mathcal{M}} := \xi_x f(x) \quad \forall \xi \in \mathfrak{X}(\mathcal{M})$$

is a covector. More generally, the differential of a smooth map $F\colon \mathcal{M} \to \mathcal{N}$ between manifolds at $x \in \mathcal{M}$ is a linear map, cf. [AMS08, Section 3.5.6]

$$F'(x) \in \mathcal{L}(T_x\mathcal{M}, T_{F(x)}\mathcal{N}), \quad (F'(x)[\xi_x])f(F(x)) := \xi_x(f \circ F)(x) \quad \forall f \in C^\infty(\mathcal{N}).$$

Let us now generalize the notion of the directional derivative of a vector field.

Definition 2.45. *(i) An* affine connection $\blacktriangledown$ *on a manifold* $\mathcal{M}$ *is an operator*

$$\blacktriangledown\colon \mathfrak{X}(\mathcal{M}) \times \mathfrak{X}(\mathcal{M}) \to \mathfrak{X}(\mathcal{M}), \ (\eta, \xi) \mapsto \blacktriangledown_\eta \xi,$$

satisfying

$$\begin{aligned}
\blacktriangledown_{f\eta+g\zeta}\,\xi &= f\,\blacktriangledown_\eta\,\xi + g\,\blacktriangledown_\zeta\,\xi && \forall \eta,\zeta,\xi \in \mathfrak{X}(\mathcal{M}) \text{ and } f,g \in C^\infty(\mathcal{M}),\\
\blacktriangledown_\eta(a\xi + b\zeta) &= a\,\blacktriangledown_\eta\,\xi + b\,\blacktriangledown_\xi\,\zeta && \forall \eta,\zeta,\xi \in \mathfrak{X}(\mathcal{M}) \text{ and } a,b \in \mathbb{R},\\
\blacktriangledown_\eta(f\xi) &= (\eta f)\xi + f\,\blacktriangledown_\eta\,\xi && \forall \eta,\xi \in \mathfrak{X}(\mathcal{M}) \text{ and } f \in C^\infty(\mathcal{M}).
\end{aligned}$$

(ii) For two vector fields $\xi, \eta \in \mathfrak{X}(\mathcal{M})$ *the* Lie bracket $[\xi,\eta] \in \mathfrak{X}(\mathcal{M})$ *is a vector field characterized by*

$$[\xi,\eta]f := \xi(\eta f) - \eta(\xi f) \quad \forall f \in C^\infty(\mathcal{M}).$$

(iii) The affine connection which satisfies

$$[\xi, \eta] = \blacktriangledown_\xi \eta - \blacktriangledown_\eta \xi \qquad \forall \xi, \eta \in \mathfrak{X}(\mathcal{M}), \tag{2.18}$$

$$\eta g(\xi, \zeta) = g(\blacktriangledown_\eta \xi, \zeta) + g(\xi, \blacktriangledown_\eta \zeta) \qquad \forall \xi, \eta, \zeta \in \mathfrak{X}(\mathcal{M}), \tag{2.19}$$

is called the Levi-Civita connection *or* Riemannian connection *and will be denoted by* $\tilde{\blacktriangledown}$.

Remark 2.46. We summarize: an affine connection is $C^\infty(\mathcal{M})$-linear in η, $\mathbb{R}$-linear in ξ, and satisfies Leibniz' law, i.e., the product rule. The Riemannian connection has the additional properties of symmetry (2.18), also called absence of torsion, and preservation of the Riemannian metric (2.19).

To introduce second order derivatives on manifolds one needs to be able to transport a tangent vector ξ_x from the tangent space $T_x\mathcal{M}$ into some other tangent space $T_y\mathcal{M}$. In particular one is interested in *parallel transport*. The specific parallel transport along a curve $\gamma\colon [0,1] \to \mathcal{M}$ depends on the chosen affine connection, and is the solution $\xi(t) \in T_{\gamma(t)}\mathcal{M}$ of the initial value problem

$$\blacktriangledown_{\dot\gamma(t)} \xi(t) = 0, \quad \xi(0) = \xi_0 \in T_{\gamma(0)}\mathcal{M}. \tag{2.20}$$

If two vectors are transported parallel with respect to the Riemannian connection along a curve γ the inner product $g_{\gamma(t)}(\xi(t), \zeta(t))$ is *constant*.

Finally, we want to be able to move on $\mathcal{M}$ from a point $x \in \mathcal{M}$ *in a direction* $\xi_x \in T_x\mathcal{M}$. This is achieved via retractions.

Definition 2.47. *A* retraction *is a smooth mapping* $\mathcal{R}_x\colon T_x\mathcal{M} \to \mathcal{M}$*, satisfying* $\mathcal{R}_x(0) = x$ *and* $\mathcal{R}'_x(0) = \mathrm{id}_x$*. Here* id_x *denotes the identity mapping on* $T_x\mathcal{M}$*, and* $\mathcal{R}'_x(\xi_x)\colon T_x\mathcal{M} \to T_{\mathcal{R}_x(\xi_x)}\mathcal{M}$ *is the differential of* $\mathcal{R}_x$*.*

For any $x \in \mathcal{M}$ and $\xi_x \in T_x\mathcal{M}$ the map $\mathbb{R} \ni t \mapsto \mathcal{R}_x(t\xi_x) \in \mathcal{M}$ defines a smooth curve. We can use a retraction to obtain the *pullback* $\hat f_x := f \circ \mathcal{R}_x$ of a function $f \in C^\infty(\mathcal{M})$ onto the tangent space $T_x\mathcal{M}$. Note that $\hat f_x$ is defined on a standard vector space, and the usual concepts of derivatives, etc. can be used. In particular, by the chain rule it holds that

$$\hat f'_x(0) = f'(x).$$

We speak of a *second order retraction* if the zero initial acceleration condition

$$\blacktriangledown_{\dot{\mathcal{R}}_x(t\xi_x)} \dot{\mathcal{R}}_x(t\xi_x)\Big|_{t=0} = 0, \quad \forall \xi_x \in T_x\mathcal{M} \tag{2.21}$$

is satisfied. It will become clear in a moment why second order retractions are of interest.

On Riemannian manifolds it is common to define the *Riemannian gradient* of a function $f \in C^\infty(\mathcal{M})$ as the Riesz representative of the differential f' with respect to the Riemannian metric, i.e.,

$$g_x(\operatorname{grad} f(x), \xi_x) = \langle f'(x), \xi_x \rangle_{T_x^*\mathcal{M}, T_x\mathcal{M}}, \quad \forall \xi_x \in T_x\mathcal{M} \text{ and } x \in \mathcal{M}.$$

The *Riemannian Hessian* of a function $f \in C^\infty(\mathcal{M})$ is then defined by

$$\operatorname{Hess} f(x) \in \mathcal{L}(T_x\mathcal{M}, T_x\mathcal{M}), \quad \operatorname{Hess} f(x)[\xi_x] := \tilde{\nabla}_{\xi_x} \operatorname{grad} f(x).$$

The Riemannian Hessian is symmetric with respect to $g(\cdot,\cdot)$, and it holds

$$g(\operatorname{Hess} f[\xi], \eta) = \xi(\eta f) - (\tilde{\nabla}_\xi \eta) f.$$

It needs to be emphasized, that this construction works only, if a Riemannian metric, and the associated Riemannian connection are available. A more general construction is the *second covariant derivative*, cf. [AMS08, Section 5.6], which only requires an affine connection. It is a generalization of the second Fréchet derivative. Consider a functional $f \in C^\infty(\mathcal{M})$. Recall that the map $x \mapsto \langle f'(x), \xi_x\rangle_{T_x\mathcal{M}^*, T_x\mathcal{M}} = \xi_x f(x)$ is an element of $C^\infty(\mathcal{M})$. We introduce for its differential the short notation $x \mapsto (\xi f)'_x \in T_x^*\mathcal{M}$. The *covariant derivative* of the covector $f'(x) \in T_x^*\mathcal{M}$ in the direction $\xi_x \in T_x\mathcal{M}$ is again a covector denoted by $\nabla_{\xi_x} f' \in T_x^*\mathcal{M}$, and defined for all $\eta \in \mathfrak{X}(\mathcal{M})$ via

$$\langle \nabla_{\xi_x} f'(x), \eta_x\rangle_{T_x^*\mathcal{M}, T_x\mathcal{M}} := \langle (\xi f)'_x, \xi_x\rangle_{T_x^*\mathcal{M}, T_x\mathcal{M}} - \langle f'(x), \nabla_{\xi_x} \eta_x\rangle_{T_x^*\mathcal{M}, T_x\mathcal{M}}.$$

The *second covariant derivative* of f at $x \in \mathcal{M}$ is then defined by

$$\nabla^2 f(x) \colon T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}, \quad \nabla^2 f(x)[\xi_x, \eta_x] := \langle \nabla_{\xi_x} f'(x), \eta_x\rangle_{T_x^*\mathcal{M}, T_x\mathcal{M}}. \tag{2.22}$$

For any *second order retraction* $\mathcal{R}$ it holds for all $x \in \mathcal{M}$

$$\nabla^2 f(x) = (f \circ \mathcal{R}_x)''(0), \tag{2.23}$$

where $(f \circ \mathcal{R}_x)''(0)$ is the second Fréchet derivative of $f \circ \mathcal{R}_x \colon T_x\mathcal{M} \to \mathbb{R}$ at $0 \in T_x\mathcal{M}$. If x is a critical point of f, i.e., $f'(x) = 0$ then the above identity holds for *any* retraction $\mathcal{R}$. Finally, if we have a Riemannian metric available, and $\tilde{\nabla}$ is the Riemann connection, then there holds

$$\tilde{\nabla}^2 f(x)[\xi_x, \eta_x] = g_x(\operatorname{Hess} f(x)[\xi_x], \eta_x).$$

Translation to shape optimization

We will now introduce some concepts in the framework of shape optimization which are inspired by their counterparts from Riemannian manifolds. However, the situation is a little bit more complicated, since we have on the one hand the group $\mathcal{F}(\Theta)$ where we know the tangent space, and on the other hand the group $\mathcal{O}_\Theta(\Omega_0)$ on which shape functionals are defined. We work in the following always with the tangent space Θ of mappings from $\mathbb{R}^d$ to $\mathbb{R}^d$. Since the images of a domain Ω are uniquely determined by the restrictions $(F - \mathrm{Id})|_\Omega$ for some $F \in \mathcal{F}(\Theta)$ it would be interesting to see whether one could also work with 'local' tangent spaces Θ_Ω. We leave this question for future research.

We consider the following setting.

Assumption 2.3. *It holds $k \geq 0$, Θ is equal to $C^{k+1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$ or $C^{k,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$, and $\Omega_0 \subset \mathbb{R}^d$ is a nonempty, bounded set that is either closed or satisfies $\Omega_0 = \operatorname{int}\overline{\Omega_0}$. It holds $\mathcal{O} = \mathcal{O}_\Theta(\Omega_0)$.*

Due to Theorem 2.5 we know that we can identify the tangent space Θ_F to $\mathcal{F}(\Theta)$ at some $F \in \mathcal{F}(\Theta)$ with Θ. Still we often prefer to make the dependence on the point F explicit. Recall the definition of $L^1([0,1],\Theta)$ in Section 2.3.1. Given some path $[0,1] \ni t \mapsto F(t) =: F_t \in \mathcal{F}(\Theta)$, a velocity field $\mathscr{V} \in L^1([0,1],\Theta)$ with $\mathscr{V}(t) \in \Theta_{F_t}$ will be our notion of a *tangent vector field* along F_t. More generally, we want to characterize the smoothness of a tangent vector field defined on the tangent bundle of $\mathcal{F}(\Theta)$. Recall the definition of the spaces $\mathcal{V}^{m,k}$ from (2.16), and that $\Theta = \mathcal{V}^k$.

Definition 2.48. *Let Assumption 2.3 be satisfied. We call a map*

$$\mathscr{V} \colon \mathcal{F}(\Theta) \ni F \mapsto \mathscr{V}_F \in \Theta_F$$

*a tangent vector field. A tangent vector field $\mathscr{V}$ is said to be (m,k)-*smooth *if, for any path $[0,\vartheta] \ni t \mapsto F_t \in \mathcal{F}(\Theta)$ such that the map $t \mapsto \partial_t F_t$ is in $\mathcal{V}^{m,k}$, the mapping $t \mapsto \mathscr{V}_{F_t}$ is also in $\mathcal{V}^{m,k}$. Furthermore, given shape functionals $f, g \colon \mathcal{O}_\Theta(\Omega_0) \to \mathbb{R}$, and tangent vector fields $\mathscr{V}, \mathscr{W}$, we define a new tangent vector field by*

$$(f\mathscr{V} + g\mathscr{W})_F := f(F(\Omega_0))\mathscr{V}_F + f(F(\Omega_0))\mathscr{W}_F \quad \text{for all } F \in \mathcal{F}(\Theta).$$

Given a tangent vector $V \in \Theta_F$ for some F there are many possibilities to define the *vector transport* of V along a path F_t with $F_0 = F$. One of them is to define

$$\mathscr{V}(t) := V \circ F_0 \circ F_t^{-1}. \tag{2.24}$$

Here and in the following '$\circ$' always denotes the composition of two mappings from $\mathbb{R}^d$ to $\mathbb{R}^d$. This choice is motivated by Theorem 2.31. Our definition of an inner product $g_F(\cdot,\cdot)$ on the tangent spaces Θ_F is closely related to (2.24).

Definition 2.49. *Let Assumption 2.3 be satisfied. For any $F \in \mathcal{F}(\Theta)$ we define the inner product*

$$g_F(\cdot,\cdot) : \Theta_F \times \Theta_F \to \mathbb{R}, \quad (V,W) \mapsto g_F(V,W) := (V \circ F, W \circ F)_{L^2(\Omega_0,\mathbb{R}^d)}.$$

Remark 2.50. Note that the norm induced by $g_F(\cdot,\cdot)$ is weaker than $\|\cdot\|_\Theta$. In fact, the choice of the $L^2(\Omega_0,\mathbb{R}^d)$-scalar product is a bit arbitrary. In the following we will only use the fact that it is a *bounded* bilinear form for all $F \in \mathcal{F}(\Theta)$ and $V, W \in \Theta$.

Lemma 2.51. *Let Assumption 2.3 be satisfied and let $F_t \colon [0,1] \to \mathcal{F}(\Theta)$ be a path in $\mathcal{F}(\Theta)$. Consider two tangent vector fields $V \circ F_0 \circ F_t^{-1}, W \circ F_0 \circ F_t^{-1}$ transported along the path, where $V, W \in \Theta_{F_0}$. Then we have the identity*

$$\forall t \in [0,1]: \quad g_{F_t}\left(V \circ F_0 \circ F_t^{-1}, W \circ F_0 \circ F_t^{-1}\right) = g_{F_0}(V,W).$$

Proof. This is satisfied by the definition of $g_F(\cdot,\cdot)$. □

We will now show that the inner product of two smooth tangent vector fields along a smooth path F_t varies smoothly.

Lemma 2.52. *Let Assumption 2.3 be satisfied and let $F_t\colon [0,1] \to \mathcal{F}(\Theta)$ be a path in $\mathcal{F}(\Theta)$ such that $t \mapsto \partial_t F_t \in \mathcal{V}^{1,1}$. Then, for any (1,1) tangent vector fields $\mathscr{V}, \mathscr{W}$ there exists $\tilde{\mathscr{V}}, \tilde{\mathscr{W}} \in \mathcal{V}^{1,1}$ such that $\tilde{\mathscr{V}}(t) = \mathscr{V}_{F_t}$ and $\tilde{\mathscr{W}}(t) = \mathscr{W}_{F_t}$. Furthermore, the map*

$$[0,1] \ni t \mapsto g_{F_t}\left(\mathscr{V}_{F_t}, \mathscr{W}_{F_t}\right) = g_{F_t}\left(\tilde{\mathscr{V}}(t), \tilde{\mathscr{W}}(t)\right) \in \mathbb{R}$$

is continuously differentiable for all $t \in (0,1)$.

Proof. The first assertion is true by the definition of a (1,1) tangent vector field. Now consider

$$g_{F_t}\left(\tilde{\mathscr{V}}(t), \tilde{\mathscr{W}}(t)\right) = \left(\tilde{\mathscr{V}}(t) \circ F_t, \tilde{\mathscr{W}}(t) \circ F_t\right)_{L^2(\Omega_0, \mathbb{R}^d)}.$$

For every $x \in \mathbb{R}^d$ the derivative of $t \mapsto \tilde{\mathscr{V}}(t) \circ F_t = \tilde{\mathscr{V}}(t, F_t(x))$ is given by

$$\partial_t \tilde{\mathscr{V}}(t, F_t(x)) + D\tilde{\mathscr{V}}(t, F_t(x)) \partial_t F_t(x).$$

An analogous computation holds for $\mathscr{W}$. Hence the derivative of $g_{F_t}\left(\tilde{\mathscr{V}}(t), \tilde{\mathscr{W}}(t)\right)$ is given by

$$\begin{aligned} &\left(\partial_t \tilde{\mathscr{V}}(t) \circ F_t + (D\tilde{\mathscr{V}}(t) \circ F_t)\partial_t F_t, \tilde{\mathscr{W}}(t) \circ F_t\right)_{L^2(\Omega_0, \mathbb{R}^d)} \\ &+ \left(\tilde{\mathscr{V}}(t) \circ F_t, \partial_t \tilde{\mathscr{W}}(t) \circ F_t + (D\tilde{\mathscr{W}}(t) \circ F_t)\partial_t F_t\right)_{L^2(\Omega_0, \mathbb{R}^d)}. \end{aligned} \tag{2.25}$$

Due to our assumptions on $F, \mathscr{V}$, and $\mathscr{W}$ this expression is continuous with respect to t. □

We now present our choice of a *connection*. We begin with an extension of the Eulerian semiderivative to tangent vector fields. Recall the notion of the flow map $T_U(t)$ from Definition 2.14.

Lemma 2.53. *Let Assumption 2.3 be satisfied and $m, k \geq 0$. Consider a $(m+1, k)$ tangent vector field $\mathscr{W}$. Then the* Eulerian semiderivative *of $\mathscr{W}$ at $F \in \mathcal{F}(\Theta)$ in the direction $U \in \Theta$ given by*

$$d\mathscr{W}(F; U) := \lim_{s \searrow 0} \frac{1}{s}\left(\mathscr{W}_{T_U(s) \circ F} - \mathscr{W}_F\right)$$

exists, and satisfies $d\mathscr{W}(F; U) \in \Theta$.

Proof. The path $s \mapsto G_s := T_U(s) \circ F$ satisfies $\partial_s G_s = U \circ F \in \Theta$ for all s, hence $s \mapsto \partial_s G_s$ is in $\mathcal{V}^{\infty,k}$. Thus, by the definition of a $(m+1, k)$ tangent vector field the map $t \mapsto \tilde{\mathscr{W}}(t) := \mathscr{W}_{T_U(s) \circ F}$ is in $\mathcal{V}^{m+1,k}$. In particular, it holds

$$d\mathscr{W}(F; U) = \lim_{s \searrow 0} \frac{1}{s}\left(\tilde{\mathscr{W}}(s) - \tilde{\mathscr{W}}(0)\right) = \partial_s \tilde{\mathscr{W}}(0) \in \Theta.$$

□

Definition 2.54. *Let Assumption 2.3 be satisfied and $m, k \geq 0$. Given two $(m+1, k+1)$ tangent vector fields $\mathscr{V}, \mathscr{W}$, we define*

$$(\blacktriangledown_{\mathscr{V}} \mathscr{W})_F := d\mathscr{W}(F; \mathscr{V}_F) + D\mathscr{W}_F \mathscr{V}_F, \quad \text{for } F \in \mathcal{F}(\Theta).$$

In particular, it holds that $(\blacktriangledown_{\mathscr{V}} \mathscr{W})_F \in \Theta$.

Theorem 2.55. *Let Assumption 2.3 be satisfied and $\mathscr{U}, \mathscr{V}, \mathscr{W}$ be $(m+1, k+1)$ vector fields, $m, k \geq 0$. Then the following properties hold for all $F \in \mathcal{F}(\Theta)$, $a, b \in \mathbb{R}$, and sufficiently smooth shape functionals $f, g \colon \mathcal{O}_\Theta(\Omega_0) \to \mathbb{R}$.*

$$\begin{aligned}
&(i) && (\blacktriangledown_{f\mathscr{V}} \mathscr{W})_F = f(F(\Omega_0)) (\blacktriangledown_{\mathscr{V}} \mathscr{W})_F, \\
&(ii) && (\blacktriangledown_{\mathscr{U}} (a\mathscr{V} + b\mathscr{W}))_F = a (\blacktriangledown_{\mathscr{U}} \mathscr{V})_F + b (\blacktriangledown_{\mathscr{U}} \mathscr{W})_F, \\
&(iii) && (\blacktriangledown_{\mathscr{V}} (f\mathscr{W}))_F = df(F(\Omega_0); \mathscr{V}_F)\mathscr{W}_F + f(F(\Omega_0)) (\blacktriangledown_{\mathscr{V}} \mathscr{W})_F.
\end{aligned}$$

If additionally $d\mathscr{W}(F; U+V) = d\mathscr{W}(F; U) + d\mathscr{W}(F; V)$, then we have also

$$(i^*) \qquad (\blacktriangledown_{f\mathscr{V}+g\mathscr{U}} \mathscr{W})_F = f(F(\Omega_0)) (\blacktriangledown_{\mathscr{V}} \mathscr{W})_F + g(F(\Omega_0)) (\blacktriangledown_{\mathscr{U}} \mathscr{W})_F,$$

i.e., $\blacktriangledown$ corresponds to an affine connection.

Proof. (i) If $\tilde{f} = f(F(\Omega_0)) \in \mathbb{R}$ is zero then the equality is trivially satisfied. Hence, suppose $\tilde{f} \neq 0$. Recall that $T_{\varepsilon V}(t) = T_V(\varepsilon t)$ for $\varepsilon > 0$, and define $T_V(-t) = T_{-V}(t)$. Then

$$\begin{aligned}
d\mathscr{W}(F; \tilde{f}\mathscr{V}_F) &= \lim_{s \searrow 0} \frac{1}{s} \left(\mathscr{W}_{T_{\tilde{f}\mathscr{V}_F}(s) \circ F} - \mathscr{W}_F \right) = \lim_{s \searrow 0} \frac{1}{s} \left(\mathscr{W}_{T_{\mathscr{V}_F}(\tilde{f}s) \circ F} - \mathscr{W}_F \right) \\
&= \tilde{f} \lim_{s \searrow 0} \frac{1}{\tilde{f}s} \left(\mathscr{W}_{T_{\mathscr{V}_F}(\tilde{f}s) \circ F} - \mathscr{W}_F \right) = \tilde{f} d\mathscr{W}(F; \mathscr{V}_F),
\end{aligned}$$

hence the first assertion follows. Property (i^*) is now implied by the stipulated linearity of $d\mathscr{W}(F; \cdot)$.

(ii) The second claim is satisfied since $d(a\mathscr{V} + b\mathscr{W})(F; U) = a(d\mathscr{V}(F; U)) + b(d\mathscr{W}(F; U))$.

(iii) We now verify that

$$d(f\mathscr{W})(F; U) = df(F(\Omega_0); U)\mathscr{W}_F + f(F(\Omega_0)) d\mathscr{W}(F; U),$$

which implies the third assertion. Recall that $(f\mathscr{W})_F = f(F(\Omega_0))\mathscr{W}_F$. Hence

$$\begin{aligned}
&d(f\mathscr{W})(F; U) = \lim_{s \searrow 0} \frac{1}{s} \left(f((T_U(s) \circ F)(\Omega_0)) \mathscr{W}_{T_U(s) \circ F} - f(F(\Omega_0)) \mathscr{W}_F \right) \\
&= \lim_{s \searrow 0} \frac{1}{s} \left((f((T_U(s) \circ F)(\Omega_0)) - f(F(\Omega_0))) \mathscr{W}_{T_U(s) \circ F} + f(F(\Omega_0)) \left(\mathscr{W}_{T_U(s) \circ F} - \mathscr{W}_F \right) \right) \\
&= df(F(\Omega_0); U)\mathscr{W}_F + f(F(\Omega_0)) d\mathscr{W}(F; U).
\end{aligned}$$

□

Remark 2.56. For the choice $\mathscr{W}_F = W \circ F^{-1}$ it holds

$$\begin{aligned} d\mathscr{W}(F;U) &= \partial_s \left(W \circ F^{-1} \circ T_U(s)^{-1} \right)\Big|_{s=0} = \left(D(W \circ F^{-1}) \circ T_U(s)^{-1} \right) \partial_s T_U(s)^{-1}\Big|_{s=0} \\ &= D(W \circ F^{-1})(-U), \end{aligned}$$

and hence indeed $d\mathscr{W}(F;U+V) = d\mathscr{W}(F;U) + d\mathscr{W}(F;V)$, as required for (i^*).

In an appropriate sense the connection is also symmetric. To formulate this we first need to make sense of the second Eulerian semiderivative of a shape functional with respect to tangent vector fields $\mathscr{V}, \mathscr{W} \colon \mathcal{F}(\Theta) \to \Theta$. Recall the usual definition in shape optimization from (2.15) and its characterization in Theorem 2.39. We obtain a compatible expression by defining for some $F \in \mathcal{F}(\Theta)$ the vector field $\tilde{\mathscr{V}} \in \mathcal{V}^{1,k}$ by $\tilde{\mathscr{V}}(t) = \mathscr{V}_{T_{\mathscr{W}_F}(t) \circ F}$, and setting

$$\begin{aligned} d^2 f(F(\Omega_0); \mathscr{V}; \mathscr{W}) &:= d^2 f(F(\Omega_0); \tilde{\mathscr{V}}; \mathscr{W}_F) \\ &= \lim_{t \searrow 0} \frac{1}{t} \left(df(T_{\mathscr{W}_F}(t, F(\Omega_0)); \tilde{\mathscr{V}}(t)) - df(F(\Omega_0); \tilde{\mathscr{V}}(0)) \right). \end{aligned}$$

Note that $d\mathscr{V}(F; \mathscr{W}_F) = \partial_t \tilde{\mathscr{V}}(0)$, hence we obtain from Theorem 2.39 for a shape functional f which is twice Hadamard differentiable that

$$d^2 f(F(\Omega_0); \mathscr{V}; \mathscr{W}) = d^2 f_{F(\Omega_0)}(0; \mathscr{V}_F; \mathscr{W}_F) + df(F; D\mathscr{V}_F \mathscr{W}_F + d\mathscr{V}(F; \mathscr{W}_F)). \tag{2.26}$$

We obtain now immediately the following symmetry result.

Theorem 2.57. *Let the conditions of Theorem 2.55 be satisfied. For any shape functional $f \colon \mathcal{O}_\Theta(\Omega_0) \to \mathbb{R}$ which is twice Hadamard differentiable with respect to Θ, and has the property that $f_{F(\Omega_0)}$ is twice Fréchet differentiable at 0, it holds*

$$df(F(\Omega_0); (\blacktriangledown_{\mathscr{V}} \mathscr{W})_F - (\blacktriangledown_{\mathscr{W}} \mathscr{V})_F) = d^2 f(F(\Omega_0); \mathscr{W}; \mathscr{V}) - d^2 f(F(\Omega_0); \mathscr{V}; \mathscr{W}),$$

i.e., the connection is symmetric.

Proof. This follows directly from the characterization (2.26), the symmetry of the second Fréchet derivative, and the linearity of the derivative $df(F;\cdot)$. □

Finally, we show that our notions of a scalar product and a connection fit together.

Theorem 2.58. *Let the conditions of Theorem 2.55 be satisfied and consider the functional $\psi \colon \mathcal{F}(\Theta) \to \mathbb{R}$*

$$\psi(F) := g_F(\mathscr{V}_F, \mathscr{W}_F), \quad \text{for } F \in \mathcal{F}(\Theta).$$

Then it holds for the Eulerian derivative of ψ at $F \in \mathcal{F}(\Theta)$ in the direction $U = \mathscr{U}_F \in \Theta$ that

$$d\psi(F; \mathscr{U}_F) = g_F((\blacktriangledown_{\mathscr{U}} \mathscr{V})_F, \mathscr{W}_F) + g_F(\mathscr{V}_F, (\blacktriangledown_{\mathscr{U}} \mathscr{W})_F),$$

i.e., $\blacktriangledown$ corresponds to the Levi Cevita connection *for the metric $g_F(\cdot,\cdot)$.*

Proof. The Eulerian derivative of ψ at $F \in \mathcal{F}(\Theta)$ in the direction $U \in \Theta$ is given by

$$d\psi(F;U) = \partial_t \psi(T_U(t) \circ F)|_{t=0}.$$

Note that $T_U(t) \circ F$ is a path satisfying the conditions of Lemma 2.52. Furthermore, it holds $T_U(0) \circ F = F$, and $\partial_t T_U(0) \circ F = U \circ F$. Hence, we obtain from (2.25) that

$$\begin{aligned} d\psi(F;U) =& \partial_t \psi(T_U(t) \circ F)|_{t=0} \\ =& \left(d\mathscr{V}(F;U) \circ F + (D\mathscr{V}_F \circ F)U \circ F, \mathscr{W}_F \circ F\right)_{L^2(\Omega_0,\mathbb{R}^d)} \\ &+ \left(\mathscr{V}_F \circ F, d\mathscr{W}(F;U) \circ F + (D\mathscr{W}_F \circ F)U \circ F\right)_{L^2(\Omega_0,\mathbb{R}^d)} \\ =& g_F((\blacktriangledown_{\mathscr{U}} \mathscr{V})_F, \mathscr{W}_F) + g_F(\mathscr{V}_F, (\blacktriangledown_{\mathscr{U}} \mathscr{W})_F). \end{aligned}$$

□

While these findings are already of interest by themselves, we would like to emphasize especially the following observation. Comparing our notion of the *shape Hessian* from Definition 2.42, with the *second covariant derivative* (2.22), as it is known from the theory of manifolds, we realize that they are *the same for our choice of a connection*.

$$\blacktriangledown^2 \mathrm{j}(\Omega)[V,W] = \langle \blacktriangledown_V \mathrm{j}'(\Omega), W\rangle_{\Theta^*,\Theta} := d^2\mathrm{j}(\Omega;W;V) - d\mathrm{j}(\Omega;\blacktriangledown_V W) = \nabla^2 \mathrm{j}(\Omega)[V,W].$$

We would also like to stress the identity

$$\nabla^2 \mathrm{j}(\Omega)[V,W] = d^2\mathrm{j}_\Omega(0;V;W)$$

from Corollary 2.43. As we will see now it is the analog of (2.23) in shape optimization.

We have so far encountered two constructions in shape optimization which can serve in the role of a *retraction*, i.e., a map from the tangent space to the group of shapes. One of them is the *perturbation of identity* $\tau_U = \mathrm{Id} + U$, the other the *flow map* T_U associated with a vector field $U \in \Theta$. We start our discussion of the merits and drawbacks of those two with the latter.

Due to Theorem 2.15 we know that $T_U(1) \in \mathcal{F}(\Theta)$ for all $U \in \Theta$. Hence the mapping

$$\mathcal{R}_F \colon \Theta_F \to \mathcal{F}(\Theta) \colon \quad U \mapsto \mathcal{R}_F(U) := T_U(1) \circ F$$

is well defined for all $F \in \mathcal{F}(\Theta)$. It can be interpreted as an *retraction*. Indeed, it holds $\mathcal{R}_F(0) = F$, and its derivative is a mapping from Θ_F to $\Theta_{\mathcal{R}_F(U)}$. It follows from [You10, Theorem 8.10] that, for $U, V \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ with support in some bounded $\mathcal{D} \subset \mathbb{R}^d$, we have

$$(\mathcal{R}_F'(U)V)(x) = \int_0^1 DT_U(1-t, T_U(t,x))V(T_U(t,x))\,\mathrm{d}t.$$

In particular, $\mathcal{R}_F'(0)V = V$ for all V, and hence $\mathcal{R}_F'(0) = \mathrm{id}_{\Theta_F}$ as required for a retraction. This can also be verified for other tangent spaces Θ not covered by [You10, Theorem 8.10]. Unfortunately, $\mathcal{R}$ is *not* a *second order retraction* for $\blacktriangledown$. The tangent vector field to the path

$\mathcal{R}_F(tU) = T_{tU}(1) = T_U(t)$ for some $U \in \Theta_F$ is given by $\partial_t \mathcal{R}_F(tU) = \partial_t T_U(t) \equiv U$. Thus, in general, the zero initial acceleration condition (2.21) is not satisfied

$$\blacktriangledown_{\partial_t \mathcal{R}_F(tU)} \partial_t \mathcal{R}_F(tU)\Big|_{t=0} = DUU \neq 0.$$

An alternative is the mapping

$$R\colon B^{\Theta}(0,1) \subset \Theta \to \mathcal{F}(\Theta)\colon U \mapsto \tau_U = \mathrm{Id} + U, \tag{2.27}$$

which we already used in Definition 2.29. A clear disadvantage compared to $\mathcal{R}$ is, that the domain of R is only the unit ball in Θ, whereas $\mathcal{R}$ is defined on the whole space. In that sense R does not satisfy the requirements of a retraction. Its advantage is, that the computation of derivatives of τ_U and the composed function $\mathrm{j}_\Omega(U) = \mathrm{j}(\tau_U(\Omega))$ is straightforward. Furthermore, the tangent vector to the path $t \mapsto R(tU) = \mathrm{Id} + tU$ for some $U \in \Theta$ at t is given by $\partial_t R(tU) = U \circ \tau_{tU}^{-1}$. Here $\tau_{tU}^{-1}\colon \mathbb{R}^d \to \mathbb{R}^d$ is the inverse of the mapping $R(tU) = \tau_{tU} : \mathbb{R}^d \to \mathbb{R}^d$, and given by

$$\tau_{tU}^{-1}(\tilde{x}) = (\mathrm{Id} - tU \circ \tau_{tU}^{-1})(\tilde{x}).$$

We conclude

$$\begin{aligned} D\tau_{tU}^{-1}(\tilde{x}) &= \mathcal{I} - tDU \circ \tau_{tU}^{-1}(\tilde{x}) D\tau_{tU}^{-1}(\tilde{x}) \Rightarrow \\ D\tau_{tU}^{-1}(\tilde{x}) &= \left(\mathcal{I} + tDU \circ \tau_{tU}^{-1}(\tilde{x})\right)^{-1}, \end{aligned}$$

and for the path $t \mapsto \tau_{tU}^{-1}(\tilde{x})$ it holds

$$\begin{aligned} \partial_t \tau_{tU}^{-1}(\tilde{x}) &= -U \circ \tau_{tU}^{-1}(\tilde{x}) - tDU \circ \tau_{tU}^{-1}(\tilde{x}) \partial_t \tau_{tU}^{-1}(\tilde{x}) \Rightarrow \\ \partial_t \tau_{tU}^{-1}(\tilde{x}) &= -\left(\mathcal{I} + tDU \circ \tau_{tU}^{-1}(\tilde{x})\right)^{-1} U \circ \tau_{tU}^{-1}(\tilde{x}). \end{aligned}$$

With the help of these formulas we calculate for some $\mathscr{V}(t) := V \circ \tau_{tU}^{-1}$ that

$$\begin{aligned} \blacktriangledown_{\partial_t R(tU)} \mathscr{V}(t) &= DV \circ \tau_{tU}^{-1} \partial_t \tau_{tU}^{-1} + DV \circ \tau_{tU}^{-1} D\tau_{tU}^{-1} U \circ \tau_{tU}^{-1} \\ &= -DV \circ \tau_{tU}^{-1} (\mathcal{I} + tDU \circ \tau_{tU}^{-1})^{-1} U \circ \tau_{tU}^{-1} \\ &\quad + DV \circ \tau_{tU}^{-1} (\mathcal{I} + tDU \circ \tau_{tU}^{-1})^{-1} U \circ \tau_{tU}^{-1} \\ &= 0. \end{aligned}$$

Hence $\mathscr{V}(t) = V \circ \tau_{tU}^{-1}$ is the *parallel vector transport* of $V \in \Theta_{\mathrm{Id}}$ along the path $t \mapsto R(tU) = \mathrm{Id} + tU$, cf. (2.20). We have already seen in Lemma 2.51 that the inner product of two such transported vectors is constant. Furthermore, we can conclude from $\partial_t R(tU) = U \circ \tau_{tU}^{-1}$, that

$$\blacktriangledown_{\partial_t R(tU)} \partial_t R(tU)\Big|_{t=0} = 0,$$

hence R corresponds to a *second order retraction* for our choice of a connection, cf. (2.21). This fits nicely together with the observation from above, i.e.,

$$\blacktriangledown^2 \mathrm{j}(\Omega)[V,W] = \nabla^2 \mathrm{j}(\Omega)[V,W] = d^2 \mathrm{j}_\Omega(0; V; W),$$

which is exactly the analog of (2.23).

From the above discussion, and considering the results of the Section 2.5 and 2.6, we are lead to the conclusion that it might be preferable to use R to map an element U of the tangent space to $\tau_U \in \mathcal{F}(\Theta)$. The theory of the next two sections is based on this choice.

2.8. A globally convergent linesearch method on the group of transformations

In this section we extend some basic concepts of optimization to the $\mathcal{O}_\Theta(\Omega_0)$ framework, cf. Section 2.2.1. In particular, we show global convergence of a suitable linesearch descent method. We consider the following setting.

Assumption 2.4. *Let $k \geq 0$, Θ be equal to $C^{k+1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$ or $C^{k,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$, and let the nonempty set $\Omega_0 \subset \mathbb{R}^d$ be closed or satisfy $\Omega_0 = \operatorname{int} \overline{\Omega_0}$. The family of admissible sets is given by $\mathcal{O} = \mathcal{O}_\Theta(\Omega_0)$ and $\mathfrak{j}\colon \mathcal{O}_\Theta(\Omega_0) \to \mathbb{R}$ is a shape functional which is Hadamard differentiable with respect to Θ at every $\Omega \in \mathcal{O}_\Theta(\Omega_0)$.*

We will work with the following concept of a ball with radius $r > 0$ around $\Omega \in \mathcal{O} = \mathcal{O}_\Theta(\Omega_0)$.

$$B^{\mathcal{O}}(\Omega, r) := \{\tau_U(\Omega) \mid \tau_U = \mathrm{Id} + U,\ U \in B^\Theta(0, r)\}. \tag{2.28}$$

The following identity shows that this makes sense.

Lemma 2.59. *Let Assumption 2.4 be satisfied. Then for every $\Omega \in \mathcal{O}_\Theta(\Omega_0)$ it holds*

$$\mathcal{O}_\Theta(\Omega) = \mathcal{O}_\Theta(\Omega_0), \text{ and } B^{\mathcal{O}}(\Omega, r) \subset \mathcal{O} = \mathcal{O}_\Theta(\Omega_0) \text{ for all } r \in (0, 1).$$

Proof. The first assertion follows directly from the group property of $\mathcal{F}(\Theta)$ for the composition. Indeed let $\Omega = F(\Omega_0) \in \mathcal{O}_\Theta(\Omega_0)$. Then for all $G \in \mathcal{F}(\Theta)$ we have $G \circ F \in \mathcal{F}(\Theta)$ and hence $G(\Omega) = G \circ F(\Omega_0) \in \mathcal{O}_\Theta(\Omega_0)$. The second assertion follows from the first and Theorem 2.5. □

Remark 2.60. We will in the following work with the above topology generated by the open balls $B^{\mathcal{O}}(\cdot, \cdot)$ on $\mathcal{O} = \mathcal{O}_\Theta(\Omega_0)$. It is equivalent to the topology defined by the metric $d_{\mathcal{F}}$ from (2.4). Recall that $d_{\mathcal{F}}(\mathrm{Id}, F) < \delta$ implies $\|F - \mathrm{Id}\|_\Theta < \delta c$, where $c > 0$ depends only on Θ, cf. Lemma 2.7 and its proof. Conversely for every $\varepsilon > 0$ there exists $\delta > 0$ such that $\|F - \mathrm{Id}\|_\Theta < \delta$ implies $d_{\mathcal{F}}(\mathrm{Id}, F) < \varepsilon$, cf. Theorem 2.5.

Definition 2.61. *Let Assumption 2.4 be satisfied, and let $(\Omega_n) \subset \mathcal{O}_\Theta(\Omega_0)$ be a sequence. The sequence has an* accumulation point *Ω^* in $\mathcal{O}_\Theta(\Omega_0)$, if $\Omega^* \in \mathcal{O}_\Theta(\Omega_0)$, and for every $r \in (0, 1)$ there are infinitely many Ω_n satisfying $\Omega_n \in B^{\mathcal{O}}(\Omega^*, r)$.*

This is equivalent to the existence of a *convergent subsequence* (Ω_{n_k}) satisfying $\Omega_{n_k} = F_{n_k}(\Omega^*)$, $F_{n_k} \in \mathcal{F}(\Theta)$, and $\|F_{n_k} - \mathrm{Id}\|_\Theta \to 0$.

Definition 2.62. *Let Assumption 2.4 be satisfied. We call $\Omega^* \in \mathcal{O} = \mathcal{O}_\Theta(\Omega_0)$ a* local solution *of the shape optimization problem*

$$\min_{\Omega \in \mathcal{O}} \mathfrak{j}(\Omega) \tag{2.29}$$

if there exists an $r \in (0, 1)$ such that

$$\mathfrak{j}(\Omega^*) \leq \mathfrak{j}(\Omega), \quad \text{for all } \Omega \in B^{\mathcal{O}}(\Omega^*, r).$$

Analogously we speak of a strict local solution *if '$<$' holds in the above formula for all $\Omega \in B^{\mathcal{O}}(\Omega^*, r) \backslash \Omega^*$, and of a* (strict) global solution *if we can replace $B^{\mathcal{O}}(\Omega^*, r)$ by $\mathcal{O}$.*

As usual a strict local solution is not necessarily isolated since it might be the accumulation point of a series of local minima. We have the following characterization of a local minimum.

Lemma 2.63. *Let Assumption 2.4 be satisfied. If $\Omega^* \in \mathcal{O}$ is a local solution of* (2.29)*, then the following* necessary optimality condition *holds:*

$$\langle \mathfrak{j}'(\Omega^*), V\rangle_{\Theta^*,\Theta} = 0 \text{ holds for all } V \in \Theta.$$

If $\mathfrak{j}_{\Omega^*}\colon B^{\Theta}(0,1) \to \mathbb{R}$ (cf. Definition 2.29) is twice Gâteaux differentiable, then additionally

$$\nabla^2 \mathfrak{j}(\Omega^*)[V,V] = d^2\mathfrak{j}_{\Omega^*}(0;V;V) \geq 0 \text{ holds for all } V \in \Theta.$$

Proof. By premise $\mathfrak{j}(\Omega) - \mathfrak{j}(\Omega^*) \geq 0$, $\forall \Omega \in B^{\mathcal{O}}(\Omega^*, r)$ for some $r > 0$. For any $V \in \Theta$ it holds $T_V(t) - \mathrm{Id} \in B^{\Theta}(0,\vartheta)$ for small enough $t > 0$. Dividing by t and taking the limit $t \searrow 0$ shows $d\mathfrak{j}(\Omega; V) \geq 0$. Inserting V and $-V$ yields the first claim. The second order necessary condition can be verified with the standard technique for Banach spaces via a Taylor expansion. □

We study now linesearch methods along paths on $\mathcal{O}_\Theta(\Omega_0)$. As we have seen, shape optimization has a lot of similarities to optimization on manifolds, and thus our analysis is inspired by [RW12]. Most results are obtained by studying the localized functionals $\mathfrak{j}_\Omega\colon B^{\Theta}(0,1)$, which places us in a standard Banach space framework, cf., e.g., [HPUU09]. Algorithm 2.1 is adapted from the linesearch minimization algorithm on manifolds of [RW12]. Recall the notations $\tau_U := \mathrm{Id} + U$ and $\mathfrak{j}_{\Omega_k}\colon B^{\Theta}(0,1) \to \mathbb{R}$ for $\Omega_k \in \mathcal{O}_\Theta(\Omega_0)$ from Definition 2.29.

Algorithm 2.1: Monotone linesearch minimization on $\mathcal{O}_\Theta(\Omega_0)$

Require: let Assumption 2.4 be satisfied for $\Omega_0 \subset \mathbb{R}^d$ and $\mathfrak{j}\colon \mathcal{O}_\Theta(\Omega_0) \to \mathbb{R}$
1: set the iteration index to $k = 0$
2: **repeat**
3: choose a descent direction $S_k \in \Theta$, i.e., satisfying $d\mathfrak{j}(\Omega_k; S_k) < 0$
4: choose a step length $\sigma_k > 0$ such that $\|\sigma_k S_k\|_\Theta < 1$ and $\mathfrak{j}(\tau_{\sigma_k S_k}(\Omega_k)) < \mathfrak{j}(\Omega_k)$
5: set $\Omega_{k+1} = \tau_{\sigma_k S_k}(\Omega_k)$
6: increment k
7: **until** $\|\mathfrak{j}'(\Omega_k)\|_{\Theta^*} = 0$

Remark 2.64. Recall that, due to Theorem 2.31, Hadamard differentiability of $\mathfrak{j}$ implies Gâteaux differentiability of $\mathfrak{j}_{\Omega_k}$ for all k. In particular it holds $\mathfrak{j}'_{\Omega_k}(0) = \mathfrak{j}'(\Omega_k)$.

As usual, to ensure convergence of the linesearch method, one needs to impose quality requirements on the choice of descent direction and step length. We employ the notion of *admissible search directions* and *admissible step lengths*:

Definition 2.65. *(i) The sequence of search directions $(S_k) \subset \Theta_{\Omega_k}$ is* admissible *if*

$$\frac{d\mathfrak{j}_{\Omega_k}(0;S_k)}{\|S_k\|_\Theta} \xrightarrow{k\to\infty} 0 \quad \text{implies} \quad \|\mathfrak{j}'_{\Omega_k}(0)\|_{\Theta^*} \xrightarrow{k\to\infty} 0.$$

(ii) The sequence of step sizes (σ_k) *is* admissible *if*

$$\mathfrak{j}_{\Omega_k}(\sigma_k S_k) < \mathfrak{j}_{\Omega_k}(0)\ \forall k \geq 0, \ \textit{and}$$
$$\mathfrak{j}_{\Omega_k}(\sigma_k S_k) - \mathfrak{j}_{\Omega_k}(0) \xrightarrow{k\to\infty} 0 \ \ \textit{implies} \ \ \frac{d\mathfrak{j}_{\Omega_k}(0; S_k)}{\|S_k\|_\Theta} \xrightarrow{k\to\infty} 0.$$

These conditions are enough to ensure global convergence of the algorithm in the following sense.

Theorem 2.66. *Let Assumption 2.4 be satisfied. Consider Algorithm 2.1 and the corresponding sequences* (Ω_k), (S_k), (σ_k). *Suppose that the sequence* $(\mathfrak{j}(\Omega_k)) \subset \mathbb{R}$ *is bounded from below, and that the search directions and step lengths are admissible. Then it holds*

$$\lim_{k\to\infty} \mathfrak{j}'(\Omega_k) = 0 \ \textit{in} \ \ \Theta^*.$$

If $\Omega^* \in \mathcal{O}$ *is an accumulation point of the sequence* $(\Omega_k) \subset \mathcal{O}$, *and the functional* $\mathfrak{j}_{\Omega^*}$ *is continuously Fréchet differentiable at* 0, *then* Ω^* *is a stationary point of* $\mathfrak{j}$.

Proof. For the convenience of the reader we recapitulate the proof of [HPUU09, Theorem 2.2] for the sequences $(\mathfrak{j}_{\Omega_k})$, (Ω_k), (S_k), and (σ_k). Setting $\mathfrak{j}^* = \inf_{k\geq 0} \mathfrak{j}(\Omega_k)$ it follows from $\mathfrak{j}(\Omega_{k+1}) < \mathfrak{j}(\Omega_k)$ that $\mathfrak{j}(\Omega_k) \ \to \mathfrak{j}^*$, and

$$\mathfrak{j}(\Omega_0) - \mathfrak{j}^* = \sum_{k=0}^{\infty} (\mathfrak{j}(\Omega_k) - \mathfrak{j}(\Omega_{k+1})) = \sum_{k=0}^{\infty} |\mathfrak{j}_{\Omega_k}(\sigma_k S_k) - \mathfrak{j}_{\Omega_k}(0)|.$$

Thus $\mathfrak{j}_{\Omega_k}(\sigma_k S_k) - \mathfrak{j}_{\Omega_k}(0) \to 0$, admissibility of the step lengths guarantees

$$\frac{d\mathfrak{j}_{\Omega_k}(0; S_k)}{\|S_k\|_\Theta} \xrightarrow{k\to\infty} 0, \ \text{and hence} \ \ \|\mathfrak{j}'_{\Omega_k}(0)\|_{\Theta^*} \xrightarrow{k\to\infty} 0,$$

since the sequence of search directions is also admissible. For the second assertion suppose that $\Omega^* \in \mathcal{O}$ is an accumulation point of (Ω_k) satisfying the differentiability condition. Then there exists a *convergent subsequence* $(\Omega_k)_K$ satisfying $\Omega_k = F_k(\Omega^*)$, $F_k \in \mathcal{F}(\Theta)$, and $\|F_k - \mathrm{Id}\|_\Theta \to 0$ for all $k \in K$. For k large enough it holds $U_k := F_k - \mathrm{Id} \in B^\Theta(0,1)$, and hence

$$\langle \mathfrak{j}'_{\Omega^*}(U_k), V\rangle_{\Theta^*,\Theta} = \langle \mathfrak{j}'(\Omega_k), V \circ F_k^{-1}\rangle_{\Theta^*,\Theta}, \quad \forall V \in \Theta,$$

cf. Theorem 2.31. Since $\mathfrak{j}_{\Omega^*}$ is continuously Fréchet differentiable the claim follows now from the first part. □

As shown in [HPUU09, Lemma 2.1] admissibility of the search directions can be ensured by the simple *angle condition*

$$\langle \mathfrak{j}'_{\Omega_k}(0), S_k\rangle_{\Theta^*,\Theta} \leq -\nu \|\mathfrak{j}'_{\Omega_k}(0)\|_{\Theta^*} \, \|S_k\|_\Theta \quad \text{for all} \ \ k \geq 0.$$

We focus on the well known *Armijo rule* to select the step length σ_k. It chooses the largest value $\sigma_k \in \{\beta^n \mid n \in \mathbb{N}\}$, $\beta \in (0,1)$ which satisfies $\sigma_k S_k \in B^\Theta(0,1)$ and

$$\mathfrak{j}(\tau_{\sigma_k S_k}(\Omega_k)) - \mathfrak{j}(\Omega_k) \leq \gamma \sigma_k d\mathfrak{j}(\Omega_k; S_k),$$

for some $\gamma \in (0,1)$. Note that the condition is equivalent to

$$\mathfrak{j}_{\Omega_k}(\sigma_k S_k) - \mathfrak{j}_{\Omega_k}(0) \leq \gamma\sigma_k \langle \mathfrak{j}'_{\Omega_k}(0), S_k\rangle_{\Theta^*,\Theta}.$$

One can formulate conditions which guarantee the existence of Armijo step sizes.

Lemma 2.67. *[HPUU09, Lemma 2.2] Let Assumption 2.4 be satisfied, and choose some $\gamma \in (0,1)$. Suppose that for all $k \in \mathbb{N}$, $\Omega_k \in \mathcal{O}$ and $\mathfrak{j}_{\Omega_k}$ is Lipschitz continuously Fréchet differentiable on $B^{\Theta}(0,1)$ for some Lipschitz constant that is independent of k. Then, for every $\varepsilon > 0$, there exists a $0 < \delta < 1$ such that, for all Ω_k and all $S_k \in \Theta(\Omega_k)$ which satisfy*

$$\langle \mathfrak{j}'_{\Omega_k}(0), S_k\rangle_{\Theta^*,\Theta} \leq -\varepsilon \left\|S_k\right\|_{\Theta},$$

there holds

$$\mathfrak{j}_{\Omega_k}(\sigma S_k) - \mathfrak{j}_{\Omega_k}(0) \leq \gamma\sigma \langle \mathfrak{j}'_{\Omega_k}(0), S_k\rangle_{\Theta^*,\Theta}, \quad \forall \sigma \in [0, \delta/\left\|S_k\right\|_{\Theta}].$$

Proof. Since Θ is a Banach space the result [HPUU09, Lemma 2.2] can be applied. We omit it here for brevity. □

One could combine the Armijo condition, which guarantees sufficient decrease, with a curvature condition in the spirit of the *Powell-Wolfe conditions*. But since we only consider trial steps with $\|\sigma_k S_k\|_{\Theta} < 1$ we would need to impose strong assumptions to guarantee the existence of such step lengths. Instead, we require that the descent directions are selected such that they are not too short in the following sense.

Lemma 2.68. *[HPUU09, Lemma 2.3] Let Assumption 2.4 be satisfied, and (Ω_k), (S_k), (σ_k) be generated by Algorithm 2.1. Suppose that, for all $k \in \mathbb{N}$, $\mathfrak{j}_{\Omega_k}$ is Lipschitz continuously Fréchet differentiable on $B^{\Theta}(0,1)$ for some Lipschitz constant that is independent of k. Let the sequence (σ_k) be chosen in accordance with the Armijo rule, and let the sequence (S_k) satisfy*

$$\|S_k\|_{\Theta} \geq \Phi\left(-\frac{\langle \mathfrak{j}'_{\Omega_k}(0), S_k\rangle_{\Theta^*,\Theta}}{\|S_k\|_{\Theta}}\right),$$

where $\Phi\colon [0,\infty) \to [0,\infty)$ is monotonically increasing and fulfills $\Phi(s) > 0$ for all $s > 0$. Then the step lengths (σ_k) are admissible.

Proof. For the convenience of the reader we recapitulate the proof of [HPUU09, Lemma 2.3]. The strict monotonicity property is guaranteed by the Armijo rule. For the second condition of admissibility we carry out an indirect proof. Hence, suppose that there exists an infinite set K and an $\varepsilon > 0$ such that

$$\langle \mathfrak{j}'_{\Omega_k}(0), S_k\rangle_{\Theta^*,\Theta} \leq -\varepsilon \left\|S_k\right\|_{\Theta}, \quad \text{for all } \ k \in K.$$

Then $\|S_k\|_{\Theta} \geq \Phi(\varepsilon) > 0$ for all $k \in K$. Due to Lemma 2.67 we have either $\sigma_k = 1$ or $\sigma_k \geq \delta/(2\left\|S_k\right\|_{\Theta})$ for all $k \in K$. Thus $\sigma_k \left\|S_k\right\|_{\Theta} \geq \min\{\delta/2, \Phi(\varepsilon)\}$ for all $k \in K$. Combining the Armijo rule and the contradictory premise we conclude

$$\mathfrak{j}_{\Omega_k}(\sigma S_k) - \mathfrak{j}_{\Omega_k}(0) \leq \gamma\sigma \langle \mathfrak{j}'_{\Omega_k}(0), S_k\rangle_{\Theta^*,\Theta} \frac{\|S_k\|_{\Theta}}{\|S_k\|_{\Theta}} \leq -\gamma\varepsilon \min\{\delta/2, \Phi(\varepsilon)\} \quad \text{for all } k \in K,$$

and hence $\mathfrak{j}_{\Omega_k}(\sigma_k S_k) - \mathfrak{j}_{\Omega_k}(0) \not\to 0$. □

Remark 2.69. (i) Let us briefly comment on how the conditions on the sequence of functionals $\mathfrak{j}_{\Omega_k}$ can be related to properties of the functional $\mathfrak{j}$. Due to Theorem 2.31 we know that $\mathfrak{j}'(\Omega) = \mathfrak{j}'_\Omega(0)$ if $\mathfrak{j}$ is Hadamard differentiable. Thus continuous differentiability of $\mathfrak{j}$ could be related to $\mathfrak{j}_\Omega$. However, to formulate continuity properties of the shape derivative, we need to have a suitable concept of *transport* from one tangent space to another. As we have seen in section Section 2.7, the canonical choice of transport leads exactly to the definition of the functionals $\mathfrak{j}_\Omega$. Hence, we refrain here from stating explicit conditions on $\mathfrak{j}$ which guarantee properties like the Lipschitz continuous differentiability of $\mathfrak{j}_{\Omega_k}$ for all $k \in \mathbb{N}$.

(ii) For all the concrete shape functionals considered in this thesis we check directly the continuous Fréchet differentiability of $\mathfrak{j}_\Omega$ for all $\Omega \in \mathcal{O}$. This is possible due to the function space parametrization approach described in Section 2.14.

(iii) We worked here with a space Θ of vector fields defined on the whole $\mathbb{R}^d$. However, due to the structure theorem, the support of the shape derivative $\mathfrak{j}'(\Omega)$ is concentrated only on the boundary of Ω. Furthermore, a transformed domain $\tau(\Omega)$ can already be determined if the transformation $\tau \in \Theta$ is only known on $\overline{\Omega}$. Thus, in practice, one will often determine the search direction S_k only on $\overline{\Omega_k}$, and work with *varying* spaces Θ_k of vector fields defined only on $\overline{\Omega_k}$. This could be related to the presented theory if suitable extension and restriction operators exist. We do not pursue this further, but note that, for example in the case of Lipschitz vector fields, the *Kirszbraun-Valentine theorem* [Sch69, Theorem 1.31] states that, for *any* set $\Omega \subset \mathbb{R}^d$ there exists an extension operator which preserves the Lipschitz constant of a vector field $V : \Omega \to \mathbb{R}^d$.

(iv) In fact, it might be desirable to work always only on the tangent spaces Θ_k of vector fields defined on $\overline{\Omega_k}$ and the associated variable norms $\|\cdot\|_{\Theta_k}$. Considering the results of the previous sections, we believe that it is possible to carry our analysis of optimization methods on the group of transformations over to such a localized setting. However, we leave this as subject of future research.

2.9. Second order methods on the group of transformations

It is well known that, in general, steepest descent methods exhibit slow convergence properties. This motivates us to study second order methods, i.e., Newton-type methods, which have the potential of fast local convergence. We proceed as follows. We begin by considering an auxiliary generalized Newton method in the vicinity of a stationary point of $\mathfrak{j}$ and check fast local convergence of the generated iterates under certain assumptions. Since it requires the a-priori knowledge of the stationary point this method is not applicable in practice. However, it serves as a convenient tool for the theoretical convergence analysis. By relating the iterates of this auxiliary method to the Newton steps obtained from Newton's equation for $\mathfrak{j}_{\Omega_k}$ at 0 we obtain an equivalent method which can be used in practice.

The usual assumption for the convergence analysis of Newton's method is continuous invertibility of the Hessian in a stationary point. Unfortunately, the shape Hessian features a nontrivial kernel, and hence the standard argument fails. Although a complete analysis of Newton's

method and second order sufficient conditions in this context is beyond the scope of this thesis, we shed some light on the issues involved. We begin by studying the shape derivative and the shape Hessian more closely with the help of the Hadamard-Zolésio structure theorem. It provides us with a candidate for the kernel of the Hessian. Since the shape gradient is orthogonal to this candidate subspace there is still some hope that Newton's equation can be solved. We sketch some possible strategies which ensure the solvability, and might lead to superlinear convergence of Newton's method *despite* the presence of a nontrivial kernel of the Hessian. We end this section by recalling the method of conjugate gradients (CG), which is the standard tool for the iterative solution of Newton's equation, and operates naturally only on the orthogonal complement of the kernel of a self-adjoint bounded linear operator. Thus, under appropriate assumptions, it can be used to solve a linear equation even if the involved linear operator has a nontrivial kernel.

The presented approach is based on the generalized Newton method in Banach space as developed for example in [HPUU09]. The authors in [RW12] take a slightly different approach to second order methods.

2.9.1. Generalized Newton's method

If we have given a stationary point Ω^* of $\mathfrak{j}$, we can analyze a generalized Newton's method for the functional $\mathfrak{j}_{\Omega^*}$ around 0 in a Banach space setting. Consider Algorithm 2.2. The non-standard termination criterion is taken from [Ulb11, Algorithm 3.10].

Algorithm 2.2: Generalized Newton's method for the functional $\mathfrak{j}_{\Omega^*}$

Require: let Assumption 2.4 be satisfied for $\Omega_0 \subset \mathbb{R}^d$ and $\mathfrak{j}\colon \mathcal{O}_\Theta(\Omega_0) \to \mathbb{R}$. Consider a stationary point Ω^* of $\mathfrak{j}$ and choose $U_0 \in B^\Theta(0,1)$

1: set the iteration index to $k = 0$
2: **repeat**
3: choose an invertible operator $M_k \in \mathcal{L}(\Theta, \Theta^*)$
4: find a solution $P_k \in \Theta$ of

$$M_k P_k = -\mathfrak{j}'_{\Omega^*}(U_k) \quad \text{in } \Theta^*$$

5: set $U_{k+1} = U_k + P_k$
6: increment k
7: **until** $U_{k+1} = U_k$

Noting that

$$M_k U_{k+1} = M_k(U_k + P_k) = M_k U_k - \mathfrak{j}'_{\Omega^*}(U_k) + \mathfrak{j}'_{\Omega^*}(0),$$

one realizes that the sequence $(U_k) \in \Theta$ converges *q-superlinearly* to 0 if and only if $\|U_k\|_\Theta \to 0$

and

$$\text{for all } \eta \in (0,1) \text{ there exists a } \delta_\eta \text{ such that: } \forall k \text{ with } \|U_k\| < \delta_\eta \text{ there holds} \\ \left\| M_k^{-1}(\mathfrak{j}'_{\Omega^*}(U_k) - \mathfrak{j}'_{\Omega^*}(0) - M_k U_k) \right\|_\Theta \leq \eta \, \|U_k\|_\Theta \,. \tag{2.30}$$

Often this requirement on the choice of the operator M_k is split into two stronger conditions.

Assumption 2.5. *The operators $M_k \in \mathcal{L}(\Theta, \Theta^*)$ satisfy*

(i) Regularity condition: There exists $C > 0$ such that

$$\|M_k^{-1}\|_{\mathcal{L}(\Theta^*,\Theta)} \leq C \quad \forall k \geq 0.$$

(ii) Approximation condition: For all $\eta \in (0,1)$ there exists a δ_η such that

$$\|\mathfrak{j}'_{\Omega^*}(U_k) - \mathfrak{j}'_{\Omega^*}(0) - M_k U_k\|_{\Theta^*} \leq \eta \, \|U_k\|_\Theta \quad \forall k \text{ with } \|U_k\| < \delta_\eta.$$

Theorem 2.70. *[HPUU09, Theorem 2.9] Let Assumption 2.4 be satisfied, Ω^* be a stationary point of $\mathfrak{j}$, and consider Algorithm 2.2. If $\|U_0\|_\Theta$ is small enough, and either (2.30) or Assumption 2.5 holds, then the algorithm either terminates with $U_k = 0$ or generates a sequence (U_k) which converges q-superlinearly to 0.*

Proof. Choosing some $\eta \in (0,1)$ and $\|U_0\|_\Theta < \delta_\eta$ one may verify $\|U_{k+1}\|_\Theta \leq \eta \, \|U_k\|_\Theta$ inductively for all $k \geq 0$. If $U_{k+1} = U_k$ then necessarily $U_k = 0$. If the algorithm generates an infinite sequence, then $\|U_k\|_\Theta < \delta_\eta$ for all $k \geq 0$. Due to (2.30) the sequence converges q-superlinearly to 0. Note that Assumption 2.5 implies (2.30). □

Remark 2.71. (i) Of course one can not use Algorithm 2.2 in practice, since it requires the a-priori knowledge of the stationary point Ω^*. Fortunately, we can relate $\mathfrak{j}_{\Omega^*}$ and its derivatives with $\mathfrak{j}_{\Omega_k}$. This will lead us to a realizable method.

(ii) We consider in the following analysis only the classical Newton method. Observe that for the superlinear convergence result we merely need to satisfy (2.30). Hence an extension to semismooth Newton methods (cf., e.g., [Ulb11]), or quasi-Newton methods like the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (cf., e.g., [RW12] for BFGS on manifolds) seems possible. An essential ingredient will be the proper choice of transport between the tangent spaces.

In particular, the above result implies fast local convergence of the classical Newton's method under certain assumptions on $\mathfrak{j}_{\Omega^*}$.

Corollary 2.72. *[HPUU09, Corollary 2.1] Let Assumption 2.4 be satisfied, and Ω^* be a stationary point of $\mathfrak{j}$. Assume that $\mathfrak{j}_{\Omega^*}$ is twice continuously differentiable on $B^\Theta(0,1)$, and that (2.30) is satisfied for the choice $M_k = \mathfrak{j}''_{\Omega^*}(U_k)$. Then Newton's method, i.e., Algorithm 2.2 with the choice $M_k = \mathfrak{j}''_{\Omega^*}(U_k)$ for all k, converges q-superlinearly if $\|U_0\|_\Theta$ is small enough. If $\mathfrak{j}_{\Omega^*}$ is twice Lipschitz-continuously differentiable near 0, then the order of convergence is quadratic.*

Proof. The superlinear convergence result follows directly from Theorem 2.70. For the quadratic convergence we refer to [HPUU09, Corollary 2.1]. □

Remark 2.73. Usually (2.30) is ensured by supposing that the Hessian is continuously invertible in the stationary point. In shape optimization such an assumption is unrealistic. We will discuss this issue in the next subsection.

We begin by translating Algorithm 2.2 with the choice $M_k = \mathfrak{j}''_{\Omega^*}(U_k)$ into an algorithm which does not require the knowledge of the solution Ω^*. We will show that Algorithm 2.3 generates the same sequence of domains as Algorithm 2.2.

Algorithm 2.3: Newton's method for the functional $\mathfrak{j}_{\Omega_k}$

Require: let Assumption 2.4 be satisfied for $\Omega_0 \subset \mathbb{R}^d$ and $\mathfrak{j}\colon \mathcal{O}_\Theta(\Omega_0) \to \mathbb{R}$

1: set the iteration index to $k = 0$
2: **repeat**
3: find a solution $S_k \in \Theta$ of

$$\mathfrak{j}''_{\Omega_k}(0)S_k = -\mathfrak{j}'_{\Omega_k}(0) \quad \text{in } \Theta^*$$

4: set $\Omega_{k+1} = \tau_{S_k}(\Omega_k)$
5: increment k
6: **until** $S_k = 0$

Relating $\mathfrak{j}''_{\Omega_k}(0)$ to $\mathfrak{j}''_{\Omega^*}(U_k)$ we verify now that under the conditions of Corollary 2.72 this algorithm is well defined.

Lemma 2.74. *Let Assumption 2.4 be satisfied and suppose that for some $\Omega^* \in \mathcal{O}_\Theta(\Omega_0)$ the functional $\mathfrak{j}_{\Omega^*}\colon B^\Theta(0,1) \to \mathbb{R}$ is twice Gâteaux differentiable with respect to Θ. Then, for any $\Omega = \tau_U(\Omega^*) \in B^\mathcal{O}(\Omega^*, 1)$, the functional $\mathfrak{j}_\Omega\colon B^\Theta(0,1) \to \mathbb{R}$ is twice Gâteaux differentiable at 0. Moreover, for all $V, W \in \Theta$ it holds*

$$\langle \mathfrak{j}''_\Omega(0)V, W\rangle_{\Theta^*,\Theta} = \langle \mathfrak{j}''_{\Omega^*}(U)V \circ \tau_U, W \circ \tau_U\rangle_{\Theta^*,\Theta}.$$

Proof. Let $\Omega = \tau_U(\Omega^*) \in B^\mathcal{O}(\Omega^*, 1)$. By definition we have for all $V, W \in \Theta$

$$\langle \mathfrak{j}''_\Omega(0)V, W\rangle_{\Theta^*,\Theta} = \lim_{t \searrow 0} \frac{1}{t}\left(\langle \mathfrak{j}'_\Omega(tV), W\rangle_{\Theta^*,\Theta} - \langle \mathfrak{j}'_\Omega(0), W\rangle_{\Theta^*,\Theta}\right).$$

Due to Hadamard differentiability of $\mathfrak{j}$ and Theorem 2.31 it holds

$$\langle \mathfrak{j}'_\Omega(0), W\rangle_{\Theta^*,\Theta} = \langle \mathfrak{j}'(\Omega), W\rangle_{\Theta^*,\Theta} = \langle \mathfrak{j}'_{\Omega^*}(U), W \circ \tau_U\rangle_{\Theta^*,\Theta},$$

and similarly for t small enough

$$\begin{aligned}\langle \mathfrak{j}'_\Omega(tV), W\rangle_{\Theta^*,\Theta} &= \langle \mathfrak{j}'(\tau_{tV}(\Omega)), W \circ \tau_{tV}^{-1}\rangle_{\Theta^*,\Theta} = \langle \mathfrak{j}'(\tau_{tV} \circ \tau_U(\Omega^*)), W \circ \tau_{tV}^{-1}\rangle_{\Theta^*,\Theta} \\ &= \langle \mathfrak{j}'_{\Omega^*}(U + tV \circ \tau_U), (W \circ \tau_{tV}^{-1}) \circ (\tau_{tV} \circ \tau_U)\rangle_{\Theta^*,\Theta} \\ &= \langle \mathfrak{j}'_{\Omega^*}(U + tV \circ \tau_U), W \circ \tau_U\rangle_{\Theta^*,\Theta}.\end{aligned}$$

Since

$$\langle \mathfrak{j}''_{\Omega^*}(U)V, W\rangle_{\Theta^*,\Theta} = \lim_{t\searrow 0} \frac{1}{t}\left(\langle \mathfrak{j}'_{\Omega^*}(U+tV), W\rangle_{\Theta^*,\Theta} - \langle \mathfrak{j}'_{\Omega^*}(U), W\rangle_{\Theta^*,\Theta}\right),$$

we conclude that $\langle \mathfrak{j}''_{\Omega^*}(U)V\circ\tau_U, W\circ\tau_U\rangle_{\Theta^*,\Theta} = \langle \mathfrak{j}''_{\Omega}(0)V, W\rangle_{\Theta^*,\Theta}$. In particular, twice Gâteaux differentiability of $\mathfrak{j}_{\Omega^*}$ implies twice Gâteaux differentiability of $\mathfrak{j}_\Omega$ at 0. □

Remark 2.75. By the same reasoning one could relate twice Fréchet differentiability of $\mathfrak{j}_{\Omega^*}$ on $B^\Theta(0,1)$ and twice Fréchet differentiability of $\mathfrak{j}_\Omega$ at 0.

Theorem 2.76. *Let Assumption 2.4 and the conditions of Corollary 2.72 be satisfied. Consider Algorithm 2.2 with $M_k = \mathfrak{j}''_{\Omega^*}(U_k)$ for $\|U_0\|_\Theta$ small enough, and Algorithm 2.3 with initial choice $\Omega_0 = \tau_{U_0}(\Omega^*)$. Then the algorithms generate the same iterates, i.e., for all $k \geq 0$ it holds*

$$\Omega_k = \tau_{U_k}(\Omega^*) \text{ and } S_k = P_k \circ \tau_{U_k}^{-1}.$$

In particular, the algorithms exhibit q-superlinear convergence $\|U_k\|_\Theta \to 0$.

Proof. Considering Lemma 2.74, and recalling from Theorem 2.31 that for $\Omega = \tau_U(\Omega^*)$

$$\langle \mathfrak{j}'_{\Omega}(0), W\rangle_{\Theta^*,\Theta} = \langle \mathfrak{j}'_{\Omega^*}(U), W\circ\tau_U\rangle_{\Theta^*,\Theta},$$

we realize that P_0 solves

$$\mathfrak{j}''_{\Omega^*}(U_0)P_0 = -\mathfrak{j}'_{\Omega^*}(U_0),$$

if and only if $S_0 = P_0 \circ \tau_{U_0}^{-1}$ solves

$$\mathfrak{j}''_{\Omega_0}(0)S_0 = -\mathfrak{j}'_{\Omega_0}(0).$$

In particular,

$$\Omega_1 = \tau_{S_0}(\Omega_0) = \tau_{P_0\circ\tau_{U_0}^{-1}}(\Omega_0) = (\tau_{U_0} + P_0)\circ\tau_{U_0}^{-1}(\Omega_0) = \tau_{U_0+P_0}(\Omega^*) = \tau_{U_1}(\Omega^*),$$

and since $\|U_1\|_\Theta < \|U_0\|_\Theta$ it holds again $\Omega_1 \in B^{\mathcal{O}}(\Omega^*, \|U_0\|_\Theta)$. Hence the claimed equalities follow by induction. The superlinear convergence property was asserted in Corollary 2.72. □

Remark 2.77. Newton's method is only locally convergent and needs to be *globalized*. It is straightforward to combine it with the globally convergent linesearch descent method Algorithm 2.1. In each step one tries to determine the Newton direction. If this is successful and the angle condition is satisfied then the Newton direction is admissible. Otherwise, the direction of steepest descent may be chosen. In combination with admissible step sizes this yields a globally convergent method, and transition to fast local convergence can be expected under appropriate conditions. However, we leave a rigorous discussion of this strategy, as well as other possible globalization techniques via trust-regions or filters to future research.

2.9.2. Solving the Newton equation

Let us study the Newton equation in more detail, i.e., for some $\Omega = \tau_U(\Omega_0)$ consider

$$\mathfrak{j}''_\Omega(0)S = -\mathfrak{j}'_\Omega(U).$$

As already mentioned, the fact that $\mathfrak{j}$ is constant along directions which leave the shape of the domain unchanged, entails a nontrivial kernel of the Hessian. Let us be more precise. Recall the definition of $\mathcal{V}^k$ in Section 2.6 and Clarke's tangent cone $C_\Omega(x)$ from Definition 2.18. As in [DZ11] we introduce for a set $\Omega \subset \mathbb{R}^d$ and an integer $k \geq 0$ the spaces

$$\begin{aligned} L^k_\Omega &:= \{V \in \mathcal{V}^k \mid V(x) \in (-C_\Omega(x)) \cap C_\Omega(x)\ \forall x \in \overline{\Omega}\}, \\ N^k_{\partial\Omega} &:= \{V \in \mathcal{V}^k \mid \partial^\alpha V = 0 \text{ on } \partial\Omega,\ \forall \alpha \text{ with } |\alpha| \leq k\} \subset L^k_\Omega. \end{aligned}$$

The *Hadamard-Zolésio structure theorem* provides some insight into the behavior of $\mathfrak{j}'(\Omega)$.

Theorem 2.78. *[DZ11, Theorem 9.3.6] Let $\mathfrak{j}$ be a real-valued shape functional which is shape differentiable at $\Omega \subset \mathbb{R}^d$. Then the following holds.*

(i) The support of the shape derivative is contained in $\partial\Omega$.

(ii) If Ω is open or closed in $\mathbb{R}^d$ and the shape derivative is of order $k \geq 0$, then there exists $[\mathfrak{j}'(\Omega)] \in (\mathcal{V}^k/L^k_\Omega)^*$ such that for all $V \in \mathcal{V}^k$

$$d\mathfrak{j}(\Omega; V) = \langle [\mathfrak{j}'(\Omega)], q_L V \rangle_{(\mathcal{V}^k/L^k_\Omega)^*, (\mathcal{V}^k/L^k_\Omega)},$$

where $q_L \colon \mathcal{V}^k \to \mathcal{V}^k/L^k_\Omega$ is the canonical quotient surjection. Moreover

$$\mathfrak{j}'(\Omega) = q_L^*[\mathfrak{j}'(\Omega)],$$

where q_L^* denotes the dual of the linear map q_L.

There exists a similar result for the second shape derivative.

Theorem 2.79. *[DZ11, Theorem 9.6.3] Let $\mathfrak{j}$ be a real-valued shape functional which is twice shape differentiable at $\Omega \subset \mathbb{R}^d$. Then the following holds.*

(i) The vector distribution associated with $d^2\mathfrak{j}(\Omega; \cdot; \cdot)$ has support in $\partial\Omega \times \partial\Omega$.

(ii) If Ω is an open or closed domain in $\mathbb{R}^d$ and $d^2\mathfrak{j}(\Omega; \cdot; \cdot)$ is of order $k \geq 0$, then there exists a continuous bilinear form

$$[H(\Omega)] \colon (\mathcal{V}^k/N^k_{\partial\Omega}) \times (\mathcal{V}^k/L^k_\Omega) \to \mathbb{R},$$

such that for all $V, W \in \mathcal{V}^k$,

$$d^2\mathfrak{j}(\Omega; V; W) = [H(\Omega)](q_N V, q_L W),$$

where $q_N \colon \mathcal{V}^k \to \mathcal{V}^k/N^k_{\partial\Omega}$, and $q_L \colon \mathcal{V}^k \to \mathcal{V}^k/L^k_\Omega$ are the canonical quotient surjections.

We are specifically interested in the second Gâteaux derivative of $\mathfrak{j}_\Omega$. Recall that, under appropriate assumptions, Theorem 2.39 provides the identity

$$d^2\mathfrak{j}_\Omega(0;V;W) = d^2\mathfrak{j}(\Omega;V;W) - d\mathfrak{j}(\Omega;DVW),$$

for $V \in \mathcal{V}^{k+1}$ and $W \in \mathcal{V}^k$. In particular, it holds for $V \in N_{\partial\Omega}^{k+1}$ and arbitrary $W \in \mathcal{V}^k$ that $DVW \in N_{\partial\Omega}^k \subset L_\Omega^k$ and hence $d^2\mathfrak{j}_\Omega(0;V;W) = 0$. On the other hand, if $W \in L_\Omega^k$ and $V \in \mathcal{V}^{k+1}$, then $d^2\mathfrak{j}_\Omega(0;V;W) = -d\mathfrak{j}(\Omega;DVW)$ is, in general, *not* zero. The reason for this is that $\tau_{tV}(\Omega) \neq \Omega$ for a tangential displacement field V, no matter how small we choose $t > 0$. This is a notable difference to the *flow* $T_V(t)$. Thus, in general,

$$N_{\partial\Omega}^{k+1} \subset \operatorname{Ker}(\mathfrak{j}_\Omega''(0)), \quad \text{but} \quad L_\Omega^k \not\subset \operatorname{Ker}(\mathfrak{j}_\Omega''(0)).$$

However, in a critical point Ω^* the Hadamard derivative vanishes, and we conclude

$$L_{\Omega^*}^{k+1} \subset \operatorname{Ker}(\mathfrak{j}_{\Omega^*}''(0)).$$

In any case, the shape Hessian has a *nontrivial kernel*, and is *not* invertible if considered as an operator from $\mathcal{V}^k$ to $(\mathcal{V}^k)^*$. However, the fact that $\mathfrak{j}_\Omega'(0) = \mathfrak{j}'(\Omega) = q_L^*[\mathfrak{j}'(\Omega)]$, feeds the hope that one can still solve the Newton equation and somehow ensure (2.30). A thorough treatment of this issue is very much beyond the scope of this thesis. In the following we only discuss some immediate observations which might point the way for future research.

Perhaps the most obvious solution to the problem is to add a coercive correction term to the Hessian. Since the Hessian is positive semidefinite in a local minimum (cf. Lemma 2.63) a strong enough correction term would provide solvability of the Newton equation in the vicinity of the minimum. To ensure superlinear convergence of such a modified Newton method, the correction may not be too strong. An investigation of this strategy will likely be based on the famous Dennis-Moré condition, cf. [DM74]. However, additional complications occur since we do not have the invertibility of the Hessian in the minimum available. Alternatively, one might add a suitable Tikhonov regularization term to the objective as is often done in optimal control with PDEs. This idea was already applied for specific examples in shape optimization, we mention, for instance, [Bur04, KU15, Lau00]. However, the solutions of the regularized problem will usually not coincide with the solutions of the original problem. Thus, one may then have to decrease the regularization parameter iteratively. These two approaches are very attractive, especially in terms of improving the reliability of the corresponding algorithms.

Alternatively, one might suppose that the Hessian is at least coercive on a suitable subspace. This line of reasoning could be combined with the study of suitable *second order sufficient conditions*. To fix some ideas, we specialize now to a Hilbert space setting. In particular, we would like to employ the method of conjugate gradients to solve the Newton equation. Note that second order sufficient conditions and associated quadratic growth conditions, might also be studied by considering suitable quotient spaces of Θ.

Assumption 2.6. *Assumption 2.4 is satisfied, in particular it holds* $\Theta = \mathcal{V}^k$. *Furthermore,* $\mathcal{H}$ *is a* Hilbert space *of mappings* $\mathbb{R}^d \to \mathbb{R}^d$ *which is* densely *embedded into* Θ*:*

$$\mathcal{H} \hookrightarrow \Theta \text{ densely},$$

with scalar product $(\cdot,\cdot)_{\mathcal{H}}$, *associated norm* $\|\cdot\|_{\mathcal{H}}$, *and* Riesz isomorphism $R\colon \mathcal{H}^* \to \mathcal{H}$.

Remark 2.80. The *Sobolev imbedding theorem* [AF03, Theorem 4.12] states conditions on m, p such that Assumption 2.6 is satisfied for the Sobolev spaces $W^{m,p}(\mathbb{R}^d, \mathbb{R}^d)$.

Now consider some domain $\Omega \in \mathcal{O}_\Theta(\Omega_0)$. The structure theorem states that $\langle \mathfrak{j}'_\Omega(0), W\rangle_{\Theta^*,\Theta} = 0$ for all $W \in L^k_\Omega$. Let us abbreviate $Y = L^k_\Omega \cap \mathcal{H}$. Then the gradient $R\mathfrak{j}'_\Omega(0)$ satisfies $R\mathfrak{j}'_\Omega(0) \in Y^{\perp_R}$, where $Y^{\perp_R} \subset \mathcal{H}$ denotes the orthogonal complement of Y with respect to R. Thus, if we suppose that the Hessian is $\mathcal{H}$-coercive on $Y^{\perp_R}$ we could find a solution of the Newton equation *restricted to* $Y^{\perp_R}$. Note however, that this is usually not a solution of the Newton equation with respect to the full space $\mathcal{H}$ if $L^k_\Omega \neq \operatorname{Ker}(\mathfrak{j}''_\Omega(0))$. Furthermore, verifying the superlinear convergence of a suitably modified version of Algorithm 2.3, respectively Algorithm 2.2 is challenging. A related question is whether one can conclude from coercivity of $\mathfrak{j}''_{\Omega^*}(0)$ in $(L^k_{\Omega^*} \cap \mathcal{H})^{\perp_R}$ the coercivity of $\mathfrak{j}''_{\tau(\Omega^*)}(0)$ with respect to $(L^k_{\tau(\Omega^*)} \cap \mathcal{H})^{\perp_R}$. Observe that, at least if enough smoothness is assumed, the unit normal fields $n, \tilde{n}$ of $\Omega, \tau(\Omega)$ are related by

$$\tilde{n} \circ \tau = \frac{D\tau^{-T} n}{|D\tau^{-T} n|}, \text{ and hence } L^k_{\tau(\Omega)} \circ \tau = D\tau L^k_\Omega.$$

A coercivity assumption with respect to the norm $\|\cdot\|_R$ might often be too restrictive. In the next section we will see an example where this is the case. We suspect that a quadratic growth condition could be obtained if, instead of an embedded Hilbert space $\mathcal{H} \hookrightarrow \Theta$, one works with the *energy space* of the Hessian. This is the weakest Hilbert space $\tilde{H}$ with $\Theta \hookrightarrow \tilde{H}$ for which the Hessian constitutes a continuous bilinear form, cf. [EHS07]. They showed that, in the case of smooth star-shaped domains, the coercivity of the Hessian with respect to the weaker norm $\|\cdot\|_{\tilde{\mathcal{H}}}$ may serve as a second order sufficient condition. We refer to the survey [CT15] for an overview of second order sufficient conditions in PDE-constrained optimization and their many uses. For instance, they are employed to ensure stability of the minimum with respect to perturbations in the data of the problem, cf., e.g., [MT99, Gri06], or for finite element error analysis, we refer to the surveys [HR12, HT10] and the references therein. In [EHS07] the second order conditions are exploited to obtain convergence of discrete solutions to a solution of the continuous problem. However, analyzing Newton's method in such a setting is an ambitious task.

We conclude this section with a brief discussion of the *method of conjugate gradients* in a Hilbert space. It is the method of choice for solving Newton's equation. We would like to especially point out that it can also be applied if the involved linear operator has a nontrivial kernel. It suffices if the operator is coercive on the orthogonal complement of its kernel.

The method of conjugate gradients

Note, that the following properties of the equation (2.31) and the method of conjugate gradients are well known in the community. However, we provide them here for completeness and the convenience of the reader. Let us begin with a brief discussion of the linear equation

$$Ax = b \text{ in } X^*, \tag{2.31}$$

where X is some Hilbert space with some scalar product $(\cdot,\cdot)_R$, and $A \in \mathcal{L}(X, X^*)$, $b \in X^*$. The *Riesz isomorphism* $R \in \mathcal{L}(X^*, X)$ associated with $(\cdot,\cdot)_R$ satisfies

$$(Rz, x)_R = \langle z, x\rangle_{X^*,X} \text{ for all } x \in X, z \in X^*.$$

The dual space X^* together with the scalar product $(y, z)_{R^{-1}} := (Ry, Rz)_R$ is also a Hilbert space. The *kernel* and *range* of A are given by

$$\begin{aligned}\operatorname{Ker}(A) &:= \{x \in X \mid Ax = 0\},\\ \operatorname{Ran}(A) &:= \{z \in X^* \mid \exists x \in X \text{ with } Ax = z\}.\end{aligned}$$

We have the following obvious relations between $\operatorname{Ker}(A)$ and $\operatorname{Ran}(A)$:

Lemma 2.81. *Let $A \in \mathcal{L}(X, X^*)$ be self-adjoint, i.e., $\langle Ax, y\rangle_{X^*,X} = \langle Ay, x\rangle_{X^*,X}$ for all $x, y \in X$. Then there holds*

$$R\overline{\operatorname{Ran}(A)} = \operatorname{Ker}(A)^{\perp_R} \quad \text{and} \quad \operatorname{Ker}(A) = R\operatorname{Ran}(A)^{\perp_{R^{-1}}},$$

where $\operatorname{Ker}(A)^{\perp_R}$ denotes the set of all orthogonal elements of X to $\operatorname{Ker}(A)$ with respect to R, and similarly $\operatorname{Ran}(A)^{\perp_{R^{-1}}}$ for the scalar product $(\cdot,\cdot)_{R^{-1}}$ on X^*.

Proof. Let $\tilde{x} \in R\operatorname{Ran}(A)$ and $y \in \operatorname{Ker}(A)$. Then there exists a $x \in X$ such that $\tilde{x} = RAx$ and

$$(\tilde{x}, y)_R = (RAx, y)_R = \langle Ax, y\rangle_{X^*,X} = \langle Ay, x\rangle_{X^*,X} = 0,$$

which implies $R\operatorname{Ran}(A) \subset \operatorname{Ker}(A)^{\perp_R}$ and hence $\operatorname{Ker}(A) \subset R\operatorname{Ran}(A)^{\perp_{R^{-1}}}$. Conversely consider $z \in \operatorname{Ran}(A)^{\perp_{R^{-1}}}$ and $y \in X$. Then

$$\langle ARz, y\rangle_{X^*,X} = \langle Ay, Rz\rangle_{X^*,X} = (Ay, z)_{R^{-1}} = 0,$$

hence $R\operatorname{Ran}(A)^{\perp_{R^{-1}}} \subset \operatorname{Ker}(A)$. From this we conclude

$$R^{-1}\operatorname{Ker}(A) \subset \left(\operatorname{Ran}(A)^{\perp_{R^{-1}}}\right)^{\perp_{R^{-1}}} = \overline{\operatorname{Ran}(A)}.$$

□

Furthermore, decomposing $X = \operatorname{Ker}(A) \oplus \operatorname{Ker}(A)^{\perp_R}$ we observe that $A\operatorname{Ker}(A)^{\perp_R} = \operatorname{Ran}(A)$.

Let us now discuss a method to find a solution of (2.31). Clearly there exists a solution if and only if $b \in \operatorname{Ran}(A)$. We want to solve the equation iteratively with the *method of conjugate gradients* (CG), which goes back to [HS52]. The convergence properties of CG in Hilbert spaces seem to go back to [Hay54], and were recently summarized and extended in [HS15]. The CG method is a very popular choice for self-adjoint problems of the form (2.31), and arguably the best-understood iterative method for this problem class. Usually, the CG method is studied in combination with a coercivity assumption on A, or a special block structure which arises from saddle point problems, cf. e.g., [SZ07]. We require the coercivity of A only with respect to $\operatorname{Ker}(A)^{\perp_R}$. Algorithm 2.4 summarizes the CG method in a Hilbert space as discussed in [GHS14].

Assumption 2.7. *The operator $A \in \mathcal{L}(X, X^*)$ is self-adjoint and $b \in \mathrm{Ran}(A)$. Furthermore, A is R-coercive on $\mathrm{Ker}(A)^{\perp_R}$, i.e., there exists a $\alpha > 0$ such that*

$$\langle Ax, x\rangle_{X^*,X} \geq \alpha \left\| P_{\perp_R} x \right\|_R^2 \quad \text{for all } x \in X,$$

where $P_{\perp_R} \colon X \to X$ denotes the projection onto $\mathrm{Ker}(A)^{\perp_R}$ with respect to $\|\cdot\|_R$.

Remark 2.82. Observe that this is equivalent to the condition $\langle Ax, x\rangle_{X^*,X} \geq \alpha \|x\|_R^2$ for all $x \in \mathrm{Ker}(A)^{\perp_R}$, since for every $x \in X$ it holds $\langle Ax, x\rangle_{X^*,X} = \langle AP_{\perp_R} x, P_{\perp_R} x\rangle_{X^*,X}$.

Algorithm 2.4: CG method in a Hilbert space

Require: let Assumption 2.7 be satisfied and choose $x_0 = 0 \in \mathrm{Ker}(A)^{\perp_R}$
set $r_0 = b - Ax_0 \in X^*$
set $p_0 = Rr_0 \in X$
set $k = 0$
repeat
  set $\alpha_k = \dfrac{\langle r_k, Rr_k\rangle_{X^*,X}}{\langle Ap_k, p_k\rangle_{X^*,X}}$
  set $x_{k+1} = x_k + \alpha_k p_k$
  set $r_{k+1} = r_k - \alpha_k Ap_k$
  set $\beta_{k+1} = \dfrac{\langle r_{k+1}, Rr_{k+1}\rangle_{X^*,X}}{\langle r_k, Rr_k\rangle_{X^*,X}}$
  set $p_{k+1} = Rr_{k+1} + \beta_{k+1} p_k$
  set $k = k + 1$
until converged

Due to Lemma 2.81 we obtain inductively for all $k \geq 1$ that

$$\begin{aligned} p_{k-1} &\in \mathrm{span}\{Rr_0, (RA)Rr_0, \dots, (RA)^{k-1}Rr_0\} \subset \mathrm{Ker}(A)^{\perp_R}, \\ x_k &\in x_0 + \mathrm{span}\{Rr_0, (RA)Rr_0, \dots, (RA)^{k-1}Rr_0\} \subset \mathrm{Ker}(A)^{\perp_R}, \\ r_k &\in r_0 + \mathrm{span}\{(AR)r_0, \dots, (AR)^k r_0\} \subset \mathrm{Ran}(A). \end{aligned}$$

Hence the CG method operates naturally only on $\mathrm{Ker}(A)^{\perp_R}$, and Assumption 2.7 suffices to carry known results concerning convergence, etc. over to our indefinite setting. We refer to [HS15] for a nice overview of q- and r-linear as well as q- and r-superlinear convergence results. Superlinear convergence of CG in Hilbert spaces was already obtained by [Hay54]. Note that the choice of R corresponds to the choice of a *preconditioner*, cf. [GHS14].

2.10. Interlude: application to a showcase problem

In this section we demonstrate how the theory of Section 2.8 and 2.9 can be applied to a concrete problem and present some numerical experiments. We consider two of the most simple

properties of a shape in two dimensions: the *area* and the *circumference*. We are looking for a domain for which these two quantities are equal. Of course this problem does *not* have a unique solution, for example, the circle with radius 2, and the square with side length 4, both satisfy this condition. However, no other circle or square is a solution of the problem. Let us be more precise in formulating and analyzing the shape optimization problem under consideration.

Assumption 2.8. *Let $\Omega_0 \subset \mathbb{R}^2$ be a bounded Lipschitz domain and consider $\Theta := C^1(\overline{\mathbb{R}^2}, \mathbb{R}^2)$. Furthermore set $\mathcal{O} = \mathcal{O}_\Theta(\Omega_0)$.*

Remark 2.83. Although this problem could easily be considered in a more general setting we restrict ourselves here to Lipschitz domains for the sake of simplicity.

We denote the *area* of a domain $\Omega \in \mathcal{O}$ by

$$\operatorname{area}(\Omega) := \int_\Omega 1 \,\mathrm{d}x,$$

the *circumference* of $\Omega \in \mathcal{O}$ by

$$\operatorname{circ}(\Omega) := \int_{\partial\Omega} 1 \,\mathrm{d}S,$$

and introduce the shape functional

$$\mathfrak{j}\colon \mathcal{O} \to \mathbb{R}, \quad \mathfrak{j}(\Omega) = \frac{1}{2}\left(\frac{\operatorname{area}(\Omega)}{\operatorname{circ}(\Omega)} - 1\right)^2.$$

The showcase shape optimization problem is given by

$$\text{minimize } \mathfrak{j}(\Omega) \quad \text{ such that } \Omega \in \mathcal{O}. \tag{2.32}$$

2.10.1. Shape derivatives

We want to apply the linesearch minimization Algorithm 2.1 for this problem. As we have seen in Section 2.8, it suffices to consider at the current iterate $\Omega_k \in \mathcal{O}$ the functional

$$\mathfrak{j}_{\Omega_k}\colon B^\Theta(0,1) \to \mathbb{R}, \quad \mathfrak{j}_{\Omega_k}(U) = \mathfrak{j}(\tau_U(\Omega_k)).$$

Recall that $\tau_U = \mathrm{Id} + U$. Before determining the derivatives of $\mathfrak{j}_{\Omega_k}$, it is advantageous to derive an explicit representation in terms of $U \in \Theta$. Consider some $\Omega \in \mathcal{O}$. Noting that

$$\operatorname{area}(\tau_U(\Omega)) = \int_\Omega \det(D\tau_U) \,\mathrm{d}x =: \mathfrak{A}_\Omega(U),$$

and

$$\operatorname{circ}(\tau_U(\Omega)) = \int_{\partial\Omega} |D\tau_U\, t| \,\mathrm{d}S =: \mathfrak{C}_\Omega(U),$$

where t denotes the unit tangent vector field to $\partial\Omega$, we obtain

$$\mathfrak{j}_\Omega(U) = \frac{1}{2}\left(\frac{\mathfrak{A}_\Omega(U)}{\mathfrak{C}_\Omega(U)} - 1\right)^2.$$

The following well known result is very helpful for the computation of derivatives in shape optimization.

Lemma 2.84. *(i) The mapping*

$$W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d) \to L^\infty(\mathbb{R}^d)\colon \quad U \mapsto \det(D(\mathrm{Id}+U)) = \det(D\tau_U)$$

is differentiable and the derivative in direction $V \in W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)$ is given by

$$\operatorname{tr}(D\tau_U^{-1}DV)\det(D\tau_U).$$

(ii) The mapping

$$W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d) \to L^\infty(\mathbb{R}^d,\mathbb{R}^{d\times d})\colon \quad U \mapsto D(\mathrm{Id}+U)^{-1} = D\tau_U^{-1}$$

is differentiable and the derivative in direction $V \in W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)$ is given by

$$-D\tau_U^{-1}DVD\tau_U^{-1}.$$

Proof. This is for example proved by Murat and Simon. Assertion *(i)* is implied by [MS76, Lemma 4.2] and *(ii)* by [MS76, Lemma 4.3]. □

Remark 2.85. Certainly, the above rules of differentiation hold also for C^1 vector fields. One could even work with Lipschitz continuous vector fields if one is only interested in the restriction of U, V, W to a bounded domain and we have Lemma 2.1 available. Note that $W^{1,\infty}(\mathbb{R}^d) \neq C^{0,1}(\overline{\mathbb{R}^d})$.

In particular, the functionals $U \mapsto \mathfrak{A}_\Omega(U)$ and $U \mapsto \mathfrak{C}_\Omega(U)$ are twice continuously Fréchet differentiable. We state both the first and second derivatives since we will also present the application of second order methods. We obtain directly from Lemma 2.84

$$\begin{aligned}
\mathfrak{A}'_\Omega(U)V &= \int_\Omega \operatorname{tr}(D\tau_U^{-1}DV)\det(D\tau_U)\,\mathrm{d}x,\\
\mathfrak{A}''_\Omega(U)(V,W) &= \int_\Omega \operatorname{tr}(D\tau_U^{-1}DW)\operatorname{tr}(D\tau_U^{-1}DV)\det(D\tau_U)\,\mathrm{d}x\\
&\quad + \int_\Omega \operatorname{tr}(-D\tau_U^{-1}DWD\tau_U^{-1}DV)\det(D\tau_U)\,\mathrm{d}x,
\end{aligned}$$

and noting that $\mathfrak{C}_\Omega(U) = \int_{\partial\Omega}(t^T D\tau_U^T D\tau_U t)^{1/2}\,\mathrm{d}S$ it holds

$$\begin{aligned}
\mathfrak{C}'_\Omega(U)V &= \int_{\partial\Omega}(t^T D\tau_U^T D\tau_U t)^{-1/2}(t^T D\tau_U^T DV t)\,\mathrm{d}S,\\
\mathfrak{C}''_\Omega(U)(V,W) &= \int_{\partial\Omega}(t^T D\tau_U^T D\tau_U t)^{-1/2}(t^T DW^T DV t)\,\mathrm{d}S\\
&\quad - \int_{\partial\Omega}(t^T D\tau_U^T D\tau_U t)^{-3/2}(t^T D\tau_U^T DW t)(t^T D\tau_U^T DV t)\,\mathrm{d}S.
\end{aligned}$$

In fact, we only require the derivatives at $U = 0$, where these expressions simplify to

$$\begin{aligned}
\mathfrak{A}'_\Omega(0)V &= \int_\Omega \operatorname{div}(V)\,\mathrm{d}x,\\
\mathfrak{A}''_\Omega(0)(V,W) &= \int_\Omega \operatorname{div}(W)\operatorname{div}(V)\,\mathrm{d}x - \int_\Omega \operatorname{tr}(DWDV)\,\mathrm{d}x,\\
\mathfrak{C}'_\Omega(0)V &= \int_{\partial\Omega} t^T DV t\,\mathrm{d}S,\\
\mathfrak{C}''_\Omega(0)(V,W) &= \int_{\partial\Omega} t^T DW^T DV t\,\mathrm{d}S - \int_{\partial\Omega} (t^T DW t)(t^T DV t)\,\mathrm{d}S.
\end{aligned}$$

The derivatives of j_Ω can now easily be determined with the chain rule. It holds

$$\langle \mathrm{j}'_\Omega(0), V\rangle_{\Theta^*,\Theta} = \left(\frac{\mathfrak{A}_\Omega(0)}{\mathfrak{C}_\Omega(0)} - 1\right)\frac{\mathfrak{C}_\Omega(0)\mathfrak{A}'_\Omega(0)V - \mathfrak{A}_\Omega(0)\mathfrak{C}'_\Omega(0)V}{\mathfrak{C}_\Omega(0)^2},$$

and

$$\begin{aligned}
\langle \mathrm{j}''_\Omega(0)V, W\rangle_{\Theta^*,\Theta} = {}& \frac{\mathfrak{C}_\Omega(0)\mathfrak{A}'_\Omega(0)V - \mathfrak{A}_\Omega(0)\mathfrak{C}'_\Omega(0)V}{\mathfrak{C}_\Omega(0)^2}\frac{\mathfrak{C}_\Omega(0)\mathfrak{A}'_\Omega(0)W - \mathfrak{A}_\Omega(0)\mathfrak{C}'_\Omega(0)W}{\mathfrak{C}_\Omega(0)^2}\\
&+ \left(\frac{\mathfrak{A}_\Omega(0)}{\mathfrak{C}_\Omega(0)} - 1\right)\frac{\mathfrak{C}_\Omega(0)\mathfrak{A}''_\Omega(0)(V,W) - \mathfrak{A}_\Omega(0)\mathfrak{C}''_\Omega(0)(V,W)}{\mathfrak{C}_\Omega(0)^2}\\
&- \left(\frac{\mathfrak{A}_\Omega(0)}{\mathfrak{C}_\Omega(0)} - 1\right)\frac{(\mathfrak{C}_\Omega(0)\mathfrak{A}'_\Omega(0)V - \mathfrak{A}_\Omega(0)\mathfrak{C}'_\Omega(0)V)2\mathfrak{C}'_\Omega(0)W}{\mathfrak{C}_\Omega(0)^3}.
\end{aligned}$$

Combining Assumption 2.8 with Assumption 2.6 we can either apply a steepest descent method or a Newton descent method to (2.32). In the first case we choose the negative gradient with respect to $(\cdot,\cdot)_R$ as search direction S_k in line 2 of Algorithm 2.1. This ensures that the search directions are *admissible*, and in combination with the Armijo rule guarantees also *admissible step sizes*. Hence, the global convergence result Theorem 2.66 is applicable. Alternatively we can employ the Newton method as described in Algorithm 2.3. However, since Newton's method is only locally convergent the need for globalization arises, cf. Remark 2.77. We do not go into detail here, but refer to Section 2.12 where we discuss a globalized Newton's method in a different setting.

Since the derivatives in this simple example are explicitly known, we can investigate the nature of the Hessian in more detail. For smooth enough domains, the directional derivatives of $\mathfrak{A}_\Omega$ and $\mathfrak{C}_\Omega$ can be characterized as

$$\mathfrak{A}'_\Omega(0)V = \int_{\partial\Omega} V^T n\,\mathrm{d}S, \quad \text{and} \quad \mathfrak{C}'_\Omega(0)V = \int_{\partial\Omega} \kappa V^T n\,\mathrm{d}S,$$

where κ denotes the *curvature* of $\partial\Omega$, cf. [DZ11, Section 9.4]. Hence we obtain for the circle with radius 2 that $\mathfrak{A}_\Omega(0) = \mathfrak{C}_\Omega(0) = 4\pi$, $\kappa = \frac{1}{2}$, and thus

$$\langle \mathrm{j}''_\Omega(0)V, V\rangle_{\Theta^*,\Theta} = \frac{(2\pi)^2}{(4\pi)^4}\left(\int_{\partial\Omega} V^T n\,\mathrm{d}S\right)^2 \geq 0. \tag{2.33}$$

The positive semidefiniteness is not surprising, cf. Lemma 2.63. However, coercivity with respect to a Hilbert space which embeds into Θ can not be expected. Indeed, we observe in our numerical experiments that a suitable coercive correction term is necessary to ensure solvability of the Newton equation and fast convergence of Newton's method.

2.10.2. Numerical examples

We implemented the proposed algorithms for this example in `MATLAB` [TM15]. For the discretization of the initial domain, the generation of the mesh, and the assembly of mass and stiffness matrices, we use the `Partial Differential Equation Toolbox` of `Matlab` with linear finite elements. Note that, for simplicity, we work on Ω_k instead of $\mathbb{R}^d$, see the discussion at the end of section Section 2.8. Furthermore, we work with linear finite elements, and choose in each iteration the Hilbert space induced by the scalar product

$$(\cdot,\cdot)_R \approx (\cdot,\cdot)_{H^1(\Omega_k)} + w\,(\Delta\cdot,\Delta\cdot)_{L^2(\Omega_k)}\,,$$

with $w = 10^{-1}$. Here we approximate the discrete bi-Laplacian scalar product matrix by $KM^{-1}K$, where K denotes the stiffness matrix, and M is the lumped mass matrix. In particular Assumption 2.6 is *not* satisfied for this choice of scalar product. However, as we will see, one can still obtain good results in practice.

The current domain is represented by the nodal coordinates of the current mesh. Once a search direction (displacement field) has been determined, we obtain the new coordinates of the nodes by displacing them according to the displacement field. Note that the topology of the mesh is not changed during the optimization. In our experiment we start with a square with side length 2. The corresponding initial mesh can be seen on the left side of Figure 2.1.

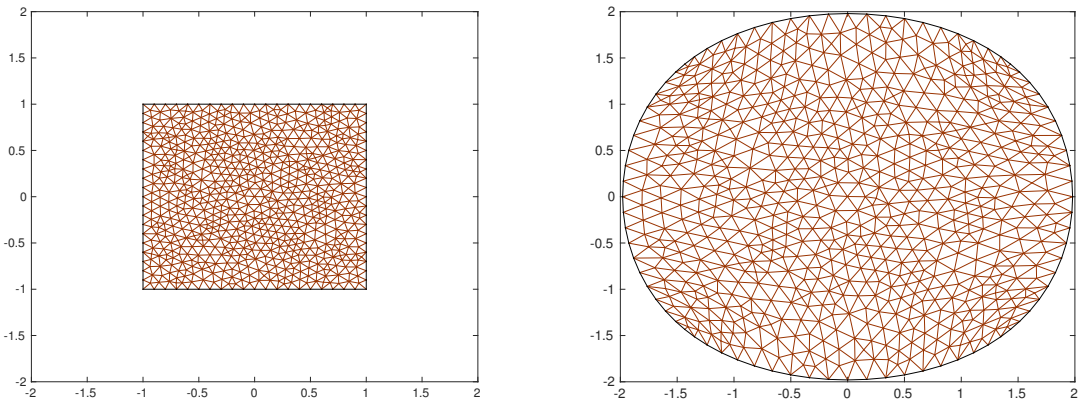


Figure 2.1.: The initial domain of the showcase example (left) and the result of the steepest descent method (right)

We begin by demonstrating the steepest descent method. On the left hand side of Table 2.1 we present the first and last steps in the progression of the algorithm. The first column shows the iteration number, the second the objective value, and the third the norm of the derivative. We observe the typical behavior of a steepest descent method which slowly drives the norm of the derivative towards zero. The result of the optimization is displayed on the right hand side of Figure 2.1. The final domain is close to the optimal circle of radius 2. As one can see from the figure, the quality of the final mesh is quite good.

Table 2.1.: Comparison of steepest descent (left) and globalized Newton (right)

k	$\mathrm{j}(\Omega_k)$	$\left\|\mathrm{j}'_{\Omega_k}(0)\right\|_{R^{-1}}$
0	$1.25{\cdot}10^{-1}$	$7.84{\cdot}10^{-2}$
1	$1.19{\cdot}10^{-1}$	$1.10{\cdot}10^{-1}$
2	$1.08{\cdot}10^{-1}$	$9.39{\cdot}10^{-2}$
3	$9.92{\cdot}10^{-2}$	$8.45{\cdot}10^{-2}$
⋮	⋮	⋮
⋮	⋮	⋮
565	$2.96{\cdot}10^{-11}$	$1.05{\cdot}10^{-6}$
566	$2.84{\cdot}10^{-11}$	$1.03{\cdot}10^{-6}$
567	$2.75{\cdot}10^{-11}$	$1.01{\cdot}10^{-6}$
568	$2.65{\cdot}10^{-11}$	$9.93{\cdot}10^{-7}$

k	$\mathrm{j}(\Omega_k)$	$\left\|\mathrm{j}'_{\Omega_k}(0)\right\|_{R^{-1}}$	step	# CG iter
0	$1.25{\cdot}10^{-1}$	$7.84{\cdot}10^{-2}$	gradient	20
1	$1.19{\cdot}10^{-1}$	$1.10{\cdot}10^{-1}$	gradient	20
⋮	⋮	⋮	⋮	⋮
14	$5.19{\cdot}10^{-2}$	$5.14{\cdot}10^{-2}$	Newton	19
15	$2.93{\cdot}10^{-2}$	$3.62{\cdot}10^{-2}$	Newton	6
⋮	⋮	⋮	⋮	⋮
19	$1.06{\cdot}10^{-4}$	$2.00{\cdot}10^{-3}$	Newton	3
20	$1.47{\cdot}10^{-6}$	$2.34{\cdot}10^{-4}$	Newton	3
21	$3.53{\cdot}10^{-10}$	$3.62{\cdot}10^{-6}$	Newton	2
22	$2.10{\cdot}10^{-17}$	$8.84{\cdot}10^{-10}$	-	-

The typical slow convergence of the steepest descent method motivates us to employ Newton's method. Not surprisingly the unmodified Newton's method diverges for the given initial domain. Thus we include an angle test and in case of failure choose the negative gradient. Furthermore, we allow a maximum of 20 CG iterations, and add the coercive correction term

$$\left\|\mathrm{j}'(\Omega_k)\right\|_{R^{-1}}(\cdot,\cdot)_R$$

to the Hessian. Finally, we choose a steps size according to the Armijo rule. On the right hand side of Table 2.1 we present the first and last steps in the progression of the corresponding algorithm. Again the first column shows the iteration number, the second the objective value, and the third the norm of the derivative. Furthermore, the fourth column indicates which kind of step was taken, and the last column shows the number of CG iterations. After a sequence of gradient steps the Newton method takes over and we clearly observe superlinear convergence. Compared to the runtime of the gradient descent method the overall speed up factor is 4.5, although each Newton iteration is much more expensive than a steepest descent iteration. The result of the optimization is displayed on the left hand side of Figure 2.2. Again the optimal circle of radius 2 is identified, and the quality of the final mesh is quite good.

The observed behavior in Table 2.1 is quite characteristic for the globalized Newton method that we described here. This suggests a simple improvement. We start with a robust steepest descent method and switch to the globalized Newton method once the norm of the derivative drops below a certain threshold. In Table 2.2 we present the last steps in the progression of the corresponding algorithm, where we chose as threshold 10^{-2}. As above we added the coercive correction term to the Hessian. We observe again fast local convergence. Compared to the gradient method the speed up factor is 5.5. The final domain is very similar to the one found by the globalized Newton's method.

If we do *not* modify the Hessian the situation is very different. In all but the very last iterations, the CG method terminates due to negative curvature of the discrete Hessian. Only once we are very close to the solution ($\left\|\mathrm{j}'_{\Omega_k}(0)\right\|_{R^{-1}} = 6.98{\cdot}10^{-5}$) the Newton equation can be solved, and a considerable reduction of the objective and the norm of the derivative is achieved. This

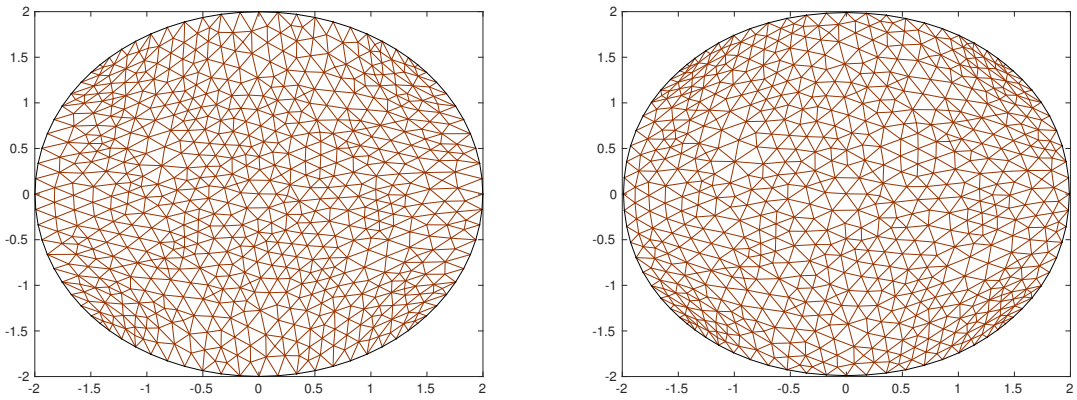


Figure 2.2.: The result of the globalized Newton method (left) and the Gauss-Newton method (right)

Table 2.2.: Progression of the Newton accelerated steepest descent method with Hessian modification

k	$j(u_k)$	$\|j'(u_k)\|_{R^{-1}}$	step	# CG iter
⋮	⋮	⋮	⋮	⋮
83	$2.59{\cdot}10^{-2}$	$1.01{\cdot}10^{-2}$	gradient	-
84	$2.49{\cdot}10^{-2}$	$9.91{\cdot}10^{-3}$	Newton	5
85	$3.82{\cdot}10^{-4}$	$3.81{\cdot}10^{-3}$	Newton	5
86	$1.56{\cdot}10^{-5}$	$7.63{\cdot}10^{-4}$	Newton	3
87	$3.70{\cdot}10^{-8}$	$3.71{\cdot}10^{-5}$	Newton	3
88	$2.29{\cdot}10^{-13}$	$9.24{\cdot}10^{-8}$	-	-

is of course not yet conclusive, but it seems that the finite dimensionality of the discretized Hessian plays a role here.

Finally, we note that our objective is of *least-squares type*. This motivates us to briefly explore a *Gauss-Newton method*, cf., e.g., [NW06, Section 10.3], for the showcase problem. The idea of Gauss-Newton is to neglect all the terms in the second derivative which feature the term

$$\left(\frac{\mathfrak{A}_\Omega(0)}{\mathfrak{C}_\Omega(0)} - 1\right).$$

If the objective is close to zero this term is very small and does not noticeably influence the Hessian of the objective. Hence, in such cases, the Gauss-Newton method mimics a Newton's method quite good, and usually displays also fast local convergence. It has two distinct advantages. For once we do not need to compute second derivatives of $\mathfrak{A}_\Omega$ and $\mathfrak{C}_\Omega$, hence we save computational effort. Furthermore, the Gauss-Newton approximation of the second derivative is always positive semi-definite, even far away from a solution. Thus, the CG method can usually solve the Gauss-Newton equation, and the Gauss-Newton method is more robust than Newton's method. For the example under consideration the benefit of Gauss-Newton

is tremendous. In Table 2.3 we present the progress of the corresponding algorithm. The method converges in three iterations and requires a total of 6 CG iterations. The speed up compared to the gradient method is a factor of 49. Of course, such a performance can *not* be expected in general. Still, it is a reminder that it often pays off to take, as much as possible, the structure of the concrete problem into account. Let us note that there is an abundance of alternative and very sophisticated optimization methods available which we did not touch upon. Interestingly, the mesh in the final domain is not as uniform as in the other examples. It is depicted on the right hand side of Figure 2.2.

Table 2.3.: Progression of the Gauss-Newton method

k	$j(u_k)$	$\left\|j'(u_k)\right\|_{R^{-1}}$	# CG iter
0	$1.25{\cdot}10^{-1}$	$7.84{\cdot}10^{-2}$	2
1	$1.67{\cdot}10^{-3}$	$8.06{\cdot}10^{-3}$	2
2	$1.38{\cdot}10^{-8}$	$2.26{\cdot}10^{-5}$	2
3	$3.95{\cdot}10^{-20}$	$3.83{\cdot}10^{-11}$	-

2.11. Alternative characterizations of shapes

So far we have encoded shapes, i.e., sets $\Omega \subset \mathbb{R}^d$, mainly as images of some reference set Ω_0. It is intuitively clear that, in fact, we only need the image of the *boundary* of the reference set to specify the shape Ω. In particular, only changes of the boundary lead to changes in the value of associated shape functionals. This is the essence of the Hadamard- Zolésio *structure theorem*, cf. Theorem 2.78. Hence, many contributions in the literature encode shapes via some boundary representation. In analogy to the setting so far, one can characterize varying boundaries as transformations of the boundary of some initial set. But there are also alternatives. The boundaries might be parametrized, e.g., by splines, cf. [NZP04, BLUU09, Lin12]. From the optimization point of view this reduces the problem to a small number of design parameters. At the same time, this approach restricts the set of possible shapes considerably. Furthermore, the value of the shape functional usually depends in a highly implicit way on the design parameters. For these reasons more and more contributions consider parametrization-free characterizations of the boundary. Depending on the application, the boundary might be described locally as the graph of a function, cf., e.g., [HM03, Lau00]. Closely related approaches consider star-shaped domains, cf., e.g., [EHS07, Kin15], or describe shapes as normal perturbations of some reference boundary, cf., e.g. [Sch10]. Usually, a change in the boundary needs to be related first to a transformation of the whole domain before the new value of the shape functional can be obtained. An alternative approach is to work only with the boundary without explicitly deforming the whole domain, as in the boundary elements method. We refer to the survey [Har08] and the references therein. However, this approach is only straightforward if Green's function can be calculated for the state equation. There are also more indirect ways to characterize shapes. For instance, in the pseudo-solid approach a shape is determined via artificial forces acting on the elastic reference boundary, cf., e.g., [THM08, Lin12]. Here a change of the design variables is directly related to a transformation of the domain without an intermediate step. A very different philosophy is followed with the homogenization approach,

cf., e.g., [All02], the fictitious domain approach, cf., e.g., [EHM08], or the phase-field approach, cf., e.g., [BGHFSS14], and the references cited in those. These methods have in common that the actual domain is not explicitly meshed, thus saving computational effort by avoiding remeshing or mesh movement procedures. Furthermore they allow for easy topology changes and are thus quite popular in structural optimization. On the other hand, compared to approaches which work with an explicit domain, these techniques suffer from an intrinsic loss of accuracy since the domain is never exactly resolved. A related and very popular approach of encoding shapes is the level set method, which goes back to [OS88].

In the rest of this thesis we concentrate mostly on the characterization of domains via transformations of a reference boundary. Assuming that the initial domain is close to a solution of the shape optimization problem, the problem can be transformed into a standard optimization problem posed in Banach spaces. Thus, all the sophisticated machinery of optimal control with PDEs is available. For instance, with a little care, exact discrete derivatives can be obtained via the continuous adjoint approach, i.e., optimization and discretization commute. We describe this approach in more detail in the next paragraph, and discuss corresponding shape optimization algorithms in Section 2.12. However, sometimes a good initial guess is not available. In particular, it might be the case that the topology of the optimal domain is not known a-priori. The level set method combines nicely an exact shape representation with the possibility of describing elegantly topologic changes of the underlying domain. We briefly recall some of its properties in Section 2.11.3 and apply it in Chapter 7 to minimize the resonance of a harbor basin.

2.11.1. Extension of boundary displacements to domain displacements

Let us describe how a transformation of the boundary can be related to a transformation of the domain. We focus only on transformations given as perturbations of the identity. So far we studied for a suitable Banach space Θ the group of transformations $\mathcal{F}(\Theta)$, and the set of images $\mathcal{O}_\Theta(\Omega)$ of some domain $\Omega \subset \mathbb{R}^d$. The transformations $\tau = \mathrm{Id} + U \in \mathcal{F}(\Theta)$ are determined by vector fields in the tangent space Θ. Analogously, we can consider transformations of the boundary $\partial\Omega$, with an underlying tangent space $\mathcal{U}$ of vector fields $\partial\Omega \to \mathbb{R}^d$. A transformed boundary is obtained as $\tau(\partial\Omega)$, where $\tau = \mathrm{id} + u \colon \partial\Omega \to \mathbb{R}^d$, $u \in \mathcal{U}$, and id is the identity mapping from $\partial\Omega \subset \mathbb{R}^d$ to $\mathbb{R}^d$, i.e., $\mathrm{id}(x) = x \in \mathbb{R}^d$ for all $x \in \partial\Omega$. To connect this approach with the results we have obtained so far for $\mathcal{F}(\Theta)$, we need to extend $u \in \mathcal{U}$ to some vector field $U \in \Theta$. This is done by introducing a suitable extension operator

$$T \colon \mathcal{U} \to \Theta, \quad u \mapsto T(u).$$

While it is of course possible to consider nonlinear extension operators, the common choice are linear ones. Most of the theory we present works for a general smooth extension operator T, but we have in mind a linear operator. As already mentioned in Remark 2.69, in practice one usually works with transformations which are not defined on the whole $\mathbb{R}^d$, but only on the current domain $\overline{\Omega}$. Thus the extension operator has to relate a displacement of the boundary $\partial\Omega$ to a displacement of the domain $\overline{\Omega}$ in some Banach space $\Theta(\overline{\Omega})$.

Remark 2.86. One has to balance between requirements posed on the spaces $\mathcal{U}, \Theta$, and the choice of the operator $T \colon \mathcal{U} \to \Theta$. Vector fields in Θ have to be smooth enough to properly

model the specific problem. In particular, they should ensure a certain regularity of the transformed domains. On the other hand one wants to pose as few restrictions as possible on the design space $\mathcal{U}$. Finally, T is supposed to transport a displacement from $\mathcal{U}$ to Θ, while being as cheap as possible for the actual implementation. Despite the delicacy of these choices, the introduction of T also offers some opportunities. One might, for example, consider only normal perturbations of the boundary, thus obtaining a one-to-one correspondence between perturbations and domains. Another possibility is to gain a compactness property by choosing an extension operator which is completely continuous, see Section 2.11.2 for an example and Section 5.9 for an application.

In simple situations, it may be possible to construct an explicit extension operator with desirable properties. In more complex settings, one chooses T usually as the solution operator of some linear elliptic PDE. An intuitive idea is to consider the domain as an elastic body deformed by the prescribed boundary displacement. Other popular choices for T are the solution operators of the Laplace equation $\Delta y = 0$, or the bi-Laplace equation $\Delta^2 y = 0$. One of the main concerns influencing the choice of the extension operator is the quality of the deformed mesh used for the discretization of a PDE constraint. We refer to the recent contribution of Wick and Wollner [WW14], where these three extension operators were compared in the context of fluid-structure interaction. Their comparison indicates that the bi-Laplacian extension operator is the best suited one, followed by the elasticity based operator. They concluded that for large deformations it might be worthwhile to invest in the additional computational effort required for the bi-Laplacian extension operator. In this thesis we always work with the solution operator of the linear elasticity equation, which offers a good compromise between computational costs and mesh quality.

2.11.2. Example: extension via linear elasticity

Let us consider a domain $\Omega \in \mathcal{O}$. We allow for the possibility that only a part of the boundary $\partial\Omega$ is allowed to be transformed, and denote this design boundary by Γ_B. Of course, the case $\Gamma_B = \partial\Omega$ is included. We want to relate a boundary displacement $u\colon \Gamma_B \to \mathbb{R}^d$ to a domain displacement $U\colon \Omega \to \mathbb{R}^d$. The idea is to prescribe u as boundary data of the linear elasticity equation without volume forces, i.e.

$$\begin{aligned} (\lambda+\mu)(\nabla \operatorname{div} U)^T + \mu \Delta U &= 0 \quad \text{in } \Omega, \\ U &= u \quad \text{on } \Gamma_B, \\ U &= 0 \quad \text{on } \partial\Omega \backslash \Gamma_B. \end{aligned} \tag{2.34}$$

Here $\lambda = \frac{\nu E}{(1+\nu)(1-2\nu)} > 0$, $\mu = \frac{E}{2(1+\nu)} > 0$ are the Lamé parameters, where E is Young's modulus and ν the Poisson ratio. For given boundary datum u we denote the solution of (2.34) by U_u.

To embed the boundary displacement approach into the developed theory we have to ensure at least that $U_u \in C^{0,1}(\overline{\Omega}, \mathbb{R}^d)$, or is even smoother. It is well known that the regularity of a solution of the linear elasticity equation depends on the domain Ω and the regularity of the boundary data. The theory is particularly intricate in the case of Lipschitz domains where corner singularities may appear. It is beyond the scope of this thesis to discuss the necessary

conditions in general. Instead, we demonstrate in the following exemplarily the necessary steps for a concrete setting in $\mathbb{R}^2$. Note that an operator T between Banach spaces X, Y is called *completely continuous* if $x_n \rightharpoonup x$ in X implies $T(x_n) \to T(x)$ in Y. We will use that property extensively in Chapter 5.

Theorem 2.87. *Suppose that Ω is convex with polygonal boundary $\partial\Omega$ and Γ_B is one of its edges. We denote the largest interior angle by $\omega \in (0, \pi)$. It is possible to choose $q < 2$ such, that the equation*

$$\sin^2(z\omega) = \frac{(\lambda+\mu)^2}{(\lambda+3\mu)^2} z^2 \sin^2(\omega) \tag{2.35}$$

has no complex roots in the strip $0 < \mathrm{Re}(z) \leq \frac{2}{q}$. Now we set $p = \frac{q}{q-1}$, and consider a space of boundary displacements $\mathcal{U} \hookrightarrow W_0^{2-1/p,p}(\Gamma_B, \mathbb{R}^2)$. We study for $0 < \alpha < 1 - \frac{2}{p}$ the operator

$$T \colon \mathcal{U} \to C^{1,\alpha}(\overline{\Omega}, \mathbb{R}^2), \quad T(u) = U_u,$$

where U_u solves (2.34). Then T is well defined, linear, and completely continuous.

Proof. The elliptic equation (2.34) admits a unique solution $U_u \in H^1(\Omega, \mathbb{R}^2)$. A classical result of Grisvard [Gri89, Theorem 6.1] states that for $u \in W_0^{2-1/p,p}(\Gamma_B, \mathbb{R}^2)$ the solution has the additional regularity $U_u \in W^{2,p}(\Omega, \mathbb{R}^2)$, if the characteristic equation (2.35) has no roots in the strip $0 < \mathrm{Re}(z) \leq \frac{2}{q}$, where $\frac{1}{p} + \frac{1}{q} = 1$. From [Gri92, Lemma 3.3.1] we know that (2.35) has no roots in $0 < \mathrm{Re}(z) \leq 1$, hence we can choose a $q < 2$ with the desired properties. Thus the auxiliary operator

$$\hat{T} \colon \mathcal{U} \to W^{2,p}(\Omega, \mathbb{R}^2), \quad \hat{T}(u) = U_u,$$

is well defined. Furthermore, it is linear and, as we will now demonstrate, continuous. For this we use the *closed graph theorem*. If the graph

$$(\mathcal{U}, \hat{T}(\mathcal{U})) \subset W^{2-1/p,p}(\Gamma_B, \mathbb{R}^2) \times W^{2,p}(\Omega, \mathbb{R}^2)$$

is closed then the operator $\hat{T}$ is continuous. Hence, consider a sequence $(u_n) \subset \mathcal{U}$ with

$$(u_n, \hat{T}(u_n)) \to (u, U) \in W^{2-1/p,p}(\Gamma_B, \mathbb{R}^2) \times W^{2,p}(\Omega, \mathbb{R}^2).$$

It is clear that

$$(\lambda+\mu)(\nabla \operatorname{div} U)^T + \mu \Delta U = \lim_{n\to\infty} (\lambda+\mu)(\nabla \operatorname{div} \hat{T}(u_n))^T + \mu \Delta \hat{T}(u_n) = 0.$$

Denote by

$$g \colon W^{2,p}(\Omega, \mathbb{R}^2) \to W^{2-1/p,p}(\Gamma_B, \mathbb{R}^2)$$

the *trace operator*. It remains to show

$$\|g(U) - u\|_{W^{2-1/p,p}(\Gamma_B, \mathbb{R}^2)} = 0.$$

Due to [Gri85, Theorem 1.5.2.1] we know that g is continuous. Furthermore

$$\begin{aligned}\|g(U) - u\|_{W^{2-1/p,p}(\Gamma_B,\mathbb{R}^2)} \leq &\left\|g(U) - g(\hat{T}(u_n))\right\|_{W^{2-1/p,p}(\Gamma_B,\mathbb{R}^2)} \\ &+ \left\|g(\hat{T}(u_n)) - u\right\|_{W^{2-1/p,p}(\Gamma_B,\mathbb{R}^2)}.\end{aligned}$$

By the continuity of g the first term goes to zero for $n \to \infty$. Since $g(\hat{T}(u_n)) = u_n$ the same holds for the second term. Hence

$$\hat{T} \in \mathcal{L}\left(W^{2-1/p,p}(\Gamma_B,\mathbb{R}^2), W^{2,p}(\Omega,\mathbb{R}^2)\right).$$

With the well known compact imbedding

$$W^{2,p}(\Omega,\mathbb{R}^2) \hookrightarrow_c C^{1,\alpha}(\overline{\Omega},\mathbb{R}^2)$$

for $p > 2$ and $0 < \alpha < 1 - \frac{2}{p}$ [Gri85, section 1.4.4.] we conclude

$$T \in \mathcal{L}\left(W^{2-1/p,p}(\Gamma_B,\mathbb{R}^2), C^{1,\alpha}(\overline{\Omega},\mathbb{R}^2)\right).$$

In particular, due to linearity, $u_n \rightharpoonup u$ implies $\hat{T}(u_n) \rightharpoonup \hat{T}(u)$ in $W^{2,p}(\Omega_{ref},\mathbb{R}^2)$. The compact imbedding finally provides $T(u_n) \to T(u)$ in $C^{1,\alpha}(\overline{\Omega},\mathbb{R}^2)$, which shows that T is completely continuous. □

Remark 2.88. (i) In particular, the choice $\mathcal{U} = H^2(\Gamma_B,\mathbb{R}^2)$ is possible, since the Sobolev imbedding theorem, cf., e.g., [HPUU09, Theorem 1.14], states that

$$H^2(\Gamma_B,\mathbb{R}^2) \hookrightarrow W^{2-1/p,p}(\Gamma_B,\mathbb{R}^2), \text{ for } 2 < p \leq 4.$$

(ii) Although the extension operator T maps the whole space $\mathcal{U}$ to $\Theta(\Omega) = C^{1,\alpha}(\overline{\Omega},\mathbb{R}^2)$, the associated transformation $\mathrm{Id} + Tu$ is not necessarily a homeomorphism if u, and hence $T(u)$, is to large. We will discuss this issue in more detail in more detail in Section 2.12. In numerical computations a domain displacement which is to large may lead to a corrupted mesh.

We proceed by introducing a very different philosophy of encoding shapes.

2.11.3. Level set representation of the shape

An alternative to the approaches introduced so far, is the representation of a domain $\Omega \subset \mathbb{R}^d$ as the *sub-zero level set* of some function $\Phi\colon \mathbb{R}^d \to \mathbb{R}$. This is the essence of the *level set method*. It was introduced in [OS88], and is widely used to describe propagating fronts, moving interfaces, image segmentation, morphing bodies and similar quantities. We refer to the monographs [FO03, Set99] for a detailed presentation of this rich topic.

Since the introduction of the level set method to shape optimization at the turn of the century, it has developed into one of the most powerful techniques in this area. Some early

works are [AJT02, AJT04, OS01, SW00, WWG03]. Many publications deal with structural optimization. Usually the Ersatz material approach in a hold-all domain $\mathcal{D}$ is used to compute the mechanical properties of the structure and the domain Ω is never explicitly resolved. But there are also approaches where the domain is exactly meshed, consult for instance [ADF14, HC08, Per04, XSLW12] and the references therein. The literature is extensive, we refer to the review paper [vDMLvK13] for an overview of level set based methods in structural topology and shape optimization. The survey [BO05] presents the level set method from the perspective of inverse problems and optimal design.

The most appealing aspects of the level set description of a shape are

(i) the easy global description of the geometry,

(ii) the possibility to encode also geometries of low regularity,

(iii) and the possibility to describe shape and/or *topology* changes of the geometry under consideration, by manipulating the corresponding level set function.

In particular the ability to handle topology changes in a natural and easy way distinguishes the level set method from many other approaches of encoding shapes.

Let us fix the notation of the *sub-zero level set* and the *zero level set*.

Definition 2.89. *Let $\Phi\colon \mathbb{R}^d \to \mathbb{R}$ be a continuous function. We denote*

$$\Omega^\Phi := \{x \in \mathbb{R}^d \mid \Phi(x) < 0\}, \text{ and } \Gamma^\Phi := \{x \in \mathbb{R}^d \mid \Phi(x) = 0\}.$$

Obviously, Ω^Φ is an open set, and $\Gamma^\Phi \supset \partial\Omega^\Phi$ is closed. On the other hand the inclusion $\Gamma^\Phi \subset \partial\Omega^\Phi$ is not necessarily true, this effect is referred to as *fattening*. Whereas we can associate unique sets Ω^Φ, Γ^Φ with a given function Φ, any set $A \subset \mathbb{R}^d$ admits arbitrarily many level set functions. Of those we want to especially point out the *oriented distance function*,

$$b_A := d_A - d_{A^c}, \tag{2.36}$$

which was introduced in Section 2.2.2. Often it is also called the *signed distance function*, but we follow here the terminology of [DZ11]. The oriented distance function features many interesting properties, and is closely linked to several geometric properties of the underlying set. We collect only a few results here, and refer to [DZ11, Chapter 7] for a more detailed presentation of this topic. Let us introduce the notion of the *set of projections* of x onto ∂A

$$\Pi_{\partial A}(x) := \{z \in \partial A \mid d_{\partial A}(x) = |z - x|\},$$

and the *skeleton* of ∂A

$$\mathrm{Sk}(\partial A) := \{x \in \mathbb{R}^d \mid \Pi_{\partial A}(x) \text{ is not a singleton}\} = \{x \in \mathbb{R}^d \mid \nexists \nabla d^2_{\partial A}(x)\}.$$

Theorem 2.90. *[DZ11, Theorem 7.2.1 and 7.3.1] Let A be a subset of $\mathbb{R}^d$ with $\partial A \neq \emptyset$. Then the following hold:*

(i) the function b_A is uniformly Lipschitz continuous on $\mathbb{R}^d$ and

$$\forall x, y \in \mathbb{R}^d\colon\ |b_A(x) - b_A(y)| \leq |x - y|.$$

Moreover, b_A is Fréchet differentiable almost everywhere, and

$$|\nabla b_A(x)| \leq 1 \text{ a.e. in } \mathbb{R}^d.$$

(ii) We have the following chain of equivalences: b_A^2 is Fréchet differentiable at $x \Leftrightarrow$ b_A^2 is Gâteaux differentiable at $x \Leftrightarrow \Pi_{\partial A}(x)$ is a singleton.

There is a close link between the regularity of the set A and the regularity of b_A.

Theorem 2.91. *[DZ11, Theorem 7.8.2] Let $A \subset \mathbb{R}^d$, $\partial A \neq \emptyset$, let $k \geq 1$ be an integer, and let $0 \leq l \leq 1$ be a real number. Denote the ball in $\mathbb{R}^d$ with center x and radius ρ by $B^{\mathbb{R}}(x, \rho)$. Then we have the following characterizations.*

(i) $k = 1, 0 \leq l < 1$: A is of class $C^{1,l}$ and $\partial A \cap \overline{\mathrm{Sk}(\partial A)} = \emptyset$, if and only if

$$\mathrm{meas}(\partial A) = 0 \text{ and } \forall x \in \partial A, \exists \rho > 0 \text{ such that } b_A \in C^{1,l}(\overline{B^{\mathbb{R}}(x, \rho)}).$$

(ii) $k = l = 1$ or $k = 2, 0 \leq l \leq 1$: A is of class $C^{k,l}$, if and only if

$$\mathrm{meas}(\partial A) = 0 \text{ and } \forall x \in \partial A, \exists \rho > 0 \text{ such that } b_A \in C^{k,l}(\overline{B^{\mathbb{R}}(x, \rho)}).$$

Moreover, in all cases, $\nabla b_A = n \circ P_{\partial A}$ in $\overline{B^{\mathbb{R}}(x, \rho)}$, where n is the unit exterior normal to A on ∂A, $P_{\partial A}$ denotes the projection onto ∂A, and ∂A is a $C^{k,l}$-submanifold of dimension $d - 1$.

Observe that, due to Theorem 2.90, we can not infer any regularity of A from a Lipschitz continuous b_A.

Let us briefly discuss the connection between the transformation of a domain and the change in the associated level set function. Since we are mainly interested in situations where the domains under consideration are at least Lipschitz we restrict ourselves here to this case. Note that it is possible to handle much more general situations in the level set framework, cf., e.g., [Kra15a] and the references cited therein. In the setting of the level set method it is more natural to work with the flowmap T than the perturbation of the identity τ. The reason for this is revealed by the following deliberation.

Consider a Lipschitz domain Ω with associated oriented distance function b_Ω. Furthermore, let Θ be equal to $C^{k,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$ or $C^{k+1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$ with $k \geq 0$. For any $V \in \Theta$ and any $T > 0$ the associated flow satisfies $T_V(T) \in \mathcal{F}(\Theta)$. Setting

$$\Phi\colon [0, T] \times \mathbb{R}^d \to \mathbb{R}, \quad \Phi(t, x) := b_\Omega \circ T_V(t)^{-1}(x),$$

it holds $T_V(t)(\Omega) = \{x \in \mathbb{R}^d \mid \Phi(t, x) < 0\}$, and $T_V(t)(\partial\Omega) = \{x \in \mathbb{R}^d \mid \Phi(t, x) = 0\}$. Thus, Φ describes a family of level set functions encoding the evolution of Ω through T_V. Furthermore,

for $x_0 \in \mathbb{R}^d$ with $b_A(x_0) = c \in \mathbb{R}$, denote the associated trajectory by $x(t) = T_V(t)(x_0)$. Obviously it holds $\Phi(t, x(t)) = c$, for all $t \in [0, T]$. Differentiating this equation with respect to t yields

$$\partial_t \Phi(t, x(t)) + \nabla \Phi(t, x(t))^T V(x(t)) = 0, \quad \forall t \in (0, T).$$

Indeed, it is a classical result that a solution of the transport equation

$$\partial_t \Psi + \nabla \Psi^T V = 0 \text{ with initial condition } \Psi(0, x) = b_\Omega(x), \tag{2.37}$$

where $\nabla \Psi^T V$ is meant in the weak sense, satisfies for a.e. $x \in \mathbb{R}^d$ and all $t \in [0, T]$

$$\Psi(t, x) = b_\Omega \circ T_V(t)^{-1}(x),$$

cf. [AC08, Proposition 3.3]. Note that this result is due to the regularity of V. In general the transport equation may lead to shocks and discontinuous solutions. Special solution concepts are required, we point out in particular the concept of *viscosity solutions*, cf. e.g., [Gig06, CIL92]. This is also the usual framework to handle the classical *level set equation*

$$\partial_t \Phi(t, x) + F(x)|\Phi(t, x)| = 0.$$

It is obtained by inserting the velocity field $V = F \nabla \Phi / |\nabla \Phi|$, which points in normal direction with respect to $\partial\Omega$, into (2.37). The level set equation was already employed in [OS88] and is very often used to describe the evolution of the level set function. The use of velocity fields which point in normal direction is motivated by the Hadamard-Zolésio structure theorem, cf. Theorem 2.78. Note that, in contrast to (2.37), Lipschitz continuity of the speed field F is not enough to guarantee well-posedness of the level set equation in the classical sense. Instead, one has to resort to generalized solution concepts like the mentioned viscosity solutions. If one relaxes the regularity of V in (2.37), or uses the classical level set equation, one can also describe topological changes of the underlying family of domains.

Summarizing, we can describe domains as sub-zero level sets of suitable functions Φ. In particular, once a suitable descent direction $V \in \Theta$ is chosen, it is possible to obtain the associated family of transformed domains via the solution of the transport equation (2.37). We employ the level set approach to shape optimization in Chapter 7 to optimize the shape of a breakwater. In particular, we will describe a numerical implementation of the method.

2.12. Shape optimization on a reference domain

For simplicity we suppose now that our initial domain Ω_0 is close to a solution of the shape optimization problem under consideration. That allows us to restrict our attention to a fixed reference domain. In many applications this is a valid assumption, since one is tasked to improve an expert design which is already good. Denote the design boundary by $\Gamma_B \subset \partial\Omega_0$. We specify the setting of this section in the following condition.

Assumption 2.9. *$\tilde{\Theta}$ is given by $C^{k+1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$ or $C^{k,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$ for some $k \geq 0$. The set $\Omega_0 \subset \mathbb{R}^d$ is nonempty, bounded, and either closed, or satisfies $\operatorname{int}\overline{\Omega_0} = \Omega_0$. The shape functional $\mathfrak{j}\colon \mathcal{O}_\Theta(\Omega_0) \to \mathbb{R}$ is Hadamard differentiable with respect to $\Theta = \{U \in \tilde{\Theta} \mid U = 0 \text{ on } \partial\Omega_0 \backslash \Gamma_B\}$. There exists a solution Ω^* of the shape optimization problem $\min_{\Omega \in \mathcal{O}_\Theta(\Omega_0)} \mathfrak{j}(\Omega)$ which satisfies*

$$\Omega^* \in B^{\mathcal{O}}(\Omega_0, 1) = \{\tau_U(\Omega_0) \mid \tau_U = \mathrm{Id} + U,\ U \in B^{\Theta}(0,1)\}.$$

The functional $\mathfrak{j}_{\Omega_0}\colon B^{\Theta}(0,1) \to \mathbb{R}$ from Definition 2.29 may be used instead of $\mathfrak{j}$ on $B^{\mathcal{O}}(\Omega_0, 1)$. Since we need to consider only transformations of $\overline{\Omega_0}$ we can restrict the tangent space to $\Theta(\overline{\Omega_0}) =: \Theta_0$. Thus it suffices to study the localized problem

$$\min_{U \in \Theta_0} \mathfrak{j}_{\Omega_0}(U) \quad \text{s.t.} \quad \|U\|_{\Theta_0} < 1.$$

In fact, we would like to consider only the displacement of the design boundary as our control. We suppose that we have a suitable linear extension operator available.

Assumption 2.10. *Assumption 2.9 is satisfied. Moreover, $\mathcal{U}$ is a Hilbert space of vector fields $\Gamma_B \to \mathbb{R}^d$, and there exists a continuous linear operator $T\colon \mathcal{U} \to \Theta_0$ satisfying*

$$B^{\mathcal{O}}(\Omega_0, 1) \subset \{\tau(\Omega_0) \mid \tau = \mathrm{Id} + Tu,\ u \in \mathcal{U}\}.$$

Introducing the set of feasible boundary displacements

$$\mathcal{U}_{feas} := \{u \in \mathcal{U} \mid \ \|Tu\|_{\Theta_0} < 1\},$$

we can now consider the functional

$$j\colon \mathcal{U}_{feas} \ \to \mathbb{R}, \quad j(u) = \mathfrak{j}_{\Omega_0}(Tu).$$

One may extend j to the whole space $\mathcal{U}$ by setting $j(u) = \infty$ for all $u \notin \mathcal{U}_{feas}$. Clearly the set $\mathcal{U}_{feas}$ is open. In particular, j is differentiable at every $u \in \mathcal{U}_{feas}$. This is due to Theorem 2.31 and the Hadamard differentiability of $\mathfrak{j}$. The derivative of j is given by

$$j'(u) = T^* \mathfrak{j}'_{\Omega_0}(Tu) \in \mathcal{U}^*,$$

where $T^* \in \mathcal{L}(\Theta_0^*, \mathcal{U}^*)$ is the dual operator of T. If T^* is not available in closed form it can be evaluated with the help of some additional adjoint equations. This is briefly described in Section A.2, see also [BLUU09, Lin12]. If $\mathfrak{j}_{\Omega_0}$ is twice continuously Fréchet differentiable, then so is j. Its second derivative can also be obtained via the adjoint approach.

Remark 2.92. In many publications a different approach is chosen to determine the derivatives of j. It relies on the structure theorem of shape optimization Theorem 2.78, which states that only the normal component of the boundary displacement affects the shape derivative. In particular, assuming enough smoothness, the shape derivative can be reformulated as a boundary integral which contains only the normal of the boundary displacement. Thus one can bypass the evaluation of T^*. However, we favor the described approach via the extension operator. The reason for this is that the mentioned reformulation of the shape derivative is *not* valid for a Lipschitz domain. In particular, while such a reformulation may be possible on

the continuous level, it is usually not justified for a finite element discretization of the problem. Moreover, the volume expression of the shape derivative requires less regular finite element functions. In our experience it is also numerically more stable than the Hadamard form. This assessment is shared in the recent papers [HPS15, LS13].

We know from Assumption 2.9 that there exists a solution of

$$\min_{u\in\mathcal{U}} j(u) \tag{2.38}$$

in the open set $\mathcal{U}_{feas}$, in particular, a necessary optimality condition for a local minimum is $j'(u^*) = 0$. Furthermore, a descent method will not leave $\mathcal{U}_{feas}$ since $j(u) = \infty$ for all $u \notin \mathcal{U}_{feas}$. Hence, as long as we start with a feasible point, e.g., $u_0 = 0$, we do not have to incorporate this constraint explicitly in our optimization algorithm.

Remark 2.93. We restrict ourselves here to the open ball $B^{\Theta_0}(0,1)$. More generally one might consider $T\colon \mathcal{U} \to \Theta$ and $\mathcal{U}_{feas} := \{u \in \mathcal{U} \mid \mathrm{Id} + T(u) \in \mathcal{F}(\Theta)\}$. Since the set $\{U \in \Theta \mid \tau_U \in \mathcal{F}(\Theta)\}$ is open in Θ, cf. [MS76, Lemma 2.4], the set $\mathcal{U}_{feas}$ is also open.

The linesearch descent method specified in Algorithm 2.5, with the two variants globalized Newton method (Algorithm 2.6), or Newton-type method (Algorithm 2.7), is standard. We refer, to [HPUU09] for the analysis of these methods in a Banach space framework. Note that the choice $R = \mathcal{A}$ in Algorithm 2.7 corresponds to the classical steepest descent method. In the execution of Newton's method the CG method is terminated early if we encounter negative curvature.

Algorithm 2.5: Monotone linesearch minimization on $\mathcal{U}$

Require: a Riesz isomorphism $\mathcal{A}\colon \mathcal{U}^* \to \mathcal{U}$ with associated dual norm $\|\cdot\|_{\mathcal{A}^{-1}}$
1: set $u_0 = 0$
2: set the iteration index to $k = 0$
3: **repeat**
4: choose a descent direction $v_k \in \mathcal{U}$, i.e., $\langle j'(u_k), v_k\rangle_{\mathcal{U}^*,\mathcal{U}} < 0$
5: employ the Armijo rule (cf. Lemma 2.67) to select a step length $\sigma_k > 0$
6: set $u_{k+1} = u_k + \sigma_k v_k$
7: increment k
8: **until** $\|j'(u_k)\|_{\mathcal{A}^{-1}} = 0$

We conclude this section with a few remarks regarding the practical implementation of these algorithms in the setting of shape optimization.

Remark 2.94. In the algorithms of this section we require the evaluation of $j(u), j'(u)$, and $j''(u)$. These are determined by the functional $\mathfrak{j}_{\Omega_0}$. In many situations, the derivatives of $\mathfrak{j}_{\Omega_0}$ can be calculated explicitly at every $U \in B^{\Theta}(0,1)$. We describe in Section 2.14 a quite general technique for this, and exemplify it in Section 3.1 for an elliptic model problem. See also Section 6.1 for an application to drag minimization in Stokes flow. Utilizing those expressions one can work on the fixed reference domain. We refer to [KV13], where this is

Algorithm 2.6: Computing a globalized Newton direction

Require: a point $u \in \mathcal{U}$ and Riesz isomorphisms $\mathcal{A}\colon \mathcal{U}^* \to \mathcal{U}$, $R\colon \mathcal{U}^* \to \mathcal{U}$

1: try to solve Newton's equation

$$j''(u)v = -j'(u) \quad \text{in } \mathcal{U}^*$$

with the CG method (Algorithm 2.4) using the preconditioner R

2: **if** the CG method exited successfully and

$$\langle j'(u), v\rangle_{\mathcal{U}^*,\mathcal{U}} \leq -\nu \|j'(u)\|_{\mathcal{A}^{-1}} \|v\|_{\mathcal{A}}$$

then

3: **return** $v \in \mathcal{U}$

4: **else**

5: **return** $v = -\mathcal{A}j'(u) \in \mathcal{U}$

Algorithm 2.7: Computing a Newton-type direction

Require: a point $u \in \mathcal{U}$ and a Riesz isomorphism $R\colon \mathcal{U}^* \to \mathcal{U}$

1: set $v = -Rj'(u) \in \mathcal{U}$

2: **return** $v \in \mathcal{U}$

done for an elliptic model problem. However, even in such a 'simple' situation the transformed state equation, as well as the other equations which determine the reduced derivative of the objective, are highly nonlinear with respect to the displacement U, and require specialized solution techniques. For more complex state equations, e.g., the instationary Navier-Stokes equation, this may be very tedious. Fortunately, it is *not necessary to do this*. Instead, the objective and its derivatives may be evaluated on the current domain $\Omega = (\mathrm{Id} + Tu_k)(\Omega_0)$. With the help of the relations provided by Theorems 2.31 and 2.39, respectively Lemma 2.74, these can then be transported to the reference domain Ω_0. We explain this in more detail in Section 3.4.

Remark 2.95. Of course the question arises, whether Newton's method exhibits fast local convergence, i.e., whether the Hessian is continuously invertible near the optimum. Compared with Newton's method in terms of the domain displacement, cf. Section 2.9, we have now excluded the pathological situation of a displacement field U which is zero on the boundary. However, if we allow for free boundary displacements, vector fields which are *tangential* to the boundary may still cause problems. We refer to the discussion in Section 2.9.2. One possible remedy is to consider only *normal* displacements of the boundary. This restriction can be easily combined with the approach described in the section at hand. While such a restriction may yield a positive definite Hessian in the optimum, the *coercivity* of the Hessian with respect to $\mathcal{U}$ is still not ensured. As it is often the case in PDE- constrained optimization, it might even be unrealistic in many situations. In Chapter 4 we analyze the Hessian of a model problem in detail. We conclude that it corresponds roughly to a differential operator of order one, which is positive but not even $H^1(\Gamma_B)$-coercive. See also [EHS07] for a related

discussion involving smooth star-shaped domains. Many optimal control problems feature a term $\frac{\beta}{2}\|u\|_{\mathcal{U}}^2$, which provides a coercive contribution to the Hessian and is either termed control cost or Tikhonov regularization. In particular, usage of such a term avoids problems with tangential displacements. In Section 3.4 we combine normal boundary displacements with such a regularization term, while in Section 6.3 we work with free boundary displacements and a similar regularization term.

2.13. Point-wise geometric constraints and a projected descent method

In this section we start our discussion of possible approaches to incorporate point-wise geometric constraints into the shape optimization procedure, i.e., constraints of the form

$$A \subset \Omega, \text{ or } L \cap \Omega = \emptyset.$$

Here $A, L \subset \mathbb{R}^d$ describe some region which should be contained in Ω, or which is forbidden. These are *point-wise constraints*, in the sense that they have to be satisfied for every point of the admissible domains. In contrast, restrictions on, e.g., the volume, or the boundary smoothness, are more global constraints which we do not address here. Sometimes the term geometric constraints is also used to describe constraints like minimum/maximum thickness of shapes, cf., e.g., [Mic14] and the references therein. However, these are not point-wise properties of the shape, but local properties.

Point-wise geometric constraints appear frequently in practical applications. They may be part of the model to obtain a sensible solution. For example, in Chapter 7 we consider a shape optimization problem involving a harbor breakwater. Naturally, it is not a feasible solution to completely enclose the harbor basin, hence the harbor basin and the harbor approach are forbidden regions for the breakwater. In other applications the geometric constraints may have nothing to do with the physical model described in the shape optimization problem. Instead, they may be outside restrictions on the available space of the component to be optimized, or one might, for instance, want to find a body with prescribed volume and minimum drag which can be stored in a certain box. One can think of many more examples. Such design constraints have been considered in various publications. However, usually they are either only considered with regard to some particular parametrization, e.g., constraints on the control points of some Bézier curve, or discretization, see for example [ABV13, BLUU09, BLUU11, Bra11, Lin12, HLA08, NZP04], or they are tacitly assumed to be inactive in the solution, see for example [Lau00, KV13].

If we consider the admissible family of domains to be transformations of an initial reference domain, then the point-wise geometric constraints can be considered as constraints on the transformations. We will pursue this line of reasoning in Section 5.9 for geometric constraints of the form $\tau(\Gamma_B) \subset C$, where Γ_B is the design part of the boundary, and $C \subset \mathbb{R}^d$ some closed, convex set. In this section we consider the more general situation of a family of admissible domains given by

$$\mathcal{O}_{ad} := \{\Omega \in \mathcal{O} \mid A \subset \Omega,\ L \cap \Omega = \emptyset\},$$

for some $A, L \subset \mathbb{R}^d$.

A very natural idea for solving the constrained optimization problem

$$\min_{\Omega \in \mathcal{O}} \mathrm{j}(\Omega) \text{ s.t. } \Omega \in \mathcal{O}_{ad}, \tag{2.39}$$

is to use appropriate *projections*. The *projected descent method* is the prototypical example of an algorithm which always stays feasible with regard to the constraints. The idea is to take a step in an appropriately chosen descent direction, and afterwards to project onto the admissible set. The method is classical in the context of optimization in Hilbert spaces, cf., e.g., [HPUU09, Section 2.2.2], and has recently also been generalized to the Banach space context, cf. [BR15]. Unfortunately, an extension to the shape optimization framework is not straightforward. It is in principle possible to define projections in the different metric shape spaces. However, the practical realization of such a projection with respect to the metric $d_{\mathcal{F}}$ is a challenging topic. We leave this option open and note that it would be a very interesting topic for further research. On the other hand, realizing the projection, for instance, with respect to the distance induced by the measure of the symmetric set difference or the Hausdorff metric is possible [Kra15b]. Unfortunately, there is no intrinsic notion of an appropriate general tangent space available. Hence the computation of derivatives which are compatible with these metrics can only be done in specific situations, but not in a canonical way.

Due to the mentioned difficulties associated with the classical projected descent method we propose to use an alternative approach. We choose again $\mathcal{O} = \mathcal{O}_\Theta(\Omega_0)$ for some initial domain $\Omega_0 \subset \mathbb{R}^d$ and suitable Banach space Θ. Suppose furthermore that Ω_0 satisfies the geometric constraints. The idea of our iterative scheme is the following. After having determined a descent direction at the current iterate, we project the descent direction onto a suitable set of vector fields which keep the deformed domain in the admissible set. The easiest way to ensure this is to define

$$\mathcal{V}_{feas} := \{V \in \Theta \mid V|_{A \cup L} = 0\}.$$

Obviously, if $\Omega \in \mathcal{O}_{ad}$ and $V \in \mathcal{V}_{feas}$, then $T_V(t)(\Omega) \in \mathcal{O}_{ad}$ for all $t > 0$. The same is true for small enough perturbations of the identity.

There is some freedom in defining the descent direction and the associated projection. Note however, that it is important to derive these two quantities with respect to the same scalar product. Otherwise one can not guarantee that the projected direction is a *descent direction*. There are already easy examples of linear-quadratic constrained optimization problems in two dimensions showing that the projected Newton direction is not necessarily a descent direction, cf., e.g., [HPUU09, Example 2.2]. We consider the following setting.

Assumption 2.11. *Assumption 2.4 is satisfied. $\mathcal{H}$ is a Hilbert space with $\mathcal{H} \hookrightarrow \Theta$ densely. Furthermore, $a(\cdot,\cdot) : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is a symmetric, continuous, and coercive bilinear form with associated norm $\|\cdot\|_a$.*

Remark 2.96. (i) We work here with an imbedded Hilbert space instead of the Banach space Θ. It might be interesting to see whether one can transfer the results of [BR15] also to Θ. Since we have to require Θ to be at least $C^{0,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$ this is not straightforward.

(ii) One may generalize the fixed bilinear form $a(\cdot,\cdot)$ to a variable one, and retain convergence of the corresponding algorithm under suitable uniformness requirements, cf., e.g., [BR15]. Thus it is possible to consider second order information in the bilinear form.

Definition 2.97. *Let Assumption 2.11 be satisfied, and denote $\mathcal{V}_{ad} := \mathcal{H} \cap \mathcal{V}_{feas}$. Define the projection onto $\mathcal{V}_{ad}$ with respect to $a(\cdot,\cdot)$ as*

$$P_a\colon \mathcal{H} \to \mathcal{H}\colon \quad P_a(U) \in \mathcal{V}_{ad},\ \text{and } a(U - P_a(U), W) = 0 \quad \forall W \in \mathcal{V}_{ad}.$$

For $\Omega_k \in \mathcal{O}$ we denote with V_k the direction

$$V_k \in \mathcal{H}\colon\ a(V_k, V) = \langle \mathrm{j}'(\Omega_k), V\rangle_{\mathcal{H}^*,\mathcal{H}} \quad \forall V \in \mathcal{H}.$$

Remark 2.98. We will in the following refer to V_k as *gradient*, since it is the Riesz representative of $\mathrm{j}'(\Omega_k)$ with respect to $a(\cdot,\cdot)$. In particular $-V_k$ is the direction of steepest descent w.r.t. $\|\cdot\|_a$. Note again, that with a suitable choice of the bilinear form one could also obtain a *Newton-type direction*.

Since $\mathcal{V}_{ad}$ is a linear subspace the projection can be implemented efficiently. In fact, instead of first computing the gradient V_k and then projecting it, we can directly compute the gradient with respect to the subspace and obtain the same vector field.

Proposition 2.99. *[KK15, Lemma 5.5] Let Assumption 2.11 be satisfied and $\Omega_k \in \mathcal{O}$. The element $U_k \in \mathcal{V}_{ad}$ satisfies*

$$a(U_k, W) = \langle \mathrm{j}'(\Omega_k), W\rangle_{\mathcal{H}^*,\mathcal{H}} \quad \forall W \in \mathcal{V}_{ad} \tag{2.40}$$

if and only if there holds $U_k = P_a(V_k)$.

Proof. By definition of the gradient it holds for all $W \in \mathcal{H}$

$$a(V_k - U_k, W) = a(V_k, W) - a(U_k, W) = \langle \mathrm{j}'(\Omega_k), W\rangle_{\mathcal{H}^*,\mathcal{H}} - a(U_k, W)\,.$$

Considering $W \in \mathcal{V}_{ad}$ shows the equivalence of the statements. □

We have the following result concerning the optimality of the current iterate Ω_k.

Proposition 2.100. *[KK15, Lemma 5.6] Let Assumption 2.11 be satisfied and $\Omega_k \in \mathcal{O}$.*

(i) If $P_a(V_k) = 0$, then it holds $\langle \mathrm{j}'(\Omega_k), W\rangle_{\mathcal{H}^*,\mathcal{H}} = 0$ for all $W \in \mathcal{V}_{ad}$.

(ii) If $P_a(V_k) \neq 0$, then $P_a(-V_k)$ is a descent direction, i.e.

$$\langle \mathrm{j}'(\Omega_k), P_a(-V_k)\rangle_{\mathcal{H}^*,\mathcal{H}} = -\,\|P_a(-V_k)\|_a^2 < 0.$$

Proof. This follows directly from the definitions, respectively from the above equivalence. □

Remark 2.101. Note that Proposition 2.100 does *not* guarantee us, that Ω_k is a local solution of (2.39) if $P_a(V_k) = 0$. We can only expect to obtain a local solution of the restricted problem

$$\min_{\Omega \in \mathcal{O}} \mathrm{j}(\Omega) \quad \text{s.t.} \quad \Omega \in \mathcal{O}_{\mathcal{V}_{ad}}(\Omega_0).$$

In fact, $P_a(-V_k)$ is an admissible descent direction with respect to the space $\mathcal{V}_{ad}$. Combined with the Armijo linesearch we obtain *global convergence with respect to $\mathcal{V}_{ad}$* of the projected descent method, cf. Theorem 2.66.

2.14. Shape optimization with a PDE-constraint

So far we did not specify how the shape functional $\mathfrak{j}$ depends on the shape Ω. From now on we focus on situations where the shape functional involves the solution of some *state equation*. We develop the theory for a state equation in abstract Banach spaces, but usually we have in mind partial differential equations (PDEs). Shape optimization problems involving PDEs which model some physical phenomena are of great practical importance. Thus, there exists a rich literature, varying from thoroughly investigated mathematical model problems, to heuristically motivated engineering applications. The solution of the state equation, called the *state*, is a dependent variable of the domain. This is similar to the situation in optimal control theory, where the state depends on some control. In many situations it is advantageous to exploit this dependency and a lot of the concepts of optimal control have been transferred to shape optimization. However, additional difficulties occur, since the state space also depends on the domain.

A shape optimization problem with a state equation can be formulated abstractly as

$$\begin{aligned} &\text{Find } \Omega^* \in \mathcal{O} \text{ and } \tilde{y}^* \in \mathcal{Y}(\Omega^*) \text{ such that} \\ &\tilde{J}(\Omega^*, \tilde{y}^*) = \inf\left\{\tilde{J}(\Omega, \tilde{y}) \mid \Omega \in \mathcal{O},\ \tilde{y} \in \mathcal{Y}(\Omega),\ \text{and}\ \ \tilde{E}(\Omega, \tilde{y}) = 0\right\}, \end{aligned} \tag{2.41}$$

where $\mathcal{Y}(\Omega), \mathcal{Z}(\Omega)$ are Banach spaces for all $\Omega \in \mathcal{O}$, and

$$\begin{aligned} &\tilde{J}\colon \{(\Omega, \tilde{y}) \mid \Omega \in \mathcal{O},\ \tilde{y} \in \mathcal{Y}(\Omega)\} \to \mathbb{R}, \\ &\tilde{E}\colon \{(\Omega, \tilde{y}) \mid \Omega \in \mathcal{O},\ \tilde{y} \in \mathcal{Y}(\Omega)\} \to \{\tilde{z} \mid \Omega \in \mathcal{O},\ \tilde{z} \in \mathcal{Z}(\Omega)\}, \end{aligned}$$

with $\tilde{E}(\Omega, \tilde{y}) \in \mathcal{Z}(\Omega)\ \forall \Omega \in \mathcal{O}$. If the *state equation*

$$\text{find}\ \ \tilde{y} \in \mathcal{Y}(\Omega) \text{ such that } \tilde{E}(\Omega, \tilde{y}) = 0 \text{ in } \mathcal{Z}(\Omega),$$

admits a unique solution for every $\Omega \in \mathcal{O}$, we can introduce the *design-to-state operator*

$$\tilde{S}\colon \mathcal{O} \to \{\tilde{y} \mid \Omega \in \mathcal{O},\ \tilde{y} \in \mathcal{Y}(\Omega)\},\ \text{with } \tilde{E}(\Omega, \tilde{S}(\Omega)) := 0\ \forall \Omega \in \mathcal{O}.$$

In that case one can study the *reduced problem*, where the dependence on the state equation is hidden in the design-to-state operator, i.e., one can introduce the *reduced objective functional*

$$\mathfrak{j}\colon \mathcal{O} \to \mathbb{R}, \quad \mathfrak{j}(\Omega) := \tilde{J}(\Omega, \tilde{S}(\Omega)).$$

In order to analyze the reduced objective one needs to study the continuity and differentiability properties of $\tilde{S}$. Let us note, that the existence of a unique design-to-state mapping is not a necessary condition for the shape differentiability of $\tilde{J}$. In [DZ11, Section 10.5] a saddle-point formulation of the Lagrangian is used. It is combined with a result concerning the differentiability of a saddle-point with respect to a parameter [DZ11, Theorem 10.5.1], which goes back to Correa and Seeger. In [LS13, Theorem 2.4] a differentiability result without saddle-point assumptions is presented. Nevertheless, in this thesis we will focus on situations where a design-to-state mapping is available. In fact, we only considered state equations in which the state enters linearly. We believe that the reduced approach offers more advantages

for such equations. However, in the case of nonlinear state equations, the effort of evaluating the reduced objective increases dramatically, and a Lagrange-Newton method might be better suited. The connection to optimization problems on manifolds, or rather to optimization over vector bundles, may point the way to suitable algorithms. A first step in this direction is taken in [SSW14].

The differentiability of $\tilde{J}, \tilde{E}$ is delicate, as is already evident by the rather cumbersome notation. The domain, and in the case of $\tilde{E}$ also the range, of these operators consist of function spaces which depend again on the shape. In principal there are two methods available to circumvent this difficulty: the *function space embedding* approach, and the *function space parametrization* approach. In the function space embedding method the variable function spaces are extended from Ω to a larger fixed holdall domain $\mathcal{D}$, and the computations are carried out in the extended framework. The function space parametrization approach transports functions living on varying domains to some fixed reference domain, where they can be compared. We will focus on the latter approach, since it fits nicely into our framework of shape optimization. For an introduction to the function space embedding technique we refer to [DZ11] and the references therein.

For convenience of presentation we will focus on transformations given as perturbations of the identity τ. The idea of function space parametrization can also be combined with other concepts like the flow map T, cf., e.g., [SZ92, DZ11].

2.14.1. Function space parametrization

The idea of function space parametrization is to transport a function $\tilde{y} \in \mathcal{Y}(\Omega)$ defined on $\Omega = \tau(\Omega_0)$ to a function $y := \tilde{y} \circ \tau$, known as *pull-back* of $\tilde{y}$, which is defined on a fixed reference domain Ω_0. Let us formalize this argument. Recall the notation $\tau_U = \mathrm{Id} + U$.

Assumption 2.12. *The set $\Omega_0 \subset \mathbb{R}^d$ is nonempty, $\mathcal{O} = \mathcal{O}_\Theta(\Omega_0)$ for some suitable Banach space Θ, and $\mathcal{Y}(\Omega), \mathcal{Z}(\Omega)$ are Banach spaces $\forall \Omega \in \mathcal{O}$. There exists a $r > 0$, such that for all $U \in B^\Theta(0, r)$*

$$\mathcal{Y}(\Omega_0) = \{\tilde{y} \circ \tau_U \mid \tilde{y} \in \mathcal{Y}(\tau_U(\Omega_0))\}, \quad \textit{and} \quad \mathcal{Z}(\Omega_0) = \{z \circ \tau_U \mid \tilde{z} \in \mathcal{Z}(\tau_U(\Omega_0))\}.$$

Furthermore, the mappings

$$\mathcal{Y}(\tau_U(\Omega_0)) \ni \tilde{y} \mapsto y := \tilde{y} \circ \tau_U \in \mathcal{Y}(\Omega_0), \quad \textit{and} \quad \mathcal{Z}(\tau_U(\Omega_0)) \ni \tilde{z} \mapsto z := \tilde{z} \circ \tau_U \in \mathcal{Z}(\Omega_0),$$

are homeomorphisms.

Remark 2.102. From Lemma 2.9 it follows, that Assumption 2.12 is satisfied for the choice $\Theta = C^{0,1}(\overline{\mathbb{R}^d}, \mathbb{R}^d)$, $r = 1$, and the domain and range spaces usually encountered in the context of PDEs, i.e., $L^p, W^{1,p}, W_0^{1,p}$.

Definition 2.103. *Let Assumption 2.12 be satisfied. The* transformed state equation operator *is given by*

$$E\colon B^\Theta(0, r) \times \mathcal{Y}(\Omega_0) \to \mathcal{Z}(\Omega_0), \quad E(U, y) := \tilde{E}(\tau_U(\Omega_0), y \circ \tau_U^{-1}) \circ \tau_U.$$

The transformed objective functional *is given by*

$$J\colon B^{\Theta}(0,r)\times\mathcal{Y}(\Omega_0)\to\mathbb{R},\quad J(U,y):=\tilde{J}(\tau_U(\Omega_0),y\circ\tau_U^{-1}).$$

Corollary 2.104. *Let Assumption 2.12 be satisfied and assume that the design-to-state operator $\tilde{S}$ is well defined for all $\tau_U(\Omega_0)$, $U\in B^0(r,)$. Then the* transformed design-to-state operator

$$S\colon B^{\Theta}(0,r)\to\mathcal{Y}(\Omega_0),\quad S(U):=\tilde{S}(\tau_U(\Omega_0))\circ\tau_U,$$

satisfies

$$E(U,S(U))=0,\quad \forall U\in B^{\Theta}(0,r).$$

Proof. Clearly $E(U,S(U))=\tilde{E}(\tau_U(\Omega_0),\tilde{S}(\tau_U(\Omega_0))\circ\tau_U\circ\tau_U^{-1})\circ\tau_U=0\circ\tau_U=0$. □

We refer to Section 3.1, where the derivation of the transformed state equation operator and the transformed objective functional is described in detail for an concrete example. The main ingredient is usually the application of the transformation rule for integrals.

The importance of the transformed quantities J,E, and S is made clear by the following characterization.

Corollary 2.105. *Let Assumption 2.12 be satisfied, and suppose that $\tilde{S}$ is well defined on $B^{\mathcal{O}}(\Omega_0,r)=\{\tau_U(\Omega_0)\mid U\in B^{\Theta}(0,r)\}$. Recall from Definition 2.29 the functional*

$$\mathfrak{j}_{\Omega_0}\colon B^{\Theta}(0,\vartheta)\to\mathbb{R},\quad \mathfrak{j}_{\Omega_0}(U):=\mathfrak{j}(\tau_U(\Omega_0)).$$

Then, for all $U\in B^{\Theta}(0,\vartheta)$, it holds

$$\mathfrak{j}_{\Omega_0}(U)=J(U,S(U)).\tag{2.42}$$

Proof. We have $\mathfrak{j}_{\Omega_0}(U)=\mathfrak{j}(\tau_U(\Omega_0))=\tilde{J}(\tau_U(\Omega_0),\tilde{S}(\tau_U(\Omega_0)))=J(U,S(U))$. □

Thus *derivatives* of $\mathfrak{j}_{\Omega_0}$ can be obtained via J,E, and S. In general one is interested in differentiability of the design-to-state map $\tilde{S}$. Since $\tilde{S}$ is not given with respect to a fixed domain and range space, there exists no canonical choice for the derivative. As usual in shape optimization there are different differentiability concepts available. First we recall the notion of a material derivative, cf. [SZ92, Definition 2.71].

Definition 2.106. *Let Assumption 2.12 be satisfied and let $\tilde{F}\colon\mathcal{O}\to\{\tilde{y}\mid\tilde{y}\in\mathcal{Y}(\Omega)\}$ be some operator. The transformed operator $F(U):=\tilde{F}(\tau_U(\Omega_0))\circ\tau_U^{-1}$ maps $B^{\Theta}(0,r)$ to $\mathcal{Y}(\Omega_0)$. We say that $\tilde{F}$ has a strong (weak) material derivative in the direction $V\in\Theta$, if the limit*

$$\dot{\tilde{F}}(\Omega_0;V):=\lim_{t\searrow 0}\frac{1}{t}\left(F(tV)-F(0)\right)\in\mathcal{Y}(\Omega_0),$$

exists in the strong (weak) topology of the Banach space $\mathcal{Y}(\Omega_0)$.

Remark 2.107. We stayed here close to [SZ92, Definition 2.71]. Since F is an operator between the Banach spaces Θ and $\mathcal{Y}(\Omega_0)$ one can also define the Gâteaux derivative of F. We work in the following with the Gâteaux respectively the Fréchet derivative of the transformed design-to-state operator S.

It is quite common in shape optimization to consider also the *local shape derivative* of an operator $\tilde{F}$. We state it here for completeness. Let $\mathcal{Y}(\Omega_0)$ be an appropriate Sobolev space. If it exists, the local shape derivative of $\tilde{F}$ is defined as the element $\tilde{F}'(\Omega_0; V) \in \mathcal{Y}(\Omega_0)$, that satisfies

$$\tilde{F}'(\Omega_0; V) = \dot{\tilde{F}}(\Omega_0; V) - D\tilde{F}(\Omega_0)V.$$

It is motivated by the fact that, under appropriate assumptions,

$$\tilde{F}'(\Omega_0; V)(x) = \lim_{t \searrow 0} \frac{1}{t} \left(\tilde{F}(\tau_{tU}(\Omega_0)) - \tilde{F}(\Omega) \right)(x), \text{ for } x \in \operatorname{int}(\Omega_0).$$

The material derivative corresponds to a Lagrangian description of the deformation process, whereas the local shape derivative corresponds to an Eulerian description, i.e., a stationary observer.

2.14.2. Differentiating the reduced objective functional

Let us now turn to the subject of calculating derivatives of $\mathfrak{j}$ and $\mathfrak{j}_{\Omega_0}$. Our approach is based on the identity in Corollary 2.105. Usually, we are only interested in the derivatives of $\mathfrak{j}_{\Omega_0}$, however recalling the connections between the derivatives of $\mathfrak{j}$ and $\mathfrak{j}_{\Omega_0}$, we can use the characterization (2.42) to obtain also the derivatives of $\mathfrak{j}$ via the derivatives of J, E and S. These are defined in a standard Banach space framework on $\Theta, \mathcal{Y}(\Omega_0), \mathcal{Z}(\Omega_0)$, and the usual results concerning differentiability properties apply. For example, the *implicit function theorem* provides us with sufficient conditions for the existence and differentiability of S.

Assumption 2.13. *Assumption 2.12 is satisfied. Furthermore*

(i) $E\colon B^{\Theta}(0,r) \times \mathcal{Y}(\Omega_0) \to \mathcal{Z}(\Omega_0)$ *is continuously Fréchet differentiable.*

(ii) There exists a $\overline{y} \in \mathcal{Y}(\Omega_0)$ *such that* $E(0,\overline{y}) = 0$ *and* $E_y(0,\overline{y}) \in \mathcal{L}(\mathcal{Y}(\Omega_0), \mathcal{Z}(\Omega_0))$ *has a bounded inverse.*

Corollary 2.108. *Let Assumption 2.13 be satisfied. Then there exists an open neighborhood* $N_\Theta(0) \times N_{\mathcal{Y}(\Omega_0)}(\overline{y}) \subset B^{\Theta}(0,r) \times \mathcal{Y}(\Omega_0)$ *of* $(0,\overline{y})$*, and a unique operator* $S\colon N_\Theta(0) \to \mathcal{Y}(\Omega_0)$*, such that* $S(0) = \overline{y}$*. For all* $U \in N_\Theta(0)$ *there exists exactly one* $y \in N_{\mathcal{Y}(\Omega_0)}(\overline{y})$ *with* $E(U,y) = 0$*, namely* $y = S(U)$*. Moreover, the operator* $S\colon N_\Theta(0) \to \mathcal{Y}(\Omega_0)$ *is continuously Fréchet differentiable, with derivative*

$$S'(U) = -E_y(U, S(U))^{-1} E_U(U, S(U)).$$

If E *is* m*-times continuously Fréchet differentiable, then so is* S*.*

Proof. This is exactly the implicit function theorem, cf., e.g., [HPUU09, Theorem 1.41]. □

We verify for all the concrete examples considered in this thesis that Assumption 2.13 is satisfied. Note that $E_y(0,\overline{y}) = \tilde{E}_y(\Omega_0,\overline{y})$, hence continuous invertibility of $\tilde{E}_y(\Omega_0,\overline{y})$ implies the same for the transformed operator at $U = 0$. The chain rule yields now differentiability of the functional $\mathfrak{j}_{\Omega_0}$.

Corollary 2.109. *Let Assumption 2.13 be satisfied, and $J\colon B^{\Theta}(0,\vartheta) \times \mathcal{Y}(\Omega_0) \to \mathbb{R}$ be continuously Fréchet differentiable. Then $\mathfrak{j}_{\Omega_0}\colon N_\Theta(0) \to \mathbb{R}$ is continuously Fréchet differentiable. If E and J are m-times continuously Fréchet differentiable, then so is $\mathfrak{j}_{\Omega_0}$.*

Proof. This follows directly from the characterization $\mathfrak{j}_{\Omega_0}(U) = J(U, S(U))$. □

Remark 2.110. It is well known that the representation

$$\mathfrak{j}'_{\Omega_0}(0)(U) = J_U(U, S(U)) + J_y(U, S(U))S'(U),$$

is not suitable for efficient numerical implementations of optimization algorithms. Instead, an equivalent formulation is used which is derived via the *adjoint approach*. The same holds for the second derivative of $\mathfrak{j}_{\Omega_0}$. We refer to Section A.1 for a brief recapitulation of the adjoint approach.

3. Model problem

In this chapter we consider a model shape optimization problem and demonstrate some of the results and techniques of Chapter 2. The model problem is inspired by *potential flow pressure matching* in inverse aerodynamic design. The task is to find a geometry which matches a given desired pressure distribution. This is amenable to an approach via potential flow, because the pressure can be linked via Bernoulli's law to the velocity of the flow. The potential flow is a simplified model for a frictionless, irrotational, and incompressible flow, where the gradient of the potential corresponds to the velocity of the flow [CK08]. Although potential flow fails to describe many physical properties of real fluids there are still several applications where it provides a reasonable approximation of the behavior of the flow. Due to its comparatively low computational costs, it is still in use in early design stage simulations, or as an approximation in regions far away from boundary layers. Potential flow pressure matching was used to fit a known good flight characteristic to a new wing design. In practice, aerodynamic panel methods are employed to simulate the potential flow around a wing or airplane, cf. [KP91]. Problems with a potential flow state equation are often used as a test case in shape optimization, cf., e.g., [Pir82, Ang83, Pir84, But93, MP01, ESSI09].

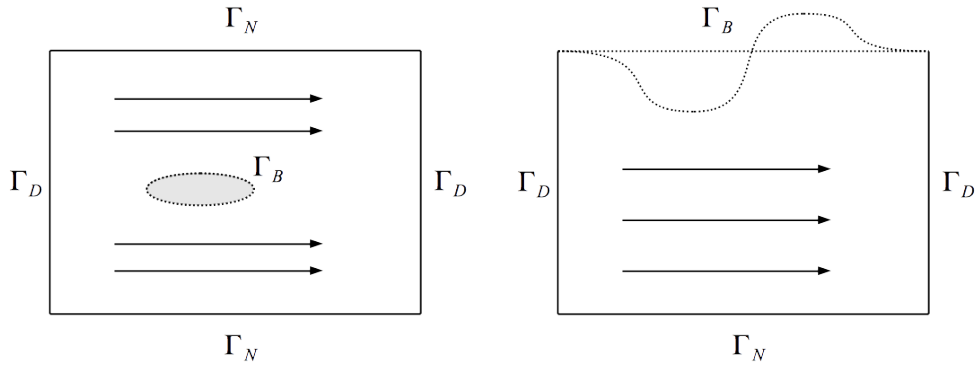
We consider the following shape optimization problem in two dimensions

$$\min_{\Omega\in\mathcal{O},\tilde{y}\in\mathcal{Y}(\Omega)} \frac{1}{2}\int_{\Gamma_B} (\nabla\tilde{y}^T t - p_d)^2 \,\mathrm{d}S$$

subject to the potential flow equation

$$\begin{aligned} -\Delta\tilde{y} &= 0 \quad \text{in } \Omega \\ \nabla\tilde{y}^T n &= 0 \quad \text{on } \Gamma_B \\ \nabla\tilde{y}^T n &= 0 \quad \text{on } \Gamma_N \\ \tilde{y} &= \overline{y}_0 \quad \text{on } \Gamma_D, \end{aligned}$$

where the boundary of the flow domain decomposes into $\partial\Omega = \Gamma_B \cup \Gamma_N \cup \Gamma_D$. The design boundary is denoted by Γ_B, Γ_N is some far away boundary, and Γ_D contains the inflow and outflow boundaries. Furthermore, n denotes the unit (outward) normal, and t the tangent vector oriented in flow direction. Finally, $\overline{y}_0$ induces some potential difference between in- and outflow and p_d encodes the desired pressure distribution. Figure 3.1 shows two possible typical configurations.



(a) A body immersed in a potential flow **(b)** Potential flow through a channel

Figure 3.1.: Two possible configurations in potential flow pressure matching

This chapter is structured as follows. We begin with specifying the problem setting in more detail, before showing how the function space parametrization approach can be employed to obtain shape derivatives of the reduced cost functional. To this end we characterize the functional j_Ω in Section 3.1. In Section 3.2 we introduce the necessary ingredients for deriving the shape derivatives in Section 3.3. At the end of the chapter we discuss a possible implementation of the model problem and present numerical experiments.

3.1. Function space parametrization

Recall the general shape optimization problem with state equation (2.41):

$$\begin{gathered}\text{Find } \Omega^* \in \mathcal{O} \text{ and } \tilde{y}^* \in \mathcal{Y}(\Omega^*) \text{ such that} \\ \tilde{J}(\Omega^*, \tilde{y}^*) = \inf\left\{\tilde{J}(\Omega, \tilde{y}) \mid \Omega \in \mathcal{O},\ \tilde{y} \in \mathcal{Y}(\Omega),\ \text{and}\ \ \tilde{E}(\Omega, \tilde{y}) = 0\right\}.\end{gathered}$$

We will embed the potential flow pressure matching problem in this framework and use then the *function space parametrization* approach, see Section 2.14. Let us specify the problem setting of this chapter in detail.

We consider a bounded, nonempty Lipschitz domain $\Omega_0 \subset \mathbb{R}^2$. Only a part of the boundary of Ω_0 is allowed to be deformed. It is denoted by Γ_B. The other two parts, Γ_D and Γ_N are fixed. Hence, given a disjoint decomposition $\partial\Omega_0 = \Gamma_B \cup \Gamma_N \cup \Gamma_D$, we set

$$\begin{aligned}\mathcal{O} &:= \{\tau(\Omega_0) \mid \tau \in \mathcal{F}(\mathcal{B}),\ \tau(\Gamma_D) = \Gamma_D,\ \tau(\Gamma_N) = \Gamma_N\} \\ &= \{\tau(\Omega_0) \mid \tau \in \mathcal{F}(\Theta)\},\end{aligned}$$

where $\mathcal{B} = C^1(\overline{\mathbb{R}^2}, \mathbb{R}^2)$, and $\Theta := \{U \in \mathcal{B} \mid U = 0 \text{ on } \Gamma_D \cup \Gamma_N\}$. For all $\Omega \in \mathcal{O}$ the design part of the boundary $\partial\Omega$ is given by $\tau(\Gamma_B) = \partial\Omega \backslash (\Gamma_D \cup \Gamma_N)$. For every $\Omega \in \mathcal{O}$ we define the space $H^1_D(\Omega)$ as the space of all functions $\tilde{y} \in H^1(\Omega)$ with trace $\mathrm{tr}(\tilde{y}) = 0$ on Γ_D. Note that, due to $\mathcal{B} \hookrightarrow C^1(\mathbb{R}^2, \mathbb{R}^2)$, the images of Ω_0 are again Lipschitz domains, so the trace is well defined.

We set $\mathcal{Y}(\Omega) = H^1_D(\Omega)$ and $\mathcal{Z}(\Omega) = \mathcal{Y}(\Omega)^*$. Given some smooth extension $\tilde{y}_0$ of the Dirichlet datum $\overline{y}_0$ onto Ω, we introduce

$$\langle \tilde{E}(\Omega, \tilde{y}), \tilde{\varphi} \rangle_{\mathcal{Y}(\Omega)^*, \mathcal{Y}(\Omega)} := (\nabla(\tilde{y} + \tilde{y}_0), \nabla\tilde{\varphi})_{L^2(\Omega)} \qquad \text{for } \tilde{y} \in \mathcal{Y}(\Omega), \tilde{\varphi} \in \mathcal{Y}(\Omega).$$

Hence, the state equation in variational form can be written as

$$\text{find } \tilde{y} \in \mathcal{Y}(\Omega) \text{ such that } \langle \tilde{E}(\Omega, \tilde{y}), \tilde{\varphi} \rangle_{\mathcal{Y}(\Omega)^*, \mathcal{Y}(\Omega)} = 0 \quad \forall \tilde{\varphi} \in \mathcal{Y}(\Omega).$$

Finally, we arrive at the setting of (2.41) by choosing

$$\tilde{J}(\Omega, \tilde{y}) := \frac{1}{2} \int_{\partial\Omega \setminus (\Gamma_D \cup \Gamma_N)} \left(t^T \nabla(\tilde{y} + \tilde{y}_0) - p_d \right)^2 \, \mathrm{d}S, \tag{3.1}$$

where $p_d \in L^2(\Gamma_B)$ encodes the desired pressure distribution which is transformation invariant, i.e., $p_d(\tau(x)) = p_d(x)$ for all $x \in \Gamma_B$. In fact, it is reasonable that the desired pressure distribution does *not* depend on the shape of the design boundary. In the concrete example later in this chapter this is achieved by choosing a function p_d which depends only on the x_1-coordinate, and is constant in x_2-direction.

For every $\Omega \in \mathcal{O}$ the state equation admits a unique solution, hence we can define a design-to-state operator $\tilde{S}$ such that $\tilde{E}(\Omega, \tilde{S}(\Omega)) = 0$. The reduced objective is given by

$$\mathfrak{j}(\Omega) := \tilde{J}(\Omega, \tilde{S}(\Omega)).$$

Let us now demonstrate how one can compute the derivatives of the shape functional $\mathfrak{j}$ at some $\Omega_{ref} \in \mathcal{O}$. As proposed in Section 2.14, we will do this via the localized functional $\mathfrak{j}_{\Omega_{ref}}$. Once we have the derivatives of $\mathfrak{j}_{\Omega_{ref}}$ available, we can use the results from Theorems 2.31 and 2.39 to compute the shape derivatives of $\mathfrak{j}$. However, as discussed in Section 2.8, 2.9 and 2.11, from an algorithmic point of view we are mainly interested in the derivatives of $\mathfrak{j}_{\Omega_{ref}}$. Recall from Corollary 2.105 that there exists an $r > 0$ such that

$$\mathfrak{j}_{\Omega_{ref}}(U) = J(U, S(U)) \quad \text{for all } \; U \in B^{\Theta}(0, r),$$

if the *function space parametrization* approach is applicable. In fact, due to Lemma 2.9, the space $\mathcal{Y} := \mathcal{Y}(\Omega_{ref})$ is equal to $\{\tilde{y} \circ \tau \mid \tilde{y} \in \mathcal{Y}(\tau(\Omega_{ref}))\}$, and the mapping

$$\mathcal{Y}(\tau(\Omega_{ref})) \ni \tilde{y} \mapsto y := \tilde{y} \circ \tau \in \mathcal{Y}$$

is a homeomorphism for all $\tau \in \mathcal{F}(\Theta)$. We now transform $\tilde{E}$ and $\tilde{J}$ to Ω_{ref} as described in Definition 2.103. Let us circumstantiate this process in detail for the state equation. The transformed obejctive can be obtained with an analogous procedure. We employ the transformation rule for integrals, i.e.,

$$\int_{\tau(\Omega_{ref})} \tilde{f}(\tilde{x}) \, \mathrm{d}\tilde{x} = \int_{\Omega_{ref}} \tilde{f}(\tau(x)) |\det(D\tau(x))| \, \mathrm{d}x.$$

Since we only work with transformations close to the identity, i.e., $\|\tau - \mathrm{Id}\|_{\Theta} < 1$, it holds $\det(D\tau) > 0$, and we can drop the absolute value in the following. Furthermore, we deduce from $y = \tilde{y} \circ \tau$, that

$$y'(x) = \tilde{y}'(\tau(x)) D\tau(x) \;\Rightarrow\; \nabla\tilde{y}(\tilde{x}) = D\tau^{-T}(\tau^{-1}(\tilde{x})) \nabla y(\tau^{-1}(\tilde{x})),$$

and similarly for $\varphi = \tilde{\varphi} \circ \tau$ and $y_0 = \tilde{y}_0 \circ \tau$. We conclude

$$\begin{aligned}
&(\nabla(\tilde{y} + \tilde{y}_0), \nabla\tilde{\varphi})_{L^2(\tau(\Omega_{ref}))} \\
&= \Big((D\tau^{-T} \circ \tau^{-1})\nabla(y \circ \tau^{-1} + y_0 \circ \tau^{-1}), (D\tau^{-T} \circ \tau^{-1})\nabla(\varphi \circ \tau^{-1})\Big)_{L^2(\tau(\Omega_{ref}))} \\
&= \Big(D\tau^{-T}\nabla(y + y_0), D\tau^{-T}\nabla\varphi \det(D\tau)\Big)_{L^2(\Omega_{ref})}.
\end{aligned}$$

Hence, we study the transformed state equation operator

$$E\colon B^{\Theta}(0,1) \times \mathcal{Y} \to \mathcal{Y}^*, \quad \langle E(U,y), \varphi\rangle_{\mathcal{Y}^*,\mathcal{Y}} = \Big(\det(D\tau_U) D\tau_U^{-1} D\tau_U^{-T} \nabla(y + y_0), \nabla\varphi\Big)_{L^2(\Omega_{ref})}.$$

For $\tilde{J}$ we use the transformation rule for boundary integrals, and note that, if t is the tangent vector to Γ_B, then the tangent vector to $\tau(\Gamma_B)$ is given by $|D\tau t|^{-1} D\tau t$. We obtain the transformed objective functional

$$J\colon B^{\Theta}(0,1) \times \mathcal{Y} \to \mathbb{R}, \quad J(U,y) = \frac{1}{2}\int_{\Gamma_B} \left(\frac{t^T \nabla(y + y_0)}{|D\tau_U t|} - p_d\right)^2 |D\tau_U t| \, \mathrm{d}S,$$

where Γ_B now denotes the design boundary of Ω_{ref}.

From the results of Section 2.14 we conclude that the transformed design-to-state operator $S(U) = \tilde{S}(\tau_U(\Omega_{ref})) \circ \tau_U$ satisfies $E(U, S(U)) = 0$, and that

$$\mathfrak{j}_{\Omega_{ref}}(U) = J(U, S(U)).$$

With the help of the implicit function theorem we can now derive a formula for the derivatives of $\mathfrak{j}_{\Omega_{ref}}$. For this we need the partial derivatives of E and J.

3.2. Partial derivatives

Recall the differentiation rules of Lemma 2.84, which we restate here for the convenience of the reader.

Lemma 3.1. *(i) The mapping*

$$W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d) \to L^\infty(\mathbb{R}^d)\colon \quad U \mapsto \det(D(\mathrm{Id} + U)) = \det(D\tau_U)$$

is differentiable and the derivative in direction $V \in W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)$ is given by

$$\mathrm{tr}(D\tau_U^{-1} DV) \det(D\tau_U).$$

(ii) The mapping

$$W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d) \to L^\infty(\mathbb{R}^d, \mathbb{R}^{d\times d})\colon \quad U \mapsto D(\mathrm{Id} + U)^{-1} = D\tau_U^{-1}$$

is differentiable and the derivative in direction $V \in W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)$ is given by

$$-D\tau_U^{-1} DV D\tau_U^{-1}.$$

By the above lemma the derivative of the mapping

$$U \mapsto A(U) := D\tau_U^{-1} D\tau_U^{-T} \det(D\tau_U) \tag{3.2}$$

in the direction V is given by

$$M^V(U) := D\tau_U^{-1} \left(-DV D\tau_U^{-1} - D\tau_U^{-T} DV^T + \mathcal{I} \operatorname{tr}(D\tau_U^{-1} DV) \right) D\tau_U^{-T} \det(D\tau_U), \tag{3.3}$$

where $\mathcal{I}$ denotes the identity matrix in $\mathbb{R}^{d\times d}$. For a clear presentation, we will employ in the following formulas the short notation

$$\varphi^* E(U,y) := \langle E(U,y), \varphi \rangle_{\mathcal{Y}^*,\mathcal{Y}}.$$

Corollary 3.2. *Consider the operator*

$$E\colon B^\Theta(0,1) \times \mathcal{Y} \to \mathcal{Y}^*, \quad \langle E(U,y), \varphi \rangle_{\mathcal{Y}^*,\mathcal{Y}} = \varphi^* E(U,y) = \left(A(U)\nabla(y+y_0), \nabla\varphi \right)_{L^2(\Omega_{ref})}.$$

The operator E is twice continuously Fréchet differentiable. The partial derivatives in the directions $z \in \mathcal{Y}$, respectively $V \in \Theta$, are given by

$$\begin{aligned} \varphi^* E_y(U,y)z &= \left(A(U)\nabla z, \nabla\varphi \right)_{L^2(\Omega_{ref})}, \\ \varphi^* E_U(U,y)V &= \left(M^V(U)\nabla(y+y_0), \nabla\varphi \right)_{L^2(\Omega_{ref})}. \end{aligned}$$

The partial second derivatives in the directions $z \in \mathcal{Y}$ and $V, W \in \Theta$ are given by

$$\begin{aligned} \varphi^* E_{yy}(U,y) &= 0, \\ \varphi^* E_{Uy}(U,y)(V,z) &= \left(M^V(U)\nabla z, \nabla\varphi \right)_{L^2(\Omega_{ref})}, \\ \varphi^* E_{yU}(U,y)(z,V) &= \varphi^* E_{Uy}(U,y)(V,z), \end{aligned}$$

and

$$\begin{aligned} \varphi^* E_{UU}(U,y)(V,W) = &- \left(D\tau_U^{-1} DW M^V(U)\nabla(y+y_0), \nabla\varphi \right)_{L^2(\Omega_{ref})} \\ &- \left(M^V(U) DW^T D\tau_U^{-T} \nabla(y+y_0), \nabla\varphi \right)_{L^2(\Omega_{ref})} \\ &+ \left(M^V(U) \operatorname{tr}(D\tau_U^{-1}\ DW)\nabla(y+y_0), \nabla\varphi \right)_{L^2(\Omega_{ref})} \\ &+ \left(D\tau_U^{-1} DV D\tau_U^{-1}\ DW A(U)\nabla(y+y_0), \nabla\varphi \right)_{L^2(\Omega_{ref})} \\ &+ \left(A(U) DW^T D\tau_U^{-T} DV^T D\tau_U^{-T} \nabla(y+y_0), \nabla\varphi \right)_{L^2(\Omega_{ref})} \\ &- \left(A(U) \operatorname{tr}(D\tau_U^{-1} DW D\tau_U^{-1} DV)\nabla(y+y_0), \nabla\varphi \right)_{L^2(\Omega_{ref})}. \end{aligned}$$

Proof. With Lemma 3.1 the differentiability of E is clear, and the first derivatives can be calculated in a straightforward manner. Since these are again composed of differentiable operators the second derivatives can be obtained by the chain rule. □

Usually we are only interested in the derivatives at $U = 0$. It holds

$$\begin{aligned} A(0) &= \mathcal{I}, \text{ and} \\ M^V(0) &= (\mathcal{I}\operatorname{div}(V) - DV - DV^T), \end{aligned}$$

hence the above expressions simplify to

$$\begin{aligned} \varphi^* E_y(0,y)z &= (\nabla z, \nabla\varphi)_{L^2(\Omega_{ref})} \\ \varphi^* E_U(0,y)V &= \Big((\mathcal{I}\operatorname{div}(V) - DV - DV^T)\nabla(y+y_0), \nabla\varphi\Big)_{L^2(\Omega_{ref})} \\ \varphi^* E_{Uy}(0,y)(V,z) &= \Big((\mathcal{I}\operatorname{div}(V) - DV - DV^T)\nabla z, \nabla\varphi\Big)_{L^2(\Omega_{ref})}, \end{aligned}$$

and

$$\begin{aligned} \varphi^* E_{UU}(0,y)(V,W) = &- \Big(DW(\mathcal{I}\operatorname{div}(V) - DV - DV^T)\nabla(y+y_0), \nabla\varphi\Big)_{L^2(\Omega_{ref})} \\ &- \Big((\mathcal{I}\operatorname{div}(V) - DV - DV^T)DW^T\nabla(y+y_0), \nabla\varphi\Big)_{L^2(\Omega_{ref})} \\ &+ \Big((\mathcal{I}\operatorname{div}(V) - DV - DV^T)\operatorname{div}(W)\nabla(y+y_0), \nabla\varphi\Big)_{L^2(\Omega_{ref})} \\ &+ (DVDW\nabla(y+y_0), \nabla\varphi)_{L^2(\Omega_{ref})} \\ &+ \Big(DW^T DV^T\nabla(y+y_0), \nabla\varphi\Big)_{L^2(\Omega_{ref})} \\ &- (\operatorname{tr}(DWDV)\nabla(y+y_0), \nabla\varphi)_{L^2(\Omega_{ref})}. \end{aligned}$$

Rewriting the objective as

$$\begin{aligned} J(U,y) &= \frac{1}{2}\int_{\Gamma_B}\left(\frac{t^T\nabla(y+y_0)}{|D\tau_U t|} - p_d\right)^2 |D\tau_U t|\,\mathrm{d}S \\ &= \frac{1}{2}\int_{\Gamma_B}\Big(t^T\nabla(y+y_0)\Big)^2 (t^T D\tau_U^T D\tau_U t)^{-1/2}\,\mathrm{d}S \\ &\quad - \frac{1}{2}\int_{\Gamma_B} 2t^T\nabla(y+y_0)p_d\,\mathrm{d}S \\ &\quad + \frac{1}{2}\int_{\Gamma_B} p_d^2 (t^T D\tau_U^T D\tau_U t)^{1/2}\,\mathrm{d}S, \end{aligned}$$

the following result is obtained by straightforward calculations.

Corollary 3.3. *Consider the functional*

$$J\colon B^{\Theta}(0,1)\times\mathcal{Y}\to\mathbb{R}, \quad J(U,y) = \frac{1}{2}\int_{\Gamma_B}\left(\frac{t^T\nabla(y+y_0)}{|D\tau_U t|} - p_d\right)^2 |D\tau_U t|\,\mathrm{d}S.$$

The functional J is twice continuously Fréchet differentiable. The partial derivatives in the directions $z \in \mathcal{Y}$, respectively $V \in \Theta$, are given by

$$\begin{aligned}
\langle J_y(U,y),z\rangle_{\mathcal{Y}^*,\mathcal{Y}} &= \int_{\Gamma_B} \left(t^T\nabla(y+y_0)\right)\left(t^T\nabla z\right)(t^T D\tau_U^T D\tau_U t)^{-1/2} - \left(t^T\nabla z\right) p_d \,\mathrm{d}S,\\
\langle J_U(U,y),V\rangle_{\Theta^*,\Theta} &= -\frac{1}{2}\int_{\Gamma_B}\left(t^T\nabla(y+y_0)\right)^2 (t^T D\tau_U^T D\tau_U t)^{-3/2}(t^T D\tau_U^T DV t)\,\mathrm{d}S\\
&\quad + \frac{1}{2}\int_{\Gamma_B} p_d^2 (t^T D\tau_U^T D\tau_U t)^{-1/2}(t^T D\tau_U^T DV t)\,\mathrm{d}S.
\end{aligned}$$

The partial second derivatives in the directions $z,\varphi \in \mathcal{Y}$ and $V,W \in \Theta$ are given by

$$\begin{aligned}
\langle J_{yy}(U,y)z,\varphi\rangle_{\mathcal{Y}^*,\mathcal{Y}} &= \int_{\Gamma_B}\left(t^T\nabla\varphi\right)\left(t^T\nabla z\right)(t^T D\tau_U^T D\tau_U t)^{-1/2}\,\mathrm{d}S,\\
\langle J_{Uy}(U,y)U,z\rangle_{\mathcal{Y}^*,\mathcal{Y}} &= -\int_{\Gamma_B}\left(t^T\nabla(y+y_0)\right)\left(t^T\nabla z\right)(t^T D\tau_U^T D\tau_U t)^{-3/2}(t^T D\tau_U^T DV t)\,\mathrm{d}S,\\
\langle J_{yU}(U,y)z,V\rangle_{\Theta^*,\Theta} &= \langle J_{Uy}(U,y)U,z\rangle_{\mathcal{Y}^*,\mathcal{Y}},
\end{aligned}$$

and

$$\begin{aligned}
&\langle J_{UU}(U,y)V,W\rangle_{\Theta^*,\Theta} =\\
&+\frac{3}{2}\int_{\Gamma_B}\left(t^T\nabla(y+y_0)\right)^2 (t^T D\tau_U^T D\tau_U t)^{-5/2}(t^T D\tau_U^T DV t)(t^T D\tau_U^T DW t)\,\mathrm{d}S\\
&-\frac{1}{2}\int_{\Gamma_B}\left(t^T\nabla(y+y_0)\right)^2 (t^T D\tau_U^T D\tau_U t)^{-3/2}(t^T DW^T DV t)\,\mathrm{d}S\\
&-\frac{1}{2}\int_{\Gamma_B} p_d^2 (t^T D\tau_U^T D\tau_U t)^{-3/2}(t^T D\tau_U^T DV t)(t^T D\tau_U^T DW t)\,\mathrm{d}S\\
&+\frac{1}{2}\int_{\Gamma_B} p_d^2 (t^T D\tau_U^T D\tau_U t)^{-1/2}(t^T DW^T DV t)\,\mathrm{d}S.
\end{aligned}$$

Proof. Differentiability follows from Lemma 3.1. The stated formulas may be verified by straightforward calculations. □

These expressions also simplify at $U = 0$. It holds

$$\begin{aligned}
\langle J_y(0,y),z\rangle_{\mathcal{Y}^*,\mathcal{Y}} &= \int_{\Gamma_B}\left(\left(t^T\nabla(y+y_0)\right) - p_d\right)\left(t^T\nabla z\right)\,\mathrm{d}S,\\
\langle J_U(0,y),V\rangle_{\Theta^*,\Theta} &= \frac{1}{2}\int_{\Gamma_B}\left(p_d^2 - \left(t^T\nabla(y+y_0)\right)^2\right)(t^T DV t)\,\mathrm{d}S,\\
\langle J_{yy}(0,y)z,\varphi\rangle_{\mathcal{Y}^*,\mathcal{Y}} &= \int_{\Gamma_B}\left(t^T\nabla\varphi\right)\left(t^T\nabla z\right)\,\mathrm{d}S,\\
\langle J_{Uy}(0,y)U,z\rangle_{\mathcal{Y}^*,\mathcal{Y}} &= \int_{\Gamma_B}\left(t^T\nabla(y+y_0)\right)\left(t^T\nabla z\right)(t^T DV t)\,\mathrm{d}S,
\end{aligned}$$

and

$$\begin{aligned}\langle J_{UU}(0,y)V,W\rangle_{\Theta^*,\Theta} = &+\frac{3}{2}\int_{\Gamma_B}\left(t^T\nabla(y+y_0)\right)^2(t^TDVt)(t^TDWt)\,\mathrm{d}S\\ &-\frac{1}{2}\int_{\Gamma_B}\left(t^T\nabla(y+y_0)\right)^2(t^TDW^TDVt)\,\mathrm{d}S\\ &-\frac{1}{2}\int_{\Gamma_B}p_d^2(t^TDVt)(t^TDWt)\,\mathrm{d}S\\ &+\frac{1}{2}\int_{\Gamma_B}p_d^2(t^TDW^TDVt)\,\mathrm{d}S.\end{aligned}$$

3.3. Shape derivatives

Recall that we would like to employ the implicit function theorem to show differentiability of $S(U)$ and hence of

$$\mathfrak{j}_{\Omega_{ref}}(U) = J(U,S(U)).$$

The implicit function theorem requires $E_y(U,y) \in \mathcal{L}(\mathcal{Y},\mathcal{Y}^*)$ to be continuously invertible, cf. Section 2.14.2. We verify this now with the help of Lax-Milgram.

Lemma 3.4. *For every $U \in B^\Theta(0,1)$ the linear operator $E_y(U,y) \in \mathcal{L}(\mathcal{Y},\mathcal{Z})$ is continuously invertible.*

Proof. Let $U \in B^\Theta(0,1)$ be arbitrary but fixed. Recall $\mathcal{Z} = \mathcal{Y}^*$. Define the bilinear form

$$b(\cdot,\cdot)\colon \mathcal{Y}\times\mathcal{Y}\to\mathbb{R},\quad b(z,\varphi) := \varphi^*E_y(U,y)z = \left(A(U)\nabla z,\nabla\varphi\right)_{L^2(\Omega_{ref})}.$$

Since $\mathcal{B} = C^1(\mathbb{R}^2,\mathbb{R}^2)$ this is a bounded and coercive bilinear form. Hence, for any $f \in \mathcal{Y}^*$ Lax-Milgram provides us with a unique solution $z \in \mathcal{Y}$ of the variational equation

$$b(z,\varphi) = \langle f,\varphi\rangle_{\mathcal{Y}^*,\mathcal{Y}}\quad \forall\varphi\in\mathcal{Y}.$$

Thus, the linear map $f \mapsto z$ is the inverse of $E_y(U,y)$. Boundedness of the inverse operator is again provided by Lax-Milgram. □

Hence, all prerequisites of the implicit function theorem are satisfied and we obtain the announced result.

Corollary 3.5. *The functional*

$$\mathfrak{j}_{\Omega_{ref}}\colon B^\Theta(0,1)\to\mathbb{R},\quad \mathfrak{j}_{\Omega_{ref}}(U) = J(U,S(U))$$

is twice continuously Fréchet differentiable.

Proof. Combining Corollaries 3.2 and 3.3 with Lemma 3.4, we see that all the conditions of Corollary 2.109 are satisfied which provides the claimed result. □

We conclude this section by showing how the derivatives of $\mathfrak{j}_{\Omega_{ref}}$ can be computed. For this we choose the *sensitivity approach* where the derivatives of S are used explicitly. Note that the sensitivity approach is *not* recommended for practical implementations. Instead, the equivalent representation of the derivatives via the *adjoint approach* is usually to be preferred. The adjoint approach is briefly recapitulated in Section A.1, we refer to [HPUU09, Section 1.6] for a more detailed discussion of the two approaches and their merits. We present here the sensitivity approach since we will use it in Chapter 4 to derive the *symbol of the Hessian* for our model problem.

By the chain rule the derivatives of $\mathfrak{j}_{\Omega_{ref}}$ in the directions $V, W \in \Theta$ are given by

$$\begin{aligned}
\langle \mathfrak{j}'_{\Omega_{ref}}(U), V\rangle_{\Theta^*,\Theta} &= \langle J_U(U, S(U)), V\rangle_{\Theta^*,\Theta} + \langle J_y(U, S(U)), S'(U)V\rangle_{\mathcal{Y}^*,\mathcal{Y}},\\
\langle \mathfrak{j}''_{\Omega_{ref}}(U)V, W\rangle_{\Theta^*,\Theta} &= \langle J_{UU}(U, S(U))V, W\rangle_{\Theta^*,\Theta}\\
&\quad + \langle J_{Uy}(U, S(U))V, S'(U)W\rangle_{\mathcal{Y}^*,\mathcal{Y}}\\
&\quad + \langle J_{yU}(U, S(U))S'(U)V, W\rangle_{\Theta^*,\Theta}\\
&\quad + \langle J_{yy}(U, S(U))S'(U)V, S'(U)W\rangle_{\mathcal{Y}^*,\mathcal{Y}}\\
&\quad + \langle J_y(U, S(U)), S''(U)(V, W)\rangle_{\mathcal{Y}^*,\mathcal{Y}}.
\end{aligned}$$

The derivatives of S are given by the implicit function theorem, cf. Corollary 2.108. Let us demonstrate this for the choice $U = 0$.

Definition 3.6. *For every $V \in \Theta$ we denote by $z_V \in \mathcal{Y}$ the solution of*

$$\begin{aligned}
\langle E_y(0, S(0))z, \varphi\rangle_{\mathcal{Y}^*,\mathcal{Y}} &= -\langle E_U(0, S(0))V, \varphi\rangle_{\mathcal{Y}^*,\mathcal{Y}} && \forall \varphi \in \mathcal{Y}\\
\textit{i.e.,}\ (\nabla z, \nabla\varphi)_{L^2(\Omega_{ref})} &= -\Big((\mathcal{I}\operatorname{div}(V) - DV - DV^T)\nabla(S(0) + y_0), \nabla\varphi\Big)_{L^2(\Omega_{ref})} && \forall \varphi \in \mathcal{Y}.
\end{aligned} \tag{3.4}$$

This equation is known as the *linearized state equation* and it holds

$$z_V = S'(0)V.$$

Definition 3.7. *For all $V, W \in \Theta$ we denote by $\mu_{VW} \in \mathcal{Y}$ the solution of*

$$\begin{aligned}
\langle E_y(0, S(0))\mu, \varphi\rangle_{\mathcal{Y}^*,\mathcal{Y}} = \quad & -\langle E_{UU}(0, S(0))(V, W), \varphi\rangle_{\mathcal{Y}^*,\mathcal{Y}}\\
& -\langle E_{Uy}(0, S(0))(V, z_W), \varphi\rangle_{\mathcal{Y}^*,\mathcal{Y}}\\
& -\langle E_{yU}(0, S(0))(z_V, W), \varphi\rangle_{\mathcal{Y}^*,\mathcal{Y}}
\end{aligned} \tag{3.5}$$

for all $\varphi \in \mathcal{Y}$. Inserting the concrete formulas yields

$$\begin{aligned}(\nabla\mu, \nabla\varphi)_{L^2(\Omega_{ref})} = &+ \Big(DW(\mathcal{I}\operatorname{div}(V) - DV - DV^T)\nabla(S(0)+y_0), \nabla\varphi\Big)_{L^2(\Omega_{ref})} \\ &+ \Big((\mathcal{I}\operatorname{div}(V) - DV - DV^T)DW^T\nabla(S(0)+y_0), \nabla\varphi\Big)_{L^2(\Omega_{ref})} \\ &- \Big(M^V(U)\operatorname{div}(W)\nabla(S(0)+y_0), \nabla\varphi\Big)_{L^2(\Omega_{ref})} \\ &- (DVDW\nabla(S(0)+y_0), \nabla\varphi)_{L^2(\Omega_{ref})} \\ &- \Big(DW^TDV^T\nabla(S(0)+y_0), \nabla\varphi\Big)_{L^2(\Omega_{ref})} \\ &+ (\operatorname{tr}(DWDV)\nabla(S(0)+y_0), \nabla\varphi)_{L^2(\Omega_{ref})} \\ &- \Big((\mathcal{I}\operatorname{div}(V) - DV - DV^T)\nabla z_W, \nabla\varphi\Big)_{L^2(\Omega_{ref})} \\ &- \Big((\mathcal{I}\operatorname{div}(W) - DW - DW^T)\nabla z_V, \nabla\varphi\Big)_{L^2(\Omega_{ref})}\end{aligned}$$

for all $\varphi \in \mathcal{Y}$.

The solution satisfies

$$\mu_{VW} = S''(0)(V,W).$$

Hence, if we want to evaluate the derivatives of $\mathrm{j}_{\Omega_{ref}}(0)$ in directions $V, W \in \Theta$, we can do this by solving (3.4), (3.5) for z_V, z_W, and μ_{VW} and inserting these into the formulas above. For example, it holds

$$\begin{aligned}\langle \mathrm{j}'(\Omega_{ref}), V\rangle_{\Theta^*,\Theta} = \langle \mathrm{j}'_{\Omega_{ref}}(0), V\rangle_{\Theta^*,\Theta} = &\frac{1}{2}\int_{\Gamma_B}\left(p_d^2 - \left(t^T\nabla(S(0)+y_0)\right)^2\right)(t^TDVt)\,\mathrm{d}S \\ &+ \int_{\Gamma_B}\left(\left(t^T\nabla(S(0)+y_0)\right) - p_d\right)\left(t^T\nabla z_V\right)\,\mathrm{d}S.\end{aligned}$$

Remark 3.8. (i) Note that for $U = 0$ the linearized state equation, as well as the equation for the second derivative of S, are again Poisson equations, and can be solved with standard methods. This is a general property of the function space parametrization approach. For every operator $\tilde{E}(\Omega, \tilde{y})$, the linearized state equation, as well as the adjoint equation, and the equations characterizing higher order derivatives at $U = 0$, correspond to the standard linearized state equation, adjoint equation, etc., but with a nonstandard right-hand side. This has the advantage that one can use established and efficient solution methods to solve these equations. See [BLUU09, BLUU11] for an application to shape optimization with the instationary Navier-Stokes equation.

(ii) It is also possible to calculate $S'(U)$ and $S''(U)$ if $U \neq 0$, but for this one has to develop a specialized solution method. We refer to [KV13] where this is done for an elliptic equation. For more complex state equations, e.g., the Navier-Stokes equation, this approach becomes very tedious. Fortunately it is *not necessary to do this*. In the

framework of Section 2.8 and 2.9, one *only* needs to evaluate $\mathfrak{j}_{\Omega_k}$ and its derivatives *at* 0 in every iteration of the proposed optimization methods. Even if one is interested in these quantities at $U \neq 0$, they can be conveniently obtained by working on $\tau_U(\Omega_k)$ and using the relations provided by Theorems 2.31 and 2.39, respectively by Lemma 2.74. In the next section we employ the framework of Section 2.11 and 2.12, and describe this technique in more detail.

3.4. Numerical examples

Let us conclude this chapter with some numerical experiments. We consider the situation in Figure 3.1(b), i.e., the initial domain Ω_0 is a rectangle and we are allowed to modify the upper boundary denoted by Γ_B. Furthermore, the height of the boundaries on the left and right side can also vary. In this setting it seems reasonable to parametrize the domains via the vertical displacement of the upper boundary. It is convenient to work here with a fixed reference domain $\Omega_{ref} = \Omega_0$ and vertical boundary displacements u in a suitable Hilbert space $\mathcal{U}$ of functions defined on Γ_B. This is exactly the framework of Section 2.11. Note that in this simple example one could easily extend the boundary displacements linearly to the whole domain. However, we want to demonstrate here a setting which works in more general situations, and extend the boundary displacement via linear elasticity to a corresponding domain displacement and an associated transformation of Ω_0. Recall that Theorem 2.87 provides the existence of a suitable extension operator from the space of boundary displacements $\mathcal{U} := H^2(\Gamma_B)$ to Θ. The setting of Figure 3.1(a) is also covered if the immersed body, i.e., the boundary Γ_B, is smooth, since then the only interior angles are the ones in the corner of the rectangle.

In Section 2.12 we described a monotone linesearch descent method (Algorithm 2.5) for the solution of

$$\min_{u\in\mathcal{U}} j(u), \text{ where } \quad j\colon \mathcal{U} \to \mathbb{R}\cup\infty, \quad j(u) = \begin{cases} \mathfrak{j}_{\Omega_{ref}}(Tu) & \text{if } u \in \mathcal{U}_{feas}, \\ \infty & \text{else.} \end{cases}$$

The crucial aspect is of course the selection of the descent direction v_k. One has different possibilities here. We choose either the classical Newton direction (Algorithm 2.6), or a Newton-type direction (Algorithm 2.7).

A special feature of our method is that we compute the state, the reduced objective $\mathfrak{j}$, and its derivatives on the current, physical domain Ω_k. With the help of the formulas in Theorem 2.31 and Lemma 2.74 the derivatives are then transported back to Ω_0, and used to determine the derivatives of j. We have detailed in Algorithms 3.1 and 3.2 how to evaluate $j(u)$ and $j'(u)$ by combining Section A.1 and A.2. Similarly, one can evaluate $j''(u)v$. Note that each application of T and T^* corresponds to a solve of the linear elasticity equation. The specific formulas for J, E, and their derivatives are given in Corollaries 3.2 and 3.3.

We implemented the proposed method in `MATLAB` [TM15]. The discretization uses piecewise linear finite elements to approximate $\mathcal{Y}$, as well as Θ and $\mathcal{U}$. In this setting optimization and discretization commute, cf. [BLUU09]. We obtain, up to computational accuracy, exact discrete derivatives by employing the continuous adjoint approach. In particular, we can check

Algorithm 3.1: Evaluating $j(u)$

Require: $u \in \mathcal{U}_{feas}$
1: set $U = Tu \in \Theta_0$
2: set $\Omega = \tau_U(\Omega_0)$
3: solve the state equation $\tilde{E}(\Omega, \tilde{y}) = 0$ to obtain $\tilde{y} \in \mathcal{Y}(\Omega)$
4: **return** $j(u) = \tilde{J}(\Omega, \tilde{y})$

Algorithm 3.2: Evaluating $j'(u)$

Require: $u \in \mathcal{U}_{feas}$
1: set $U = Tu \in \Theta$
2: set $\Omega = \tau_U(\Omega_0)$
3: solve the state equation $\tilde{E}(\Omega, \tilde{y}) = 0$ to obtain $\tilde{y} \in \mathcal{Y}(\Omega)$
4: solve the adjont equation $\tilde{p}^* \tilde{E}_{\tilde{y}}(\Omega, \tilde{y}) = 0$ to obtain $\tilde{p} \in \mathcal{Z}(\Omega)$
5: evaluate $\mathrm{j}'_\Omega(0) = J_U(0, \tilde{y}) + \tilde{p}^* E_U(0, \tilde{y}) \in \Theta(\Omega)^*$
6: transport the derivative to obtain $\mathrm{j}'_{\Omega_0}(U)$ as in Theorem 2.31, i.e.,

$$\langle \mathrm{j}'_{\Omega_0}(U), V \rangle_{\Theta_0^*, \Theta_0} = \langle \mathrm{j}'_\Omega(0), V \circ \tau_U^{-1} \rangle_{\Theta(\Omega)^*, \Theta(\Omega)}$$

7: **return** $j'(u) = T^* \mathrm{j}'_{\Omega_0}(U) \in \mathcal{U}^*$

the correct implementation of derivatives via finite differences. Furthermore, this ensures that the coefficient vectors of the discrete derivatives $\mathrm{j}'_{\Omega_0,h}(U_h)$ and $\mathrm{j}'_{\Omega,h}(0)$ are equal, i.e., the step 6 in Algorithm 3.2 does not need to be executed explicitly. Note that, although we choose $\mathcal{U} = H^2(\Gamma_B)$ in theory, we do not employ a conforming discretization of $H^2(\Gamma_B)$. Instead, we experimented with different choices for $\mathcal{A}$. We use either

$$(v, u)_{\mathcal{A}} = (v, u)_{L^2(\Gamma_B)} + w(v', u')_{L^2(\Gamma_B)}, \text{ or}$$
$$(v, u)_{\mathcal{A}} \approx (v, u)_{L^2(\Gamma_B)} + w(v'', u'')_{L^2(\Gamma_B)},$$

where $w > 0$ is a weighting parameter. The bi-Laplacian scalar product $(v'', u'')_{L^2(\Gamma_B)}$ is approximated by $KM^{-1}K$. Here K is the stiffness matrix and M the lumped mass matrix. If not stated otherwise we chose the weighted bi-Laplacian scalar product with $w = 1$. In Section 4.5 we discuss different Riesz-isomorphisms R for Algorithms 2.6 and 2.6, and compare them for both scalar products and different weights.

The pure pressure matching objective (3.1) without additional constraints does in general *not* guarantee uniqueness or even existence of a solution of the optimization problem. It is immediately clear that a non-physical p_d will cause problems. But even an exactly realizable pressure distribution may admit infinitely many global minimizers. Consider, for example, a straight channel of arbitrary height and fixed length L. Suppose that the potential difference between left- and right-hand side is P. The potential induces a flow parallel to the channel with velocity $v = P/L$, *independent* of the height of the channel, i.e., if this is the desired

velocity then there are infinitely many global minimizers. Another issue is that coercivity of the Hessian with respect to $\mathcal{U}$ is not ensured, cf. Remark 2.95. In fact, as we will see in Chapter 4, for the current model problem the Hessian resembles rather a differential operator of order one which is not H^1-coercive. For these reasons, we add a cost term to the objective which promotes smooth solutions close to the initial domain, i.e., we consider

$$j(u) = \tilde{j}(u) + \frac{\beta}{2} \|u\|_{\mathcal{A}}^2 .$$

Here $\tilde{j}$ denotes the reduced tracking term functional we studied so far. If the existence of a solution is guaranteed by additional constraints, the cost term can be viewed as a *Tikhonov regularization*, and the parameter β can be driven to 0 in a continuation scheme. Tikhonov regularization is a standard technique in optimization with partial differential equations and inverse problems. Efficient algorithms iteratively adapt the regularization term along with the discretization, cf., e.g., [Kir14]. In Section 5.10 we investigate the addition of geometric constraints to the problem setting and present an example where β is iteratively decreased in the progression of a path following scheme. In the mean time, if not stated otherwise, we choose $\beta = 10^{-2}$ fixed.

Let us start with an example demonstrating the features of the employed algorithm. The initial domain is a 3-by-1 rectangle, and the outside potential induces a parallel flow with velocity one. The finite element mesh consists of 9221 nodes. The termination tolerance for the norm of the gradient is set to $\mathrm{TOL} = 10^{-6}$.

Example 3.1. In the first example the desired tangential velocity profile p_d increases linearly from $\frac{3}{4}$ to $\frac{5}{4}$. Table 3.1 recounts the iteration history of the globalized Newton method. The first column shows the iteration count, the second the objective value of the current iterate, the third the value of the tracking term, and the fourth column the norm of the gradient. The fifth column indicates whether the Newton step was accepted (Newton), or if the negative gradient was used as a search direction instead (gradient). The last column shows the number of iterations of the CG method. We observe fast local convergence of Newton's method. The final domain is depicted in Figure 3.2. For better visibility of the mesh the result of a run with coarser discretization is displayed.

Table 3.1.: Detailed history of Newton's method for example 3.1

k	$j(u_k)$	$\tilde{j}(u_k)$	$\|j'(u_k)\|_{\mathcal{A}^{-1}}$	step	# CG iter
0	$3.12{\cdot}10^{-2}$	$3.12{\cdot}10^{-2}$	$2.24{\cdot}10^{-1}$	Newton	5
1	$7.03{\cdot}10^{-3}$	$1.49{\cdot}10^{-3}$	$8.30{\cdot}10^{-2}$	Newton	7
2	$3.02{\cdot}10^{-3}$	$1.38{\cdot}10^{-3}$	$6.09{\cdot}10^{-2}$	Newton	5
3	$9.77{\cdot}10^{-4}$	$1.30{\cdot}10^{-4}$	$1.56{\cdot}10^{-2}$	Newton	5
4	$5.21{\cdot}10^{-4}$	$7.24{\cdot}10^{-5}$	$4.91{\cdot}10^{-3}$	Newton	5
5	$4.99{\cdot}10^{-4}$	$7.62{\cdot}10^{-5}$	$1.97{\cdot}10^{-4}$	Newton	5
6	$4.99{\cdot}10^{-4}$	$7.69{\cdot}10^{-5}$	$5.64{\cdot}10^{-7}$	-	-

If we employ a gradient descent strategy instead of Newton's method, 568 iterations are required to achieve a norm of the gradient of $9.99{\cdot}10^{-7}$. Although each iteration of the steepest

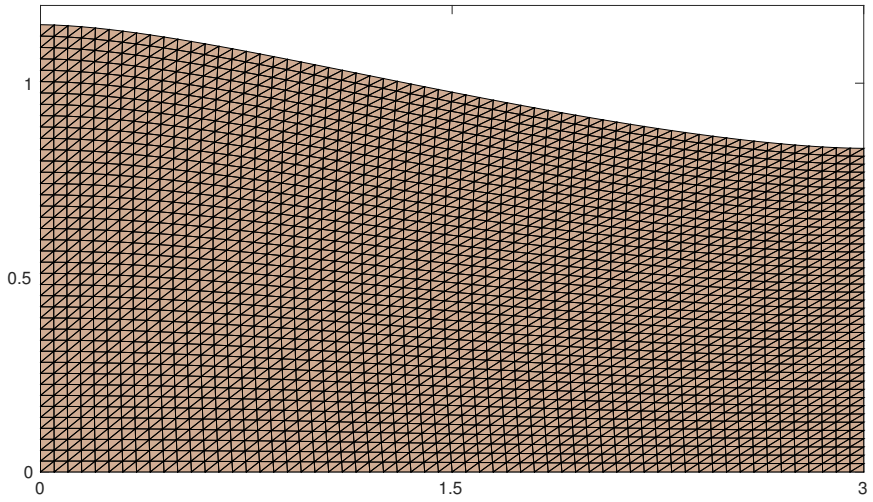


Figure 3.2.: The final domain of Example 3.1

descent method is much cheaper than in Newton's method, the overall runtime is still more than nine times as long as for Newton's method.

In Table 3.2 we compare the results of Newton's method for different mesh sizes. Clearly, a mesh-independent behavior can be observed. Finally, we present in Table 3.3 a comparison of the results for different choices of the $\mathcal{A}$-scalar product and the cost parameter β. While the effort of solving Newton's equation increases with decreasing w and β values, the overall behavior of the algorithm is not affected.

Table 3.2.: Comparison of different mesh sizes for Example 3.1

# nodes	$j(u_k)$	$\tilde{j}(u_k)$	$\|j'(u_k)\|_{\mathcal{A}^{-1}}$	# iter	# CG iter total
4941	$4.98{\cdot}10^{-4}$	$7.69{\cdot}10^{-5}$	$5.61{\cdot}10^{-7}$	6	32
9221	$4.98{\cdot}10^{-4}$	$7.69{\cdot}10^{-5}$	$5.64{\cdot}10^{-7}$	6	32
19521	$4.98{\cdot}10^{-4}$	$7.69{\cdot}10^{-5}$	$5.49{\cdot}10^{-7}$	6	32
50601	$4.98{\cdot}10^{-4}$	$7.69{\cdot}10^{-5}$	$5.42{\cdot}10^{-7}$	6	31

Let us conclude this section with two further examples.

Example 3.2. In the second example large deformations of the initial domain can be observed. The desired velocity profile is given by

$$p_d(x_1) = 1 + \frac{1}{4}\arctan\left(4x_1 - 3\right),$$

i.e., a steep increase around $x_1 = \frac{3}{4}$ which levels out later. Table 3.4 recounts the iteration history of the globalized Newton method. The first column shows the iteration count, the second the objective value of the current iterate, the third the value of the tracking term, and

Table 3.3.: Comparison of different choices for $\mathcal{A}$ and β for Example 3.1

$\mathcal{A}$-scalar product	β	$j(u_k)$	$\tilde{j}(u_k)$	$\|j'(u_k)\|_{\mathcal{A}^{-1}}$	# iter	# CG iter total
H^1, $w = 10^0$	10^{-2}	$4.22{\cdot}10^{-4}$	$6.34{\cdot}10^{-5}$	$8.03{\cdot}10^{-7}$	6	35
H^1, $w = 10^{-1}$	10^{-2}	$2.51{\cdot}10^{-4}$	$6.10{\cdot}10^{-5}$	$6.78{\cdot}10^{-9}$	7	67
H^1, $w = 10^{-2}$	10^{-2}	$2.34{\cdot}10^{-4}$	$6.08{\cdot}10^{-5}$	$2.65{\cdot}10^{-8}$	7	69
H^1, $w = 10^0$	10^{-4}	$4.15{\cdot}10^{-5}$	$2.25{\cdot}10^{-5}$	$2.71{\cdot}10^{-8}$	7	67
H^1, $w = 10^{-1}$	10^{-4}	$4.04{\cdot}10^{-5}$	$2.30{\cdot}10^{-5}$	$1.68{\cdot}10^{-8}$	7	69
H^1, $w = 10^{-2}$	10^{-4}	$4.03{\cdot}10^{-5}$	$2.30{\cdot}10^{-5}$	$5.40{\cdot}10^{-8}$	7	68
biLap, $w = 10^0$	10^{-2}	$4.99{\cdot}10^{-4}$	$7.69{\cdot}10^{-5}$	$5.64{\cdot}10^{-7}$	6	32
biLap, $w = 10^{-1}$	10^{-2}	$2.61{\cdot}10^{-4}$	$6.16{\cdot}10^{-5}$	$1.51{\cdot}10^{-10}$	7	37
biLap, $w = 10^{-2}$	10^{-2}	$2.35{\cdot}10^{-4}$	$6.09{\cdot}10^{-5}$	$2.36{\cdot}10^{-10}$	7	37
biLap, $w = 10^0$	10^{-4}	$4.24{\cdot}10^{-5}$	$2.25{\cdot}10^{-5}$	$1.34{\cdot}10^{-8}$	7	45
biLap, $w = 10^{-1}$	10^{-4}	$4.05{\cdot}10^{-5}$	$2.30{\cdot}10^{-5}$	$1.63{\cdot}10^{-9}$	7	58
biLap, $w = 10^{-2}$	10^{-4}	$4.03{\cdot}10^{-5}$	$2.30{\cdot}10^{-5}$	$2.03{\cdot}10^{-8}$	7	60

the fourth column the norm of the gradient. The fifth column indicates whether the Newton step was accepted (Newton), or if the negative gradient was used as a search direction instead (gradient). The last column shows the number of iterations of the CG method. In this example we can nicely observe the globalization strategy at work. After several gradient descent steps Newton's method takes over in the end and displays fast local convergence. Note that we stop the CG method early if negative curvature of the Hessian is detected. Of course, simply taking a negative gradient step after computing several CG iterations is a bit wasteful. The performance can be improved if the last viable CG direction is chosen instead, as it is done in the truncated CG method cf. [Ste83]. The final domain is depicted in Figure 3.3. For better visibility of the mesh the result of a run with coarser discretization is displayed.

Table 3.4.: Detailed history of Newton's method for Example 3.2

k	$j(u_k)$	$\tilde{j}(u_k)$	$\|j'(u_k)\|_{\mathcal{A}^{-1}}$	step	# CG iter
0	$1.31{\cdot}10^{-1}$	$1.31{\cdot}10^{-1}$	$3.39{\cdot}10^{-1}$	Newton	6
1	$8.35{\cdot}10^{-2}$	$7.70{\cdot}10^{-2}$	$2.24{\cdot}10^{-1}$	gradient	3
2	$7.87{\cdot}10^{-2}$	$7.18{\cdot}10^{-2}$	$3.28{\cdot}10^{-1}$	gradient	3
⋮	⋮	⋮	⋮	⋮	⋮
20	$5.79{\cdot}10^{-2}$	$5.06{\cdot}10^{-2}$	$1.98{\cdot}10^{-2}$	gradient	6
21	$5.79{\cdot}10^{-2}$	$5.06{\cdot}10^{-2}$	$2.34{\cdot}10^{-2}$	Newton	11
22	$5.78{\cdot}10^{-2}$	$5.26{\cdot}10^{-2}$	$2.87{\cdot}10^{-2}$	Newton	7
23	$5.71{\cdot}10^{-2}$	$5.17{\cdot}10^{-2}$	$2.25{\cdot}10^{-3}$	Newton	7
24	$5.71{\cdot}10^{-2}$	$5.27{\cdot}10^{-2}$	$3.58{\cdot}10^{-3}$	Newton	7
25	$5.71{\cdot}10^{-2}$	$5.25{\cdot}10^{-2}$	$5.03{\cdot}10^{-5}$	Newton	6
26	$5.71{\cdot}10^{-2}$	$5.25{\cdot}10^{-2}$	$5.59{\cdot}10^{-6}$	Newton	7
27	$5.71{\cdot}10^{-2}$	$5.25{\cdot}10^{-2}$	$1.06{\cdot}10^{-10}$	-	-

Example 3.3. Finally we present an example where the solution is known analytically. Recall

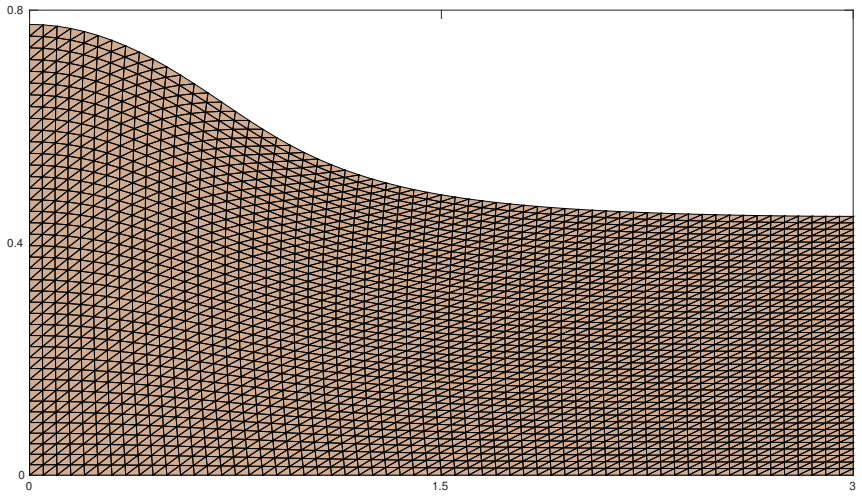


Figure 3.3.: The final domain of Example 3.2

from above that for $p_d \equiv 1$ any straight channel is a global solution. We start with the perturbed domain depicted in Figure 3.2. In this example we choose $\beta = 10^{-4}$. The iteration history is recounted in Table 3.5. Again we observe several gradient steps before Newton's method takes over. We recover a flat domain, i.e., a global solution. Although the convergence is quite fast it does not seem to be quadratic. This observation is in line with the fact that in the optimum the Hessian is *not* coercive in all directions which are *normal* to the boundary.

Table 3.5.: Detailed history of Newton's method for Example 3.3

k	$j(u_k)$	$\tilde{j}(u_k)$	$\|j'(u_k)\|_{\mathcal{A}^{-1}}$	step	# CG iter
0	$3.03{\cdot}10^{-2}$	$3.03{\cdot}10^{-2}$	$2.23{\cdot}10^{-1}$	Newton	5
1	$2.34{\cdot}10^{-3}$	$2.29{\cdot}10^{-3}$	$1.09{\cdot}10^{-1}$	gradient	2
2	$4.84{\cdot}10^{-4}$	$4.36{\cdot}10^{-4}$	$4.15{\cdot}10^{-2}$	gradient	3
⋮	⋮	⋮	⋮	⋮	⋮
7	$1.13{\cdot}10^{-4}$	$2.99{\cdot}10^{-5}$	$1.63{\cdot}10^{-2}$	gradient	6
8	$1.04{\cdot}10^{-4}$	$2.08{\cdot}10^{-5}$	$1.25{\cdot}10^{-2}$	Newton	8
9	$7.53{\cdot}10^{-5}$	$6.79{\cdot}10^{-5}$	$7.98{\cdot}10^{-3}$	Newton	5
10	$7.80{\cdot}10^{-6}$	$3.20{\cdot}10^{-7}$	$2.78{\cdot}10^{-4}$	Newton	6
11	$4.15{\cdot}10^{-6}$	$7.80{\cdot}10^{-9}$	$4.48{\cdot}10^{-5}$	Newton	6
12	$4.24{\cdot}10^{-6}$	$1.61{\cdot}10^{-9}$	$4.54{\cdot}10^{-6}$	Newton	6
13	$4.14{\cdot}10^{-6}$	$1.52{\cdot}10^{-9}$	$4.76{\cdot}10^{-7}$	-	-

4. The symbol of the Hessian in potential flow pressure matching

The numerical efficiency of many optimization methods is greatly influenced by the availability of a reasonably good approximation of the Hessian. In this chapter we want to exemplify how such an approximation can be obtained using the *symbol of the Hessian*. We consider the model problem of Chapter 3, and study the operator $j''(u) \in \mathcal{L}(\mathcal{U},\mathcal{U}^*)$. The term *symbol of an operator* originates in Fourier analysis. Denote the Fourier transform of a function by $\hat{\cdot}$ and consider an operator P. If there exists a function m such that

$$\widehat{Pf}(\alpha) = m(\alpha)\hat{f}(\alpha), \quad \forall\alpha,$$

then the multiplier function m is called the *symbol* of the operator P. For example it is well known that the symbol of the differential operator of order one is $i\alpha$. Note that the symbol depends on the employed definition of the Fourier transform. The symbol $i\alpha$ corresponds to a Fourier transform with angular frequency. To determine the symbol of the Hessian in shape optimization, one usually concentrates on a single Fourier mode v with frequency α as input, and tries to characterize $j''(u)v$ in terms of v and α. We will work with the real-valued Fourier mode $\cos(\alpha x)$ and obtain the symbol $C\alpha^2$, i.e., the Hessian of our model problem corresponds roughly to a differential operator of order two.

In our derivation of the symbol we exploit the fact that the Hessian is symmetric which implies

$$\langle j''(u)v, w\rangle_{\mathcal{U}^*,\mathcal{U}} = \frac{1}{2}\left(\langle j''(u)(v+w), v+w\rangle_{\mathcal{U}^*,\mathcal{U}} - \langle j''(u)v, v\rangle_{\mathcal{U}^*,\mathcal{U}} - \langle j''(u)w, w\rangle_{\mathcal{U}^*,\mathcal{U}}\right).$$

Hence, it suffices to characterize the mapping $v \mapsto \langle j''(u)v, v\rangle_{\mathcal{U}^*,\mathcal{U}}$. To be of practical use, the obtained approximation should be much cheaper to evaluate than the true Hessian. Once one has a suitable approximation of the Hessian available it can be used for several applications. The two most obvious choices are to either use the approximation *instead* of the true Hessian in a Newton-type method, or to use it as a *preconditioner* in a Krylov-subspace method applied to the full Newton equation.

Traditionally, the symbol of the shape Hessian has been used in Newton-type descent methods known as *preconditioned gradient methods*. It is particularly popular in the field of shape optimization problems arising in fluid dynamics. Usually, Fourier analysis is employed to study the highest order terms of the shape Hessian. Prominent early examples are [AT95, AT96, AV99]. Later this approach was pursued, for example, in [ESSI09, Sch10], and implemented for various flow problems.

We consistently derive the symbol of the shape Hessian for the elliptic model problem of Chapter 3. Following the usual approach to operator symbols in shape optimization, we consider a localized problem, and quantify the image of the Hessian for a smooth periodic perturbation. We use the exact shape derivative and shape Hessian of the localized problem, and a mapping from boundary displacement to domain displacement via an extension operator. We obtain a simple expression as approximation of the Hessian. It has the additional charm of being symmetric and positive semidefinite. In particular, our expression differs from the one obtained in [ESSI09], and yields superior results when used in a preconditioned gradient method. An exemplary numerical comparison of the effect of the approximations with the exact Hessian indicates that our version is more accurate.

Just as interesting as using the approximation in a Newton-type strategy is the second application mentioned above. The practical performance of Newton's method is vastly dominated by the effort of solving the Newton equation in each iteration. Since the computation of the full Hessian is infeasible in PDE constrained problems, a matrix-free, iterative Krylov-subspace method has to be employed, for instance, the CG method (Algorithm 2.4). For this it is only necessary to evaluate Hessian-times-vector products, which can be achieved by solving two additional PDEs. The efficient evaluation of first and second order derivatives via the adjoint approach is briefly summarized in Section A.1. For a more detailed discussion we refer to [HPUU09]. The performance of the CG method, and hence of Newton's method, is highly depended on the availability of a good preconditioner, i.e., an good approximation of the Hessian. Our numerical experiments indicate that using the approximation based on the symbol of the Hessian leads to a consistent, significant reduction of the number of necessary CG iterations.

4.1. Localized problem

We consider the model problem of Chapter 3. Recall that for $\Omega_{ref} \in \mathcal{O}$ it holds

$$\mathrm{j}_{\Omega_{ref}}(U) = J(U, S(U)),$$

where $\tau_U = \mathrm{Id} + U$, the functional

$$J\colon B^{\Theta}(0,1) \times \mathcal{Y} \to \mathbb{R}, \quad J(U,y) = \frac{1}{2}\int_{\Gamma_B} \left(\frac{t^T \nabla(y + y_0)}{|D\tau_U\, t|} - p_d \right)^2 |D\tau_U\, t| \,\mathrm{d}x,$$

corresponds to a pressure-tracking objective on the design boundary, and S is the design-to-solution operator associated with the potential flow, and characterized by the operator

$$E\colon B^{\Theta}(0,1) \times \mathcal{Y} \to \mathcal{Y}^*, \quad \langle E(U,y), \varphi \rangle_{\mathcal{Y}^*,\mathcal{Y}} = \left(A(U)\nabla(y + y_0), \nabla\varphi\right)_{L^2(\Omega_{ref})},$$

Furthermore, recall the map (3.2)

$$U \mapsto A(U) := D\tau_U^{-1} D\tau_U^{-T} \det(D\tau_U)$$

from Section 3.2 and its derivative in a direction V

$$M^V(U) := D\tau_U^{-1}\left(-DVD\tau_U^{-1} - D\tau_U^{-T}DV^T + \mathcal{I}\,\mathrm{tr}(D\tau_U^{-1}DV)\right)D\tau_U^{-T}\det(D\tau_U),$$

where $\mathcal{I}$ denotes the identity matrix in $\mathbb{R}^{d\times d}$. As in Section 3.4 we focus on the design boundary, i.e., we control only the transformation of the boundary directly, and obtain the associated domain transformation by a suitable extension operator $T\colon \mathcal{U} \to \Theta$, cf. Section 2.11 and 2.12.

We will use the following notation in this chapter: if a is a vector, then $[a]_i$ denotes the i-th component of a. Similarly $\nabla_i f$ denotes the i-th component of the gradient of a function f.

We want to characterize the map

$$v \mapsto \langle j''(u)v, v\rangle_{\mathcal{U}^*,\mathcal{U}},$$

where $j\colon \mathcal{U} \to \mathbb{R}, j = \mathfrak{j} \circ T$ is the reduced objective in terms of the boundary displacement. For this purpose, we study an auxiliary *localized shape optimization problem*. It is formally derived by the following reasoning. We *zoom* in on some point of the design boundary of the current object, until the design boundary *appears flat*, and the other boundaries are *far away*, i.e., at infinity. Thus, in the localized problem Ω_{ref} corresponds to the half-plane $\mathbb{R}^2_- := \{x \in \mathbb{R}^2 \mid x_2 < 0\}$, and the design boundary Γ_B to $\partial\mathbb{R}^2_-$. For a given small displacement of the boundary we set $\tau_U = \mathrm{Id} + U = \mathrm{Id} + T(u)$, and the state equation reads

$$\begin{aligned} -\operatorname{div}(A(U)\nabla y) &= 0 \text{ in } \mathbb{R}^2_- \\ n^T A(U)\nabla y &= 0 \text{ on } \partial\mathbb{R}^2_-, \end{aligned}$$

with additional boundary conditions at infinity.

The boundary conditions at infinity are assumed to induce a potential difference between left and right, which implies a *constant flow parallel to the boundary*. Thus $y(x) = cx_1$ is the analytical solution $y = S(0)$ of the state equation.

For simplicity, we consider only vertical displacements of the boundary described by a function $u\colon \partial\mathbb{R}^2_- \to \mathbb{R}$, and choose the extension operator

$$T \in \mathcal{L}(\mathcal{U}, \Theta), \quad Tu(x) = (0, u(x_1))^T \ \forall x \in \mathbb{R}^2.$$

Remark 4.1. This can be thought of as an approximation for a setting with normal boundary displacements and an extension operator as discussed in Section 2.11.2. Close to the design boundary such an extension operator will resemble the T from above in normal direction.

We study

$$\langle j''(0)v, v\rangle_{\mathcal{U}^*,\mathcal{U}} = \langle \mathfrak{j}''_{\Omega_{ref}}(0)Tv, Tv\rangle_{\Theta^*,\Theta}.$$

for a specific periodic perturbation v given by

$$v(x_1) = \cos(\alpha x_1), \text{ for some } \alpha \in \mathbb{R}.$$

Let us abbreviate $V := Tv$ and recall the representation of the second derivative of $\mathfrak{j}_{\Omega_{ref}}$

$$\begin{aligned}\langle \mathfrak{j}''_{\Omega_{ref}}(0)V,V\rangle_{\Theta^*,\Theta} = &\langle J_{UU}(0,S(0))V,V\rangle_{\Theta^*,\Theta}\\ &+\langle J_{Uy}(0,S(0))V,S'(0)V\rangle_{\mathcal{Y}^*,\mathcal{Y}}\\ &+\langle J_{yU}(0,S(0))S'(0)V,V\rangle_{\Theta^*,\Theta}\\ &+\langle J_{yy}(0,S(0))S'(0)V,S'(0)V\rangle_{\mathcal{Y}^*,\mathcal{Y}}\\ &+\langle J_{y}(0,S(0)),S''(0)(V,V)\rangle_{\mathcal{Y}^*,\mathcal{Y}}.\end{aligned}$$

As a first step we characterize $S'(0)Tv$ and $S''(0)(Tv,Tv)$ in the next section. This is the crucial part when deriving the symbol of the Hessian.

For later use we briefly note that, for our choice of $V = Tv$, and using the notation $v' = Dv$, it holds

$$DV(x) = \begin{pmatrix} 0 & 0 \\ v'(x_1) & 0 \end{pmatrix}.$$

Hence, we find $\operatorname{div}(V) = 0$, $DVDV = 0$, and

$$M^V(0)(x) = (\mathcal{I}\operatorname{div}(V) - DV - DV^T)(x) = \begin{pmatrix} 0 & -v'(x_1) \\ -v'(x_1) & 0 \end{pmatrix}.$$

4.2. Characterization of the design-to-state operator

In this section we examine the derivatives of S more closely. Recall from Section 3.3 that $z_V = S'(0)V$, and $\mu_{VV} = S''(0)(V,V)$ can be obtained by solving (3.4), respectively (3.5). Transferred to the localized problem the linearized state equation in strong form reads

$$\begin{aligned} -\Delta z &= \operatorname{div}(M^V(0)\nabla y) && \text{in } \mathbb{R}^2_-,\\ \nabla_2 z &= -\left[M^V(0)\nabla y\right]_2 && \text{on } \partial\mathbb{R}^2_-.\end{aligned}$$

Inserting our knowledge about y and V we obtain for $x \in \mathbb{R}^2_-$

$$-\left[M^V(0)\nabla y\right]_2(x) = v'(x_1)\nabla_1 y(x) + v'(x_1)\nabla_2 y(x) = -\alpha\nabla_1 y \sin(\alpha x_1),$$

and

$$\operatorname{div}(M^V(0)\nabla y)(x) = -(v''(x_1)\nabla_2 y(x) + v'(x_1)\nabla^2_{21} y(x) + v'(x_1)\nabla^2_{12} y(x)) = 0,$$

where $\nabla_1 y = c = \nabla_1 y(x)$ for all $x \in \mathbb{R}^2_-$. Hence the function $z_V = S'(0)V$ solves

$$\begin{aligned} -\Delta z(x) &= 0 && \text{in } \mathbb{R}^2_-,\\ \nabla_2 z(x) &= -\alpha\nabla_1 y \sin(\alpha x_1) && \text{on } \partial\mathbb{R}^2_-.\end{aligned} \tag{4.1}$$

Turning to (3.5), we first note that for our specific choices of v and y it holds $\operatorname{div}(V) = 0$, $DVDV = 0$, $M^V(0)DV^T\nabla y = 0$, and $DVM^V(0)\nabla y = 0$. Hence, several terms on the right-hand side of (3.5) drop out and the strong form for our setting reads

$$\begin{aligned} -\Delta\mu &= \operatorname{div}(2M^V(0)\nabla z_V) && \text{in } \mathbb{R}^2_-, \\ \nabla_2\mu &= -\left[2M^V(0)\nabla z_V\right]_2 && \text{on } \partial\mathbb{R}^2_-. \end{aligned} \tag{4.2}$$

4.2.1. The first derivative

We would like to find a simple, explicit expression for z_V which solves (4.1). Taking the right-hand side of (4.1) into account we are led to the ansatz

$$z_V(x) = \sin(\alpha x_1) f(x_2).$$

Here $f\colon \mathbb{R} \to \mathbb{R}$ is some suitable function yet to be determined. Inserting this ansatz into (4.1) we obtain an initial value problem for f:

$$f''(x_2) - \alpha^2 f(x_2) = 0 \quad \forall x_2 < 0, \qquad f'(0) = -\alpha\nabla_1 y.$$

It is easily checked that the initial value problem admits the solutions

$$f(x_2) = \mp\nabla_1 y \exp(\pm\alpha x_2).$$

To obtain a unique solution we require additionally that $f(x_2) \to 0$ as $x_2 \to -\infty$. This is motivated by the reasoning that the influence of a perturbation of the boundary should dwindle with the distance to the boundary. Thus, we obtain the linearized state as

$$z_V(x) = -\operatorname{sgn}(\alpha)\nabla_1 y \sin(\alpha x_1) \exp(|\alpha| x_2).$$

4.2.2. The second derivative

We are now ready to specify the right-hand side of (4.2). It holds

$$\begin{aligned} \nabla z_V(x) &= -\operatorname{sgn}(\alpha)\nabla_1 y \begin{pmatrix} \alpha\cos(\alpha x_1)\exp(|\alpha| x_2) \\ |\alpha|\sin(\alpha x_1)\exp(|\alpha| x_2) \end{pmatrix}, \\ M^V(0)\nabla z_V(x) &= -\operatorname{sgn}(\alpha)\nabla_1 y \begin{pmatrix} |\alpha|\sin^2(\alpha x_1)\exp(|\alpha| x_2) \\ \alpha^2\sin(\alpha x_1)\cos(\alpha x_1)\exp(|\alpha| x_2) \end{pmatrix}, \\ \operatorname{div}(M^V(0)\nabla z_V(x) &= -3\alpha^3\nabla_1 y \sin(\alpha x_1)\cos(\alpha x_1)\exp(|\alpha| x_2). \end{aligned}$$

Hence, the function $\mu_{VV} = S''(0)(V,V)$ is the solution of

$$\begin{aligned} -\Delta\mu &= -6\alpha^3\nabla_1 y \sin(\alpha x_1)\cos(\alpha x_1)\exp(|\alpha| x_2) && \text{in } \mathbb{R}^2_- \\ \nabla_2\mu &= 2\operatorname{sgn}(\alpha)\alpha^2\nabla_1 y \sin(\alpha x_1)\cos(\alpha x_1)\exp(|\alpha| x_2) && \text{on } \partial\mathbb{R}^2_-. \end{aligned} \tag{4.3}$$

We are led to the ansatz

$$\mu_{VV}(x) = \sin(\alpha x_1)\cos(\alpha x_1) f(x_2),$$

where $f\colon \mathbb{R} \to \mathbb{R}$ is again some suitable function yet to be determined. Inserting this ansatz into (4.3) we obtain this time an inhomogenuous initial value problem for f:

$$f''(x_2) - 4\alpha^2 f(x_2) = 6\alpha^3 \nabla_1 y \exp(|\alpha| x_2) \quad \forall x_2 < 0, \qquad f'(0) = \operatorname{sgn}(\alpha) 2\alpha^2 \nabla_1 y.$$

The solutions of the homogenous equation

$$f''(x) - 4\alpha^2 f(x) = 0$$

are given by $f_{1,2}(x) = \exp(\pm 2\alpha x)$. Now we look for a solution of the inhomogeneous problem using the method of variation of parameters. Setting $g(x) = 6\alpha^3 \nabla_1 y \exp(|\alpha| x)$, the solution has the form

$$f(x) = A(x) f_1(x) + B(x) f_2(x),$$

with A, B given by

$$\begin{aligned} A(x) &= -\int W^{-1} f_2(x) g(x) \,\mathrm{d}x \\ B(x) &= \int W^{-1} f_1(x) g(x) \,\mathrm{d}x \\ W &= f_1(x) f_2'(x) - f_1'(x) f_2(x). \end{aligned}$$

We calculate

$$\begin{aligned} W &= -2\alpha \cdot 1 - 2\alpha \cdot 1 = -4\alpha \\ A(x) &= \frac{1}{4\alpha} \int \exp(-2\alpha x) 6\alpha^3 \nabla_1 y \exp(|\alpha| x_2) \,\mathrm{d}x \\ B(x) &= -\frac{1}{4\alpha} \int \exp(2\alpha x) 6\alpha^3 \nabla_1 y \exp(|\alpha| x_2) \,\mathrm{d}x. \end{aligned}$$

Hence, it holds for $\alpha \geq 0$

$$\begin{aligned} A(x) &= \frac{6\nabla_1 y}{4\alpha} \int \alpha^3 \exp(-\alpha x) \,\mathrm{d}x = -\frac{3}{2} \nabla_1 y \alpha \exp(-\alpha x) + C_1 \\ B(x) &= -\frac{6\nabla_1 y}{4\alpha} \int \alpha^3 \exp(3\alpha x) \,\mathrm{d}x = -\frac{1}{2} \nabla_1 y \alpha \exp(3\alpha x) + C_2 \\ \Rightarrow f(x) &= \left(-\frac{3}{2} \nabla_1 y \alpha \exp(-\alpha x) + C_1 \right) \exp(2\alpha x) \\ &\quad + \left(-\frac{1}{2} \nabla_1 y \alpha \exp(3\alpha x) + C_2 \right) \exp(-2\alpha x) \\ &= -2\nabla_1 y \alpha \exp(\alpha x) + C_1 \exp(2\alpha x) + C_2 \exp(-2\alpha x), \end{aligned}$$

where $C_1, C_2 \in \mathbb{R}$, and similarly for $\alpha < 0$

$$\begin{aligned}
A(x) &= \frac{6\nabla_1 y}{4\alpha} \int \alpha^3 \exp(-3\alpha x)\,\mathrm{d}x = -\frac{1}{2}\nabla_1 y\alpha \exp(-3\alpha x) + C_1 \\
B(x) &= -\frac{6\nabla_1 y}{4\alpha} \int \alpha^3 \exp(\alpha x)\,\mathrm{d}x = -\frac{3}{2}\nabla_1 y\alpha \exp(\alpha x) + C_2 \\
\Rightarrow f(x) &= \left(-\frac{1}{2}\nabla_1 y\alpha \exp(-3\alpha x) + C_1\right)\exp(2\alpha x) \\
&\quad + \left(-\frac{3}{2}\nabla_1 y\alpha \exp(\alpha x) + C_2\right)\exp(-2\alpha x_2) \\
&= -2\nabla_1 y\alpha \exp(-\alpha x) + C_1 \exp(2\alpha x) + C_2 \exp(-2\alpha x).
\end{aligned}$$

Requiring again $f(x) \to 0$ for $x \to -\infty$ we obtain

$$f(x) = \begin{cases} -2\nabla_1 y\alpha \exp(\alpha x) + C_1 \exp(2\alpha x) & \text{for } \alpha \geq 0 \\ -2\nabla_1 y\alpha \exp(-\alpha x) + C_2 \exp(-2\alpha x) & \text{for } \alpha < 0, \end{cases}$$

and taking the condition on $f'(0)$ into account yields finally

$$f(x) = -2\nabla_1 y\alpha \exp(|\alpha| x).$$

To summarize we found

$$\mu_{VV}(x) = -2\nabla_1 y\alpha \sin(\alpha x_1)\cos(\alpha x_1)\exp(|\alpha| x_2).$$

4.3. Derivation of the symbol

After characterizing $z_V = S'(0)V$ and $\mu_{VV} = S''(0)(V,V)$, we can now derive the *symbol of the Hessian*. Inserting the specific expressions from Section 3.2 into

$$\begin{aligned}
\langle j''(0)v, v\rangle_{\mathcal{U}^*,\mathcal{U}} = \langle \mathrm{j}''_{\Omega_{ref}}(0)V, V\rangle_{\Theta^*,\Theta} = \quad & \langle J_{UU}(0, S(0))V, V\rangle_{\Theta^*,\Theta} \\
& + \langle J_{Uy}(0, S(0))V, S'(0)V\rangle_{\mathcal{Y}^*,\mathcal{Y}} \\
& + \langle J_{yU}(0, S(0))S'(0)V, V\rangle_{\Theta^*,\Theta} \\
& + \langle J_{yy}(0, S(0))S'(0)V, S'(0)V\rangle_{\mathcal{Y}^*,\mathcal{Y}} \\
& + \langle J_y(0, S(0)), S''(0)(V,V)\rangle_{\mathcal{Y}^*,\mathcal{Y}},
\end{aligned} \tag{4.4}$$

we obtain

$$\begin{aligned}
\langle j''(0)v, v\rangle_{\mathcal{U}^*,\mathcal{U}} = & \int_{\Gamma_B} \left(\frac{3}{2}\left(t^T\nabla y\right)^2 - \frac{1}{2}p_d^2\right)(t^T DV\, t)(t^T DV\, t)\,\mathrm{d}x \\
& + \int_{\Gamma_B} \left(-\frac{1}{2}\left(t^T\nabla y\right)^2 + \frac{1}{2}p_d^2\right)(t^T DV^T DV\, t)\,\mathrm{d}x \\
& - 2\int_{\Gamma_B} \left(t^T\nabla y\right)\left(t^T\nabla z_V\right)(t^T DV\, t)\,\mathrm{d}x \\
& + \int_{\Gamma_B} \left(t^T\nabla z_V\right)^2\,\mathrm{d}x \\
& + \int_{\Gamma_B} \left(\left(t^T\nabla y\right) - p_d\right)\left(t^T\nabla \mu_{VV}\right)\,\mathrm{d}x.
\end{aligned}$$

For the specific choice of $\Gamma_B = \partial\mathbb{R}^2_-$, $t = (1,0)^T$, and V, we realize that $t^T DV t = 0$ and $t^T DV^T DV t = (v')^2$. Finally with

$$\begin{aligned} z_V(x) &= -\operatorname{sgn}(\alpha)\nabla_1 y \sin(\alpha x_1)\exp(|\alpha| x_2), \text{ and} \\ \mu_{VV}(x) &= -2\nabla_1 y \alpha \sin(\alpha x_1)\cos(\alpha x_1)\exp(|\alpha| x_2), \end{aligned}$$

it holds

$$\begin{aligned} \langle j''(0)v, v\rangle_{\mathcal{U}^*,\mathcal{U}} &= \int_{\partial\mathbb{R}^2_-} \left(-\frac{1}{2}\left(t^T\nabla y\right)^2 + \frac{1}{2}p_d^2\right)(t^T DV^T DV t)\,\mathrm{d}x \\ &\quad + \int_{\partial\mathbb{R}^2_-} \left(t^T\nabla z_V\right)^2\,\mathrm{d}x \\ &\quad + \int_{\partial\mathbb{R}^2_-} \left(\left(t^T\nabla y\right) - p_d\right)\left(t^T\nabla\mu_{VV}\right)\,\mathrm{d}x \\ &= \int_{\partial\mathbb{R}^2_-} \left(-\frac{1}{2}\left(t^T\nabla y\right)^2 + \frac{1}{2}p_d^2\right)(v')^2\,\mathrm{d}x \\ &\quad + \int_{\partial\mathbb{R}^2_-} \left(-\operatorname{sgn}(\alpha)\alpha\nabla_1 y\cos(\alpha x)\right)^2\,\mathrm{d}x \\ &\quad + \int_{\partial\mathbb{R}^2_-} \left(\left(t^T\nabla y\right) - p_d\right)\left(-2\nabla_1 y\alpha^2(\cos^2(\alpha x_1) - \sin^2(\alpha x_1))\right)\,\mathrm{d}x \\ &= \int_{\partial\mathbb{R}^2_-} \left(-\frac{1}{2}\left(t^T\nabla y\right)^2 + \frac{1}{2}p_d^2\right)v'v'\,\mathrm{d}x \\ &\quad + \int_{\partial\mathbb{R}^2_-} \left(t^T\nabla y\right)^2(-vv'')\,\mathrm{d}x \\ &\quad + \int_{\partial\mathbb{R}^2_-} \left(t^T\nabla y - p_d\right)2t^T\nabla y\left(vv'' + v'v'\right)\,\mathrm{d}x. \end{aligned}$$

We have arrived at an expression which depends explicitly only on y, p_d, and on v. Supposing that y and p_d vary much slower than v it can be simplified further. Formally performing integration by parts only with respect to v yields

$$\begin{aligned} \langle j''(0)v, v\rangle_{\mathcal{U}^*,\mathcal{U}} &= \int_{\partial\mathbb{R}^2_-} \left(-\frac{1}{2}\left(t^T\nabla y\right)^2 + \frac{1}{2}p_d^2\right)v'v'\,\mathrm{d}x \\ &\quad + \int_{\partial\mathbb{R}^2_-} \left(t^T\nabla y\right)^2(-vv'')\,\mathrm{d}x \\ &\quad + \int_{\partial\mathbb{R}^2_-} \left(t^T\nabla y - p_d\right)2t^T\nabla y\left(vv'' + v'v'\right)\,\mathrm{d}x \\ &\approx \int_{\partial\mathbb{R}^2_-} \left(-\frac{1}{2}\left(t^T\nabla y\right)^2 + \frac{1}{2}p_d^2\right)v'v'\,\mathrm{d}x \\ &\quad + \int_{\partial\mathbb{R}^2_-} \left(t^T\nabla y\right)^2 v'v'\,\mathrm{d}x \\ &\quad + \int_{\partial\mathbb{R}^2_-} \left(t^T\nabla y - p_d\right)2t^T\nabla y\left(-v'v' + v'v'\right)\,\mathrm{d}x \\ &= \int_{\partial\mathbb{R}^2_-} \frac{1}{2}\left(\left(t^T\nabla y\right)^2 + p_d^2\right)v'v'\,\mathrm{d}x. \end{aligned}$$

Hence the symbol of the Hessian is approximately

$$\frac{1}{2}\left(\left(t^T\nabla y\right)^2 + p_d^2\right)\alpha^2,$$

corresponding to a differential operator of order two, and we obtain the following approximation of the Hessian

$$\langle j''(0)v, w\rangle_{\mathcal{U}^*,\mathcal{U}} \approx \int_{\Gamma_B} \frac{1}{2}\left(\left(t^T\nabla y\right)^2 + p_d^2\right) v'w' \,\mathrm{d}x. \tag{KU}$$

4.4. Comparison to previous results

Let us put the presented approach and our result in context with the literature. To the best of our knowledge, the analysis of the symbol of the shape Hessian was pioneered by Arian and coworkers in [AT95, Ari95, AT96, AV99]. The applications range from potential and Euler flow to a coupled aeroelastic problem. In order to determine the nature of the Hessian in a local solution, they use a Taylor expansion approach and study the resulting *small disturbance problem*. However, they derive the symbol of the Hessian in a quite pragmatic way, and often numerical tests are missing.

Eppler et al. [ESSI09] studied the same model problem as we did in this chapter. They calculate first and second shape derivatives for a star-shaped domain. Claiming certain properties for the linearized state and linearized adjoint state they obtain the approximation

$$\langle j''(0)v, w\rangle_{\mathcal{U}^*,\mathcal{U}} \approx \int_{\Gamma_B}\left(t^T\nabla y\right)v'w' \,\mathrm{d}x + \int_{\Gamma_B}\left(t^T\nabla p\right)v'w \,\mathrm{d}x + \int_{\Gamma_B}\left(t^T\nabla y\right)\left(t^T\nabla p\right)vw \,\mathrm{d}x, \tag{ESSI}$$

corresponding also to a differential operator of order two. Note however, that this expression is not symmetric, reflecting the use of the second Eulerian semiderivative instead of the shape Hessian from Definition 2.42. Moreover, it does not define a positive (semi-) definite bilinear form on $\mathcal{U}$. Finally (ESSI) does not depend explicitly on the desired velocity profile p_d, i.e., one can expect it to be not very sensitive with respect to different profiles p_d.

In his dissertation [Sch10] Schmidt extensively uses the symbol of the Hessian to design preconditioned gradient methods for shape optimization problems in aerodynamics. For potential and Stokes flow he gives analytical formulas. For Navier-Stokes and Euler flow he resorts to studying numerically the response of a finite difference approximation of the Hessian for different periodic perturbations. Approximating the shape optimization problem with potential flow by an optimal control problem he obtains also the symbol α^2, i.e., a differential operator of order two. For the other applications he obtains the symbol $|\alpha|$, corresponding to a *pseudo-differential operator* of order one. Finally, for the actual implementation, he always uses a weighted H^1-scalar product on the boundary to compute the search direction.

In our approach we consistently derive the symbol of the Hessian for the localized shape optimization problem. The applied technique appears to be easily adapted for other shape optimization problems. Note however, that additional complications occur if the objective is not given as a boundary integral on the design boundary. In this case the identification of the

symbol from the formula (4.4) is challenging. An alternative might be to reformulate $\mathrm{j}''_{\Omega_{ref}}$ with the help of a boundary integral on the design boundary. Due to Theorem 2.79, this is usually possible.

Let us conclude this section with a numerical comparison of our approximation (KU) with (ESSI) and the *true* Hessian. As test example we choose a setting which is close to the localized shape optimization problem, but we consider also domains which are quite different from the ideal half plane setting.

4.5. Numerical examples

In this section we first present a qualitatively comparison between the different approximations of the Hessian and the *true* Hessian. We then proceed to exploit the approximations algorithmically. As we will see, it is worthwhile to go through the above analysis. The approximations lead to a considerable speed up, both if used in a Newton-type setting or as preconditioner in the CG method. However, depending on the application, some modifications may be necessary. If the obtained approximations (KU) and (ESSI) are to be used to determine a Newton-type descent direction they should be modified such that one obtains *coercive* bilinear forms. If one applies these as a preconditioner in the CG method, one needs to ensure additionally their *symmetry*.

Qualitative comparison of the approximations

We work with the setting presented in Section 3.4, and begin with a qualitative check of the approximation quality of (KU), respectively (ESSI). For this we consider a periodic perturbation

$$v(x) = 1/10 \sin\left(10\pi(1 + x/3)\right)$$

and compare $\tilde{j}''(0)v \in \mathcal{U}^*$, i.e., the true Hessian applied to v, with the respective approximations (KU), (ESSI). Furthermore, we plot the H^1-Riesz map $(v, \cdot)_{H^1(\Gamma_B)} \in \mathcal{U}^*$ of v. Note that we study here only the second derivative of the tracking term, i.e., we set the regularization parameter to $\beta = 0$.

We start our investigation with Example 3.1 from Section 3.4. Recall that the desired tangential velocity profile increases linearly from $\frac{3}{4}$ to $\frac{5}{4}$, and the initial domain is a 3-by-1 rectangle. As one can see from Figure 4.1, all the outputs are very similar. A closer look reveals that the approximation given by (KU) captures the slow increase from left to right of the true Hessian output. On the other hand, the approximation (ESSI) can hardly be discerned from the simple H^1-scalar product output. This is not surprising if one recalls that the outside potential induces a constant flow with velocity one through the rectangle, i.e., $t^T \nabla y \equiv 1$. In particular, if p_d is close to one, the approximation (ESSI) is similar to the H^1-scalar product.

In Example 3.2 we considered a velocity profile which deviates farther from one, i.e.

$$p_d(x_1) = 1 + \frac{1}{4} \arctan(4x - 1 - 3).$$

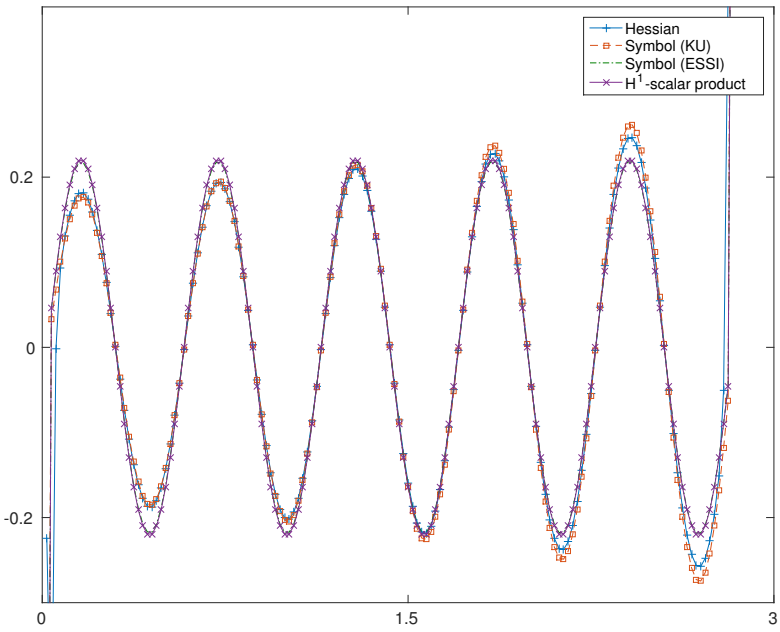


Figure 4.1.: Plot of different output signals for Example 3.1

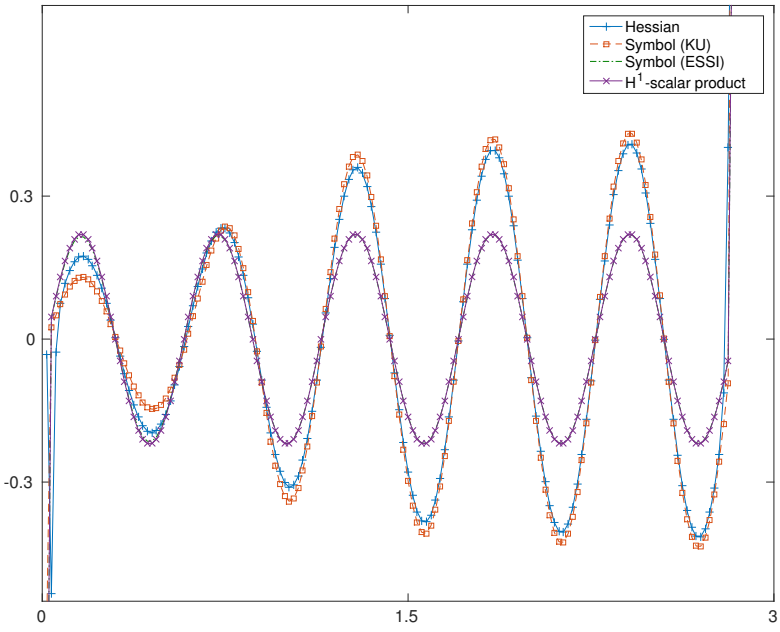


Figure 4.2.: Plot of different output signals for Example 3.2

Figure 4.2 shows that our approximation (KU) mimics the true Hessian again very closely. As predicted, the approximation (ESSI) does not capture the changed velocity profile p_d, and is again almost indistinguishable from the H^1-scalar product. However, it needs to be noted that, during the derivation of (ESSI) in [ESSI09], it was assumed that $t^T \nabla y - p_d$ is small, and corresponding terms were neglected. Thus, the situation of example 2 does not quite fit into their setting.

Example 4.1. We consider a different outside potential which induces a flow with velocity two through the initial domain. The desired tangential velocity profile increases linearly from $\frac{7}{4}$ to $\frac{9}{4}$. In particular, as in Example 3.1, this is a situation where $t^T \nabla y - p_d$ is relatively small. However, Figure 4.3 reveals that the approximation (ESSI) is again not able to mimic the true Hessian as good as (KU).

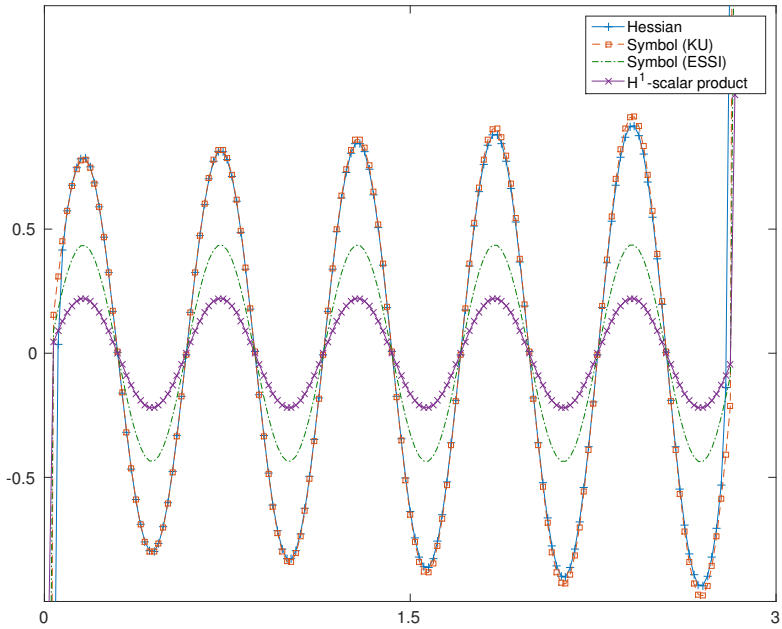


Figure 4.3.: Plot of different output signals for Example 4.1

Since the domain considered so far is very close to the ideal half-plane setting, the question arises whether the observed behavior also manifests itself on other domains. As an example we plot in Figure 4.4 the different quantities for the same periodic perturbation on the final domain of Example 3.2, cf. Figure 3.3. Again we observe a good match between the approximation (KU) and the true Hessian.

Similar observations as above can be made if one varies the frequency α.

Let us now see how the different approximations influence the performance of our algorithms.

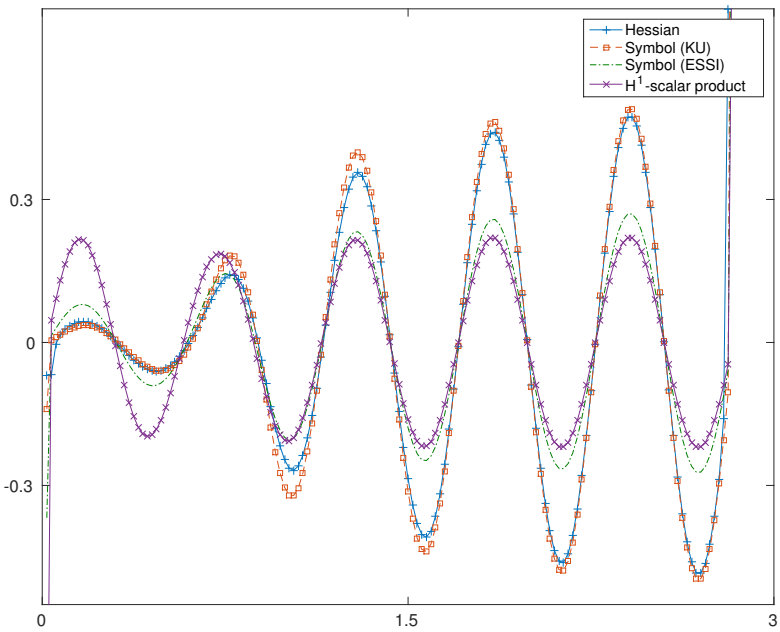


Figure 4.4.: Plot of different output signals for Example 3.2, final domain

Application to Newton-type methods

We begin with the Newton-type method, i.e., we employ Algorithm 2.5 in combination with Algorithm 2.7. Recall that this approach is also referred to as preconditioned gradient method. In accordance with this, we use the term preconditioner for the Riesz isomorphism in Algorithm 2.7. As already mentioned, the preconditioner should induce a coercive bilinear form. Otherwise we are not guaranteed to obtain a descent direction. If the objective functional consists only of the tracking term, the expressions (KU) would need to be modified before it could be applied successfully as a preconditioner. However, as in Section 3.4, we include a cost term in the objective, i.e., we consider $j(u) = \tilde{j}(u) + \frac{\beta}{2} \|u\|_{\mathcal{A}}^2$. Hence, the Hessian of the combined objective is given by

$$\langle j''(0)v, w\rangle_{\mathcal{U}^*,\mathcal{U}} = \langle \tilde{j}''(0)v, w\rangle_{\mathcal{U}^*,\mathcal{U}} + \beta\,(v, w)_{\mathcal{A}}\,.$$

In particular, the additional term should be considered in the preconditioner as well. Since our approximation (KU) is already positive semidefinite, no further modifications are necessary to obtain a coercive bilinear form

$$P\colon \mathcal{U} \times \mathcal{U} \to \mathbb{R}, \quad P(v, w) = \int_{\Gamma_B} \frac{1}{2} \left(\left(t^T \nabla y \right)^2 + p_d^2 \right) v' w' \,\mathrm{d}x + \beta\,(v, w)_{\mathcal{A}}\,. \tag{4.5}$$

Unfortunately, this is not the case if the approximation (ESSI) is used. Eppler et al. proposed already in [ESSI09, Section 5] to add a small positive correction $\delta > 0$ to the first term yielding

the preconditioner

$$\tilde{P}\colon \mathcal{U}\times\mathcal{U}\to\mathbb{R},\quad \tilde{P}(v,w)=\int_{\Gamma_B}\left(\left(t^T\nabla y\right)^2+\delta\right)^{1/2}v'w'\,\mathrm{d}x+\int_{\Gamma_B}\left(t^T\nabla p\right)v'w\,\mathrm{d}x \\ +\int_{\Gamma_B}\left(t^T\nabla y\right)\left(t^T\nabla p\right)vw\,\mathrm{d}x+\beta\,(v,w)_{\mathcal{A}}\,.$$

However, $\tilde{P}$ is still not a coercive bilinear form, and one is not guaranteed to obtain a descent direction. In fact, in most of our tests this occurred in one the first iterations. Thus, we decided to modify also the other two terms of (ESSI), and tested the preconditioner

$$\hat{P}\colon \mathcal{U}\times\mathcal{U}\to\mathbb{R},\ \ \hat{P}(v,w)=\int_{\Gamma_B}\left(\left(t^T\nabla y\right)^2+\delta\right)^{1/2}v'w'\,\mathrm{d}x+\int_{\Gamma_B}\left(\left(t^T\nabla p\right)^2+\delta\right)^{1/2}v'w\,\mathrm{d}x \\ +\int_{\Gamma_B}\left(\left(t^T\nabla y\right)^2\left(t^T\nabla p\right)^2+\delta\right)^{1/2}vw\,\mathrm{d}x+\beta\,(v,w)_{\mathcal{A}}\,. \tag{4.6}$$

In our experiments we always chose $\delta=10^{-4}$. We compare the performance of P and $\hat{P}$ with the standard steepest descent method. Recall our different choices for $\mathcal{A}$ from Section 3.4

$$(v,u)_{\mathcal{A}}=(v,u)_{L^2(\Gamma_B)}+w(v',u')_{L^2(\Gamma_B)},\ \text{or}$$
$$(v,u)_{\mathcal{A}}\approx(v,u)_{L^2(\Gamma_B)}+w(v'',u'')_{L^2(\Gamma_B)}.$$

The approximation of the bi-Laplacian scalar product is described in Section 3.4. For completeness we test once both scalar products with different weighting parameters $w>0$. Afterwards we choose always $w=1$. The cost parameter is set to $\beta=10^{-2}$.

We start again with Example 3.1, and report the results of our experiment in Table 4.1. The first column displays the particular choice for $\mathcal{A}$ and the second column the employed preconditioner. The third column shows the value of the objective, the fourth the tracking functional, the fifth the final norm of the gradient, and the last the number of iterations. Note that we stopped the algorithms if 1000 iterations were exceeded. Evidently the Newton-type search directions accelerate the convergence of the algorithm significantly.

Repeating the experiment with the other examples yields essentially the same result, cf. Tables 4.2, 4.3 and 4.4. Again the Newton-type search directions accelerate the convergence significantly. The best performance is obtained with the choice P, i.e., the approximation proposed in this chapter.

Application as preconditioner in the CG method

Let us now discuss the application of the obtained approximations as preconditioners in the CG method, i.e., Algorithm 2.5 in combination with Algorithm 2.6 and Algorithm 2.4. For this task the preconditioner should not only be coercive, but also *symmetric*. In particular, the asymmetric expression (ESSI) is not appropriate for such a purpose. Hence we concentrate only on (KU).

Table 4.1.: Comparison of different scalar products, Algorithm 2.7, Example 3.1

$\mathcal{A}$-scalar product	preconditioner	$j(u_k)$	$\tilde{j}(u_k)$	$\|j'(u_k)\|_{\mathcal{A}^{-1}}$	# iter
H^1, $w = 10^0$	$\mathcal{A}$	$4.22{\cdot}10^{-4}$	$6.34{\cdot}10^{-5}$	$9.98{\cdot}10^{-7}$	566
H^1, $w = 10^0$	P	$4.22{\cdot}10^{-4}$	$6.34{\cdot}10^{-5}$	$9.03{\cdot}10^{-7}$	39
H^1, $w = 10^0$	$\hat{P}$	$4.22{\cdot}10^{-4}$	$6.34{\cdot}10^{-5}$	$9.45{\cdot}10^{-7}$	300
H^1, $w = 10^{-1}$	$\mathcal{A}$	$2.52{\cdot}10^{-4}$	$6.14{\cdot}10^{-5}$	$8.36{\cdot}10^{-5}$	1000
H^1, $w = 10^{-1}$	P	$2.51{\cdot}10^{-4}$	$6.10{\cdot}10^{-5}$	$9.60{\cdot}10^{-7}$	95
H^1, $w = 10^{-1}$	$\hat{P}$	$2.51{\cdot}10^{-4}$	$6.10{\cdot}10^{-5}$	$1.08{\cdot}10^{-6}$	655
H^1, $w = 10^{-2}$	$\mathcal{A}$	$2.35{\cdot}10^{-4}$	$6.18{\cdot}10^{-5}$	$2.88{\cdot}10^{-4}$	1000
H^1, $w = 10^{-2}$	P	$2.34{\cdot}10^{-4}$	$6.08{\cdot}10^{-5}$	$9.98{\cdot}10^{-7}$	225
H^1, $w = 10^{-2}$	$\hat{P}$	$2.34{\cdot}10^{-4}$	$6.08{\cdot}10^{-5}$	$5.09{\cdot}10^{-6}$	1000
biLap, $w = 10^0$	$\mathcal{A}$	$4.98{\cdot}10^{-4}$	$7.69{\cdot}10^{-5}$	$9.99{\cdot}10^{-7}$	568
biLap, $w = 10^0$	P	$4.98{\cdot}10^{-4}$	$7.69{\cdot}10^{-5}$	$1.63{\cdot}10^{-7}$	20
biLap, $w = 10^0$	$\hat{P}$	$4.98{\cdot}10^{-4}$	$7.69{\cdot}10^{-5}$	$7.68{\cdot}10^{-8}$	298
biLap, $w = 10^{-1}$	$\mathcal{A}$	$2.61{\cdot}10^{-4}$	$6.16{\cdot}10^{-5}$	$9.92{\cdot}10^{-7}$	608
biLap, $w = 10^{-1}$	P	$2.61{\cdot}10^{-4}$	$6.16{\cdot}10^{-5}$	$1.95{\cdot}10^{-7}$	20
biLap, $w = 10^{-1}$	$\hat{P}$	$2.61{\cdot}10^{-4}$	$6.16{\cdot}10^{-5}$	$2.67{\cdot}10^{-7}$	274
biLap, $w = 10^{-2}$	$\mathcal{A}$	$2.35{\cdot}10^{-4}$	$6.09{\cdot}10^{-5}$	$6.47{\cdot}10^{-6}$	1000
biLap, $w = 10^{-2}$	P	$2.35{\cdot}10^{-4}$	$6.09{\cdot}10^{-5}$	$2.13{\cdot}10^{-7}$	20
biLap, $w = 10^{-2}$	$\hat{P}$	$2.61{\cdot}10^{-4}$	$6.16{\cdot}10^{-5}$	$5.50{\cdot}10^{-7}$	272

Table 4.2.: Comparison of different scalar products, Algorithm 2.7, Example 3.2

$\mathcal{A}$-scalar product	preconditioner	$j(u_k)$	$\tilde{j}(u_k)$	$\|j'(u_k)\|_{\mathcal{A}^{-1}}$	# iter
H^1, $w = 1$	$\mathcal{A}$	$5.61{\cdot}10^{-2}$	$5.29{\cdot}10^{-2}$	$7.71{\cdot}10^{-5}$	1000
H^1, $w = 1$	P	$5.61{\cdot}10^{-2}$	$5.28{\cdot}10^{-2}$	$9.62{\cdot}10^{-7}$	158
H^1, $w = 1$	$\hat{P}$	$5.61{\cdot}10^{-2}$	$5.28{\cdot}10^{-2}$	$3.18{\cdot}10^{-6}$	1000
biLap, $w = 1$	$\mathcal{A}$	$5.71{\cdot}10^{-2}$	$5.26{\cdot}10^{-2}$	$7.47{\cdot}10^{-5}$	1000
biLap, $w = 1$	P	$5.71{\cdot}10^{-2}$	$5.25{\cdot}10^{-2}$	$6.71{\cdot}10^{-7}$	244
biLap, $w = 1$	$\hat{P}$	$5.71{\cdot}10^{-2}$	$5.25{\cdot}10^{-2}$	$1.65{\cdot}10^{-5}$	1000

Table 4.3.: Comparison of different scalar products, Algorithm 2.7, Example 3.3

$\mathcal{A}$-scalar product	preconditioner	$j(u_k)$	$\tilde{j}(u_k)$	$\|j'(u_k)\|_{\mathcal{A}^{-1}}$	# iter
H^1, $w = 1$	$\mathcal{A}$	$3.53{\cdot}10^{-6}$	$4.11{\cdot}10^{-10}$	$1.76{\cdot}10^{-6}$	1000
H^1, $w = 1$	P	$3.51{\cdot}10^{-6}$	$4.22{\cdot}10^{-10}$	$5.17{\cdot}10^{-7}$	17
H^1, $w = 1$	$\hat{P}$	$3.54{\cdot}10^{-6}$	$5.43{\cdot}10^{-10}$	$2.95{\cdot}10^{-6}$	1000
biLap, $w = 1$	$\mathcal{A}$	$4.16{\cdot}10^{-6}$	$2.39{\cdot}10^{-9}$	$2.20{\cdot}10^{-6}$	1000
biLap, $w = 1$	P	$4.14{\cdot}10^{-6}$	$1.54{\cdot}10^{-9}$	$4.95{\cdot}10^{-7}$	17
biLap, $w = 1$	$\hat{P}$	$4.17{\cdot}10^{-6}$	$1.53{\cdot}10^{-9}$	$2.18{\cdot}10^{-6}$	1000

Table 4.4.: Comparison of different scalar products, Algorithm 2.7, Example 4.1

$\mathcal{A}$-scalar product	preconditioner	$j(u_k)$	$\tilde{j}(u_k)$	$\left\|j'(u_k)\right\|_{\mathcal{A}^{-1}}$	# iter
H^1, $w=1$	$\mathcal{A}$	$1.06{\cdot}10^{-4}$	$1.55{\cdot}10^{-5}$	$9.95{\cdot}10^{-7}$	953
H^1, $w=1$	P	$1.06{\cdot}10^{-4}$	$1.55{\cdot}10^{-5}$	$9.97{\cdot}10^{-7}$	91
H^1, $w=1$	$\hat{P}$	$1.06{\cdot}10^{-4}$	$1.55{\cdot}10^{-5}$	$9.33{\cdot}10^{-7}$	328
biLap, $w=1$	$\mathcal{A}$	$1.27{\cdot}10^{-4}$	$1.72{\cdot}10^{-5}$	$9.99{\cdot}10^{-7}$	796
biLap, $w=1$	P	$1.27{\cdot}10^{-4}$	$1.72{\cdot}10^{-5}$	$2.13{\cdot}10^{-7}$	20
biLap, $w=1$	$\hat{P}$	$1.27{\cdot}10^{-4}$	$1.72{\cdot}10^{-5}$	$7.94{\cdot}10^{-7}$	316

We note again that if the objective functional consists only of the tracking term, then the expression (KU) would need to be modified before one could employ it as a preconditioner. However, for the objective under consideration $j(u) = \tilde{j}(u) + \frac{\beta}{2}\left\|u\right\|_{\mathcal{A}}^2$ this is not necessary since the cost term should contribute to the preconditioner as well. Thus we can employ again the preconditioner P from (4.5).

As alternative we simply approximate the Hessian of the tracking term with the H^1-scalar product, and test the preconditioner

$$P_{H^1} : \mathcal{U} \times \mathcal{U} \to \mathbb{R}, \quad P_{H^1}(v,u) = (v,u)_{H^1(\Omega)} + \beta\,(v,u)_{\mathcal{A}}\,. \tag{4.7}$$

This choice corresponds also to a differential operator of order two. It is further motivated by the observation in the tests above that the true Hessian of the tracking term can be roughly approximated by this term. Note that P_{H^1} performs already better than the simple preconditioner $\beta\,(v,u)_{\mathcal{A}}$ which is only based on the regularization term.

To be able to compare different preconditioners in a fair manner, we choose a stopping criterion for the CG method which is *preconditioner independent*. It is based on an estimation of the *discrete energy error*, cf. [DW12, Section 5.3.3]. In our experience, the termination tolerance 10^{-3} proved sufficient to obtain good Newton directions. Furthermore, we only consider examples where the optimization algorithm generated the same iterates, i.e., where the only difference is the number of CG iterations till a Newton direction is determined. This is not guaranteed in the early stage of the optimization where the globalization strategy is active. For this reason, we always start with a steepest descent method, and only switch to the Newton method once the norm of the derivative drops below 10^{-2}. In our tests this ensures that the different Newton algorithms start with the same initial point close to the optimum.

We test again both the H^1-scalar product and the approximated bi-Laplacian scalar product as choices for $\mathcal{A}$. The results of the comparison are presented in Table 4.5. The first column indicates which example was tested, the second shows the choice of the $\mathcal{A}$-scalar product, the third the final value of the objective, the fourth the final value of the tracking term, and the fifth the norm of the derivative. In the sixth, and seventh column the total number of iterations, and the number of globalized Newton iterations are depicted. Finally, we compare the total number of CG iterations generated by the different preconditioners. As one can see, our investigation of the symbol of the Hessian pays off. The total number of CG iterations required by P was consistently below the one required by P_{H^1}, often even less than half.

Table 4.5.: Comparison of different preconditioners, Algorithm 2.6

ex.	$\mathcal{A}$-scalar product	$j(u_k)$	$\tilde{j}(u_k)$	$\|j'(u_k)\|_{\mathcal{A}^{-1}}$	# iter	# iter Newton	# CG iter total P_{H^1}	# CG iter total P
3.1	H^1, $w=1$	$4.22{\cdot}10^{-4}$	$6.34{\cdot}10^{-5}$	$2.70{\cdot}10^{-7}$	4	2	42	21
3.1	biLap, $w=1$	$4.99{\cdot}10^{-4}$	$7.69{\cdot}10^{-5}$	$9.72{\cdot}10^{-8}$	5	2	16	10
3.2	H^1, $w=1$	$5.61{\cdot}10^{-2}$	$5.29{\cdot}10^{-2}$	$6.81{\cdot}10^{-7}$	7	4	60	23
3.2	biLap, $w=1$	$5.71{\cdot}10^{-2}$	$5.24{\cdot}10^{-2}$	$8.12{\cdot}10^{-9}$	18	5	52	30
3.3	H^1, $w=1$	$3.51{\cdot}10^{-6}$	$4.07{\cdot}10^{-10}$	$2.00{\cdot}10^{-7}$	7	4	142	42
3.3	biLap, $w=1$	$4.14{\cdot}10^{-6}$	$1.52{\cdot}10^{-9}$	$2.03{\cdot}10^{-7}$	9	5	38	28
4.1	H^1, $w=1$	$1.06{\cdot}10^{-4}$	$1.55{\cdot}10^{-5}$	$1.13{\cdot}10^{-8}$	7	2	75	48
4.1	biLap, $w=1$	$1.27{\cdot}10^{-4}$	$1.72{\cdot}10^{-5}$	$3.34{\cdot}10^{-9}$	11	2	24	10

Finally, let us briefly comment on the comparison Newton-type versus Newton's method. No general statement can be given here, since it depends on the concrete example which method is faster. The Newton-type method iterations are significantly cheaper than full Newton iterations. However, fast local convergence to the optimum is not guaranteed, and in the end it may take the Newton-type method very long to drive the norm of the derivative below a given tolerance. Usually, the best results will be obtained by combining the two strategies, i.e., starting with the robust Newton-type method, and switching to full Newton close to the solution.

5. Moreau-Yosida path following

In this chapter we study a nonlinear optimal control problem with a specific class of control constraints. We are motivated by a shape optimization problem with point-wise geometric shape constraints. Nevertheless, since our results may be useful also in other applications, we present them here in a more general framework. In Section 5.9 we demonstrate on a concrete example that the following analysis is applicable for shape optimization problems. Our notation reflects the origin of our work in shape optimization, and may in parts be slightly non-standard for optimal control problems. In particular, we call the specific control constraint under consideration a *geometric constraint* to distinguish it from the standard L^2-control constraints.

The results obtained in this chapter were published in [KU15] in the context of shape optimization. As announced, we derive them here in a slightly more general setting, but the overall arrangement is kept, and the presentation follows closely the one in [KU15].

We are motivated by geometric constraints of the form

$$\tau(\Gamma_B) \subset C,$$

where $C \subset \mathbb{R}^d$ is some set, $\Gamma_B \subset \partial\Omega$ denotes the design part of the boundary $\partial\Omega$ of a reference domain Ω, and $\tau = \mathrm{id} + u$ is a transformation. Such design constraints appear in many applications and have been considered in various publications. Usually, they are either considered only with regard to some particular parametrization, e.g., constraints on the control points of some Bézier curve, or discretization, see for example [ABV13, BLUU09, BLUU11, Bra11, Lin12, HLA08, NZP04], or they are tacitly assumed to be inactive in the solution, see for example [Lau00, KV13]. To the best of our knowledge, so far our work in [KU15] is the only algorithmic treatment of such point-wise geometric shape constraints in a function space setting. The difficulty here is, that the Lagrange multiplier associated with the constraint has a-priori only a low regularity. This is a similar setting as in the case of state constraints in optimal control problems, where the multiplier associated with the state constraint is in general only a measure. Basically three approaches have been proposed in the literature on state constraints to deal with the associated difficulties. Inexact primal-dual path following techniques based on Moreau–Yosida regularization were first investigated in [IK03, HK06a, HK06b], Lavrentiev regularization methods and the related concept of virtual control were proposed in [Trö05, MRT06, PTW08, KR09], and barrier methods were studied in [Sch09, SG11, Kru14]. The Lavrentiev regularization concept relaxes the state constraints to mixed control and state constraints which feature Lagrange multipliers with higher regularity. However, in our setting the smoothness of the control causes the problems. Therefore, this approach is not applicable here. The theory of barrier methods is only available for convex optimal control problems. Since our optimization problem contains a highly nonlinear state

equation, we decide to follow the approach taken in [HK06a]. We introduce a Moreau-Yosida type penalty term, and study the properties of the solutions to the associated subproblems. Facing a nonlinear problem we will assume some strong second order conditions to hold. Note, that the study of second order necessary and sufficient conditions, especially in the context of semismooth derivatives is even in finite dimensions a quite involved topic. It is out of the scope of this thesis to investigate this aspect in more detail. We will show local Lipschitz continuity of the regularized solutions, and prove convergence rates estimates similar to [HSW14]. The subproblems can be solved efficiently by a semismooth Newton method, cf., e.g., [Ulb11].

This chapter is organized as follows. In Section 5.1 we introduce the setting under consideration, and briefly discuss existence of a solution as well as differentiability of the reduced objective. We derive the regularized problems employing a Moreau-Yosida type penalty term in Section 5.2, and discuss the convergence of regularized solutions towards a solution of the original problem in Section 5.3. We state strong second order conditions in Section 5.4, and show superlinear convergence of a semismooth Newton method for the regularized problems. We exploit the second order condition in Section 5.5 to show local Lipschitz continuity of the regularized solutions. The value function, which maps the regularization parameter to the optimal objective value of the associated problem, is studied in Section 5.6, and a model function in the spirit of [HK06a] is proposed. In Section 5.7 we use the optimality conditions to show convergence of the approximate Lagrange multipliers to the Lagrange multiplier associated with the geometric constraint. Convergence rate estimates are derived in Section 5.8. In Section 5.9 we specify conditions which make the results applicable for a shape optimization problem. Finally we present some numerical tests.

If not stated otherwise, we will denote in this chapter with $c > 0$ some generic constant which may change its value in the computations.

5.1. A nonlinear optimal control problem with point-wise geometric constraints

Let us specify the setting of this chapter. We study the optimal control problem

$$\min_{u\in\mathcal{U},y\in\mathcal{Y}} J(T(u),y) + \frac{\beta}{2}\,\|u\|_{\mathcal{U}}^2 \qquad \text{s.t. } E(T(u),y) = 0,\ u\in\mathcal{U}_{ad}, \tag{5.1}$$

where $\mathcal{U}$ is a Hilbert space with $\mathcal{U}_{ad}\subset\mathcal{U}$, furthermore $\mathcal{V},\mathcal{Y},\mathcal{Z}$ are Banach spaces, and $T\colon\mathcal{U}\to\mathcal{V}$, $E\colon\mathcal{V}\times\mathcal{Y}\to\mathcal{Z}$, $J\colon\mathcal{V}\times\mathcal{Y}\to\mathbb{R}$. Depending on the application the term $\frac{\beta}{2}\,\|u\|_{\mathcal{U}}^2$ with $\beta>0$ may be viewed as a control cost or Tikhonov regularization term. This term guarantees boundedness of minimizing sequences, which we need to ensure the existence of a minimizer. Furthermore, it generates a coercive contribution to the Hessian. Thus, it is reasonable to assume *second order conditions*, which are indispensable for the analysis of nonlinear optimization problems. The operator T is slightly non-standard in optimal control. We will refer to T as the *extension operator*, and have in mind, that T maps a boundary displacement u to a domain displacement U. In other applications, it might only be the compact embedding of some stronger space $\mathcal{U}$ into L^p, or some more involved mapping as it is often found in inverse problems. We refer to

[BU15] for an application in seismic tomography. As usual, we assume the existence of the *design-to-state mapping*

$$S\colon \mathcal{V} \to \mathcal{Y}, \quad E(U, S(U)) = 0 \quad \forall U \in T(\mathcal{U}_{ad}),$$

and introduce the *reduced objective functionals*

$$\begin{aligned} \mathfrak{j}\colon \mathcal{V} &\to \mathbb{R}, \quad \mathfrak{j}(V) := J(V, S(V)), \\ j\colon \mathcal{U} &\to \mathbb{R}, \quad j(u) := \mathfrak{j}(T(u)) + \frac{\beta}{2} \|u\|_{\mathcal{U}}^2 . \end{aligned}$$

Most of the time, we will work with the reduced problem

$$\min_{u \in \mathcal{U}} j(u) \quad \text{s.t. } u \in \mathcal{U}_{ad}, \tag{P}$$

which is equivalent to (5.1).

Let us briefly discuss the existence of minimizers. It is a standard result that (P) admits a solution if

(i) $\mathcal{U}$ is a reflexive Banach space, and $\emptyset \neq \mathcal{U}_{ad} \subset \mathcal{U}$ is a closed and convex set,

(ii) j is coercive (in the sense that $j(u) \to \infty$ if $\|u\|_{\mathcal{U}} \to \infty$), and weakly lower semicontinuous.

We specify a setting which is appropriate for the shape optimization context, but can also be motivated from an optimal control perspective. In many cases it is unreasonable to assume weak continuity of the operator S. Instead, we place stronger assumptions on T. We require *complete continuity* of T, i.e.,

$$u_n \rightharpoonup u \text{ implies } T(u_n) \to T(u).$$

This can, for example, be ensured if T is linear, continuous, and maps to a space $\tilde{\mathcal{V}}$ which is compactly embedded into $\mathcal{V}$. We refer to Theorem 2.87 for an example of a completely continuous extension operator in shape optimization.

Assumption 5.1. *It holds*

1. *$\mathcal{U}$ is a reflexive Banach space, $\mathcal{V}, \mathcal{Y}, \mathcal{Z}$ are Banach spaces.*
2. *The extension operator $T\colon \mathcal{U} \to \mathcal{V}$ is completely continuous.*
3. *There exists a design-to-state operator $S\colon \mathcal{V} \to \mathcal{Y}$ that is continuous.*
4. *The objective J is bounded from below and is continuous.*
5. *$\mathcal{U}_{ad} \subset \mathcal{U}$ is nonempty, closed, and convex.*

These conditions ensure that j is coercive and weakly lower semicontinuous. Indeed, the coercivity of j follows from J being bounded from below and the coercivity of the norm. Furthermore, $u \mapsto J(T(u), S(T(u)))$ is weakly continuous because T is completely continuous, and S, J are continuous. Finally, $u \mapsto \|u\|_{\mathcal{U}}^2$ is weakly lower semicontinuous, since it is a convex

and continuous functional. Hence we obtain the existence of a solution if Assumption 5.1 is satisfied.

A sufficient condition for the twice continuous differentiability of the reduced objective functional j is given by the following assumption.

Assumption 5.2. *It holds*

1. $T\colon \mathcal{U} \to \mathcal{V}$, $E : \mathcal{V} \times \mathcal{Y} \to \mathcal{Z}$ *and* $J\colon \mathcal{V} \times \mathcal{Y} \to \mathbb{R}$ *are twice continuously Fréchet differentiable.*
2. *For any bounded set* $A \subset T(\mathcal{U})$*, there exists a neighborhood* $\hat{A} \subset \mathcal{V}$ *of* $A \subset \mathcal{V}$ *such that, for every* $V \in \hat{A}$*, there exists a unique solution* $y \in \mathcal{Y}$ *of the state equation. Hence, the design-to-state operator* $S\colon \hat{A} \to \mathcal{Y}$ *is well-defined.*
3. $E_y(V, S(V)) \in \mathcal{L}(\mathcal{Y}, \mathcal{Z})$ *is continuously invertible for all* $V \in \hat{A}$*, with* $\hat{A}$ *from (ii).*

The implicit function theorem yields that $S\colon T(\mathcal{U}) \to \mathcal{Y}$ is twice continuously Fréchet differentiable. Hence the same holds for the reduced objective j due to the chain rule. Finally, we recall necessary optimality conditions for the reduced problem (P) and the full problem (5.1).

Lemma 5.1. *Let* $\mathcal{U}$ *be a Hilbert space with a closed, convex subset* $\mathcal{U}_{ad} \neq \emptyset$*, and let* $\overline{u} \in \mathcal{U}$ *be a local solution of* (P) *in which* j *is Gâteaux-differentiable. Then the following optimality conditions hold and are equivalent:*

$$\overline{u} \in \mathcal{U}_{ad}, \quad \langle j'(\overline{u}), u - \overline{u} \rangle_{\mathcal{U}^*,\mathcal{U}} \geq 0 \quad \forall u \in \mathcal{U}_{ad} \tag{5.2}$$

$$\overline{u} = P_{\mathcal{U}_{ad}}(\overline{u} - c\nabla j(\overline{u})). \tag{5.3}$$

Here $P_{\mathcal{U}_{ad}}\colon \mathcal{U} \to \mathcal{U}$ *denotes the projection onto* $\mathcal{U}_{ad}$*,* $c > 0$ *is arbitrary but fixed and* $\nabla j(u) \in \mathcal{U}$ *denotes the Riesz-representation of* $j'(u) \in \mathcal{U}^*$*.*

Proof. Compare [HPUU09, Corollary 1.2]. □

An alternative formulation of the necessary optimality conditions uses the tangent cone of $\mathcal{U}_{ad}$. If $\mathcal{U}_{ad}$ is closed and convex, the tangent cone in $\overline{u} \in \mathcal{U}_{ad}$ is given by

$$\mathcal{T}(\mathcal{U}_{ad}, \overline{u}) := \operatorname{cl}\{w \in \mathcal{U} \mid w = \mu(v - \overline{u}), \text{ where } \mu > 0,\ v \in \mathcal{U}_{ad}\}.$$

If j is continuously Fréchet differentiable, and $\overline{u}$ is a local solution, then (cf. [HPUU09, Theorem 1.52])

$$\overline{u} \in \mathcal{U}_{ad}, \text{ and } \langle j'(\overline{u}), w \rangle_{\mathcal{U}^*,\mathcal{U}} \geq 0 \quad \forall w \in \mathcal{T}(\mathcal{U}_{ad}, \overline{u}).$$

We also use the tangent cone for characterizing a minimizer of (5.1). We abbreviate again

$$z^* p := \langle p, z \rangle_{\mathcal{Z}^*,\mathcal{Z}}, \text{ for } z \in \mathcal{Z} \text{ and } p \in \mathcal{Z}^*.$$

There holds the following standard necessary optimality condition.

Theorem 5.2. *[KU15, Theorem 8] Let Assumption 5.1 and Assumption 5.2 hold, and let $(\overline{u}, \overline{y}) \in \mathcal{U} \times \mathcal{Y}$ be a local solution of (5.1). Then there exists a unique adjoint state $\overline{p} \in \mathcal{Z}^*$, and a Lagrange multiplier $\overline{\lambda} \in \mathcal{U}^*$, such that the following optimality conditions hold*

$$\begin{aligned} J_U(T(\overline{u}), \overline{y})T'(\overline{u}) + \beta u + \left(E_U(T(\overline{u}), \overline{y})T'(\overline{u})\right)^* \overline{p} + \overline{\lambda} = &\ 0 \text{ in } \mathcal{U}^*, \\ J_y(T(\overline{u}), \overline{y}) + E_y(T(\overline{u}), \overline{y})^* \overline{p} = &\ 0 \text{ in } \mathcal{Y}^*, \\ E(T(\overline{u}), \overline{y}) = &\ 0 \text{ in } \mathcal{Z}, \\ \overline{u} \in &\ \mathcal{U}_{ad}, \\ \overline{\lambda} \in &\ \mathcal{T}(\mathcal{U}_{ad}, \overline{u})^\circ, \end{aligned} \tag{5.4}$$

where $\mathcal{T}(\mathcal{U}_{ad}, \overline{u})^\circ$ is the polar cone of $\mathcal{T}(\mathcal{U}_{ad}, \overline{u})$.

Proof. Assumption 5.2 implies surjectivity of $E_y(T(u), S(T(u)))$ for all $u \in \mathcal{U}_{ad}$. Hence *Robinson's constraint qualification* holds, cf. [HPUU09, Lemma 1.14], and the system can be derived in a standard way, compare for example [HPUU09, Section 1.7]. The uniqueness of $\overline{p}$ follows from Assumption 5.2 as well. □

If one can choose $\mathcal{U}$ as an L^p-space, and if the projection $P_{\mathcal{U}_{ad}}$ can be written as a superposition operator, the problem is very well understood. In particular, there exist efficient semismooth Newton-type algorithms. The situation changes if $\mathcal{U}$ is required to be a stronger space, e.g., H^k. In this case the semismoothness of the projection in $\mathcal{U}$ is an open question, and the Lagrange multiplier $\overline{\lambda} \in \mathcal{U}^*$ has a-priori only a low regularity. One possibility to overcome this difficulty is to study a sequence of regularized problems. But first let us specify $\mathcal{U}$ and the admissible set $\mathcal{U}_{ad}$ in more detail.

Assumption 5.3. *$\Gamma_B \subset \mathbb{R}^d$ is a C^1-manifold. $\mathcal{U}$ is a Hilbert space of functions $u\colon \Gamma_B \to \mathbb{R}^d$, with compact embedding $\mathcal{U} \hookrightarrow_c L^2(\Gamma_B, \mathbb{R}^d)$. Furthermore, $C \subset \mathbb{R}^d$ is a nonempty, closed, convex set. Finally, the admissible set is given by*

$$\mathcal{U}_{ad} := \{u \in \mathcal{U} \mid x + u(x) \in C \text{ for a.e. } x \in \Gamma_B\}.$$

We will often use the abbreviation $\tau := \mathrm{id} + u$, hence $\mathcal{U}_{ad} = \{u \in \mathcal{U} \mid \tau(\Gamma_B) \subset C\}$.

Remark 5.3. (i) Clearly $\mathcal{U}_{ad}$ is closed and convex. The same holds for its L^2-relaxation

$$\mathcal{U}_{ad}^L := \{u \in L^2(\Gamma_B, \mathbb{R}^d) \mid \tau(x) \subset C \text{ for a.e. } x \in \Gamma_B\}.$$

(ii) We require a compact embedding $\mathcal{U} \hookrightarrow_c L^2(\Gamma_B, \mathbb{R}^d)$. On the one hand this is obviously a restriction. On the other hand in many applications $\mathcal{U}$ is either an L^p-space where standard semismooth Newton methods are applicable, or some Sobolev space $W^{k,p}$, $k \geq 1$, which is compactly embedded into L^2.

(iii) We specified here $u\colon \Gamma_B \to \mathbb{R}^d$ and $\tau = \mathrm{id} + u$. However, we suspect that it is possible to extend the following analysis to other settings. In particular, one might exchange Γ_B with some other (sub)set of Ω, or consider functions $u\colon \Gamma_B \to \mathbb{R}^m$, $m \neq d$.

5.2. The regularized problem

As we pointed out in the last section, it is difficult to treat the optimal control problem (P) with the constraint $u \in \mathcal{U}_{ad}$, which can be written equivalently as

$$\min_{u \in \mathcal{U}} j(u) + \iota_{\mathcal{U}_{ad}}(u). \tag{5.5}$$

Recall the notion of the indicator function of some set A

$$\iota_A(v) := \begin{cases} 0 & \text{if } v \in A, \\ \infty & \text{if } v \notin A. \end{cases}$$

The indicator function of a closed, convex, nonempty set is proper, lower semicontinuous, and convex. Instead of the hard constraint $u \in \mathcal{U}_{ad}$, one can try to satisfy this constraint only approximately, and then drive the constraint violation to zero in an iterative scheme. Such infeasible methods usually rely on some functional which measures the constraint violation, and which is used in the role of $\iota_{\mathcal{U}_{ad}}(u)$. It is well known that the *Moreau envelope* (cf., e.g., [BC11, CW05]) of $\iota_{\mathcal{U}_{ad}}(\cdot)\colon \mathcal{U} \to [0,\infty]$ in $\mathcal{U}$ is given by $\frac{1}{2}(d_{\mathcal{U}_{ad}})^2$ where $d_{\mathcal{U}_{ad}}\colon \mathcal{U} \to \mathbb{R}$ denotes the distance functional to $\mathcal{U}_{ad}$ w.r.t. $\|\cdot\|_{\mathcal{U}}$. The *proximity operator* of $\iota_{\mathcal{U}_{ad}}(\cdot)$ corresponds to the projection $P_{\mathcal{U}_{ad}}$ in $\mathcal{U}$ onto $\mathcal{U}_{ad}$. In particular, it holds

$$d_{\mathcal{U}_{ad}}(u) = \|u - P_{\mathcal{U}_{ad}}(u)\|_{\mathcal{U}}.$$

Recall the L^2-relaxation $\mathcal{U}_{ad}^L$ of $\mathcal{U}_{ad}$. Since $u \in \mathcal{U} \cap \mathcal{U}_{ad}^L$ implies $u \in \mathcal{U}_{ad}$, we propose to use the Moreau envelope of $\iota_{\mathcal{U}_{ad}^L}(\cdot)$ in $L^2(\Gamma_B, \mathbb{R}^d)$ as regularization term. Note, that the projection in $L^2(\Gamma_B, \mathbb{R}^d)$ onto $\{v \in L^2(\Gamma_B, \mathbb{R}^d) \mid v(\Gamma_B) \subset C\}$ is given by the superposition operator

$$P_C\colon P_C(v)(x) := \tilde{P}_C(v(x)),$$

where $\tilde{P}_C\colon \mathbb{R}^d \to \mathbb{R}^d$ is the projection onto $C \subset \mathbb{R}^d$. We approximate (5.5) by the regularized problem

$$\min_{u \in \mathcal{U}} j(u) + \frac{\gamma}{2} \left\| \mathrm{id} + u - P_C(\mathrm{id} + u) \right\|^2_{L^2(\Gamma_B, \mathbb{R}^d)}, \tag{MY}$$

and abbreviate

$$\begin{aligned} \delta(u) &:= \mathrm{id} + u - P_C(\mathrm{id} + u), \\ e_\gamma(u) &:= \frac{\gamma}{2} \left\| \mathrm{id} + u - P_C(\mathrm{id} + u) \right\|^2_{L^2(\Gamma_B, \mathbb{R}^d)}, \\ j_\gamma(u) &:= j(u) + e_\gamma(u). \end{aligned}$$

As we will see now the term $\frac{\gamma}{2} \left\| \mathrm{id} + u - P_C(\mathrm{id} + u) \right\|^2_{L^2(\Gamma_B, \mathbb{R}^d)} = \frac{\gamma}{2} \left\| \tau - P_C(\tau) \right\|^2_{L^2(\Gamma_B, \mathbb{R}^d)}$ is the Moreau envelope of $\iota_{\mathcal{U}_{ad}^L}(\cdot)$ in $L^2(\Gamma_B, \mathbb{R}^d)$ with regularization parameter γ.

Lemma 5.4. *[KU15, Lemma 10] The projection in $L^2(\Gamma_B, \mathbb{R}^d)$ onto the closed, convex set $\mathcal{U}_{ad}^L \subset L^2(\Gamma_B, \mathbb{R}^d)$ is given by the mapping*

$$u \mapsto u^C := P_C(\mathrm{id} + u) - \mathrm{id}.$$

Proof. Recall $\tau = \mathrm{id} + u$. By construction we have $u^C = P_C(\tau) - \mathrm{id} \in \mathcal{U}_{ad}^L$. It remains to check whether

$$\left(u - u^C, v - u^C\right)_{L^2(\Gamma_B, \mathbb{R}^d)} \leq 0 \text{ holds for all } v \in \mathcal{U}_{ad}^L.$$

Let $v \in \mathcal{U}_{ad}^L$ and $x \in \Gamma_B$. Since $\tilde{P}_C$ is the projection onto C in $\mathbb{R}^d$ and $v(x) + x \in C$ we have

$$\left(\tau(x) - \tilde{P}_C(\tau(x))\right)^T \left(v(x) + x - \tilde{P}_C(\tau(x))\right) \leq 0.$$

Thus, it holds

$$\begin{aligned}\left(u - u^C, v - u^C\right)_{L^2(\Gamma_B, \mathbb{R}^d)} &= \int_{\Gamma_B} (u - u^C)^T (v - u^C)\,\mathrm{d}x \\ &= \int_{\Gamma_B} (\tau - P_C(\tau))^T (v + \mathrm{id} - P_C(\tau))\,\mathrm{d}x \leq 0.\end{aligned}$$

□

Corollary 5.5. *For all $\gamma > 0$ the Moreau envelope $e_\gamma(\cdot)\colon L^2(\Gamma_B, \mathbb{R}^d) \to \mathbb{R}$ is convex and Fréchet differentiable. Its derivative is given by*

$$e_\gamma'(u) = \gamma(\tau - P_C(\tau)) \in L^2(\Gamma_B, \mathbb{R}^d) \simeq L^2(\Gamma_B, \mathbb{R}^d)^*,$$

and is Lipschitz continuous.

Proof. These are general properties of Moreau envelopes, cf. [BC11, Propositions 12.15 and 12.29]. □

Remark 5.6. Noting that $|x - \tilde{P}_C(x)|^2 = d_C^2(x)$ where d_C is the distance function of the set C the result can also be obtained by a more geometric argumentation. In particular, [DZ11, Theorem 6.8.1] states that d_C is convex if and only if $\overline{C}$ is convex and in this case it holds $d_C^2 \in C^{1,1}_{\mathrm{loc}}(\mathbb{R}^d)$.

If j is differentiable, the chain rule and Corollary 5.5 yield for all $v \in \mathcal{U}$

$$\langle j_\gamma'(u), v\rangle_{\mathcal{U}^*,\mathcal{U}} = \langle j'(u), v\rangle_{\mathcal{U}^*,\mathcal{U}} + \gamma\left(\tau - P_C(\tau), v\right)_{L^2(\Gamma_B, \mathbb{R}^d)}.$$

Finally, we suppose that the constraint $u \in \mathcal{U}_{ad}$ is not trivially satisfied, and that the reference configuration is admissible.

Assumption 5.4. *It holds*

1. *If $\overline{u}$ solves* (P) *it is* not *a solution of the unconstrained problem* $\min_{u \in \mathcal{U}} j(u)$.
2. $u = 0 \in \mathcal{U}_{ad}$.

We collect our working assumptions for ease of reference.

Assumption 5.5. *Assumption 5.1.1.-4., Assumption 5.2, Assumption 5.3, and Assumption 5.4 are satisfied.*

Remark 5.7. It is clear from the above analysis, that (P) admits a global solution if Assumption 5.5 is satisfied. The same holds for the regularized problem (MY) for any $\gamma \geq 0$.

Let us furthermore list some notational conventions used in the following.

- We write $\overline{\tau}$ for $\mathrm{id} + \overline{u}$ and similarly $\tau^*, \hat{\tau}, \tau_n$ for $u^*, \hat{u}, u_n$.
- We write $(\mathrm{MY})_{\overline{\gamma}}$ for the problem (MY) with $\gamma = \overline{\gamma}$ and similarly $(\mathrm{MY})_{\gamma^*}, (\mathrm{MY})_{\hat{\gamma}}, (\mathrm{MY})_{\gamma_n}$.
- We denote by $\overline{u} \in \mathcal{U}_{ad}$ a local solution of $(\mathrm{MY})_{\overline{\gamma}}$, u_n is a local solution of $(\mathrm{MY})_{\gamma_n}$, etc.

5.3. Properties of the regularized solutions

In this section we show that any strict local solution of (P) is a strong accumulation point of a sequence of local solutions of $(\mathrm{MY})_\gamma$ for $\gamma \to \infty$. This convergence result and its proof are inspired by the ideas presented in [NT08] and go back to [CT02]. The same ideas are used in [MY09] to obtain similar results. The result is easily extended to show that any weakly convergent sequence of global solutions of $(\mathrm{MY})_\gamma$ converges strongly towards a global solution of (P). Furthermore, we show that any accumulation point of a sequence of local solutions (u_γ), with $\gamma \to \hat{\gamma}$, is a local solution of $(\mathrm{MY})_{\hat{\gamma}}$.

We begin by observing that $\|\delta(u_\gamma)\|_{L^2(\Gamma_B,\mathbb{R}^d)} \to 0$ for $\gamma \to \infty$ if j is locally Lipschitz continuous.

Lemma 5.8. *[KU15, Lemma 3] Let j be locally Lipschitz continuous. For all $\overline{u} \notin \mathcal{U}_{ad}$ there exists a $\gamma_0 > 0$, such that $\overline{u}$ is* not *a local solution of $(\mathrm{MY})_\gamma$ for all $\gamma > \gamma_0$.*

Proof. Consider an arbitrary $\overline{u} \notin \mathcal{U}_{ad}$ and an $0 < \varepsilon < 1$. By assumption there exists a local Lipschitz constant $L > 0$ of j on the ball $B^{\mathcal{U}}(\overline{u}, \varepsilon)$. Let $v = P_{\mathcal{U}_{ad}}(\overline{u})$ and consider the convex combination $u = (1 - \frac{\varepsilon}{2})\overline{u} + \frac{\varepsilon}{2} v$. Then $u \in B^{\mathcal{U}}(\overline{u}, \varepsilon)$, and by convexity (see Corollary 5.5)

$$\|\delta(u)\|^2_{L^2(\Gamma_B,\mathbb{R}^d)} \leq (1 - \frac{\varepsilon}{2}) \|\delta(\overline{u})\|^2_{L^2(\Gamma_B,\mathbb{R}^d)} + \frac{\varepsilon}{2} \|\delta(v)\|^2_{L^2(\Gamma_B,\mathbb{R}^d)} = (1 - \frac{\varepsilon}{2}) \|\delta(\overline{u})\|^2_{L^2(\Gamma_B,\mathbb{R}^d)},$$

since $v \in \mathcal{U}_{ad}$. Now we choose $\gamma > \gamma_0 := 4L \|\delta(\overline{u})\|^{-2}_{L^2(\Gamma_B,\mathbb{R}^d)}$ and calculate

$$\begin{aligned} j_\gamma(u) - j_\gamma(\overline{u}) &= j(u) - j(\overline{u}) + \frac{\gamma}{2} \left(\|\delta(u)\|^2_{L^2(\Gamma_B,\mathbb{R}^d)} - \|\delta(\overline{u})\|^2_{L^2(\Gamma_B,\mathbb{R}^d)} \right) \\ &< L \|u - \overline{u}\|_{\mathcal{U}} + 2L \|\delta(\overline{u})\|^{-2}_{L^2(\Gamma_B,\mathbb{R}^d)} \left(-\frac{\varepsilon}{2} \|\delta(\overline{u})\|^2_{L^2(\Gamma_B,\mathbb{R}^d)} \right) \\ &\leq L\varepsilon - L\varepsilon = 0. \end{aligned}$$

Hence for any $\overline{u} \notin \mathcal{U}_{ad}$ we find a γ_0 such that for all $\gamma > \gamma_0$, $\overline{u}$ is not a local minimum of $(\mathrm{MY})_\gamma$. □

Remark 5.9. In particular, for a sequence of local solutions (u_γ) of $(\mathrm{MY})_\gamma$, this implies that $\|\delta(u_\gamma)\|_{L^2(\Gamma_B,\mathbb{R}^d)} \to 0$ for $\gamma \to \infty$, if j is locally Lipschitz continuous.

We proceed with the derivation of the announced result regarding *local* solutions. Adopting an idea from [CT02, NT08] we consider the following auxiliary problem. For a $\overline{u} \in \mathcal{U}_{ad}$ and $r > 0$ consider

$$\min_{u\in\mathcal{U}} j_\gamma(u) \quad \text{s.t. } u \in \overline{B^{\mathcal{U}}(\overline{u}, r)}. \tag{5.6}$$

Since $\overline{B^{\mathcal{U}}(\overline{u}, r)}$ is convex, closed, bounded, and non-empty there exists at least one (global) solution $u_\gamma^r \in \mathcal{U}$ of (5.6). We begin by studying the properties of u_γ^r for $\gamma \to \infty$.

Lemma 5.10. *[KU15, Lemma 4] Let Assumption 5.5 hold and $(\gamma_n) \subset \mathbb{R}_{>0}$ tend to infinity. Furthermore, let $\overline{u} \in \mathcal{U}_{ad}$ be a local solution of* (P) *on $B^{\mathcal{U}}(\overline{u}, \delta)$. Consider a sequence of global solutions (u_n^r) of* (5.6) *with $\gamma = \gamma_n$ and $r = \delta/2$. Then there exists a weak accumulation point of u_n^r. Moreover, any weakly convergent subsequence $u_k^r \rightharpoonup u^* \in \mathcal{U}$ converges strongly, and $u^* \in \mathcal{U}_{ad}$ is a local solution of* (P).

Proof. (i) The sequence $(u_n^r) \subset \overline{B^{\mathcal{U}}(\overline{u}, r)}$ is bounded, hence there exists a weakly convergent subsequence.

(ii) Now consider an arbitrary weakly convergent subsequence $u_k^r \rightharpoonup u^* \in \overline{B^{\mathcal{U}}(\overline{u}, r)}$. We first argue that $u^* \in \mathcal{U}_{ad}$. If $\|\delta(u_k^r)\|_{L^2(\Gamma_B,\mathbb{R}^d)} \not\to 0$, then we find arbitrarily large γ_k with $u_k^r \notin \mathcal{U}_{ad}$. Note that

$$u := (1 - \frac{\varepsilon}{2})u_k^r + \frac{\varepsilon}{2} P_{\mathcal{U}_{ad}}(u_k^r) \in \overline{B^{\mathcal{U}}(\overline{u}, r)}$$

since $\overline{u} \in \mathcal{U}_{ad}$ and $\mathcal{U}_{ad}$ is convex. Copying the argument of Lemma 5.8 we conclude $j_{\gamma_k}(u) < j_{\gamma_k}(u_k^r)$ for γ_k large enough which contradicts the optimality of u_k^r. Hence it holds $\|\delta(u_k^r)\|_{L^2(\Gamma_B,\mathbb{R}^d)} \to 0$. Since $\mathcal{U} \hookrightarrow_c L^2(\Gamma_B, \mathbb{R}^d)$ we conclude $\|\delta(u^*)\|_{L^2(\Gamma_B,\mathbb{R}^d)} = 0$ and thus $u^* \in \mathcal{U}_{ad}$.

(iii) We now show that u^* is a local minimum of (P). Obviously $\overline{u} \in \mathcal{U}_{ad}$ is feasible for (5.6). From $u_k^r \rightharpoonup u^*$ in $\mathcal{U}$ and $j\colon \mathcal{U} \to \mathbb{R}$ being weakly lower semicontinuous we conclude

$$j(u^*) \le \liminf_{k\to\infty} j(u_k^r) \le \liminf_{k\to\infty} j_{\gamma_k}(u_k^r) \le \limsup_{k\to\infty} j_{\gamma_k}(u_k^r) \le \limsup_{k\to\infty} j_{\gamma_k}(\overline{u}) = j(\overline{u}). \tag{5.7}$$

We used the optimality of u_k^r in the last inequality. By $u^* \in \overline{B^{\mathcal{U}}(\overline{u}, \frac{\delta}{2})} \cap \mathcal{U}_{ad}$ we also have $j(\overline{u}) \le j(u^*)$ which implies $j(\overline{u}) = j(u^*)$. By assumption it holds

$$j(u) \ge j(\overline{u}) = j(u^*),\ \forall u \in \mathcal{U}_{ad} \cap \overline{B^{\mathcal{U}}(\overline{u}, \delta)},$$

and by construction $u^* \in \overline{B^{\mathcal{U}}(\overline{u}, \frac{\delta}{2})} \subset B^{\mathcal{U}}(\overline{u}, \delta)$. We conclude that u^* is a local minimum of (P).

(iv) Finally, we address the strong convergence of (u_k^r). Due to the optimality of u_k^r it holds $j(u_k^r) \leq j(\overline{u}) = j(u^*)$ for all k. Combined with the lower semicontinuity of j, i.e., $\liminf_{k\to\infty} j(u_k^r) \geq j(u^*)$ we obtain

$$\lim_{k\to\infty} j(u_k^r) = j(u^*) = J(T(u^*), S(T(u^*))) + \frac{\beta}{2} \|u^*\|_{\mathcal{U}}^2 .$$

On the other hand, T is completely continuous, S, J are continuous, hence $u_k^r \rightharpoonup u^*$ implies

$$J(T(u_k^r), S(T(u_k^r))) \to J(T(u^*), S(T(u^*))).$$

We conclude $\|u_k^r\|_{\mathcal{U}} \to \|u^*\|_{\mathcal{U}}$. Weak convergence plus convergence in the norm imply the strong convergence $u_k^r \to u^*$ in $\mathcal{U}$. □

The above result is not quite satisfactory in two aspects. For once, it might be that the global solutions u_k^r are situated on the boundary of the auxiliary admissible set. In this case we can not infer whether they are also local solutions of $(\mathrm{MY})_{\gamma_k}$. Secondly, we would like to obtain convergence to $\overline{u}$, and not to some other nearby local solution. As the next theorem shows it suffices to require that $\overline{u}$ is a *strict* local solution to address both points. Of course, we could require some sufficient second order condition in $\overline{u}$ to guarantee this.

Theorem 5.11. *[KU15, Theorem 1] Let Assumption 5.5 hold and $\overline{u} \in \mathcal{U}_{ad}$ be a* strict *local solution of* (P) *on $B^{\mathcal{U}}(\overline{u}, \delta)$. Then for every $\gamma_n \to \infty$, every sequence of global solutions $(u_n^r) \subset \mathcal{U}$ of* (5.6) *with $\gamma = \gamma_n$ and $r < \delta$, converges strongly in $\mathcal{U}$ to $\overline{u}$. Furthermore, there exists a $\hat{n} > 0$, such that for all $n \geq \hat{n}$ the u_n^r are local solutions of* $(\mathrm{MY})_{\gamma_n}$.

Proof. In Lemma 5.10 we showed that for every sequence (u_n^r) there exists a weakly convergent subsequence, and that every such subsequence converges to a local solution u^* of (P). In particular, we proved $j(u^*) = j(\overline{u})$. Since $\overline{u}$ is a strict local minimum this implies $u^* = \overline{u}$. Hence, $\overline{u}$ is the only weak accumulation point of the bounded sequence (u_n^r), which implies that the whole sequence converges weakly $u_n^r \rightharpoonup \overline{u}$ in $\mathcal{U}$. Lemma 5.10 implies that the convergence is strong. Finally, since $u_n^r \to \overline{u}$, there exists $\hat{n} > 0$ such that for all $n \geq \hat{n}$ it holds $u_n^r \in B^{\mathcal{U}}(\overline{u}, r)$. Hence, the u_n^r are local solutions of $(\mathrm{MY})_{\gamma_n}$. □

Let us briefly consider global solutions $\overline{u}^g, u_\gamma^g$ of (P), and $(\mathrm{MY})_\gamma$. Due to the coercivity of j we can find an r large enough such that $u_\gamma^g \in \overline{B^{\mathcal{U}}(\overline{u}^g, r)}$, in particular $u_\gamma^r = u_\gamma^g$, for all $\gamma > 0$. Hence Lemma 5.10 implies

Corollary 5.12. *[KU15, Corollary 2] Let Assumption 5.5 hold and $(\gamma_n) \subset \mathbb{R}_{>0}$ tend to infinity. Then there exists a weakly convergent subsequence of (u_n^g). Furthermore, any weakly convergent subsequence $u_k^g \rightharpoonup u^* \in \mathcal{U}$ converges strongly, and u^* is a global solution of* (P).

We show now that any accumulation point of a sequence of local solutions (u_γ), with $\gamma \to \hat{\gamma}$, is a local solution of $(\mathrm{MY})_{\hat{\gamma}}$. Consider the auxiliary problem

$$\min_{u\in\mathcal{U}} \quad j_\gamma(u), \qquad \text{s.t. } u \in A, \tag{5.8}$$

for some fixed, closed, convex, and nonempty set $A \subset \mathcal{U}$.

Theorem 5.13. *[KU15, Theorem 2] Let Assumption 5.5 hold, $\hat{\gamma} > 0$, and consider for $\gamma_n \to \hat{\gamma}$ a sequence of global solutions (u_n) of (5.8) with $\gamma = \gamma_n$. Then there exists a weakly convergent subsequence (u_k). Furthermore, any weakly convergent subsequence $u_k \rightharpoonup u^* \in \mathcal{U}$ converges strongly, and u^* solves (5.8) with $\gamma = \hat{\gamma}$.*

Proof. (i) Since j is coercive the sequence (u_n) is bounded. Hence there exists a weakly convergent subsequence.

(ii) Now let (u_k) be a weakly convergent subsequence $u_k \rightharpoonup u^*$ for some $u^* \in \mathcal{U}$. Since A is weakly closed it holds $u^* \in A$. We start by showing that u^* solves $(5.8)_{\hat{\gamma}}$. Denote by $\hat{u} := u_{\hat{\gamma}}$ a global solution of $(5.8)_{\hat{\gamma}}$. Since $u_k \rightharpoonup u^*$ in $\mathcal{U}$ we have by compact embedding $u_k \to u^*$ in $L^2(\Gamma_B, \mathbb{R}^d)$. Hence, continuity of the Moreau envelope and lower semicontinuity of j imply

$$j(u^*) \leq \liminf_{k\to\infty} j(u_k) \text{ and } e_{\gamma_k}(u_k) \to e_{\hat{\gamma}}(u^*).$$

Furthermore, for all k it holds

$$j(u_k) + e_{\gamma_k}(u_k) = j_{\gamma_k}(u_k) \leq j_{\gamma_k}(\hat{u}) = j(\hat{u}) + e_{\gamma_k}(\hat{u}).$$

Thus we see that

$$j_{\hat{\gamma}}(u^*) = j(u^*) + e_{\hat{\gamma}}(u^*) \leq \liminf_{k\to\infty} j(u_k) + e_{\gamma_k}(u_k) \leq \liminf_{k\to\infty} j(\hat{u}) + e_{\gamma_k}(\hat{u}) = j_{\hat{\gamma}}(\hat{u}),$$

which implies that u^* solves $(5.8)_{\hat{\gamma}}$.

(iii) The strong convergence $u_k \to u^*$ follows as in the proof of Lemma 5.10 part (iv). □

Remark 5.14. Obviously $A = \mathcal{U}$ is possible and yields the result for global solutions of (MY).

As in the case $\gamma \to \infty$ the result from Theorem 5.13 alone does not provide enough information about the behavior of local solutions of (MY). Again this is remedied by considering a strict local solution $u_{\hat{\gamma}}$ of $(\mathrm{MY})_{\hat{\gamma}}$, which might be guaranteed by a sufficient second order condition.

Corollary 5.15. *[KU15, Corollary 3] For $\hat{\gamma} > 0$ let Assumption 5.5 be satisfied. Denote by $\hat{u}$ a strict local solution of $(\mathrm{MY})_{\hat{\gamma}}$ on $B^{\mathcal{U}}(\overline{u}, \delta)$. Set $0 < r < \delta$ and $A = \overline{B^{\mathcal{U}}(\hat{u}, r)}$. Then, for any $\gamma_n \to \hat{\gamma}$, any sequence of global solutions (u_n^r) of (5.8) with $\gamma = \gamma_n$ converges strongly in $\mathcal{U}$ to $\hat{u}$ and, for γ_n close enough to $\hat{\gamma}$, the u_n^r are local solutions of $(\mathrm{MY})_{\gamma_n}$.*

Proof. Using Theorem 5.13 we obtain a subsequence (γ_k) with $u_k^r \to u^*$ in $\mathcal{U}$, where u^* solves (5.8) with $\gamma = \hat{\gamma}$. Furthermore, any weakly convergent subsequence converges towards such a solution. Since $\hat{u}$ is a strict local solution of $(\mathrm{MY})_{\hat{\gamma}}$ and $r < \delta$ it follows, that $\hat{u}$ is the unique solution of (5.8) with $\gamma = \hat{\gamma}$, which implies $u^* = \hat{u}$. Hence, $\hat{u}$ is the only weak accumulation point of the bounded sequence u_n^r, therefore the whole sequence converges weakly: $u_n^r \rightharpoonup \hat{u}$. Theorem 5.13 shows that the convergence is strong. Finally for γ_n close enough to $\hat{\gamma}$ we have $u_n^r \in B^{\mathcal{U}}(\hat{u}, r)$, since $u_n^r \to \hat{u}$. Hence the u_n^r are local solutions of $(\mathrm{MY})_{\gamma_n}$. □

We established two important properties of the family of regularized problems (MY): We can approximate any strict local solution of (P) with a sequence of local solutions u_γ of $(\mathrm{MY})_\gamma$ for $\gamma \to \infty$. Furthermore, we have a continuity property of u_γ in the sense, that a strict local solution of $(\mathrm{MY})_{\hat\gamma}$ for fixed $\hat\gamma > 0$ can be approximated by a sequence of local solutions u_γ of $(\mathrm{MY})_\gamma$ with $\gamma \to \hat\gamma$.

5.4. Solving the regularized problem

In Section 5.3 we showed that a strict local solution of (P) can be found by solving a sequence of relaxed problems of the form $(\mathrm{MY})_{\gamma_n}$ with regularization parameter γ_n tending to infinity. Of course, this is only a practical strategy if the regularized problems can be solved efficiently. We will apply a *semismooth Newton method* to solve the first order optimality condition:

$$j'_{\gamma_n}(u_n) = 0 \quad \text{in } \mathcal{U}^*. \tag{5.9}$$

Remember $\langle j'_\gamma(u), v\rangle_{\mathcal{U}^*,\mathcal{U}} = \langle j'(u), v\rangle_{\mathcal{U}^*,\mathcal{U}} + \gamma\,(\delta(u), v)_{L^2(\Gamma_B,\mathbb{R}^d)}$, with $\delta(u) = \tau - P_C(\tau)$. The superposition operator P_C is not differentiable but *semismooth*. For a thorough introduction to semismoothness in Banach spaces we refer to the monograph [Ulb11]. In particular semismoothness of superposition operators is discussed, see also [Sch08], or [HPUU09, Chapter 2] for a compact overview.

We start by defining the generalized differential of $P_C : \mathcal{U} \to L^q(\Gamma_B, \mathbb{R}^d)$, $q \geq 1$. We need

Assumption 5.6. *The projection $\tilde P_C\colon \mathbb{R}^d \to \mathbb{R}^d$ is $\partial\tilde P_C$-semismooth, where $\partial\tilde P_C$ denotes Clarke's generalized Jacobian [Cla83].*

Remark 5.16. This is a condition on the set $C \subset \mathbb{R}^d$. In general, the projection $\tilde P_C$ is not even directionally differentiable. See [Sha13] for conditions guaranteeing directional differentiability. If the set C is of the form $C = \{x \in \mathbb{R}^d \mid g(x) \leq 0\}$, where $g\colon \mathbb{R}^d \to \mathbb{R}^m$, $g_i \in C^2(\mathbb{R}^d, \mathbb{R})$ is convex for all i, and the *constant rank constraint qualification* is satisfied, then [FP03, Theorem 4.5.2] states that the projection is piecewise smooth, in particular semismooth. An alternative, very general approach to finite-dimensional semismoothness is the concept of *tameness*, cf. [BDL09].

Definition 5.17. *For $q \geq 1$ we introduce the set-valued mapping $\partial P_C\colon \mathcal{U} \rightrightarrows \mathcal{L}(\mathcal{U}, L^q(\Gamma_B, \mathbb{R}^d))$,*

$$\partial P_C(w) := \left\{ M \in \mathcal{L}(\mathcal{U}, L^q(\Gamma_B, \mathbb{R}^d)) \;\middle|\; \begin{array}{l} Mv(x) = K(x)v(x)\ \forall x \in \Gamma_B, \text{with} \\ K \in L^\infty(\Gamma_B, \mathbb{R}^{d\times d}),\ K(x) \in \partial\tilde P_C(w(x))\ \forall x \in \Gamma_B \end{array} \right\}.$$

Theorem 5.18. *[KU15, Theorem 3] Let Assumption 5.6 hold, and $\mathcal{U} \hookrightarrow L^p(\Gamma_B, \mathbb{R}^d)$ for $p > q \geq 1$. Then ∂P_C is well defined. Furthermore, $P_C : \mathcal{U} \to L^q(\Gamma_B, \mathbb{R}^d)$, is ∂P_C-semismooth.*

Proof. The first statement is immediate. Regarding the claimed semismoothness, it suffices to study the components $\tilde P_C^i\colon \mathbb{R}^d \to \mathbb{R}$ and the associated superposition operator P_C^i, cf. [Ulb11, Proposition 3.6]. The result [HPUU09, Theorem 2.13] states sufficient conditions for semismoothness of superposition operators. Let us check the conditions. The Lipschitz

continuity of the projection $\tilde{P}_C\colon \mathbb{R}^d \to \mathbb{R}^d$, for $C \subset \mathbb{R}^d$ closed and convex, is well known. Semismoothness of $\tilde{P}_C^i$ follows from Assumption 5.6. Furthermore, since $\mathcal{U} \hookrightarrow L^p(\Gamma_B, \mathbb{R}^d)$, the mapping $\mathcal{U} \ni u \mapsto \mathrm{id} + u \in L^p(\Gamma_B, \mathbb{R}^d)$ is continuously Fréchet-differentiable and Lipschitz continuous. Hence, we are able to employ [HPUU09, Theorem 2.13] and the claim follows. □

Thus, if we want to solve (5.9) with a Newton-type method, we need to employ semismooth calculus. The chain rule implies that j_γ' is $\partial j_\gamma'$-semismooth, where $\partial j_\gamma'\colon \mathcal{U} \rightrightarrows \mathcal{L}(\mathcal{U}, \mathcal{U}^*)$ and

$$H_\gamma^u \in \partial j_\gamma'(u) \Leftrightarrow \exists M^u \in \partial P_C(\mathrm{id} + u)\colon$$
$$\langle H_\gamma^u v, w\rangle_{\mathcal{U}^*,\mathcal{U}} = \langle j''(u)v, w\rangle_{\mathcal{U}^*,\mathcal{U}} + \gamma\left(v - M^u v, w\right)_{L^2(\Gamma_B,\mathbb{R}^d)},\ \forall v, w \in \mathcal{U}.$$

We are now equipped to solve (5.9). Consider Algorithm 5.1.

Algorithm 5.1: Semismooth Newton's method

Require: let Assumptions 5.5 and 5.6 be satisfied, and $u^0 \in \mathcal{U}$ be given
1: set the iteration index to $k = 0$
2: **repeat**
3: choose a $H_\gamma^{u^k} \in \partial j_\gamma'(u^k)$
4: solve the semismooth Newton equation $H_\gamma^{u^k} v = -j_\gamma'(u^k)$ in $\mathcal{U}^*$
5: set $u^{k+1} = u^k + v$,
6: increment k
7: **until** $u^{k+1} = u^k$

For the well-posedness and superlinear convergence of the semismooth Newton method one may require a uniform regularity condition like $\exists\, c, r > 0$ such that

$$\left\|(H_\gamma^u)^{-1}\right\|_{\mathcal{L}(\mathcal{U},\mathcal{U}^*)} \le c \quad \forall H_\gamma^u \in \partial j_\gamma'(u), \quad \forall u \in B^{\mathcal{U}}(u_\gamma, r). \tag{5.10}$$

The following strong assumption assures this regularity condition in a point $u \in \mathcal{U}$:

Assumption 5.7. *For $\gamma > 0$ and $u \in \mathcal{U}$ there exist $\tilde{\alpha}, r > 0$:*

$$\langle H_\gamma^w v, v\rangle_{\mathcal{U}^*,\mathcal{U}} \ge \tilde{\alpha}\,\|v\|_{\mathcal{U}}^2 \quad \forall v \in \mathcal{U},\ \forall H_\gamma^w \in \partial j_\gamma'(w), \quad \forall w \in B^{\mathcal{U}}(u, r).$$

Remark 5.19. In fact, already (5.10) is a stronger assumption than necessary. It would suffice if, in every iteration, there exists a solution of the semismooth Newton equation for the choice $H_\gamma^{u^k}$, and if

$$\left\|\left(H_\gamma^{u^k}\right)^{-1}\left(j_\gamma'(u^k) - j_\gamma'(u_\gamma) - H_\gamma^{u^k}(u^k - u_\gamma)\right)\right\|_{\mathcal{U}} = o(\left\|u^k - u_\gamma\right\|_{\mathcal{U}}).$$

However, guaranteeing these conditions without a strong assumption like Assumption 5.7 is challenging.

Theorem 5.20. *[KU15, Theorem 4] Let Assumptions 5.5 and 5.6 hold, and let Assumption 5.7 be satisfied for $\gamma > 0$ and a local solution u_γ of $(\mathrm{MY})_\gamma$. Then there exists an $r > 0$ such that for all initial points*

$$u^0 \in \mathcal{U} \quad \textit{with} \quad \left\| u^0 - u_\gamma \right\|_{\mathcal{U}} < r,$$

the semismooth Newton method (Algorithm 5.1) converges q-superlinearly to u_γ.

Proof. The superlinear convergence result can, for instance, be found in [HPUU09, Theorem 2.12]. The non-standard termination criterion is discussed in [Ulb11, Section 3.2.3], where semismooth Newton methods are studied in more detail. □

Usually, the following coercivity assumption suffices.

Assumption 5.8. *Let $u \in \mathcal{U}$ and suppose there exists an $\alpha > 0$ such that*

$$\langle j''(u)v, v\rangle_{\mathcal{U}^*,\mathcal{U}} \geq \alpha \left\|v\right\|_{\mathcal{U}}^2 \quad \forall v \in \mathcal{U}.$$

Theorem 5.21. *Let Assumptions 5.5 and 5.6 hold, furthermore suppose that Assumption 5.8 is satisfied for some $u \in \mathcal{U}$. Finally let $j'' \colon \mathcal{U} \to \mathcal{L}(\mathcal{U}, \mathcal{U}^*)$ be locally Lipschitz continuous in u, i.e., for some $\delta > 0$ there exists an $L > 0$, such that for all $w \in B^{\mathcal{U}}(u, \delta)$ we have*

$$\left\| j''(u)v - j''(w)v \right\|_{\mathcal{U}^*} \leq L \left\|u - w\right\|_{\mathcal{U}} \left\|v\right\|_{\mathcal{U}}.$$

Then, for all $w \in B^{\mathcal{U}}(u, r)$ with $r \leq \delta$, it holds

$$\langle j''(w)v, v\rangle_{\mathcal{U}^*,\mathcal{U}} \ \geq (\alpha - Lr) \left\|v\right\|_{\mathcal{U}}^2 \quad \forall v \in \mathcal{U}.$$

In particular, Assumption 5.7 holds in u for $r < \min(\delta, \alpha/L)$, $\tilde{\alpha} = \alpha - Lr > 0$, and any *$\gamma > 0$.*

Proof. The first claim follows directly from the Lipschitz continuity of j'' and Assumption 5.8. The second assertion follows as in [KU15, Theorem 5]. Let $\gamma > 0$ be arbitrary. The norm of Clarke's generalized Jacobian is bounded by the Lipschitz constant of the respective function. In our case, the projection $\tilde{P}_C$ has Lipschitz constant one. Hence, recalling Definition 5.17 we have for all $w \in \mathcal{U}$ and $v \in \mathcal{U}$:

$$\forall M^w \in \partial P_C(w) \colon \qquad \left\|M^w v\right\|_{L^2(\Gamma_B, \mathbb{R}^d)} \leq \left\|v\right\|_{L^2(\Gamma_B, \mathbb{R}^d)}. \tag{5.11}$$

Thus, for all $w \in B^{\mathcal{U}}(u, r)$, and any $H_\gamma^w \in \partial j_\gamma'(w)$:

$$\begin{aligned}\langle H_\gamma^w v, v\rangle_{\mathcal{U}^*,\mathcal{U}} &= \langle j''(w)v, v\rangle_{\mathcal{U}^*,\mathcal{U}} + \gamma \left(v - M^w v, v\right)_{L^2(\Gamma_B, \mathbb{R}^d)} \\ &\geq \tilde{\alpha} \left\|v\right\|_{\mathcal{U}}^2 + \gamma \left(\left\|v\right\|_{L^2(\Gamma_B, \mathbb{R}^d)}^2 - \left\|v\right\|_{L^2(\Gamma_B, \mathbb{R}^d)}^2 \right) = \tilde{\alpha} \left\|v\right\|_{\mathcal{U}}^2.\end{aligned}$$

This shows the claim. □

Of course, Assumption 5.8 is still a quite restrictive condition. If we consider it in the context of (P) it would be more natural to require the coercivity only with respect to a critical cone. Instead, we require coercivity for any $v \in \mathcal{U}$. Furthermore, the coercivity is required in the strong norm $\|\cdot\|_{\mathcal{U}}$. However, since our objective is given as $j(u) = \mathfrak{j}(T(u)) + \frac{\beta}{2}\|u\|_{\mathcal{U}}^2$, this is satisfied if the Hessian of the functional $u \mapsto \mathfrak{j}(T(u))$ is 'not too negative definite'. This depends very much on the concrete application, we refer for instance to [HK01] for an analysis of this issue for optimal control of Navier- Stokes flow. Finally, in the case of a Tikhonov-regularized objective the coercivity assumption is equivalent to a positivity assumption of j'' under certain conditions (compare [CT12, KV13]). Let us briefly discuss this in our setting. We require that the mappings $v \mapsto T'(u)v$ and $v \mapsto (T''(u)v)v$ are completely continuous. For linear extension operators this follows immediately from our requirements on T.

Assumption 5.9. *Let $u \in \mathcal{U}$ and suppose that*

$$\langle j''(u)v, v\rangle_{\mathcal{U}^*,\mathcal{U}} > 0 \quad \forall v \in \mathcal{U}\backslash\{0\}.$$

Lemma 5.22. *[KU15, Lemma 5] Let Assumption 5.5 hold. Furthermore suppose Assumption 5.9 is satisfied, and $T'(u) \in \mathcal{L}(\mathcal{U}, \mathcal{V})$ as well as $v \mapsto (T''(u)v)v$ are completely continuous. Then there exists $\alpha > 0$ such that*

$$\langle j''(u)v, v\rangle_{\mathcal{U}^*,\mathcal{U}} \geq \alpha \|v\|_{\mathcal{U}}^2, \quad \forall v \in \mathcal{U}.$$

Proof. By [CT12, Remark 2.7] the assertion is true if $j''(u)$ is a *Legendre form*. For this it needs to satisfy the following two conditions.

(i) if $v_k \rightharpoonup v$ as $k \to \infty$, then $\langle j''(u)v, v\rangle_{\mathcal{U}^*,\mathcal{U}} \leq \liminf_{n\to\infty}\langle j''(u)v_n, v_n\rangle_{\mathcal{U}^*,\mathcal{U}}$

(ii) if additionally $\langle\ j''(u)v_k, v_k\rangle_{\mathcal{U}^*,\mathcal{U}} \to \langle j''(u)v, v\rangle_{\mathcal{U}^*,\mathcal{U}}$, then $\|v - v_k\|_{\mathcal{U}} \to 0$.

Recall $j(u) = \mathfrak{j}(T(u)) + \frac{\beta}{2}\|u\|_{\mathcal{U}}^2$. The chain rule yields

$$\langle j''(u)v, v\rangle_{\mathcal{U}^*,\mathcal{U}} = \langle \mathfrak{j}''(T(u))T'(u)v, T'(u)v\rangle_{\mathcal{V}^*,\mathcal{V}} + \langle \mathfrak{j}'(T(u)), (T''(u)v)v\rangle_{\mathcal{V}^*,\mathcal{V}} + \beta\|v\|_{\mathcal{U}}^2.$$

Using the complete continuity of T' and T'' for the first two terms, and the weak lower semicontinuity of the squared Hilbert space norm $\|\cdot\|_{\mathcal{U}}^2$ condition (i) follows. Furthermore, $\langle j''(u)v_k, v_k\rangle_{\mathcal{U}^*,\mathcal{U}} \to \langle j''(u)v, v\rangle_{\mathcal{U}^*,\mathcal{U}}$ for $v_k \rightharpoonup v$ implies $\|v_k\|_{\mathcal{U}} \to \|v\|_{\mathcal{U}}$. Weak convergence plus convergence of the norm yields $\|v - v_k\|_{\mathcal{U}} \to 0$, hence (ii) is satisfied as well. □

5.5. Some implications from second order conditions

In this section we study the second order conditions of Assumptions 5.7 and 5.8 more thoroughly. In a standard manner we obtain a quadratic growth condition from Assumption 5.7, i.e., it may serve as a *sufficient second order condition*. Exploiting Assumption 5.8 we show that there exist neighborhoods of $\hat{\gamma}$ and $u_{\hat{\gamma}}$, such that the map $\gamma \mapsto u_\gamma$, restricted to those neighborhoods, is locally Lipschitz continuous.

Remark 5.23. As already mentioned, we require here a quite strong second order condition. We suspect that many results which depend on the quadratic growth condition remain true under weaker assumptions. The current status of the theory of second order conditions for optimal control problems is very nicely summarized in the recent survey [CT15]. It would also be very interesting to derive *necessary second order optimality conditions* for a local solution u_γ of $(\mathrm{MY})_\gamma$. Note that, even in finite dimensional optimization, the statement of necessary second order optimality conditions for problems with semismooth derivatives is a quite involved topic. It is out of the scope of this thesis to investigate this aspect.

We have the following *quadratic growth property* if Assumption 5.7 is satisfied:

Lemma 5.24. *[KU15, Lemma 6] For $\gamma > 0$ let Assumption 5.7 hold in a stationary point u_γ of $(\mathrm{MY})_\gamma$, and let Assumptions 5.5 and 5.6 be satisfied. Then there exist $\overline{\alpha}, \overline{r} > 0$ such that*

$$j_\gamma(u_\gamma) + \frac{\overline{\alpha}}{2} \left\|u - u_\gamma\right\|_{\mathcal{U}}^2 \leq j_\gamma(u) \quad \text{for all } u \in B^{\mathcal{U}}(u_\gamma, \overline{r}). \tag{5.12}$$

In particular u_γ is a strict local solution.

Proof. Consider a $u \in B^{\mathcal{U}}(u_\gamma, r)$ with $r > 0$ as in Assumption 5.7, and the convex combination

$$u_\gamma^t := (1-t)u_\gamma + tu.$$

Since j_γ is continuously Fréchet differentiable it holds

$$j_\gamma(u) - j_\gamma(u_\gamma) = \int_0^1 j_\gamma'(u_\gamma + t(u-u_\gamma))(u-u_\gamma)\ \mathrm{d}t = \int_0^1 j_\gamma'(u_\gamma^t)(u-u_\gamma)\ \mathrm{d}t.$$

Using the semismoothness of j_γ', we find

$$\begin{aligned}
&\int_0^1 \langle j_\gamma'(u_\gamma^t, u-u_\gamma,\ \mathrm{d}t\rangle_{\mathcal{U}^*,\mathcal{U}} \\
&= \int_0^1 \langle j_\gamma'(u_\gamma), u-u_\gamma\rangle_{\mathcal{U}^*,\mathcal{U}} + \langle tH_\gamma^u(u_\gamma^t)(u-u_\gamma), u-u_\gamma\rangle_{\mathcal{U}^*,\mathcal{U}} + to(\left\|u-u_\gamma\right\|_{\mathcal{U}}^2)\ \mathrm{d}t \\
&= \langle j_\gamma'(u_\gamma), u-u_\gamma\rangle_{\mathcal{U}^*,\mathcal{U}} + o(\left\|u-u_\gamma\right\|_{\mathcal{U}}^2) + \int_0^1 t\langle H_\gamma^u(u_\gamma^t)(u-u_\gamma), u-u_\gamma\rangle_{\mathcal{U}^*,\mathcal{U}}\ \mathrm{d}t.
\end{aligned}$$

Due to stationarity the first term drops out. By Assumption 5.7 we finally obtain

$$j_\gamma(u) \geq j_\gamma(u_\gamma) + o(\left\|u-u_\gamma\right\|_{\mathcal{U}}^2) + \int_0^1 t\tilde{\alpha}\left\|u-u_\gamma\right\|_{\mathcal{U}}^2\ \mathrm{d}t \geq j_\gamma(u_\gamma) + \frac{\overline{\alpha}}{2}\left\|u-u_\gamma\right\|_{\mathcal{U}}^2,$$

for some suitable $0 < \overline{\alpha} \leq \tilde{\alpha}$ and $\left\|u-u_\gamma\right\|_{\mathcal{U}}$ small enough. The last claim is clear. □

We now want to use Assumption 5.8 to show local Lipschitz continuity of the map $\gamma \mapsto u_\gamma$. First we state the following auxiliary lemma. Recall the notation $\delta(u) = \mathrm{id} + u - P_C(\mathrm{id} + u)$.

Lemma 5.25. *[KU15, Lemma 7] For any $u, v \in \mathcal{U}$ it holds $(\delta(u) - \delta(v), u-v)_{L^2(\Gamma_B, \mathbb{R}^d)} \geq 0$.*

Proof. The projection $\tilde{P}_C$ is non-expansive, i.e., $|\tilde{P}_C(x_1)-\tilde{P}_C(x_2)| \leq |x_1-x_2|$ for all $x_1, x_2 \in \Gamma_B$. Transferring this to the superposition operator P_C we obtain

$$\begin{aligned}
&(\delta(u)-\delta(v), u-v)_{L^2(\Gamma_B,\mathbb{R}^d)} \\
&= (\mathrm{id}+u-P_C(\mathrm{id}+u)-\mathrm{id}-v+P_C(\mathrm{id}+v), u-v)_{L^2(\Gamma_B,\mathbb{R}^d)} \\
&\geq \|u-v\|^2_{L^2(\Gamma_B,\mathbb{R}^d)} - \|P_C(\mathrm{id}+u)-P_C(\mathrm{id}+v)\|_{L^2(\Gamma_B,\mathbb{R}^d)} \|u-v\|_{L^2(\Gamma_B,\mathbb{R}^d)} \\
&\geq \|u-v\|^2_{L^2(\Gamma_B,\mathbb{R}^d)} - \|\mathrm{id}+u-\mathrm{id}-v\|_{L^2(\Gamma_B,\mathbb{R}^d)} \|u-v\|_{L^2(\Gamma_B,\mathbb{R}^d)} = 0.
\end{aligned}$$

□

Theorem 5.26. *[KU15, Theorem 6] For a $\hat{\gamma} > 0$ let the conditions of Theorem 5.21 hold in a local solution $\hat{u}$ of $(\mathrm{MY})_{\hat{\gamma}}$. Then there exist neighborhoods $\mathcal{G} \subset \mathbb{R}$ of $\hat{\gamma}$ and $\mathcal{N} \subset \mathcal{U}$ of $\hat{u}$ such that, for all $\gamma \in \mathcal{G}$, there exists a unique strict local solution u_γ of $(\mathrm{MY})_\gamma$ in $\mathcal{N}$. The map*

$$\mathcal{G} \ni \gamma \mapsto u_\gamma \in \mathcal{N}$$

is Lipschitz continuous on $\mathcal{G}$.

Proof. By Corollary 5.15 we know that for γ close enough to $\hat{\gamma}$ there exists a local solution u_γ of $(\mathrm{MY})_\gamma$ which lies close to $\hat{u}$. From Theorem 5.21 we know that there exists a $R > 0$ such that Assumption 5.7 is satisfied in $\hat{u}$ with $r = R$. Hence, for all $u \in B^{\mathcal{U}}(\hat{u}, R/2)$ Assumption 5.7 is satisfied as well with $r = R/2$. In particular, Lemma 5.24 tells us that any local solution u_γ of $(\mathrm{MY})_\gamma$ in $B^{\mathcal{U}}(\hat{u}, R/2)$ is also strict, and, due to the quadratic growth property, we obtain a unique local solution. This shows the first claim. The first order optimality conditions yield

$$j'_{\hat{\gamma}}(\hat{u}) = 0 \quad \text{and} \quad j'_\gamma(u_\gamma) = 0.$$

Testing with $u_\gamma - \hat{u}$ and subtracting those two equations yields

$$\begin{aligned}
0 &= \langle j'(u_\gamma) - j'(\hat{u}), u_\gamma - \hat{u}\rangle_{\mathcal{U}^*,\mathcal{U}} + \gamma \left(\delta(u_\gamma), u_\gamma - \hat{u}\right)_{L^2(\Gamma_B,\mathbb{R}^d)} - \hat{\gamma} \left(\delta(\hat{u}), u_\gamma - \hat{u}\right)_{L^2(\Gamma_B,\mathbb{R}^d)} \\
&= \langle j'(u_\gamma) - j'(\hat{u}), u_\gamma - \hat{u}\rangle_{\mathcal{U}^*,\mathcal{U}} + \gamma \left(\delta(u_\gamma) - \delta(\hat{u}), u_\gamma - \hat{u}\right)_{L^2(\Gamma_B,\mathbb{R}^d)} \\
&\quad + (\gamma - \hat{\gamma}) \left(\delta(\hat{u}), u_\gamma - \hat{u}\right)_{L^2(\Gamma_B,\mathbb{R}^d)}.
\end{aligned}$$

By Lemma 5.25 the second term can be bounded from below by zero. We will now use Assumption 5.8 to estimate the first term. Since j' is continuously Fréchet differentiable it holds

$$\begin{aligned}
\langle j'(u_\gamma) - j'(\hat{u}), u_\gamma - \hat{u}\rangle_{\mathcal{U}^*,\mathcal{U}} &= \langle j''(\hat{u})(u_\gamma - \hat{u}), u_\gamma - \hat{u}\rangle_{\mathcal{U}^*,\mathcal{U}} + o(\|u_\gamma - \hat{u}\|^2_{\mathcal{U}}) \\
&\geq \alpha \|u_\gamma - \hat{u}\|^2_{\mathcal{U}} + o(\|u_\gamma - \hat{u}\|^2_{\mathcal{U}}).
\end{aligned}$$

In the last step we used Assumption 5.8. Hence we have

$$\begin{aligned}
\alpha \|u_\gamma - \hat{u}\|^2_{\mathcal{U}} + o(\|u_\gamma - \hat{u}\|^2_{\mathcal{U}}) &\leq (\hat{\gamma} - \gamma) \left(\delta(\hat{u}), u_\gamma - \hat{u}\right)_{L^2(\Gamma_B,\mathbb{R}^d)} \\
&\leq |\gamma - \hat{\gamma}| \, \|\delta(\hat{u})\|_{L^2(\Gamma_B,\mathbb{R}^d)} \|u_\gamma - \hat{u}\|_{L^2(\Gamma_B,\mathbb{R}^d)}.
\end{aligned}$$

Since $u_\gamma \to \hat{u}$ for $\gamma \to \hat{\gamma}$ we can choose a $r(\beta) > 0$ such that for all $\gamma > 0$ with $|\gamma - \hat{\gamma}| < r(\beta)$:

$$\overline{\alpha} \left\| u_\gamma - \hat{u} \right\|_{\mathcal{U}}^2 \leq |\gamma - \hat{\gamma}| \left\| \delta(\hat{u}) \right\|_{L^2(\Gamma_B, \mathbb{R}^d)} \left\| u_\gamma - \hat{u} \right\|_{L^2(\Gamma_B, \mathbb{R}^d)}$$

for some $\overline{\alpha} > 0$. Boundedness of $\hat{u}$ implies boundedness of $\delta(\hat{u})$. Finally, using the embedding $\mathcal{U} \hookrightarrow L^2(\Gamma_B, \mathbb{R}^d)$ we arrive at

$$\left\| u_\gamma - \hat{u} \right\|_{\mathcal{U}} \leq C|\gamma - \hat{\gamma}|,$$

for some constant $C > 0$. □

5.6. The value function and its model

Usually the value function is defined as $V : \mathbb{R} \to \mathbb{R}$, $\gamma \mapsto \min_{u \in \mathcal{U}} j_\gamma(u)$. In [HK06a] it is shown that the value function is differentiable in the linear-quadratic setting. Its derivative is given as $\frac{1}{2} \left\| u_\gamma^g \right\|_{L^2(\Gamma_B, \mathbb{R}^d)}$, where u_γ^g denotes the unique global solution. In the nonlinear setting this is not necessarily true if the global solutions of the regularized problems are not unique. Hence, we have to differ between local and global solutions, and restrict ourselves to a local analysis for the differentiability.

Lemma 5.27. *[KU15, Lemma 8] Let Assumption 5.5 hold. Denote a* global *solutions of* $(\mathrm{MY})_\gamma$ *with* u_γ^g. *The map* $\gamma \mapsto V^g(\gamma) := j_\gamma(u_\gamma^g)$ *is globally Lipschitz continuous for all* $\gamma > 0$.

Proof. Due to optimality it holds for any $\gamma_1, \gamma_2 > 0$

$$\begin{aligned} j_{\gamma_2}(u_{\gamma_2}^g) - j_{\gamma_1}(u_{\gamma_1}^g) \leq j_{\gamma_2}(u_{\gamma_1}^g) - j_{\gamma_1}(u_{\gamma_1}^g) = \tfrac{\gamma_2 - \gamma_1}{2} \left\| \delta(u_{\gamma_1}^g) \right\|_{L^2(\Gamma_B, \mathbb{R}^d)}^2, \\ j_{\gamma_2}(u_{\gamma_2}^g) - j_{\gamma_1}(u_{\gamma_1}^g) \geq j_{\gamma_2}(u_{\gamma_2}^g) - j_{\gamma_1}(u_{\gamma_2}^g) = \tfrac{\gamma_2 - \gamma_1}{2} \left\| \delta(u_{\gamma_2}^g) \right\|_{L^2(\Gamma_B, \mathbb{R}^d)}^2. \end{aligned} \tag{5.13}$$

This implies

$$|j_{\gamma_2}(u_{\gamma_2}^g) - j_{\gamma_1}(u_{\gamma_1}^g)| \leq \frac{|\gamma_2 - \gamma_1|}{2} \max \left(\|\delta(u_{\gamma_1}^g)\|_{L^2(\Gamma_B, \mathbb{R}^d)}^2, \|\delta(u_{\gamma_2}^g)\|_{L^2(\Gamma_B, \mathbb{R}^d)}^2 \right).$$

The solutions u_γ^g are uniformly bounded in $\mathcal{U}$ for all $\gamma > 0$. To see this consider an admissible $u \in \mathcal{U}_{ad}$. Then we have for all $\gamma > 0$ the estimate $\frac{\beta}{2} \left\| u_\gamma^g \right\|_{\mathcal{U}}^2 \leq j_\gamma(u_\gamma^g) \leq j_\gamma(u) \leq j(u)$. Since $\mathcal{U} \hookrightarrow L^2(\Gamma_B, \mathbb{R}^d)$ the term $\left\| \delta(u_\gamma^g) \right\|_{L^2(\Gamma_B, \mathbb{R}^d)}^2$ is also uniformly bounded, and the claim follows. □

Remark 5.28. Interestingly, we can show the next result concerning the differentiability of a local value function without employing explicitly the differentiability of j_γ. However, the required conditions will usually only be satisfied if a suitable second order condition holds. In particular, this would necessitate twice differentiability. Theorem 5.26 shows that Assumption 5.8 is sufficient to ensure the conditions of Theorem 5.29.

Theorem 5.29. *[KU15, Theorem 7] Let Assumption 5.5 hold. Let $\hat{\gamma} > 0$ be arbitrary and $\hat{u}$ be a strict local solution of $(\mathrm{MY})_{\hat{\gamma}}$. Assume that there exists a neighborhood $\hat{\mathcal{U}}$ of $\hat{u}$ such that for any sequence $\gamma \to \hat{\gamma}$ a sequence of local solution u_γ of $(\mathrm{MY})_{\gamma_n}$ lies (for γ close enough to $\hat{\gamma}$) in $\hat{\mathcal{U}}$, and the u_γ are* unique *local solutions in $\hat{\mathcal{U}}$. Then the* local value function $V\colon \gamma \mapsto j_\gamma(u_\gamma)$ *is differentiable at $\hat{\gamma}$ with*

$$V'(\hat{\gamma}) = \frac{1}{2} \left\|\delta(\hat{u})\right\|^2_{L^2(\Gamma_B,\mathbb{R}^d)}.$$

Proof. Due to the local uniqueness assumption we can repeat the estimation (5.13) for u_γ and $\hat{u}$ if γ is close enough to $\hat{\gamma}$. This implies

$$\frac{1}{2} \left\|\delta(\hat{u})\right\|^2_{L^2(\Gamma_B,\mathbb{R}^d)} \leq \frac{j_{\hat{\gamma}}(\hat{u}) - j_\gamma(u_\gamma)}{\hat{\gamma} - \gamma} \leq \frac{1}{2} \left\|\delta(u_\gamma)\right\|^2_{L^2(\Gamma_B,\mathbb{R}^d)},$$

if $\hat{\gamma} \geq \gamma$ and the reverse inequality if $\hat{\gamma} < \gamma$. Using Corollary 5.15 we see that $u_\gamma \to \hat{u}$ as $\gamma \to \hat{\gamma}$, in particular $\left\|\delta(u_\gamma)\right\|^2_{L^2(\Gamma_B,\mathbb{R}^d)} \to \left\|\delta(\hat{u})\right\|^2_{L^2(\Gamma_B,\mathbb{R}^d)}$. Thus, we obtain the claimed result $V'(\hat{\gamma}) = \frac{1}{2} \left\|\delta(\hat{u})\right\|^2_{L^2(\Gamma_B,\mathbb{R}^d)}$ in the limit. □

Corollary 5.30. *Let the conditions of Theorem 5.29 be satisfied. The local value function V is monotonically increasing. If the local solutions satisfy $u_\gamma \notin \mathcal{U}_{ad}$ it is strictly monotonically increasing. The map $\gamma \mapsto V'(\gamma)$ is strictly monotonically decreasing.*

Proof. It holds $V'(\gamma) \geq 0$. $V'(\gamma) = 0$ would imply $\delta(u_\gamma) = 0$ which is only the case if $u_\gamma \in \mathcal{U}_{ad}$. Thus we conclude that the value function V is (strictly) monotonically increasing. Furthermore,

$$\gamma_2 > \gamma_1 \;\Rightarrow\; \left\|\delta(u_{\gamma_2})\right\|^2_{L^2(\Gamma_B,\mathbb{R}^d)} < \left\|\delta(u_{\gamma_1})\right\|^2_{L^2(\Gamma_B,\mathbb{R}^d)},$$

hence the mapping $\gamma \mapsto V'(\gamma)$ is monotonically decreasing. □

In [HK06a] it was proposed to use the value function V in an algorithmic setting to steer the γ-update and the termination criterion. Since the function is not available explicitly, one has to approximate it with a model function. Following [HK06a] one may define the model

$$m(\gamma) = C_1 - \frac{C_2}{(D + \gamma)^r},$$

for some real constants C_1, C_2, D, r, with $C_2, D, r > 0$. Note that $m' > 0$, $m'' < 0$ corresponding to the properties of V stated in Corollary 5.30.

Remark 5.31. In [HK06b] the same class of model functions was used in inexact Moreau-Yosida path following for the obstacle problem. The authors reported good results with their (partially heuristic) strategy. However, in our experience, the performance of such a scheme is very much dependent on the choice of the various parameters and the concrete example.

5.7. Optimality conditions and properties of the Lagrange multipliers

In this section we study the optimality conditions of the full problem (5.1) and its regularization. In particular, we extend the results of Lemma 5.10 and Theorem 5.26 to the state, the adjoint state, and the Lagrange multiplier associated with the geometric constraint. We do not explicitly use the second order condition in this section. However, it could be used to check some of the prerequisites of the following results. The findings in this section were partly inspired by the ideas presented in [HK06a] and [Ulb11].

The optimality conditions of (5.1) were already derived in Theorem 5.2. We repeat them here for the convenience of the reader:

$$\begin{aligned} J_U(T(\overline{u}),\overline{y})T'(\overline{u}) + \beta u + \Big(E_U(T(\overline{u}),\overline{y})T'(\overline{u})\Big)^* \overline{p} + \overline{\lambda} &= 0 \text{ in } \mathcal{U}^*, \\ J_y(T(\overline{u}),\overline{y}) + E_y(T(\overline{u}),\overline{y})^* \overline{p} &= 0 \text{ in } \mathcal{Y}^*, \\ E(T(\overline{u}),\overline{y}) &= 0 \text{ in } \mathcal{Z}, \\ \overline{u} &\in \mathcal{U}_{ad}, \\ \overline{\lambda} &\in \mathcal{T}(\mathcal{U}_{ad},\overline{u})^\circ. \end{aligned} \tag{5.14}$$

The regularized problem is given by

$$\min_{u\in\mathcal{U}, y\in\mathcal{Y}} J(T(u),y) + \frac{\beta}{2}\|u\|_{\mathcal{U}}^2 + e_\gamma(u) \quad \text{s.t. } E(T(u),y) = 0 \text{ in } \mathcal{Z}. \tag{5.15}$$

Under Assumption 5.5 we obtain for every $\gamma > 0$ and a local solution $(u_\gamma, y_\gamma) \in \mathcal{U}\times\mathcal{Y}$ the existence of a unique adjoint state $p_\gamma \in \mathcal{Z}^*$ such that

$$\begin{aligned} J_U(T(u_\gamma),y_\gamma) + \beta u_\gamma + (E_U(T(u_\gamma),y_\gamma)T'(u_\gamma))^* p_\gamma + \lambda_\gamma &= 0 \quad \text{in } \mathcal{U}^*, \\ J_y(T(u_\gamma),y_\gamma) + E_y(T(u_\gamma),y_\gamma)^* p_\gamma &= 0 \quad \text{in } \mathcal{Y}^*, \\ E(T(u_\gamma),y_\gamma) &= 0 \quad \text{in } \mathcal{Z}, \\ \gamma(\tau_\gamma - P_C(\tau_\gamma)) &= \lambda_\gamma \quad \text{in } L^2(\Gamma_B,\mathbb{R}^d). \end{aligned} \tag{5.16}$$

Lemma 5.32. *[KU15, Lemma 9] Let Assumption 5.5 hold and suppose additionally that the extension operator T is linear. If a sequence of local solutions u_γ is uniformly bounded in $\mathcal{U}$ with respect to γ, then the associated sequence $(y_\gamma, p_\gamma, \lambda_\gamma)$, as determined by* (5.16)*, is also bounded in $\mathcal{Y}\times\mathcal{Z}\times\mathcal{U}^*$.*

Proof. (i) If u_γ is uniformly bounded we can find a bounded, closed set $\tilde{\mathcal{U}} \subset \mathcal{U}$ with $(u_\gamma) \subset \tilde{\mathcal{U}}$ for all γ. Due to complete continuity of T the image $T(\tilde{\mathcal{U}})$ is relatively compact and $T(u_\gamma)$ is contained in the compact set $\tilde{\mathcal{V}} := \operatorname{cl} T(\tilde{\mathcal{U}})$. Since S is continuous $S(\tilde{\mathcal{V}})$ is also compact. In particular we obtain that $(y_\gamma) = (S(T(u_\gamma))) \subset S(\tilde{\mathcal{V}})$ is uniformly bounded.

(ii) Since J is continuously differentiable $J_y(T(u_\gamma),y_\gamma) \subset J_y(\tilde{\mathcal{V}}, S(\tilde{\mathcal{V}}))$ is also uniformly bounded. The same holds for $E_y(T(u_\gamma),y_\gamma)$ and its inverse is likewise bounded. Hence using the adjoint equation in (5.16) we can bound p_γ uniformly

$$\|p_\gamma\|_{\mathcal{Z}} \le \left\|E_y(T(u_\gamma),y_\gamma)^{-1}\right\|_{\mathcal{L}(\mathcal{Z},\mathcal{Y})} \|J_y(T(u_\gamma),y_\gamma)\|_{\mathcal{Y}^*} \le C.$$

(iii) Analogously to (ii) we obtain from the first equation of (5.16) and the boundedness of $(u_\gamma, y_\gamma, p_\gamma)$:

$$\|\lambda_\gamma\|_{\mathcal{U}^*} \leq \|J_U(T(u_\gamma), y_\gamma) + \beta u_\gamma\|_{\mathcal{U}^*} + \left\|E_U(T(u_\gamma), y_\gamma)T'(u_\gamma)\right\|_{\mathcal{L}(\mathcal{U},\mathcal{Z})} \|p_\gamma\|_{\mathcal{Z}} \leq C.$$

□

Remark 5.33. By assumption $\mathcal{U}$ (and thus also $\mathcal{U}^*$) are *reflexive* Banach spaces. Usually this applies also to $\mathcal{Y}$ and $\mathcal{Z}$. In that case the above result provides us with weakly convergent subsequences. The linearity condition on T may be replaced by suitable conditions on its derivatives. However, for simplicity we will always require it in this section.

We will now show that $\lambda_\gamma \in \mathcal{U}^*$ approximates $\overline{\lambda}$ from (5.14) for $\gamma \to \infty$. First, recall from Lemma 5.4 that

$$u \mapsto u^C = P_C(\tau) - \mathrm{id}$$

describes the $L^2(\Gamma_B, \mathbb{R}^d)$-projection onto

$$\mathcal{U}_{ad}^L = \{u \in L^2(\Gamma_B, \mathbb{R}^d) \mid \tau(x) \subset C \text{ for a. e. } x \in \Gamma_B\}.$$

For the convenience of the reader we summarize [Ulb11, Lemma 8.2 and Lemma 8.20] in the following auxiliary result.

Lemma 5.34. *(i) If $M \in \mathcal{L}(Z, X)$ is a surjective operator between Banach spaces, then there exists a constant $c > 0$ such that $\left\|x'\right\|_{X^*} \leq c \left\|M^* x'\right\|_{Z^*}$ for all $x' \in X^*$, where M^* denotes the adjoint operator of M.*

(ii) The linear operator

$$F\colon \mathcal{U} \times \mathcal{Y} \to \mathcal{U} \times \mathcal{Z}, \quad F\begin{pmatrix} v \\ z \end{pmatrix} = \begin{pmatrix} v \\ E_U(T(u), y)T'(u)v + E_y(T(u), y)z \end{pmatrix}$$

is surjective if and only if $E_y(T(u), y) \in \mathcal{L}(\mathcal{Y}, \mathcal{Z})$ is surjective. Its dual is given by

$$F^*\colon \mathcal{U}^* \times \mathcal{Z}^* \to \mathcal{U}^* \times \mathcal{Y}^*, \quad F^*\begin{pmatrix} \lambda \\ p \end{pmatrix} = \begin{pmatrix} \lambda + (E_U(T(u), y)T'(u))^* p \\ E_y(T(u), y)^* p \end{pmatrix}.$$

Recall that Assumption 5.2 implies the surjectivity of $E_y(T(u), S(T(u)))$ for all $u \in \mathcal{U}_{ad}$. We are now ready to prove the announced convergence result. For $\gamma_n > 0$ we denote by $(u_n, y_n, p_n, \lambda_n) \in \mathcal{U} \times \mathcal{Y} \times \mathcal{Z}^* \times \mathcal{U}^*$ a solution of the optimality system (5.16).

Theorem 5.35. *[KU15, Theorem 9] Let Assumption 5.5 hold and $\gamma_n \to \infty$. Furthermore, suppose that the extension operator T is linear. Then any weakly convergent subsequence $(u_k, y_k, p_k, \lambda_k) \rightharpoonup (\overline{u}, \overline{y}, \overline{p}, \overline{\lambda})$ in $\mathcal{U} \times \mathcal{Y} \times \mathcal{Z}^* \times \mathcal{U}^*$ converges strongly, and $(\overline{u}, \overline{y}, \overline{p}, \overline{\lambda})$ solves the optimality system* (5.14).

Proof. Consider a weakly convergent subsequence $(u_k, y_k, p_k, \lambda_k)$ with limit $(\overline{u}, \overline{y}, \overline{p}, \overline{\lambda})$. The plan of the proof is the following. In (i) we show $y_k \to \overline{y} = S(T(\overline{u}))$. We can associate with $(\overline{u}, \overline{y})$ a unique pair $(\hat{p}, \hat{\lambda}) \in \mathcal{Z}^* \times \mathcal{U}^*$ such that the first two equations of (5.14) are satisfied for $(\overline{u}, \overline{y}, \hat{p}, \hat{\lambda})$. We proceed to show in (ii) that $(p_k, \lambda_k) \to (\hat{p}, \hat{\lambda})$. Uniqueness of the limit implies then $(\hat{p}, \hat{\lambda}) = (\overline{p}, \overline{\lambda})$. As next step we prove $u_k \to \overline{u}$ in (iii). Finally we check that $\overline{u} \in \mathcal{U}_{ad}$ and $\overline{\lambda} \in \mathcal{T}(\mathcal{U}_{ad}, \overline{u})^\circ$ in (iv) and (v).

(i) The strong convergence $y_k = S(T(u_k)) \to S(T(\overline{u}))$ follows from the complete continuity of T and the continuity of the design-to-state operator S (compare Assumption 5.1). Hence $\overline{y} = S(T(\overline{u}))$ solves the state equation of (5.14).

(ii) As announced we associate now with $(\overline{u}, \overline{y})$ a unique pair $(\hat{p}, \hat{\lambda}) \in \mathcal{Z} \times \mathcal{U}^*$ such that the first two equations of (5.14) are satisfied for $(\overline{u}, \overline{y}, \hat{p}, \hat{\lambda})$. We claim that $(p_k, \lambda_k) \to (\hat{p}, \hat{\lambda})$. Setting $F_{\overline{u}}$ to be the operator defined in Lemma 5.34 with $u = \overline{u}$ and similarly F_k with $u = u_k$ we can write the first two equations in (5.4) and (5.16) with the help of the dual operators as

$$F_{\overline{u}}^* \begin{pmatrix} \hat{\lambda} \\ \hat{p} \end{pmatrix} = - \begin{pmatrix} J_U(T(\overline{u}), \overline{y}) T'(\overline{u}) + \beta \overline{u} \\ J_y(T(\overline{u}), \overline{y}) \end{pmatrix} \text{ and}$$
$$F_k^* \begin{pmatrix} \lambda_k \\ p_k \end{pmatrix} = - \begin{pmatrix} J_U(T(u_k), y_k) T'(u_k) + \beta u_k \\ J_y(T(u_k), y_k) \end{pmatrix}.$$

Using Lemma 5.34 we see that

$$\left\| \begin{pmatrix} \hat{\lambda} - \lambda_k \\ \hat{p} - p_k \end{pmatrix} \right\|_{\mathcal{U}^* \times \mathcal{Z}^*} \leq c \left\| F_{\overline{u}}^* \begin{pmatrix} \hat{\lambda} - \lambda_k \\ \hat{p} - p_k \end{pmatrix} \right\|_{\mathcal{U}^* \times \mathcal{Y}^*} \tag{5.17}$$
$$\leq c \left\| F_{\overline{u}}^* \begin{pmatrix} \hat{\lambda} \\ \hat{p} \end{pmatrix} - F_k^* \begin{pmatrix} \lambda_k \\ p_k \end{pmatrix} \right\|_{\mathcal{U}^* \times \mathcal{Y}^*} + c \left\| (F_{\overline{u}}^* - F_k^*) \begin{pmatrix} \lambda_k \\ p_k \end{pmatrix} \right\|_{\mathcal{U}^* \times \mathcal{Y}^*}. \tag{5.18}$$

Let us deal with those two terms separately. Using $y_k \to \overline{y}$ and the complete continuity of T we see that

$$\begin{aligned} J_U(T(u_k), y_k) &\to J_U(T(\overline{u}), \overline{y}), \\ J_y(T(u_k), y_k) &\to J_y(T(\overline{u}), \overline{y}), \end{aligned} \tag{5.19}$$

Since T is linear it holds $T'(u_k) = T'(\overline{u})$ for all k. Finally, by the definition of weak convergence we obtain

$$u_k \rightharpoonup \overline{u} \text{ in } \mathcal{U} \Rightarrow u_k \to \overline{u} \text{ in } \mathcal{U}^*.$$

Hence, we have

$$\left\| F_{\overline{u}}^* \begin{pmatrix} \hat{\lambda} \\ \hat{p} \end{pmatrix} - F_k^* \begin{pmatrix} \lambda_k \\ p_k \end{pmatrix} \right\|_{\mathcal{U}^* \times \mathcal{Y}^*}$$
$$= \left\| \begin{pmatrix} J_U(T(u_k), y_k) T'(u_k) + \beta u_k - J_U(T(\overline{u}), \overline{y}) T'(\overline{u}) - \beta \overline{u} \\ J_y(T(u_k), y_k) - J_y(T(\overline{u}), \overline{y}) \end{pmatrix} \right\|_{\mathcal{U}^* \times \mathcal{Y}^*} \to 0.$$

Addressing the second term we note that also

$$\begin{aligned} E_U(T(u_k), y_k) &\to E_U(T(\overline{u}), \overline{y}), \\ E_y(T(u_k), y_k) &\to E_y(T(\overline{u}), \overline{y}), \end{aligned} \tag{5.20}$$

and hence

$$\|F_k - F_{\overline{u}}\|_{\mathcal{L}(\mathcal{U}\times\mathcal{Y},\mathcal{U}\times\mathcal{Z})} \to 0.$$

Using the properties of the dual operator and boundedness of (λ_k, p_k) we obtain

$$\begin{aligned} \left\| (F_k^* - F_{\overline{u}}^*) \begin{pmatrix} \lambda_k \\ p_k \end{pmatrix} \right\|_{\mathcal{U}^*\times\mathcal{Y}^*} &\leq \|F_k^* - F_{\overline{u}}^*\|_{\mathcal{L}(\mathcal{U}^*\times\mathcal{Z}^*,\mathcal{U}^*\times\mathcal{Y}^*)} \left\| \begin{pmatrix} \lambda_k \\ p_k \end{pmatrix} \right\|_{\mathcal{U}^*\times\mathcal{Y}^*} \\ &= \|F_k - F_{\overline{u}}\|_{\mathcal{L}(\mathcal{U}\times\mathcal{Y},\mathcal{U}\times\mathcal{Z})} \left\| \begin{pmatrix} \lambda_k \\ p_k \end{pmatrix} \right\|_{\mathcal{U}^*\times\mathcal{Y}^*}. \end{aligned}$$

Since $\begin{pmatrix} \lambda_k \\ p_k \end{pmatrix}$ is uniformly bounded we conclude from (5.17) that $\left\| \begin{pmatrix} \hat{\lambda} - \lambda_k \\ \hat{p} - p_k \end{pmatrix} \right\|_{\mathcal{U}^*\times\mathcal{Z}^*} \to 0$. Uniqueness of the limit now provides us with

$$\begin{aligned} &\lambda_k \to \hat{\lambda} = \overline{\lambda} \text{ in } \mathcal{U}^* \text{ and} \\ &p_k \to \hat{p} = \overline{p} \text{ in } \mathcal{Z}^*. \end{aligned}$$

(iii) Testing the first equation in (5.14) and (5.16) with $\overline{u} - u_k$, and subtracting the two equations we see that

$$\begin{aligned} \beta \|\overline{u} - u_k\|_{\mathcal{U}}^2 =& \langle J_U(T(u_k), y_k)T'(u_k) - J_U(T(\overline{u}), \overline{y})T'(\overline{u}), \overline{u} - u_k\rangle_{\mathcal{U}^*,\mathcal{U}} \\ &+ \langle \Big(E_U(T(u_k), y_k)T'(u_k)\Big)^* p_k - \Big(E_U(T(\overline{u}), \overline{y})T'(\overline{u})\Big)^* \overline{p}, \overline{u} - u_k\rangle_{\mathcal{U}^*,\mathcal{U}} \\ &+ \langle \lambda_k - \overline{\lambda}, \overline{u} - u_k\rangle_{\mathcal{U}^*,\mathcal{U}}. \end{aligned}$$

Combing now $u_k \rightharpoonup \overline{u}$, $p_k \to \overline{p}$, $\lambda_k \to \overline{\lambda}$, (5.19), and (5.20) shows that the right hand side tends to zero and hence $u_k \to \overline{u}$.

(iv) From $\lambda_k \rightharpoonup \overline{\lambda}$ we know that λ_k in $\mathcal{U}^*$ is bounded. Thus

$$\gamma_k \left(\tau_k - P_C\left(\tau_k\right), v\right)_{L^2(\Gamma_B,\mathbb{R}^d)} \leq C \|v\|_{\mathcal{U}} \text{ for all } v \in \mathcal{U}.$$

Since $\gamma_k \to \infty$ we conclude that $\tau_k - P_C(\tau_k) \to 0$ in $L^2(\Gamma_B, \mathbb{R}^d)$. Furthermore $u_k \to \overline{u}$ in $\mathcal{U}$ implies $u_k \to \overline{u}$ in $L^2(\Gamma_B, \mathbb{R}^d)$. Combining these findings yields $\overline{\tau} - P_C(\overline{\tau}) = 0$ in $L^2(\Gamma_B, \mathbb{R}^d)$, i.e., $\overline{u} \in \mathcal{U}_{ad}$.

(v) Let us now check if $\overline{\lambda} \in \mathcal{T}(\mathcal{U}_{ad}, \overline{u})^\circ$. Recall $u^C = P_C(\tau) - \mathrm{id}$. Note that

$$\left\| u_k^C - \overline{u} \right\|_{L^2(\Gamma_B,\mathbb{R}^d)} \leq \|u_k - \overline{u}\|_{L^2(\Gamma_B,\mathbb{R}^d)}$$

by the projection property, and $u_k \to \overline{u}$ in $L^2(\Gamma_B, \mathbb{R}^d)$, hence $u_k^C \to \overline{u}$ in $L^2(\Gamma_B, \mathbb{R}^d)$. Now let $\mu(v - \overline{u}) \in \mathcal{T}(\mathcal{U}_{ad}, \overline{u})$ for some $\mu > 0$ and $v \in \mathcal{U}_{ad}$. We have

$$\begin{aligned}
\langle \overline{\lambda}, \mu(v - \overline{u}) \rangle_{\mathcal{U}^*, \mathcal{U}} &= \lim_{k \to \infty} \langle \lambda_k, \mu(v - \overline{u}) \rangle_{\mathcal{U}^*, \mathcal{U}} \\
&= \lim_{k \to \infty} \left(\gamma_k(\tau_k - P_C(\tau_k)), \mu(v - \overline{u}) \right)_{L^2(\Gamma_B, \mathbb{R}^d)} \\
&= \lim_{k \to \infty} \left(\gamma_k(\tau_k - P_C(\tau_k)), \mu(v - u_k^C) \right)_{L^2(\Gamma_B, \mathbb{R}^d)} \\
&= \lim_{k \to \infty} \gamma_k \mu \left(u_k - u_k^C, v - u_k^C \right)_{L^2(\Gamma_B, \mathbb{R}^d)} \leq 0.
\end{aligned}$$

In the last inequality we used again the projection property. Hence for

$$K := \{ w \in \mathcal{U} \mid w = \mu(v - \overline{u}) \text{ for some } \ \mu > 0 \text{ and } v \in \mathcal{U}_{ad} \}$$

it holds $\overline{\lambda} \in K^\circ$. Since $\mathcal{T}(\mathcal{U}_{ad}, \overline{u}) = \operatorname{cl} K$, and the polar cone of a cone is the same as the polar cone of its closure, we conclude that $\overline{\lambda} \in \mathcal{T}(\mathcal{U}_{ad}, \overline{u})^\circ$.

□

We can immediately carry Theorem 5.35 over to the case $\gamma_n \to \hat{\gamma} > 0$:

Corollary 5.36. *[KU15, Corollary 4] Let Assumption 5.5 hold, $\hat{\gamma} > 0$, and $\gamma_n \to \hat{\gamma}$. Furthermore, suppose that the extension operator T is linear. Then any weakly convergent subsequence $(u_k, y_k, p_k, \lambda_k) \rightharpoonup (u^*, y^*, p^*, \lambda^*)$ converges strongly, and the limit $(u^*, y^*, p^*, \lambda^*)$ solves $(5.16)_{\hat{\gamma}}$.*

Proof. λ^* solves the last equation of (5.16), since $u_k \rightharpoonup \overline{u}$ in $\mathcal{U}$ implies $u_k \to \overline{u}$ in $L^2(\Gamma_B, \mathbb{R}^d)$. Replacing $(\overline{u}, \overline{y}, \overline{p}, \overline{\lambda})$ by $(u^*, y^*, p^*, \lambda^*)$ in the steps (i)-(iii) of the proof of Theorem 5.35, we see that $(u^*, y^*, p^*, \lambda^*)$ solves also the other equations of (5.16) and the convergence is strong. □

Finally, we extend the results of Theorem 5.26.

Theorem 5.37. *[KU15, Theorem 10] Let Assumption 5.5 hold, $\hat{\gamma} > 0$ and $(u_\gamma, y_\gamma, p_\gamma, \lambda_\gamma)$ be a solution of $(5.16)_\gamma$. Furthermore, suppose that the extension operator T is linear. If the map $\gamma \mapsto u_\gamma$ is locally Lipschitz continuous for γ close enough to $\hat{\gamma}$, then the maps $\gamma \mapsto y_\gamma$, $\gamma \mapsto p_\gamma$, and $\gamma \mapsto \lambda_\gamma$ are also locally Lipschitz continuous.*

Proof. By assumption the solution operator S is twice continuously differentiable, in particular locally Lipschitz continuous. Hence, the mapping $\gamma \mapsto y_\gamma = S(T(u_\gamma))$ is locally Lipschitz continuous if $\gamma \mapsto u_\gamma$ is locally Lipschitz. Using again the operators $F_\gamma, F_{\hat{\gamma}}$ as defined in Lemma 5.34 with $u = u_\gamma$, respectively $u = u_{\hat{\gamma}}$ and copying the calculations in the proof of Theorem 5.35(ii) we arrive at

$$\begin{aligned}
\left\| \begin{pmatrix} \lambda_{\hat{\gamma}} - \lambda_\gamma \\ p_{\hat{\gamma}} - p_\gamma \end{pmatrix} \right\|_{\mathcal{U}^* \times \mathcal{Z}^*} &\leq c \left\| F_{\hat{\gamma}}^* \begin{pmatrix} \lambda_{\hat{\gamma}} - \lambda_\gamma \\ p_{\hat{\gamma}} - p_\gamma \end{pmatrix} \right\|_{\mathcal{U}^* \times \mathcal{Y}^*} \\
&\leq c \left\| \begin{pmatrix} J_U(T(u_{\hat{\gamma}}), y_{\hat{\gamma}}) T'(u_{\hat{\gamma}}) + \beta u_{\hat{\gamma}} - J_U(T(u_\gamma), y_\gamma) T'(u_\gamma) - \beta u_\gamma \\ J_y(T(u_{\hat{\gamma}}), y_{\hat{\gamma}}) - J_y(T(u_\gamma), y_\gamma) \end{pmatrix} \right\|_{\mathcal{U}^* \times \mathcal{Y}^*} \\
&\quad + c \left\| (F_\gamma^* - F_{\hat{\gamma}}^*) \begin{pmatrix} \lambda_\gamma \\ p_\gamma \end{pmatrix} \right\|_{\mathcal{U}^* \times \mathcal{Y}^*}.
\end{aligned}$$

Since J is twice continuously differentiable its first derivatives are Lipschitz continuous. Recall that T is linear, hence $T'(u_{\hat\gamma}) = T'(u_\gamma)$. Thus, the first term can be bounded for γ close to $\hat\gamma$ by $c(\|u_{\hat\gamma} - u_\gamma\|_{\mathcal{U}} + \|y_{\hat\gamma} - y_\gamma\|_{\mathcal{Y}})$, and analogously it holds

$$\|F_\gamma - F_{\hat\gamma}\|_{\mathcal{L}(\mathcal{U}\times\mathcal{Y},\mathcal{U}\times\mathcal{Z})} \le c(\|u_{\hat\gamma} - u_\gamma\|_{\mathcal{U}} + \|y_{\hat\gamma} - y_\gamma\|_{\mathcal{Y}}),$$

for γ close to $\hat\gamma$. Copying the calculations from the proof of Theorem 5.35(ii), we see that

$$\begin{aligned}\left\|(F_\gamma^* - F_{\hat\gamma}^*)\begin{pmatrix}\lambda_\gamma\\ p_\gamma\end{pmatrix}\right\|_{\mathcal{U}^*\times\mathcal{Y}^*} &\le \|F_\gamma - F_{\hat\gamma}\|_{\mathcal{L}(\mathcal{U}\times\mathcal{Y},\mathcal{U}\times\mathcal{Z})}\left\|\begin{pmatrix}\lambda_\gamma\\ p_\gamma\end{pmatrix}\right\|_{\mathcal{U}^*\times\mathcal{Z}^*}\\ &\le c(\|u_{\hat\gamma} - u_\gamma\|_{\mathcal{U}} + \|y_{\hat\gamma} - y_\gamma\|_{\mathcal{Y}})\left\|\begin{pmatrix}\lambda_\gamma\\ p_\gamma\end{pmatrix}\right\|_{\mathcal{U}^*\times\mathcal{Z}^*}.\end{aligned}$$

Recall that $(\lambda_\gamma, p_\gamma)$ is (uniformly) bounded. Summarizing, we found

$$\left\|\begin{pmatrix}\lambda_{\hat\gamma} - \lambda_\gamma\\ p_{\hat\gamma} - p_\gamma\end{pmatrix}\right\|_{\mathcal{U}^*\times\mathcal{Z}^*} \le c(\|u_{\hat\gamma} - u_\gamma\|_{\mathcal{U}} + \|y_{\hat\gamma} - y_\gamma\|_{\mathcal{Y}}).$$

We conclude, that the local Lipschitz continuity of $\gamma \mapsto u_\gamma$ and $\gamma \mapsto y_\gamma$, implies the local Lipschitz continuity of the mappings $\gamma \mapsto p_\gamma$ and $\gamma \mapsto \lambda_\gamma$. □

5.8. Convergence rate estimates

In this section we present some estimates on the rate of convergence towards feasibility, as well as on the distance between a solution of (P) and a solution of $(\mathrm{MY})_\gamma$. This section has been adopted from [KU15, Section 3.7] with minor changes.

Recall the notation $\delta(u)(x) = x + u(x) - \tilde P_C(x + u(x))$ for $x \in \Gamma_B$, and the properties of the oriented distance function

$$b_C = d_C - d_{C^c},$$

of the convex set $C \subset \mathbb{R}^d$, cf. Section 2.11.3 and [DZ11, Chapter 7]. In particular, for all $\tilde x \notin C$ we have

$$\nabla b_C(\tilde x) = \frac{\tilde x - \tilde P_C(\tilde x)}{|\tilde x - \tilde P_C(\tilde x)|} \text{ and } |\nabla b_C(\tilde x)| = 1.$$

Introducing the set of feasible boundary points

$$Z(u) := \{x \in \Gamma_B \mid \delta(u)(x) = 0\},$$

we note that

$$\nabla b_C(x + u(x)) = \frac{\delta(u)(x)}{|\delta(u)(x)|} \text{ for all } x \in \Gamma_B \backslash Z(u).$$

We now strengthen our assumptions on C and $\mathcal{U}$. We need to be able to smoothen out jumps of ∇b_C originating from corners of C while staying close to the original vector field.

Assumption 5.10. *We have the embedding* $\mathcal{U} \hookrightarrow C^{1,\alpha_1}(\Gamma_B, \mathbb{R}^d)$ *for some* $\alpha_1 > 0$. *Furthermore, there exists* $c_1, c_2 > 0$ *and a vector field* $V_C \colon \mathbb{R}^d \to \mathbb{R}^d$ *such that*

(i) $V_C(x)^T \nabla b_C(x) \geq c_1$ *and* $|V_C(x)| \leq c_2 |\nabla b_C(x)|$ *for all* $x \in \mathbb{R}^d \backslash C$

(ii) $V_C \circ (\mathrm{id} + u) \in \mathcal{U}$, *for all* $u \in \mathcal{U}$.

Remark 5.38. The first condition implies that V_C cannot differ to far from ∇b_C, in particular the angle between those vectors is always smaller than π. The second condition requires a certain smoothness of V_C.

We are now ready to estimate the quantity $\delta(u_\gamma)$ in the $L^1(\Gamma_B, \mathbb{R}^d)$-norm. Compare [HSW14] in the state-constrained setting.

Lemma 5.39. *[KU15, Lemma 12] Let Assumptions 5.5 and 5.10 hold. If a family of local solutions* (u_γ) *of* $(\mathrm{MY})_\gamma$ *is uniformly bounded, then there exists a constant* $c > 0$ *such that*

$$\|\delta(u_\gamma)\|_{L^1(\Gamma_B, \mathbb{R}^d)} \leq c\gamma^{-1}.$$

Proof. Let $\gamma > 0$. A local solution u_γ satisfies the optimality condition $j'_\gamma(u_\gamma) = 0$. Testing with $v_\gamma := V_C \circ (\mathrm{id} + u_\gamma) \in \mathcal{U}$ we obtain

$$0 = \langle j'_\gamma(u_\gamma), v_\gamma \rangle_{\mathcal{U}^*, \mathcal{U}} = \langle j'(u_\gamma), v_\gamma \rangle_{\mathcal{U}^*, \mathcal{U}} + \gamma \left(\delta(u_\gamma), v_\gamma\right)_{L^2(\Gamma_B, \mathbb{R}^d)}.$$

Boundedness of u_γ implies boundedness of $j'(u_\gamma)$ and by Assumption 5.10 also of v_γ. Hence, $\exists c > 0$ such that

$$\gamma \left(\delta(u_\gamma), v_\gamma\right)_{L^2(\Gamma_B, \mathbb{R}^d)} \leq \left\|j'(u_\gamma)\right\|_{\mathcal{U}^*} \|v_\gamma\|_{\mathcal{U}} \leq c, \quad \text{for all } \gamma > 0.$$

Furthermore, for all $i = 1, \ldots, d$ it holds $\|(\delta(u_\gamma))_i\|_{L^1(\Gamma_B, \mathbb{R}^d)} \leq \int_{\Gamma_B} |\delta(u_\gamma)(x)| \, \mathrm{d}x$. Thus, using the properties of b_C and V_C we find

$$\begin{aligned}
\gamma \|\delta(u_\gamma)\|_{L^1(\Gamma_B, \mathbb{R}^d)} &\leq \gamma d \int_{\Gamma_B} |\delta(u_\gamma)(x)| \, \mathrm{d}x \\
&= \gamma d \int_{\Gamma_B \backslash Z(u_\gamma)} |\delta(u_\gamma)(x)| \, \mathrm{d}x \\
&\leq \frac{\gamma d}{c_1} \int_{\Gamma_B \backslash Z(u_\gamma)} V_C(x + u_\gamma(x))^T \nabla b_C(x + u_\gamma(x)) |\delta(u_\gamma)(x)| \, \mathrm{d}x \\
&= \frac{\gamma d}{c_1} \int_{\Gamma_B \backslash Z(u_\gamma)} v_\gamma(x)^T \delta(u_\gamma)(x) \, \mathrm{d}x \\
&= \frac{\gamma d}{c_1} \int_{\Gamma_B} v_\gamma(x)^T \delta(u_\gamma)(x) \, \mathrm{d}x \\
&= \frac{d}{c_1} \gamma \left(v_\gamma, \delta(u_\gamma)\right)_{L^2(\Gamma_B, \mathbb{R}^d)} \leq \frac{dc}{c_1}.
\end{aligned}$$

This shows $\|\delta(u_\gamma)\|_{L^1(\Gamma_B, \mathbb{R}^d)} \leq c\gamma^{-1}$ for some $c > 0$. □

Now we want to have an estimate on the point-wise constraint violation, which we measure with $\|\delta(u_\gamma)\|_{L^\infty(\Gamma_B,\mathbb{R}^d)}$. The idea is to use the following result from [HSW14]. For convenience we abbreviate $C^{k+\alpha}(A) := C^{k,\alpha}(A)$ for $k \in \mathbb{N}$ and $\alpha \in (0,1]$.

Proposition 5.40. *[HSW14, Proposition 2.4] Consider a bounded, open set $A \subset \mathbb{R}^n$, and $z \in C^\beta(\overline{A}) \cap L^1(A)$, with $0 < \beta \leq 2$, and $z \geq 0$. Moreover, assume that $z = 0$ on ∂A. Then*

$$\|z\|_{L^\infty(A)} \leq c \, \|z\|^{1-\theta}_{C^\beta(A)} \, \|z\|^\theta_{L^1(A)}, \tag{5.21}$$

with $\theta = \beta/(\beta+n)$. The constant $c > 0$ is indepentdent of A.

Theorem 5.41. *[KU15, Theorem 11] Suppose Assumptions 5.5 and 5.10 hold, and consider a bounded family of local solutions $(u_\gamma) \subset C^{1,\alpha_1}(\Gamma_B, \mathbb{R}^d)$ of $(\mathrm{MY})_\gamma$ for $\gamma > 0$. Choose for all $\gamma > 0$ a point $x_\gamma \in \Gamma_B$ such that*

$$\|\delta(u_\gamma)\|_{L^\infty(\Gamma_B,\mathbb{R}^d)} = |\delta(u_\gamma)(x_\gamma)|.$$

If there exists $\hat{\gamma}$ such that $\forall \gamma \geq \hat{\gamma}$ there exists such a point $x_\gamma \notin \partial\Gamma_B$, then $\exists c > 0$:

(i) $\|\delta(u_\gamma)\|_{L^\infty(\Gamma_B,\mathbb{R}^d)} \leq c\gamma^{-\frac{1}{d}}$.

(ii) If additionally there exist $\alpha_2, \alpha_3 > 0$ such that Γ_B is a C^{1,α_2}-manifold, and C is of class C^{1,α_3}, satisfying $\partial C \cap \overline{\mathrm{Sk}(\partial C)} = \emptyset$, then there holds

$$\|\delta(u_\gamma)\|_{L^\infty(\Gamma_B,\mathbb{R}^d)} \leq c\gamma^{-\frac{\alpha+1}{\alpha+d}} \quad \forall \gamma \geq \hat{\gamma},$$

where $\alpha = \min(\alpha_1, \alpha_2, \alpha_3)$.

Proof. (i) Recall the notation $\tau_\gamma = \mathrm{id} + u_\gamma$ and that Γ_B is a C^1-manifold, cf. Assumption 5.3. Hence, there exists a neighborhood $N_\gamma \subset \mathbb{R}^d$ of x_γ, $r > 0$ and a $C^{1,0}$-diffeomorphism $h_\gamma \colon \mathbb{R}^{d-1} \supset B^{\mathbb{R}^{d-1}}(0,r) \to N_\gamma \cap \Gamma_B$. The idea of the proof is to employ Proposition 5.40 for the composed function

$$f_\gamma \colon B^{\mathbb{R}^{d-1}}(0,r) \to \mathbb{R}, \; f_\gamma(y) = (b_C \circ \tau_\gamma \circ h_\gamma)(y).$$

Note that b_C is at least Lipschitz continuous. Furthermore, $\mathcal{U} \hookrightarrow C^{1,\alpha}(\Gamma_B, \mathbb{R}^d)$ by Assumption 5.10, hence $\tau_\gamma \in C^{1,\alpha}(\Gamma_B, \mathbb{R}^d)$. Thus, we obtain for the composition $f_\gamma \in C^{0,1}(B^{\mathbb{R}^{d-1}}(0,r))$. For any $x \in \Gamma_B$ it holds

$$|\delta(u_\gamma)(x)| = |\tau_\gamma(x) - \tilde{P}_C(\tau_\gamma(x))| = (b_C \circ \tau_\gamma)(x),$$

in particular, this implies $f_\gamma \geq 0$. Note that we do not have to satisfy $f_\gamma = 0$ on $\partial A = \partial B^{\mathbb{R}^{d-1}}(0,r)$ since by assumption the maximizer lies already in the interior, cf.

[HSW14, Remark 2.5]. Thus, we can employ Proposition 5.40, and noting that in this case $\theta = 1/(1+(d-1)) = 1/d$, we obtain

$$\|f_\gamma\|_{L^\infty(B^{\mathbb{R}}(0,r))} \leq c \, \|f_\gamma\|^{1-\frac{1}{d}}_{C^{0,1}(B^{\mathbb{R}}(0,r))} \, \|f_\gamma\|^{\frac{1}{d}}_{L^1(B^{\mathbb{R}}(0,r))} . \tag{5.22}$$

By the choice of x_γ it holds

$$\|\delta(u_\gamma)\|_{L^\infty(\Gamma_B,\mathbb{R}^d)} = |\delta(u_\gamma)(x_\gamma)| = \|f_\gamma\|_{L^\infty(B^{\mathbb{R}}(0,r))} .$$

Furthermore, we find constants $c_1(h_\gamma), c_2(h_\gamma)$ depending only on h_γ such that

$$\begin{aligned} \|f_\gamma\|_{L^1(B^{\mathbb{R}}(0,r))} &\leq c_1(h_\gamma) \, \|\delta(u_\gamma)\|_{L^1(\Gamma_B \cap N,\mathbb{R}^d)} &&\leq c_1(h_\gamma) \, \|\delta(u_\gamma)\|_{L^1(\Gamma_B,\mathbb{R}^d)} , \\ \|f_\gamma\|_{C^{0,1}(B^{\mathbb{R}}(0,r))} &\leq c_2(h_\gamma) \, \|\delta(u_\gamma)\|_{C^{0,1}(\Gamma_B \cap N,\mathbb{R}^d)} &&\leq c_2(h_\gamma) \, \|\delta(u_\gamma)\|_{C^{0,1}(\Gamma_B,\mathbb{R}^d)} . \end{aligned}$$

Combining these estimates with (5.22) we arrive at

$$\|\delta(u_\gamma)\|_{L^\infty(\Gamma_B,\mathbb{R}^d)} \leq c_1(h_\gamma)c_2(h_\gamma) \, \|\delta(u_\gamma)\|^{1-\frac{1}{d}}_{C^{0,1}(\Gamma_B,\mathbb{R}^d)} \, \|\delta(u_\gamma)\|^{\frac{1}{d}}_{L^1(\Gamma_B,\mathbb{R}^d)} .$$

Since h_γ depends only on $x_\gamma \in \Gamma_B$ and Γ_B is a fixed C^1-manifold we find an upper bound for $c \geq \max(c_1(h_\gamma), c_2(h_\gamma))$ for all γ. Furthermore, the boundedness of u_γ in $\mathcal{U} \hookrightarrow C^{1,\alpha_1}(\Gamma_B, \mathbb{R}^d)$ implies boundedness of $\|\delta(u_\gamma)\|_{C^{0,1}(\Gamma_B,\mathbb{R}^d)}$, and we conclude

$$\|\delta(u_\gamma)\|_{L^\infty(\Gamma_B,\mathbb{R}^d)} \leq c \, \|\delta(u_\gamma)\|^{\frac{1}{d}}_{L^1(\Gamma_B,\mathbb{R}^d)} .$$

Finally, we invoke Lemma 5.39 to obtain

$$\|\delta(u_\gamma)\|_{L^\infty(\Gamma_B,\mathbb{R}^d)} \leq c\gamma^{-\frac{1}{d}}.$$

(ii) We argue as in the first part of the proof. However, we can now exploit the higher regularity of Γ_B and C to obtain a C^{1,α_2}-diffeomorphism h_γ, and due to Theorem 2.91 we know that (at least locally) $b_C \in C^{1,\alpha_3}$ due to Theorem 2.91. Thus, we obtain for the composition

$$f_\gamma \in C^{1,\alpha}(B^{\mathbb{R}^{d-1}}(0,r)),$$

with $\alpha = \min(\alpha_1, \alpha_2, \alpha_3) > 0$. Using again Proposition 5.40 this leads us to

$$\|f_\gamma\|_{L^\infty(B^{\mathbb{R}}(0,r))} \leq c \, \|f_\gamma\|^{1-\frac{\alpha+1}{\alpha+d}}_{C^{1,\alpha}(B^{\mathbb{R}}(0,r))} \, \|f_\gamma\|^{\frac{\alpha+1}{\alpha+d}}_{L^1(B^{\mathbb{R}}(0,r))} .$$

Analogously to the above we arrive via

$$\|\delta(u_\gamma)\|_{L^\infty(\Gamma_B,\mathbb{R}^d)} \leq c_1(h_\gamma)c_2(h_\gamma) \, \|\delta(u_\gamma)\|^{1-\frac{\alpha+1}{\alpha+d}}_{C^{1,\alpha}(\Gamma_B,\mathbb{R}^d)} \, \|\delta(u_\gamma)\|^{\frac{\alpha+1}{\alpha+d}}_{L^1(\Gamma_B,\mathbb{R}^d)}$$

and Lemma 5.39 at

$$\|\delta(u_\gamma)\|_{L^\infty(\Gamma_B,\mathbb{R}^d)} \leq c\gamma^{-\frac{\alpha+1}{\alpha+d}}.$$

□

Remark 5.42. (i) We expect that in many cases the worst case estimate of Theorem 5.41 is not sharp. See the discussion in [HSW14] on convergence rates in Moreau-Yosida path following, and their dependence on the structure of the Lagrange multiplier in the optimum.

(ii) The assumption that the maximum is attained in the interior of Γ_B, i.e., $x_\gamma \notin \partial\Gamma_B$ is of technical nature to expedite the proof. In our numerical tests we did not experience difficulties if it was not satisfied, and we suspect that it can be weakened or even dropped. It is violated in the examples presented at the end of this chapter.

Finally, we show an estimate on the distance between a solution $\overline{u}$ of (P), and a solution u_γ of the regularized problem.

Theorem 5.43. *[KU15, Theorem 12] Let Assumptions 5.5, 5.6 and 5.10 hold, and $\overline{u} \in \mathcal{U}_{ad}$ be a local solution of* (P) *in which the second order condition Assumption 5.8 is satisfied with $\alpha, \varepsilon > 0$. Further suppose that there exists a family of local solutions (u_γ) of $(\mathrm{MY})_\gamma$ with $u_\gamma \to \overline{u}$ in $\mathcal{U}$, and*

$$\|\delta(u_\gamma)\|_{L^\infty(\Gamma_B,\mathbb{R}^d)} \leq c\gamma^{-s} \text{ for some } s > 0 \text{ and } \gamma \to \infty. \tag{5.23}$$

Finally, denote $V : \gamma \mapsto j_\gamma(u_\gamma)$. Then there exists $\hat{\gamma} > 0$ such that for all $\gamma \geq \hat{\gamma}$ we have

$$0 \leq j(\overline{u}) - V(\gamma) \leq c\gamma^{-s}.$$

Furthermore, it holds

$$\|\overline{u} - u_\gamma\|_{\mathcal{U}} \leq c\gamma^{-s/2}.$$

Proof. We obtain from Theorems 5.21 and 5.29 that

$$V'(\gamma) = \frac{1}{2}\|\delta(u_\gamma)\|^2_{L^2(\Gamma_B,\mathbb{R}^d)} \text{ for all } \gamma \text{ with } \|\overline{u} - u_\gamma\|_{\mathcal{U}} < \varepsilon.$$

Using Lemma 5.39 and (5.23) we obtain

$$V'(\gamma) \leq \frac{1}{2}\|\delta(u_\gamma)\|_{L^\infty(\Gamma_B,\mathbb{R}^d)}\|\delta(u_\gamma)\|_{L^1(\Gamma_B,\mathbb{R}^d)} \leq c\gamma^{-1-s}.$$

Now let $\gamma_2 > \gamma_1$ be large enough. Since $V(\cdot)$ is differentiable for such γ we obtain

$$V(\gamma_2) - V(\gamma_1) = \int_{\gamma_1}^{\gamma_2} V'(t)\,\mathrm{d}t \leq \int_{\gamma_1}^{\gamma_2} ct^{-1-s}\,\mathrm{d}t = c(-\gamma_2^{-s} + \gamma_1^{-s}),$$

with c independent of γ. Hence, passing to the limit $\gamma_2 \to \infty$, it holds (recall (5.7))

$$j(\overline{u}) - V(\gamma_1) = \lim_{\gamma_2\to\infty} j_{\gamma_2}(u_{\gamma_2}) - V(\gamma_1) = \lim_{\gamma_2\to\infty} V(\gamma_2) - V(\gamma_1) \leq c\gamma_1^{-s}.$$

This shows the first claim.

By Theorem 5.21, Assumption 5.7 is satisfied in $\overline{u}$ for any $\gamma > 0$. Choose $0 < r_0 < \frac{\varepsilon}{2}$. For all $\gamma > 0$ such that $\|u_\gamma - \overline{u}\|_{\mathcal{U}} \leq r_0$ it holds

$$\langle H_\gamma^u v, v\rangle_{\mathcal{U}^*,\mathcal{U}} \geq \alpha \|v\|_{\mathcal{U}}^2, \quad \forall v \in \mathcal{U},\ \forall H_\gamma^u \in \partial j_\gamma'(u), \quad \forall u \in B^{\mathcal{U}}(u_\gamma, r_0).$$

As we showed in Lemma 5.24 this implies

$$j_\gamma(u) \geq j_\gamma(u_\gamma) + \frac{\alpha}{2}\|u_\gamma - u\|_{\mathcal{U}}^2 + o(\|u - u_\gamma\|_{\mathcal{U}}^2),\ \forall u \in B^{\mathcal{U}}(u_\gamma, r_0),$$

In particular, it holds

$$\frac{\alpha}{2}\|\overline{u} - u_\gamma\|_{\mathcal{U}}^2 + o(\|u_\gamma - \overline{u}\|_{\mathcal{U}}^2) \leq j_\gamma(\overline{u}) - j_\gamma(u_\gamma) = j(\overline{u}) - V(\gamma) \leq c\gamma^{-s}.$$

Thus, for γ big enough, i.e., $\|\overline{u} - u_\gamma\|_{\mathcal{U}}$ small enough, there exists a $c > 0$ such that

$$\|\overline{u} - u_\gamma\|_{\mathcal{U}}^2 \leq c\gamma^{-s}.$$

□

5.9. Application to shape optimization with point-wise geometric constraints

Let us now show how the above analysis can be applied to a shape optimization problem with geometric constraints. We demonstrate this for the concrete model problem of Chapter 3. However, the setting can be extended to quite general shape optimization problems. For simplicity we restrict ourselves to a situation where the sought-for optimal solution is close to our initial geometry, such that we can work with a fixed reference domain Ω_{ref}.

Recall the pressure tracking shape optimization problem of Chapter 3. To be more precise, we consider the situation in Figure 3.1(b), i.e., the initial domain Ω_0 is a rectangle, and we are allowed to modify the upper boundary denoted by Γ_B. Furthermore, the height of the boundaries on the left and right side can also vary. In this setting it seems reasonable to parametrize the domains via the vertical displacement of the upper boundary. As in Section 3.4, we work with a fixed reference domain $\Omega_{ref} = \Omega_0$, and concentrate on displacements of the boundary as our control, cf. Section 2.11 and 2.12. We choose the Hilbert space

$$\mathcal{U} := H^2(\Gamma_B),$$

and extended the boundary displacements via linear elasticity to domain displacements

$$U = Tu \in C^{1,\alpha}(\Omega_{ref}, \mathbb{R}^2) =: \mathcal{V},$$

cf. Theorem 2.87. In particular, the extension operator $T\colon \mathcal{U} \to \mathcal{V}$ is linear and completely continuous. Working with a fixed reference domain we are only interested in displacements $Tu \in B^{\mathcal{V}}(0,1)$. This can be enforced by restricting the boundary displacements to a suitable

set $\mathcal{U}_{feas} \subset \mathcal{U}$, cf. Section 2.12. We consider geometric constraints in the form of a minimum and maximum diameter of the channel, i.e., $0 < a < b$ and define

$$\mathcal{U}_{ad} := \{u \in \mathcal{U}_{feas} \mid a \leq x_2 + u(x_1) \leq b \text{ for a.e. } x \in \Gamma_B\}.$$

As in Section 3.1, we set $\mathcal{Y} = H^1_D(\Omega_{ref})$, $\mathcal{Z} = \mathcal{Y}^*$, and consider the state equation operator

$$E\colon B^{\mathcal{V}}(0,1) \times \mathcal{Y} \to \mathcal{Z}, \quad \langle E(U,y), \varphi\rangle_{\mathcal{Y}^*,\mathcal{Y}} = (A(U)\nabla(y+y_0), \nabla\varphi)_{L^2(\Omega_{ref})},$$

and the objective

$$J\colon B^{\mathcal{V}}(0,1) \times \mathcal{Y} \to \mathbb{R}, \quad J(U,y) = \frac{1}{2}\int_{\Gamma_B} \left(\frac{t^T\nabla(y+y_0)}{|D\tau_U t|} - p_d\right)^2 |D\tau_U t| \,\mathrm{d}S.$$

Due to Corollaries 3.2 and 3.3, we know that E and J are twice continuously differentiable on $B^{\mathcal{V}}(0,1) \times \mathcal{Y}$. Lemma 3.4 states that $E_y(U,y) \in \mathcal{L}(\mathcal{Y},\mathcal{Z})$ is continuously invertible on $B^{\mathcal{V}}(0,1)\times\mathcal{Y}$. Summarizing, the Assumptions 5.1, 5.2 and 5.3 are satisfied on $\mathcal{U}_{feas}$ respectively $T(\mathcal{U}_{feas})$. In particular, the reduced objective

$$j\colon \mathcal{U}_{feas} \to \mathbb{R}, \quad j(u) = J(T(u), S(T(u))) + \frac{\beta}{2}\|u\|^2_{\mathcal{U}}$$

is twice continuously differentiable. As in Section 3.4, we introduce here a control cost/Tikhonov regularization term $\frac{\beta}{2}\|u\|^2_{\mathcal{U}}$. Although we supposed in this chapter that β is fixed, in practice it can often be iteratively decreased during the course of the penalty method. We present a related numerical experiment in the next section. Note however, that we used $\beta > 0$, i.e., the presence of the term $\frac{\beta}{2}\|u\|^2_{\mathcal{U}}$, in many essential arguments. Hence, most of the obtained results can, in general, not be expected to hold if $\beta \to 0$.

The conditions of Assumption 5.4 can be checked during the setup of a concrete shape optimization problem. For the considered example the projection $\tilde{P}_C$ can be represented by the maximum and minimum operator. These are known to be semismooth, hence Assumption 5.6 is satisfied as well. The second order sufficient condition Assumption 5.8 can not be guaranteed a-priori. Finally, Assumption 5.10 is trivially satisfied for $C = \{x \in \mathbb{R} \mid a \leq x \leq b\}$.

Thus, the pressure tracking shape optimization problem fits into the setting considered in this chapter. We conclude with some numerical experiments in the next section.

5.10. Numerical examples

We implemented the proposed Moreau-Yosida penalty method in `MATLAB` [TM15]. For this we choose a sequence $\gamma_k \to \infty$. The subproblems $(\mathrm{MY})_{\gamma_k}$ are solved with the globalized (semismooth) Newton method described in Section 2.12. The solution of subproblem $(\mathrm{MY})_{\gamma_k}$ is used as initial iterate of the next subproblem. As it is usually done in path following strategies, we do not solve each subproblem exactly. Instead, we iteratively decrease the termination tolerance of the globalized Newton method by setting $\mathrm{TOL}_k = \gamma_k^{-1}$.

In the examples presented below, we employ a fixed factor to increase γ_k. An alternative would be to use the model function proposed in Section 5.6 to steer the γ-update and the termination tolerance TOL_k, cf. [HK06a, HK06b]. However, the efficiency of such a strategy is highly dependent on the concrete problem and various parameters. In our experiments, we did not identify a parameter set which performed consistently better than the fixed γ-update.

Example 5.1. Recall Example 3.1, where the desired tangential velocity profile p_d increases linearly from $\frac{3}{4}$ to $\frac{5}{4}$. We again choose the bi-Laplacian scalar product with weight $w = 1$ for $\mathcal{A}$, and set $\beta = 10^{-2}$. The penalty method is terminated as soon as $\|\delta(u_k)\|_{L^\infty(\Gamma_B,\mathbb{R}^d)} < 10^{-6}$. The optimal domain without geometric constraints is depicted in Figure 3.2. As discussed in the last section, we add now constraints regarding the diameter of the channel to the optimization problem, and set $a = 0.9, b = 1.2$. The progression of the penalty method applied to this problem is presented in Table 5.1. The first column shows the iteration count of the penalty method, the second column the final objective value of each subproblem $(\mathrm{MY})_k$, the third the associated value of the tracking term, and the fourth column the associated norm of the derivative. The sixth column shows the number of iterations the globalized Newton method required to solve $(\mathrm{MY})_{\gamma_k}$ up to TOL_k. Finally, the last two columns show the value of the penalty parameter γ_k, and the infeasibility of the current iterate of the penalty method. It is interesting to note that the infeasibility $\|\delta(u_k)\|_{L^\infty(\Gamma_B,\mathbb{R}^d)}$ decreases approximately with the rate γ_k^{-1}, see also Figure 5.1. Furthermore, we would like to emphasize the small number of iterations the globalized Newton method needed for each subproblem. In particular, in the last iterations one Newton step suffices already.

Table 5.1.: History of the penalty method, Example 5.1

k	$j(u_k)$	$\tilde{j}(u_k)$	$\|j'(u_k)\|_{\mathcal{A}^{-1}}$	# iter descent	γ_k	$\|\delta(u_k)\|_{L^\infty(\Gamma_B,\mathbb{R}^d)}$
1	$5.21{\cdot}10^{-4}$	$7.24{\cdot}10^{-5}$	$4.91{\cdot}10^{-3}$	4	0	$9.27{\cdot}10^{-2}$
2	$6.24{\cdot}10^{-4}$	$1.53{\cdot}10^{-4}$	$3.35{\cdot}10^{-3}$	3	10^1	$6.86{\cdot}10^{-3}$
3	$6.68{\cdot}10^{-4}$	$2.45{\cdot}10^{-4}$	$2.38{\cdot}10^{-3}$	2	10^2	$1.74{\cdot}10^{-3}$
4	$6.83{\cdot}10^{-4}$	$2.78{\cdot}10^{-4}$	$6.32{\cdot}10^{-4}$	3	10^3	$3.54{\cdot}10^{-4}$
5	$6.86{\cdot}10^{-4}$	$2.85{\cdot}10^{-4}$	$1.88{\cdot}10^{-8}$	2	10^4	$6.89{\cdot}10^{-5}$
6	$6.87{\cdot}10^{-4}$	$2.87{\cdot}10^{-4}$	$4.54{\cdot}10^{-9}$	2	10^5	$1.15{\cdot}10^{-5}$
7	$6.87{\cdot}10^{-4}$	$2.87{\cdot}10^{-4}$	$6.85{\cdot}10^{-11}$	1	10^6	$1.15{\cdot}10^{-6}$
8	$6.87{\cdot}10^{-4}$	$2.87{\cdot}10^{-4}$	$1.21{\cdot}10^{-12}$	1	10^7	$1.15{\cdot}10^{-7}$

Example 5.2. Our next example is based on Example 3.2, i.e., the desired velocity profile is given by

$$p_d(x_1) = 1 + \frac{1}{4}\arctan(4x_1 - 3).$$

We impose the diameter constraints $a = 0.9, b = 1.1$. This renders the final domain without geometric constraints completely infeasible. The final domains are compared in Figure 5.2. The progression of the penalty method applied to this problem is presented in Table 5.2. The first column shows again the iteration count of the penalty method, the second column the final objective value of each subproblem $(\mathrm{MY})_k$, the third the associated value of the tracking term, and the fourth column the associated norm of the derivative. The sixth column

shows the number of iterations the globalized Newton method required to solve $(\mathrm{MY})_k$ up to TOL_k. Finally, the last two columns show the value of the penalty parameter γ_k, and the infeasibility of the current iterate of the penalty method. We again observe an approximately linear convergence rate of the infeasibility measure, and very few iterations of the globalized Newton method.

Table 5.2.: History of the penalty method, Example 5.2

k	$j(u_k)$	$\tilde{j}(u_k)$	$\|j'(u_k)\|_{\mathcal{A}^{-1}}$	# iter descent	γ_k	$\|\delta(u_k)\|_{L^\infty(\Gamma_B,\mathbb{R}^d)}$
1	$5.77{\cdot}10^{-2}$	$4.70{\cdot}10^{-2}$	$4.43{\cdot}10^{-3}$	5	0	$7.61{\cdot}10^{-1}$
2	$7.04{\cdot}10^{-2}$	$6.69{\cdot}10^{-2}$	$8.55{\cdot}10^{-3}$	3	10^1	$3.51{\cdot}10^{-2}$
3	$7.33{\cdot}10^{-2}$	$7.19{\cdot}10^{-2}$	$7.89{\cdot}10^{-3}$	2	10^2	$6.04{\cdot}10^{-3}$
4	$7.38{\cdot}10^{-2}$	$7.28{\cdot}10^{-2}$	$3.46{\cdot}10^{-4}$	3	10^3	$7.17{\cdot}10^{-4}$
5	$7.39{\cdot}10^{-2}$	$7.29{\cdot}10^{-2}$	$2.13{\cdot}10^{-8}$	3	10^4	$1.03{\cdot}10^{-4}$
6	$7.39{\cdot}10^{-2}$	$7.29{\cdot}10^{-2}$	$1.29{\cdot}10^{-8}$	2	10^5	$1.49{\cdot}10^{-5}$
7	$7.39{\cdot}10^{-2}$	$7.29{\cdot}10^{-2}$	$3.61{\cdot}10^{-10}$	1	10^6	$2.08{\cdot}10^{-6}$
8	$7.39{\cdot}10^{-2}$	$7.29{\cdot}10^{-2}$	$3.53{\cdot}10^{-11}$	1	10^7	$2.34{\cdot}10^{-7}$

Example 5.3. Finally, we present an example where we iteratively decrease the Tikhonov parameter β. We consider the setting of Example 5.1, but decrease β in each iteration of the penalty method by the factor 10. The progression of the penalty method applied to this problem is presented in Table 5.3. As one can see, the behavior of the algorithm is essentially the same as for the fixed β value. We again observe that the infeasibility $\|\delta(u_k)\|_{L^\infty(\Gamma_B,\mathbb{R}^d)}$ decreases approximately with the rate γ_k^{-1}, see also Figure 5.1. The number of iterations of the globalized Newton method is still very low, in particular we still observe fast local convergence of Newton's method. Of course, the number of overall CG iterations increases compared to Example 5.1 It is also interesting to note that the value of the tracking term is quite close to the one in Example 5.1.

Table 5.3.: History of the penalty method, Example 5.3

k	$j(u_k)$	$\tilde{j}(u_k)$	$\|j'(u_k)\|_{\mathcal{A}^{-1}}$	# iter descent	γ_k	$\|\delta(u_k)\|_{L^\infty(\Gamma_B,\mathbb{R}^d)}$	β
1	$1.72{\cdot}10^{-2}$	$9.67{\cdot}10^{-3}$	$4.83{\cdot}10^{-3}$	1	0	$0.00{\cdot}10^{-0}$	1
2	$3.94{\cdot}10^{-3}$	$8.26{\cdot}10^{-4}$	$1.39{\cdot}10^{-3}$	3	10^1	$6.83{\cdot}10^{-3}$	10^{-1}
3	$6.70{\cdot}10^{-4}$	$2.39{\cdot}10^{-4}$	$6.04{\cdot}10^{-3}$	3	10^2	$1.86{\cdot}10^{-3}$	10^{-2}
4	$3.20{\cdot}10^{-4}$	$2.76{\cdot}10^{-4}$	$4.10{\cdot}10^{-8}$	3	10^3	$3.63{\cdot}10^{-4}$	10^{-3}
5	$2.88{\cdot}10^{-4}$	$2.83{\cdot}10^{-4}$	$3.58{\cdot}10^{-8}$	2	10^4	$5.34{\cdot}10^{-5}$	10^{-4}
6	$2.84{\cdot}10^{-4}$	$2.84{\cdot}10^{-4}$	$2.06{\cdot}10^{-8}$	1	10^5	$5.80{\cdot}10^{-6}$	10^{-5}
7	$2.84{\cdot}10^{-4}$	$2.84{\cdot}10^{-4}$	$6.63{\cdot}10^{-8}$	5	10^6	$5.84{\cdot}10^{-7}$	10^{-6}

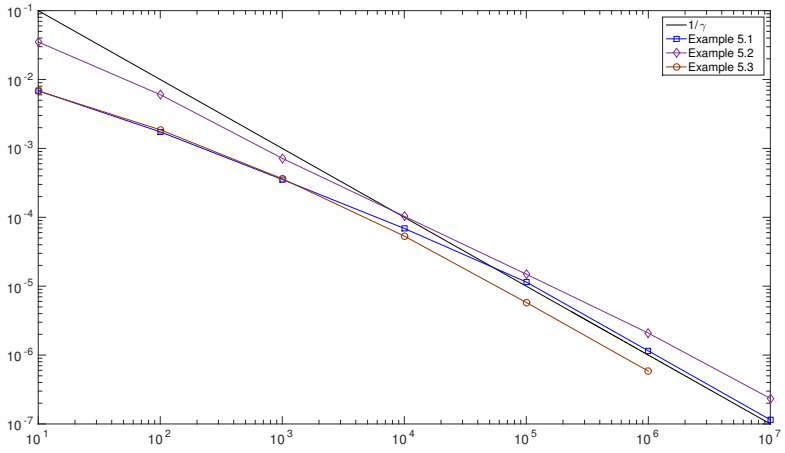
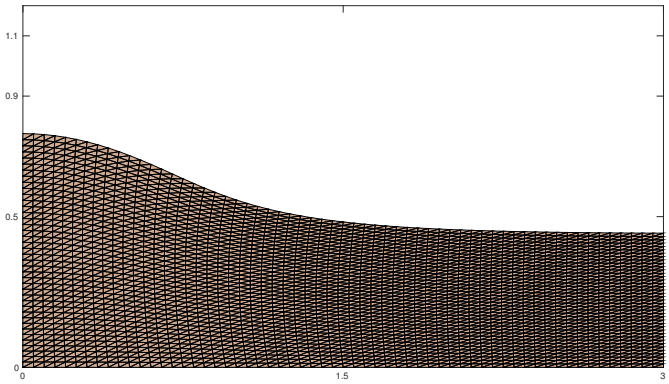
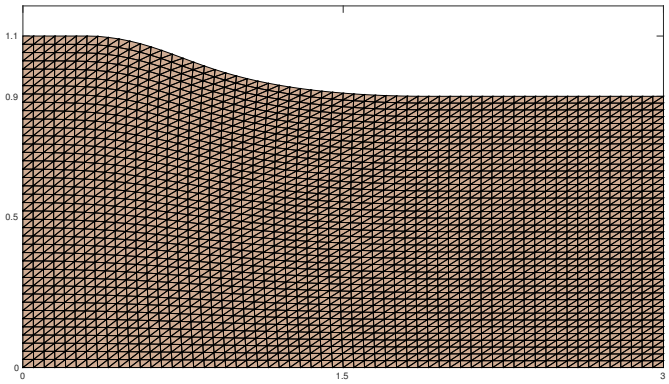


Figure 5.1.: Infeasibility $\|\delta(u_k)\|_{L^\infty(\Gamma_B,\mathbb{R}^d)}$ for the three examples



(a) Example 3.2



(b) Example 5.2

Figure 5.2.: Comparison final domains Example 3.2 and Example 5.2

6. Drag minimization in Stokes flow

In this chapter we consider the shape optimization problem of minimizing the drag of a body B in a viscous incompressible fluid. To be more precise, we suppose that the fluid is described by the stationary Stokes equations. We focus here on this simple flow problem to demonstrate our method. However, our technique can be easily extended to the instationary Navier-Stokes equations.

We denote the velocity of the flow by v, the pressure by p and the kinematic viscosity by $\nu > 0$. The Stokes equations on a bounded domain Ω in absence of body forces are given by

$$\begin{aligned} -\nu\Delta\tilde{v} + \nabla\tilde{p} &= 0 \quad \text{in } \Omega, \\ \nabla\cdot\tilde{v} &= 0 \quad \text{in } \Omega, \end{aligned} \tag{6.1}$$

coupled with suitable boundary conditions. For small Reynolds numbers, i.e., relatively large viscosity, the functional

$$\tilde{J}(\Omega,\tilde{v}) := \tfrac{\nu}{2}\int_\Omega \sum_{i=1}^{d} \nabla\tilde{v}_i^T\nabla\tilde{v}_i \,\mathrm{d}\tilde{x}$$

coincides with the usual hydrodynamical drag of a body B immersed in the fluid, cf., e.g., [BFCLS97, Section 2]. Under suitable conditions there exists a unique solution $(\tilde{v},\tilde{p}) = \tilde{S}(\Omega)$ of (6.1), and one can define the shape functional

$$\mathrm{j}\colon \mathcal{O}\to\mathbb{R}, \quad \mathrm{j}(\Omega) = \tilde{J}(\Omega,\tilde{S}(\Omega))$$

for an appropriate set of domains $\mathcal{O}\subset\mathcal{P}(\mathbb{R}^d)$. The associated shape optimization problem reads

$$\min_{\Omega\in\mathcal{O}} \mathrm{j}(\Omega) \quad \text{s.t.} \quad \Omega\in\mathcal{O}_{ad}, \tag{6.2}$$

where $\mathcal{O}_{ad}$ is the set of admissible domains and may present additional constraints, e.g., a volume or center of mass constraint.

We discuss in the next section a general setting in which the drag functional j is well defined and shape differentiable. As usual, we employ the function space parametrization approach, and exploit the formula (2.42)

$$\mathrm{j}(\tau_U(\Omega)) = \mathrm{j}_\Omega(U) = J(U,S(U)),$$

where J, S are the transformed objective, respectively the transformed design-to-state operator, cf. Section 2.14. We then proceed in Section 6.2 by discussing the constraint $\Omega\in\mathcal{O}_{ad}$. Finally, we present a concrete numerical experiment in Section 6.3.

Shape differentiability of the drag for the Stokes and Navier-Stokes equations has been discussed by various authors. Formal computations were, for example, carried out in [Pir73]. Simon rigorously obtained shape differentiability of the drag for Stokes flow in a $W^{2,\infty}$ domain in [Sim91]. Later, together with his coauthors, he studied stationary Navier-Stokes flows, and obtained shape differentiability of the drag in a Lipschitz domain with respect to Lipschitz displacements in [BFCLS97]. Shape differentiability of instationary flows is quite intricate, we mention [Lin12], where shape differentiability of the instationary, incompressible Navier-Stokes equations is discussed.

6.1. The Stokes equations and function space parametrization

Let us describe the problem setting in detail. We consider a nonempty, bounded Lipschitz domain $\Omega_0 \subset \mathbb{R}^d$ and all its images under C^1 transformations, i.e.

$$\mathcal{O} := \mathcal{O}_\Theta(\Omega_0) \quad \text{where } \Theta = C^1(\overline{\mathbb{R}^d}, \mathbb{R}^d). \tag{6.3}$$

Recall the notations $\mathcal{O}_\Theta(\Omega_0) = \{\Omega \subset \mathbb{R}^d \mid \Omega = \tau(\Omega_0), \tau \in \mathcal{F}(\Theta)\}$ from (2.5) and $\mathcal{F}(\Theta)$ see (2.3). In particular, all domains in $\mathcal{O}$ are Lipschitz domains. Note that we could also work with small enough Lipschitz deformations as in [BFCLS97].

We refrained so far from stating boundary conditions for the Stokes equations. The reason for this is that, while this choice influences the spaces in which the variational Stokes equations are posed, it has no direct impact on the considerations we will lay out in this section. Thus, we decided to work in an abstract setting, and assume that the associated equation is well-posed and uniquely solvable. For $\Omega \in \mathcal{O}$, we consider two Hilbert spaces $X(\Omega), M(\Omega)$, and set $\mathcal{Y}(\Omega) = X(\Omega) \times M(\Omega)$. Let us introduce the operators $\tilde{E}^\Omega \colon \mathcal{Y}(\Omega) \to \mathcal{Y}(\Omega)^*$ defined as

$$\langle \tilde{E}^\Omega(\tilde{v}, \tilde{p}), (\tilde{\varphi}, \tilde{\psi}) \rangle_{\mathcal{Y}(\Omega)^*, \mathcal{Y}(\Omega)} = \nu \sum_{i=1}^{d} (\nabla \tilde{v}_i, \nabla \tilde{\varphi}_i)_{L^2(\Omega)} - (\tilde{p}, \operatorname{div}(\tilde{\varphi}))_{L^2(\Omega)} + \left(\operatorname{div}(\tilde{v}), \tilde{\psi}\right)_{L^2(\Omega)}.$$

This corresponds to the variational velocity-pressure formulation of the Stokes equations.

Assumption 6.1. *Let $\mathcal{O}$ be given by* (6.3)*. For every $\Omega \in \mathcal{O}$ the operator $\tilde{E}^\Omega \colon \mathcal{Y}(\Omega) \to \mathcal{Y}(\Omega)^*$ is well defined, and for every $f \in \mathcal{Y}(\Omega)^*$ there exists a unique solution $(\tilde{v}, \tilde{p}) \in \mathcal{Y}(\Omega)$ of*

$$\tilde{E}^\Omega(\tilde{v}, \tilde{p}) = f \quad \text{in } \mathcal{Y}(\Omega)^*$$

which satisfies $\|(\tilde{v}, \tilde{p})\|_{\mathcal{Y}(\Omega)} \leq c \|f\|_{\mathcal{Y}(\Omega)^*}$ for some constant $c > 0$ independent of f.

Remark 6.1. It is a classical result that this assumption is satisfied in the case of Dirichlet conditions on the whole boundary $\partial\Omega$ where $X(\Omega) = H_0^1(\Omega, \mathbb{R}^2)$, $M(\Omega) = L_0^2(\Omega)$, see for example [GR87, Section 1.5], or [Tem77, Gal11]. If one is interested in a situation where Ω is only a part of a much larger (or even infinite) domain, e.g., a channel, than it is more natural not to describe Dirichlet data on the outflow boundary. Instead, a very popular choice are free outflow boundary conditions, often referred to as 'do-nothing' conditions, see for example

[HRT96]. For this the boundary is decomposed into $\partial\Omega = \Gamma_D \cup \Gamma_{out}$, i.e., a Dirichlet part and an outflow boundary. On Γ_{out} one requires

$$\tilde{p}n - \nu\partial_n\tilde{v} = 0. \tag{6.4}$$

The corresponding spaces are $X(\Omega) = H^1_D(\Omega)$, $M(\Omega) = L^2(\Omega)$, where $H^1_D(\Omega)$ denotes the space of vector fields in H^1 with zero trace on Γ_D. Although the free outflow boundary conditions work very well in practice, only few theoretical results concerning them are available. We refer to the recent publication [BM14] where solvability of the stationary Navier-Stokes equations with a 'directional do-nothing' boundary condition is discussed. For the case of an outflow boundary this condition coincides with the classical 'do-nothing' condition.

Hence, supposing some given (partial) Dirichlet datum is smoothly extended to $\tilde{v}^D \in X(\Omega)$ we can formulate the state equation on $\Omega \in \mathcal{O}$

$$\text{find } (\tilde{v},\tilde{p}) \in \mathcal{Y}(\Omega) \text{ satisfying } \tilde{E}^\Omega(\tilde{v},\tilde{p}) = -\tilde{E}^\Omega(\tilde{v}^D, 0). \tag{6.5}$$

As announced in the introduction, the objective functional is given by

$$\tilde{J}\colon \{(\Omega,\tilde{v}) \mid \Omega \in \mathcal{O}, \tilde{v} \in X(\Omega)\} \to \mathbb{R}, \quad \tilde{J}(\Omega,\tilde{v}) = \tfrac{\nu}{2}\sum_{i=1}^{d} (\nabla\tilde{v}_i, \nabla\tilde{v}_i)_{L^2(\Omega)}.$$

Assumption 6.1 allows us to employ the standard machinery of Section 2.14. There exists a design-to-state operator

$$\tilde{S}\colon \mathcal{O} \ni \Omega \mapsto S(\Omega) \in \mathcal{Y}(\Omega) \text{ with } \tilde{E}^\Omega(\tilde{S}) = -\tilde{E}^\Omega(\tilde{v}^D, 0) \text{ for all } \Omega \in \mathcal{O}.$$

Thus the shape functional

$$\mathfrak{j}\colon \mathcal{O} \to \mathbb{R}, \quad \mathfrak{j}(\Omega) = \tilde{J}(\Omega, \tilde{S}(\Omega))$$

is well defined. Recall for $\Omega \in \mathcal{O}$ the localized functional

$$\mathfrak{j}_\Omega\colon B^\Theta(0,1) \to \mathbb{R}, \quad \mathfrak{j}_\Omega(U) = \mathfrak{j}(\tau_U(\Omega))$$

and the relationship between the derivatives of these two functionals, see Theorems 2.31 and 2.39. As usual, we now want to use the characterization (2.42) to compute the derivatives of $\mathfrak{j}_\Omega$, and thus of $\mathfrak{j}$. For this we suppose further that for all $\Omega \in \mathcal{O}$ it holds

$$\mathcal{Y}(\Omega) = \{(\tilde{v},\tilde{p}) \circ \tau \mid (\tilde{v},\tilde{p}) \in \mathcal{Y}(\tau(\Omega))\}$$

for all $\tau \in \mathcal{F}(\Theta)$, and that the mapping

$$\mathcal{Y}(\tau(\Omega)) \ni (\tilde{v},\tilde{p}) \mapsto (v,p) := (\tilde{v},\tilde{p}) \circ \tau \in \mathcal{Y}(\Omega)$$

is a homeomorphism. Note that Lemma 2.9 asserts this property for the spaces $L^p, W^{1,p}, W^{1,p}_0$.

Consider now a domain $\Omega \in \mathcal{O}$ for which we would like to evaluate the shape derivatives of $\mathfrak{j}$, and abbreviate $\mathcal{Y}(\Omega) = \mathcal{Y}$. We obtain the transformed state equation operator as

$$E\colon B^\Theta(0,1) \times \mathcal{Y} \to \mathcal{Y}^*, \quad E(U,v,p) := \tilde{E}^{\tau_U(\Omega)}(v \circ \tau_U^{-1}, p \circ \tau_U^{-1}) \circ \tau_U^{-1},$$

and the transformed objective as

$$J\colon B^\Theta(0,1) \times X(\Omega) \to \mathbb{R}, \quad J(U,v) := J(\tau_U(\Omega), v \circ \tau_U^{-1}).$$

For completeness and the convenience of the reader we spell out the detailed formulas of E, J, and their partial derivatives in the next subsection.

6.1.1. Partial derivatives

Recall the rules for computing derivatives from Lemma 2.84, as well as the map (3.2)

$$U \mapsto A(U) := D\tau_U^{-1} D\tau_U^{-T} \det(D\tau_U),$$

and its derivative in a direction $V \in \Theta$ given by (3.3)

$$M^V(U) := D\tau_U^{-1} \left(-DV D\tau_U^{-1} - D\tau_U^{-T} DV^T + \mathcal{I} \operatorname{tr}(D\tau_U^{-1} DV) \right) D\tau_U^{-T} \det(D\tau_U),$$

where $\mathcal{I}$ denotes the identity matrix in $\mathbb{R}^{d\times d}$.

Partial derivatives of E

We begin with the partial derivatives of the state equation operator $E\colon B^\Theta(0,1) \times \mathcal{Y} \to \mathcal{Y}^*$ given by

$$\begin{aligned}\langle E(U,v,p),(\varphi,\psi)\rangle_{\mathcal{Y}^*,\mathcal{Y}} = \nu \sum_{i=1}^{d} \big(& A(U)\nabla v_i, \nabla\varphi_i)_{L^2(\Omega)} \\ & - \left(\operatorname{tr}(D\tau_U^{-T}\nabla\varphi)\det(D\tau_U), p\right)_{L^2(\Omega)} \\ & + \left(\operatorname{tr}(D\tau_U^{-T}\nabla v)\det(D\tau_U), \psi\right)_{L^2(\Omega)}.\end{aligned}$$

For clarity of presentation we employ again the short notation

$$(\varphi,\psi)^* E(U,v,p) := \langle E(U,v,p),(\varphi,\psi)\rangle_{\mathcal{Y}^*,\mathcal{Y}}.$$

It holds

$$\begin{aligned}(\varphi,\psi)^* E_{(v,p)}(U,v,p)(w,q) = \nu \sum_{i=1}^{d} \big(& A(U)\nabla w_i, \nabla\varphi_i)_{L^2(\Omega)} \\ & - \left(\operatorname{tr}(D\tau_U^{-T}\nabla\varphi)\det(D\tau_U), q\right)_{L^2(\Omega)} \\ & + \left(\operatorname{tr}(D\tau_U^{-T}\nabla w)\det(D\tau_U), \psi\right)_{L^2(\Omega)}, \\ (\varphi,\psi)^* E_U(U,v,p)V = \nu \sum_{i=1}^{d} & \left(M^V(U)\nabla v_i, \nabla\varphi_i\right)_{L^2(\Omega)} \\ & + \left(\operatorname{tr}(D\tau_U^{-T} DV^T D\tau_U^{-T}\nabla\varphi)\det(D\tau_U), p\right)_{L^2(\Omega)} \\ & - \left(\operatorname{tr}(D\tau_U^{-T}\nabla\varphi)\operatorname{tr}(D\tau_U^{-1} DV)\det(D\tau_U), p\right)_{L^2(\Omega)} \\ & - \left(\operatorname{tr}(D\tau_U^{-T} DV^T D\tau_U^{-T}\nabla v)\det(D\tau_U), \psi\right)_{L^2(\Omega)} \\ & + \left(\operatorname{tr}(D\tau_U^{-T}\nabla v)\operatorname{tr}(D\tau_U^{-1} DV)\det(D\tau_U), \psi\right)_{L^2(\Omega)}.\end{aligned}$$

Evaluated at $U = 0$ the expressions simplify to

$$\begin{aligned}
(\varphi,\psi)^* E_{(v,p)}(0,v,p)(w,q) &= \nu \sum_{i=1}^{d} (\ \nabla w_i, \nabla \varphi_i)_{L^2(\Omega)} - \big(\operatorname{div}(\varphi), q\big)_{L^2(\Omega)} + (\operatorname{div}(w), \psi)_{L^2(\Omega)}\,, \\
(\varphi,\psi)^* E_U(0,y)V &= \nu \sum_{i=1}^{d} \Big(\ (\mathcal{I}\operatorname{div}(V) - DV - DV^T)\nabla v_i, \nabla \varphi_i\Big)_{L^2(\Omega)} \\
&\quad + \Big(\operatorname{tr}(DV^T \nabla \varphi), p\Big)_{L^2(\Omega)} - \big(\operatorname{div}(\varphi)\operatorname{div}(V), p\big)_{L^2(\Omega)} \\
&\quad - \Big(\operatorname{tr}(DV^T \nabla v), \psi\Big)_{L^2(\Omega)} + (\operatorname{div}(v)\operatorname{div}(V), \psi)_{L^2(\Omega)}\,.
\end{aligned}$$

The second partial derivatives are given by

$$\begin{aligned}
(\varphi,\psi)^* E_{(v,p),(v,p)}(U,v,p) &= 0, \\
(\varphi,\psi)^* E_{(v,p),U}(U,y)((w,q),V) &= (\varphi,\psi)^* E_{U,(v,p)}(U,y)(V,(w,q)) \\
&= \nu \sum_{i=1}^{d} \Big(\ M^V(U)\nabla w_i, \nabla \varphi_i\Big)_{\mathbf{L}^2_{ref}} \\
&\quad + \Big(\operatorname{tr}(D\tau_U^{-T} DV^T D\tau_U^{-T} \nabla \varphi)\det(D\tau_U), q\Big)_{L^2(\Omega_{ref})} \\
&\quad - \Big(\operatorname{tr}(D\tau_U^{-T} \nabla \varphi)\operatorname{tr}(D\tau_U^{-1} DV)\det(D\tau_U), q\Big)_{L^2(\Omega_{ref})} \\
&\quad - \Big(\operatorname{tr}(D\tau_U^{-T} DV^T D\tau_U^{-T} \nabla w)\det(D\tau_U), \psi\Big)_{L^2(\Omega_{ref})} \\
&\quad + \Big(\operatorname{tr}(D\tau_U^{-T} \nabla w)\operatorname{tr}(D\tau_U^{-1} DV)\det(D\tau_U), \psi\Big)_{L^2(\Omega_{ref})}\,,
\end{aligned}$$

which simplifies to

$$\begin{aligned}
(\varphi,\psi)^* E_{(v,p),U}(0,v,p)((w,q),V) &= (\varphi,\psi)^* E_{U,(v,p)}(0,v,p)(V,(w,q)) \\
&= \nu \sum_{i=1}^{d} \Big(\ (\mathcal{I}\operatorname{div}(V) - DV - DV^T)\nabla w_i, \nabla \varphi_i\Big)_{L^2(\Omega)} \\
&\quad + \Big(\operatorname{tr}(DV^T \nabla \varphi), q\Big)_{L^2(\Omega)} - \big(\operatorname{div}(\varphi)\operatorname{div}(V), q\big)_{L^2(\Omega)} \\
&\quad - \Big(\operatorname{tr}(DV^T \nabla w), \psi\Big)_{L^2(\Omega)} + (\operatorname{div}(w)\operatorname{div}(V), \psi)_{L^2(\Omega)}\,.
\end{aligned}$$

The second derivative with respect to U is a bit lengthy. It holds

$$
\begin{aligned}
&(\varphi,\psi)^* E_{UU}(U,y)(V,W) = \\
&\quad - \nu \sum_{i=1}^{d} \Big(D\tau_U^{-1} DW M^V(U) \nabla v_i, \nabla \varphi_i \Big)_{L^2(\Omega)} \\
&\quad - \nu \sum_{i=1}^{d} \Big(M^V(U) DW^T D\tau_U^{-T} \nabla v_i, \nabla \varphi_i \Big)_{L^2(\Omega)} \\
&\quad + \nu \sum_{i=1}^{d} \Big(M^V(U) \operatorname{tr}(D\tau_U^{-1} DW) \nabla v_i, \nabla \varphi_i \Big)_{L^2(\Omega)} \\
&\quad + \nu \sum_{i=1}^{d} \Big(D\tau_U^{-1} DV D\tau_U^{-1} DW A(U) \nabla v_i, \nabla \varphi_i \Big)_{L^2(\Omega)} \\
&\quad + \nu \sum_{i=1}^{d} \Big(A(U) DW^T D\tau_U^{-T} DV^T D\tau_U^{-T} \nabla v_i, \nabla \varphi_i \Big)_{L^2(\Omega)} \\
&\quad - \nu \sum_{i=1}^{d} \Big(A(U) \operatorname{tr}(D\tau_U^{-1} DW D\tau_U^{-1} DV) \nabla v_i, \nabla \varphi_i \Big)_{L^2(\Omega)} \\
&\quad - \Big(\operatorname{tr}(D\tau_U^{-T} DW^T D\tau_U^{-T} DV^T D\tau_U^{-T} \nabla \varphi) \det(D\tau_U), p \Big)_{L^2(\Omega)} \\
&\quad - \Big(\operatorname{tr}(D\tau_U^{-T} DV^T D\tau_U^{-T} DW^T D\tau_U^{-T} \nabla \varphi) \det(D\tau_U), p \Big)_{L^2(\Omega)} \\
&\quad + \Big(\operatorname{tr}(D\tau_U^{-T} DV^T D\tau_U^{-T} \nabla \varphi) \operatorname{tr}(D\tau_U^{-1} DW) \det(D\tau_U), p \Big)_{L^2(\Omega)} \\
&\quad + \Big(\operatorname{tr}(D\tau_U^{-T} DW^T D\tau_U^{-T} \nabla \varphi) \operatorname{tr}(D\tau_U^{-1} DV) \det(D\tau_U), p \Big)_{L^2(\Omega)} \\
&\quad + \Big(\operatorname{tr}(D\tau_U^{-T} \nabla \varphi) \operatorname{tr}(D\tau_U^{-1} DW D\tau_U^{-1} DV) \det(D\tau_U), p \Big)_{L^2(\Omega)} \\
&\quad - \Big(\operatorname{tr}(D\tau_U^{-T} \nabla \varphi) \operatorname{tr}(D\tau_U^{-1} DV) \operatorname{tr}(D\tau_U^{-1} DW) \det(D\tau_U), p \Big)_{L^2(\Omega)} \\
&\quad + \Big(\operatorname{tr}(D\tau_U^{-T} DW^T D\tau_U^{-T} DV^T D\tau_U^{-T} \nabla v) \det(D\tau_U), \psi \Big)_{L^2(\Omega)} \\
&\quad + \Big(\operatorname{tr}(D\tau_U^{-T} DV^T D\tau_U^{-T} DW^T D\tau_U^{-T} \nabla v) \det(D\tau_U), \psi \Big)_{L^2(\Omega)} \\
&\quad - \Big(\operatorname{tr}(D\tau_U^{-T} DV^T D\tau_U^{-T} \nabla v) \operatorname{tr}(D\tau_U^{-1} DW) \det(D\tau_U), \psi \Big)_{L^2(\Omega)} \\
&\quad - \Big(\operatorname{tr}(D\tau_U^{-T} DW^T D\tau_U^{-T} \nabla v) \operatorname{tr}(D\tau_U^{-1} DV) \det(D\tau_U), \psi \Big)_{L^2(\Omega)} \\
&\quad - \Big(\operatorname{tr}(D\tau_U^{-T} \nabla v) \operatorname{tr}(D\tau_U^{-1} DW D\tau_U^{-1} DV) \det(D\tau_U), \psi \Big)_{L^2(\Omega)} \\
&\quad + \Big(\operatorname{tr}(D\tau_U^{-T} \nabla v) \operatorname{tr}(D\tau_U^{-1} DV) \operatorname{tr}(D\tau_U^{-1} DW) \det(D\tau_U), \psi \Big)_{L^2(\Omega)} .
\end{aligned}
$$

Evaluated at $U = 0$ the expression simplifies to

$$\begin{aligned}
(\varphi,\psi)^* E_{UU}(0,y)(V,W) = &- \nu \sum_{i=1}^{d} \Big(DW(\mathcal{I}\operatorname{div}(V) - DV - DV^T)\nabla v_i, \nabla\varphi_i \Big)_{L^2(\Omega)} \\
&- \nu \sum_{i=1}^{d} \Big((\mathcal{I}\operatorname{div}(V) - DV - DV^T) DW^T \nabla v_i, \nabla\varphi_i \Big)_{L^2(\Omega)} \\
&+ \nu \sum_{i=1}^{d} \Big((\mathcal{I}\operatorname{div}(V) - DV - DV^T) \operatorname{div}(W) \nabla v_i, \nabla\varphi_i \Big)_{L^2(\Omega)} \\
&+ \nu \sum_{i=1}^{d} (DV\,DW\nabla v_i, \nabla\varphi_i)_{L^2(\Omega)} \\
&+ \nu \sum_{i=1}^{d} \Big(DW^T DV^T \nabla v_i, \nabla\varphi_i \Big)_{L^2(\Omega)} \\
&- \nu \sum_{i=1}^{d} (\operatorname{tr}(DW\,DV)\nabla v_i, \nabla\varphi_i)_{L^2(\Omega)} \\
&- \Big(\operatorname{tr}(DW^T DV^T \nabla\varphi), p \Big)_{L^2(\Omega)} \\
&- \Big(\operatorname{tr}(DV^T DW^T \nabla\varphi), p \Big)_{L^2(\Omega)} \\
&+ \Big(\operatorname{tr}(DV^T \nabla\varphi) \operatorname{div}(W), p \Big)_{L^2(\Omega)} \\
&+ \Big(\operatorname{tr}(DW^T \nabla\varphi) \operatorname{div}(V), p \Big)_{L^2(\Omega)} \\
&+ (\operatorname{div}\varphi \operatorname{tr}(DW\,DV), p)_{L^2(\Omega)} \\
&- (\operatorname{div}\varphi \operatorname{div}(V) \operatorname{div}(W), p)_{L^2(\Omega)} \\
&+ \Big(\operatorname{tr}(DW^T DV^T \nabla v), \psi \Big)_{L^2(\Omega)} \\
&+ \Big(\operatorname{tr}(DV^T DW^T \nabla v), \psi \Big)_{L^2(\Omega)} \\
&- \Big(\operatorname{tr}(DV^T \nabla v) \operatorname{div}(W), \psi \Big)_{L^2(\Omega)} \\
&- \Big(\operatorname{tr}(DW^T \nabla v) \operatorname{div}(V), \psi \Big)_{L^2(\Omega)} \\
&- (\operatorname{div}(v) \operatorname{tr}(DW\,DV), \psi)_{L^2(\Omega)} \\
&+ (\operatorname{div}(v) \operatorname{div}(V) \operatorname{div}(W), \psi)_{L^2(\Omega)} .
\end{aligned}$$

Partial derivatives of J

The partial derivatives of the functional

$$J(U,v) = \tfrac{\nu}{2} \sum_{i=1}^{d} (A(U)\nabla v_i, \nabla v_i)_{L^2(\Omega)}$$

are given by

$$\begin{aligned}\langle J_v(U,v),\varphi\rangle_{X^*,X} &= \nu\sum_{i=1}^{d}\left(A(U)\nabla v_i,\nabla\varphi_i\right)_{L^2(\Omega)},\\ \langle J_U(U,v),V\rangle_{\Theta^*,\Theta} &= \tfrac{\nu}{2}\sum_{i=1}^{d}\Big(M^V(U)\nabla v_i,\nabla v_i\Big)_{L^2(\Omega)},\end{aligned}$$

where we abbreviated $X = X(\Omega)$. Evaluated at $U = 0$ the expressions simplify to

$$\begin{aligned}\langle J_v(0,v),\varphi\rangle_{X^*,X} &= \nu\sum_{i=1}^{d}\left(\nabla v_i,\nabla\varphi_i\right)_{L^2(\Omega)},\\ \langle J_U(0,v),V\rangle_{\Theta^*,\Theta} &= \tfrac{\nu}{2}\sum_{i=1}^{d}\Big((\mathcal{I}\operatorname{div}(V)-DV-DV^T)\nabla v_i,\nabla v_i\Big)_{L^2(\Omega)}.\end{aligned}$$

The second partial derivatives are given by

$$\begin{aligned}\langle J_{vv}(U,v)\varphi,w\rangle_{X^*,X} &= \nu\sum_{i=1}^{d}\left(A(U)\nabla\varphi_i,\nabla w_i\right)_{L^2(\Omega)},\\ \langle J_{vU}(U,v)\varphi,V\rangle_{\Theta^*,\Theta} &= \langle J_{Uv}(U,v)V,\varphi\rangle_{X^*,X}\\ &= \nu\sum_{i=1}^{d}\Big(M^V(U)\nabla v_i,\nabla\varphi_i\Big)_{L^2(\Omega)},\end{aligned}$$

and

$$\begin{aligned}\langle\, J_{UU}(U,v)V,W\rangle_{\Theta^*,\Theta} = &-\tfrac{\nu}{2}\sum_{i=1}^{d}\Big(D\tau_U^{-1}DWM^V(U)\nabla v_i,\nabla v_i\Big)_{L^2(\Omega)}\\ &-\tfrac{\nu}{2}\sum_{i=1}^{d}\Big(M^V(U)DW^TD\tau_U^{-T}\nabla v_i,\nabla v_i\Big)_{L^2(\Omega)}\\ &+\tfrac{\nu}{2}\sum_{i=1}^{d}\Big(M^V(U)\operatorname{tr}(D\tau_U^{-1}DW)\nabla v_i,\nabla v_i\Big)_{L^2(\Omega)}\\ &+\tfrac{\nu}{2}\sum_{i=1}^{d}\Big(D\tau_U^{-1}DVD\tau_U^{-1}DWA(U)\nabla v_i,\nabla v_i\Big)_{L^2(\Omega)}\\ &+\tfrac{\nu}{2}\sum_{i=1}^{d}\Big(A(U)DW^TD\tau_U^{-T}DV^TD\tau_U^{-T}\nabla v_i,\nabla v_i\Big)_{L^2(\Omega)}\\ &-\tfrac{\nu}{2}\sum_{i=1}^{d}\Big(A(U)\operatorname{tr}(D\tau_U^{-1}DWD\tau_U^{-1}DV)\nabla v_i,\nabla v_i\Big)_{L^2(\Omega)}.\end{aligned}$$

Evaluated at $U = 0$ the expressions simplify to

$$\begin{aligned}
\langle J_{vv}(0,v)\varphi, w\rangle_{X^*,X} &= \nu \sum_{i=1}^{d} (\nabla\varphi_i, \nabla w_i)_{L^2(\Omega)}\,,\\
\langle J_{vU}(0,v)\varphi, V\rangle_{\Theta^*,\Theta} &= \langle J_{Uv}(0,v)V, \varphi\rangle_{X^*,X}\\
&= \nu \sum_{i=1}^{d} \Big((\mathcal{I}\operatorname{div}(V) - DV - DV^T)\nabla v_i, \nabla\varphi_i\Big)_{L^2(\Omega)}\,,\\
\langle\, J_{UU}(0,v)V, W\rangle_{\Theta^*,\Theta} &= \tfrac{\nu}{2} \sum_{i=1}^{d} \Big(M^V(0)\left(\mathcal{I}\operatorname{div}(W) - 2DW^T\right)\nabla v_i, \nabla v_i\Big)_{L^2(\Omega)}\\
&\quad + \tfrac{\nu}{2} \sum_{i=1}^{d} \left((2DVDW - \operatorname{tr}(DWDV))\nabla v_i, \nabla v_i\right)_{L^2(\Omega)}.
\end{aligned}$$

6.1.2. Shape differentiability of the drag

Let us briefly discuss how shape derivatives of $\mathfrak{j}$ can be obtained. We begin by observing that, due to Assumption 6.1, the operator $E_{(v,p)}(0,v,p) = \tilde{E}^{\Omega}$ is continuously invertible. In particular, all the conditions of Assumption 2.13 are satisfied. Thus, Corollary 2.108 yields continuous Fréchet differentiability of the transformed design-to-state operator

$$S\colon B^{\Theta}(0,1) \to \mathcal{Y}, \quad S(U) = \tilde{S}(\tau_U(\Omega)) \circ \tau_U,$$

in $B^{\Theta}(0,\varepsilon)$ for some $\varepsilon > 0$. Hence the localized functional

$$\mathfrak{j}_{\Omega}(U) = J(U, S(U))$$

is also continuously Fréchet differentiable on $B^{\Theta}(0,\varepsilon)$. In particular, due to Theorem 2.31, the shape functional $\mathfrak{j}$ is shape differentiable at $\Omega \in \mathcal{O}$. Usually, e.g., in the case of full Dirichlet boundary conditions, one can easily extend the above argument to obtain continuous Fréchet differentiability of S and $\mathfrak{j}_{\Omega}$ on $B^{\Theta}(0,1)$. The first and second derivatives of $\mathfrak{j}_{\Omega}$ can be conveniently computed via the adjoint approach, see Section A.1. The necessary partial derivatives of E and J are stated in the previous subsection. For example, if $(v,p) \in \mathcal{Y}$ solve the state equations (6.5), and $(w,q) \in \mathcal{Y}$ solve the adjoint equations (A.1), then the shape derivative of $\mathfrak{j}$ at Ω in a direction $V \in \Theta$ is given by

$$\begin{aligned}
\langle \mathfrak{j}'(\Omega), V\rangle_{\Theta^*,\Theta} = \langle \mathfrak{j}'_{\Omega}(0), V\rangle_{\Theta^*,\Theta} &= J_U(0,v,p)V + (w,q)^* E_U(v,p)V\\
&= \tfrac{\nu}{2} \sum_{i=1}^{d} \Big((\mathcal{I}\operatorname{div}(V) - DV - DV^T)\nabla v_i, \nabla v_i\Big)_{L^2(\Omega)}\\
&\quad + \nu \sum_{i=1}^{d} \Big(\ (\mathcal{I}\operatorname{div}(V) - DV - DV^T)\nabla v_i, \nabla w_i\Big)_{L^2(\Omega)}\\
&\quad + \Big(\operatorname{tr}(DV^T\nabla w), p\Big)_{L^2(\Omega)} - \big(\operatorname{div}(w)\operatorname{div}(V), p\big)_{L^2(\Omega)}\\
&\quad - \Big(\operatorname{tr}(DV^T\nabla v), q\Big)_{L^2(\Omega)} + \big(\operatorname{div}(v)\operatorname{div}(V), q\big)_{L^2(\Omega)}.
\end{aligned}$$

6.2. The set of admissible domains and optimization aspects

In this section we present our choices for the set of admissible domains $\mathcal{O}_{ad}$. Moreover, we discuss a possible strategy to handle these constraints in an algorithmic setting. Let us begin by describing the geometric layout of the considered problem in more detail. We consider a bounded open holdall domain $\mathcal{D} \subset \mathbb{R}^d$ which is Lipschitz. Here $d = 2$ or 3, and $\mathcal{D}$ describes a section of a channel. The fluid domain Ω_0 is given as $\Omega_0 = \mathcal{D} \backslash B_0$, where $B_0 \subset \mathcal{D}$ is a body immersed in the fluid. We suppose that B_0 and hence Ω_0 are Lipschitz. The task is now to change the shape of the body B_0 such that the associated drag is minimized. Of course the holdall $\mathcal{D}$ should not be modified. A trivial solution would be to shrink B_0 to a single point, hence usually a constraint involving the volume of the immersed body is incorporated in the set of admissible domains. Furthermore, the drag can be reduced by placing the body in a part of the channel where the velocity of the flow is low, e.g., close to a no-slip boundary. To avoid such undesired behavior one can additionally fix the center of mass c^M of the body. Finally it might be, that additional geometric constraints require the body to lie inside some closed convex set $C \subset \mathcal{D}$. Thus, we consider the following family of admissible domains

$$\begin{aligned}\mathcal{O}_{ad} = \{\Omega = \tau(\Omega_0) \mid \tau \in \mathcal{F}(\Theta),\ \tau(\partial\mathcal{D}) = \partial\mathcal{D},\ \mathrm{Vol}(\tau(B_0)) = \mathrm{Vol}(B_0),\\ c^M(\tau(B_0)) = c^M(B_0),\ \tau(B_0) \subset C\}.\end{aligned}$$

Note that most of the constraints concern only $\tau(B_0)$. Furthermore, usually the initial shape B_0 will already be a good guess. Hence, it is often justified to assume that the solution of the shape optimization problem will be close to B_0. Thus, it is convenient to work again with shapes characterized via the transformation of the design boundary $\Gamma_B = \partial B_0$, see Section 2.11, 2.12 and 3.4. Recall that in this setting a displacement $u \in \mathcal{U}$ of the design boundary Γ_B is extended to a displacement $U = Tu$ of Ω_0 and hence to a transformed domain $\tau_U(\Omega_0)$. In particular, one can easily satisfy the requirement $\tau(\partial\mathcal{D}) = \partial\mathcal{D}$ by imposing zero Dirichlet boundary conditions on $\partial\mathcal{D}$ for the extension operator T. For our two dimensional numerical example we will use an extension via linear elasticity as described in Section 2.11.2. In this setting we can choose again $\mathcal{U} = H^2(\Gamma_B, \mathbb{R}^2)$ as space for the boundary displacements for a smooth Γ_B. As discussed in more detail in Section 2.12, we restrict ourselves to a set $\mathcal{U}_{feas} \subset \mathcal{U}$ such that $\|Tu\|_\Theta < 1$ for all $u \in \mathcal{U}_{feas}$, and assume that there exists a solution Ω^* of (6.2) which satisfies

$$\Omega^* = \tau_{T(u^*)}(\Omega_0) \quad \text{for some } u^* \in \operatorname{int}\mathcal{U}_{feas}.$$

The volume of the body $\tau(B_0)$ is given by

$$\mathrm{Vol}(\tau(B_0)) = \int_{\tau(B_0)} 1 \,\mathrm{d}\tilde{x} = \int_{B_0} \det(D\tau(x)) \,\mathrm{d}x,$$

and the center of mass $c^M(\tau(B_0))$ by

$$c_i^M(\tau(B_0)) = \frac{1}{\mathrm{Vol}(\tau(B_0))} \int_{\tau(B_0)} \tilde{x}_i \,\mathrm{d}\tilde{x} = \frac{1}{\mathrm{Vol}(\tau(B_0))} \int_{B_0} x_i \det(D\tau(x)) \,\mathrm{d}x,\ i = 1, 2.$$

In particular, derivatives with respect to domain displacements $U \in \Theta$ can easily be obtained and related to a boundary displacement via the adjoint extension operator T^*. An alternative is to rewrite the volume integrals as boundary integrals over $\tau(\Gamma_B)$. After discretization, Γ_B and the transformed boundaries $\tau(\Gamma_B)$ are polygons, and there are explicit formulas describing the volume and the center of mass in terms of the coordinates of the boundary nodes. In particular, one can calculate exact discrete derivatives of these functions with respect to displacements of the boundary nodes, cf. [Lin12, Section 5.2]. Summarizing, the volume and center of mass constraints are nonlinear smooth equality constraints depending on the boundary displacement $u \in \mathcal{U}$. As such they could be treated as explicit constraints, for instance, in a Lagrange-Newton/SQP method. Instead, we employ the *Augmented Lagrangian method*, cf., e.g., [CGT00, Chapter 14] or [NW06, Chapter 17] for a thorough treatment of this strategy. The idea is to incorporate the constraints in the objective. More precisely, the *Lagrangian*

$$\mathfrak{j}_{\Omega_0}(Tu) + \lambda_1\big(\operatorname{Vol}(\tau_{Tu}(B_0)) - \operatorname{Vol}(B_0)\big) + \lambda_2^T\big(c^M(\tau_{Tu}(B_0)) - c^M(B_0)\big),$$

where $\lambda_1 \in \mathbb{R}, \lambda_2 \in \mathbb{R}^2$ are the *Lagrange multipliers*, is augmented with quadratic penalty terms. By iteratively updating the Lagrange multipliers and the penalty parameters the constraint violation can be driven to zero and an admissible solution may be found. In contrast to pure penalty methods, the Augmented Lagrangian method will not necessarily drive the penalty parameters to ∞.

Finally, the condition $\tau(B_0) \subset C$ is equivalent to $\tau(\Gamma_B) \subset C$, which is a constraint exactly of the form treated in Chapter 5. In particular, we may repeat the arguments of Section 5.9, and realize that the shape optimization problem in terms of the boundary displacement fits into the setting considered in Chapter 5. Thus, we propose to study a series of regularized objective functionals of the form

$$\begin{aligned} f_\gamma(u,\lambda) = {} & \mathfrak{j}_{\Omega_0}(Tu) + \lambda_1\big(\operatorname{Vol}(\tau_{Tu}(B_0)) - \operatorname{Vol}(B_0)\big) + \lambda_2^T\big(c^M(\tau_{Tu}(B_0)) - c^M(B_0)\big) \\ & + \frac{\gamma_1}{2}\big(\operatorname{Vol}(\tau_{Tu}(B_0)) - \operatorname{Vol}(B_0)\big)^2 + \frac{\gamma_2}{2}\big|c^M(\tau_{Tu}(B_0)) - c^M(B_0)\big|^2 \\ & + \frac{\gamma_3}{2}\,\big\|\mathrm{id} + u - P_C\,(\mathrm{id} + u)\big\|^2_{L^2(\Gamma_B,\mathbb{R}^d)}\,. \end{aligned}$$

The functional $f_\gamma(\cdot,\lambda)\colon \mathcal{U} \to \mathbb{R}$ is a continuously Fréchet differentiable functional on $\mathcal{U}_{feas}$. If $\gamma_3 = 0$ then it is also twice differentiable, otherwise $f'_\gamma(\cdot,\lambda)$ is only semismooth, cf. Section 5.4. We do not elaborate further on the concrete optimization algorithm, instead we refer to [CGT00, Section 14.4] for a detailed discussion of the Augmented Lagrangian method. In the presence of geometric constraints we additionally increase the penalty parameter γ_3 during the progression of the Augmented Lagrangian method. We choose a trust-region globalized (possibly semismooth) Newton method to solve the subproblems of the Augmented Lagrangian method. We refer to the comprehensive monograph [CGT00] for a thorough treatment of theoretical and practical aspects of trust region methods. For the step computation we employ the truncated conjugate gradients method, cf., e.g., [Ste83] or [CGT00, Section 7.5]. To encourage fast local convergence of Newton's method we incorporate an additional control cost/Tikhonov regularization term $\frac{\beta}{2}\,\|u\|^2_{\mathcal{U}}$ in the objective, see the discussion in Remark 2.95. This term favors smooth solutions close to the current reference domain.

6.3. Numerical examples

Our implementation is based on the `C++` software package `FlowOpt` [Lin12], developed by Florian Lindemann at the chair of mathematical optimization at the TU München. It provides an object oriented framework for solving shape optimization problems where the PDE constraint is given by some variant of the Navier-Stokes equations. In particular, it separates the design object representation, the PDE solvers, and the optimization algorithm in independent blocks. The PDE solver block uses the finite element library `Sundance` [LBvBW12], which is part of the `Trilinos` project [HWH03]. We refer to [Lin12] for a detailed description of `FlowOpt` and its capabilities.

We extended `FlowOpt` in several directions.

(i) We incorporated a new design object class which makes it possible to characterize domains via the displacement $u \in \mathcal{U}$ of the reference design boundary. In our concrete implementation, we choose a nonconforming discretization and approximate $\mathcal{U} = H^2(\Gamma_B, \mathbb{R}^2)$ by continuous piecewise linear finite elements. We allow for a free displacement of each boundary node, and choose again the following scalar product $\mathcal{A}$:

$$(v,u)_{\mathcal{A}} \approx (v,u)_{L^2(\Gamma_B)} + w(v'',u'')_{L^2(\Gamma_B)},$$

where $w > 0$ is a weighting parameter, and the bi-Laplacian scalar product $(v'',u'')_{L^2(\Gamma_B)}$ is approximated by $KM^{-1}K$. Here K denotes the stiffness matrix, and M the lumped mass matrix on Γ_B.

(ii) So far `FlowOpt` provided only first order derivatives. We implemented the necessary methods to obtain also second order derivatives of the drag in stationary Stokes flow. The discretization of the Stokes flow uses Taylor-Hood finite elements, i.e., continuous, piecewise quadratic elements for the velocity and continuous, piecewise linear elements for the pressure. The extension operator T is based on the linearized elasticity equation, which is discretized with continuous, piecewise linear finite elements.

(iii) Finally, we implemented the optimization algorithm which we briefly sketched in Section 6.2.

The setting of our numerical tests is based on a DFG benchmark for two dimensional Navier-Stokes flow in a channel [STD$^+$96]. The channel is a $2.2m$ by $0.41m$ rectangle. The initial body is a ball with radius $0.05m$ and center $c^M(B_0) = (0.2m, 0.2m)$, where the origin is in the lower left corner of the channel. In the following we omit the unit m. On the surface of the body, as well as on the top and bottom of the rectangle, no-slip boundary conditions are imposed. The left hand boundary is an inflow boundary with prescribed Dirichlet data

$$\tilde{v}^D(0,x_2) = \begin{pmatrix} 6x_2(0.41 - x_2)/(0.41)^2 \\ 0 \end{pmatrix}.$$

On the right hand side of the rectangle we impose the outflow condition (6.4). The computations were carried out on a Linux cluster that was partially funded by the grant DFG INST 95/919-1 FUGG.

Let us specify some of our concrete choices for the parameters of the algorithm. If the Augmented Lagrangian method decides to increase the penalty parameter, we increase γ_1 by a factor of 10. We always set $\gamma_2 = \gamma_1/10$. In the presence of geometric constraints we choose always $\gamma_3 = \gamma_1$. As in Section 5.10, we solve the arising subproblems only inexactly, and decrease the termination tolerance as γ increases. To speed up the convergence, we start with a large regularization parameter β which is also iteratively decreased. Furthermore, if the number of iterations which are necessary to solve a subproblem increases too much, we change the reference domain to the current domain, and reset β. Note that the contribution of the regularization term to the overall objective is very small in the end, for Example 6.1 it is seven orders of magnitude smaller than the drag of the final object.

The reference ball B_0 with radius 0.05 and center of mass $(0.2, 0.2)$ defines the constraints, the corresponding drag is 1.10189958918. The discretization of Γ_B features 200 boundary nodes, the mesh of the whole domain consists of 16006 nodes. The flow around the ball, as well as the underlying mesh are depicted in Figure 6.1. Note that the mesh is particularly fine in the vicinity of the submerged object. In our experiments it proved often to be beneficial for the performance of the algorithm to choose an infeasible initial configuration, namely a slightly larger ball with radius 0.06.

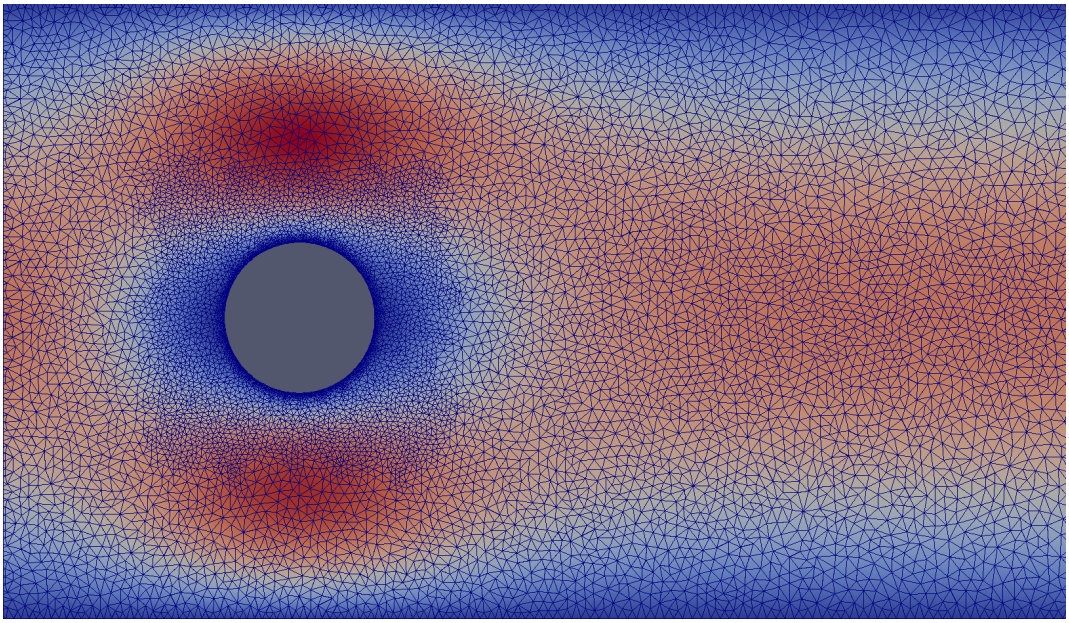


Figure 6.1.: The reference configuration, a ball with radius 0.05

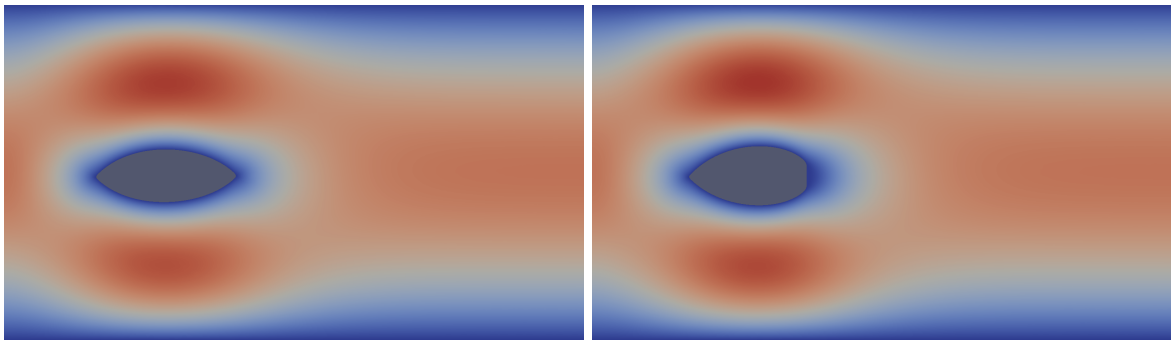
We abbreviate the various constraint violation indicators by

$$\begin{aligned}
\delta_V^k &:= \mathrm{Vol}(\tau_{Tu_k}(B_0)) - \mathrm{Vol}(B_0) \\
\delta_{c_1^M}^k &:= c_1^M(\tau_{Tu_k}(B_0)) - c_1^M(B_0) \\
\delta_{c_2^M}^k &:= c_2^M(\tau_{Tu_k}(B_0)) - c_2^M(B_0) \\
\delta_{MY}^k &:= \|\tau_{u_k} - P_C(\tau_{u_k})\|_{L^\infty(\Gamma_B, \mathbb{R}^d)}.
\end{aligned}$$

Example 6.1. We consider first a situation without geometric constraints. The progression of the Augmented Lagrangian method for Example 6.1 is presented in Table 6.1. The first column counts the iterations of the Augmented Lagrangian method, the second shows the corresponding value of the penalty parameter γ_1, and the third the current regularization parameter. The norm of the derivative of the overall objective $f_\gamma(u)$ (for brevity we suppress the dependency on λ) is presented in the fourth column. The columns five to seven show the values of the different constraint violation indicators $\delta_V^k, \delta_{c_1^M}^k, \delta_{c_2^M}^k$, and the last column gives the number of iterations required to solve the subproblem to the specified tolerance. Note that in most iterations only one or two Hessian evaluations are performed by the truncated conjugate gradients method, i.e., the computational effort per iteration is moderate. After the third iteration we updated our reference domain. The optimized body exhibits a drag of 1.03726859748. As already mentioned, the value of the regularization term for the final iterate is seven orders of magnitude smaller. It is well known that the optimal shape of a body submerged in a Stokes flow is a prolate pointed spheroid, cf. [Pir73]. Indeed, our final object has this shape, it is depicted on the left hand side of Figure 6.2.

Table 6.1.: History of the Augmented Lagrangian method, Example 6.1

k	γ_1	β_k	$\left\|f'_\gamma(u_k)\right\|_{\mathcal{A}^{-1}}$	δ_V^k	$\delta_{c_1^M}^k$	$\delta_{c_2^M}^k$	# iter
1	10^4	10^{-1}	$1.77{\cdot}10^{-4}$	$-2.83{\cdot}10^{-3}$	$1.87{\cdot}10^{-4}$	$-1.59{\cdot}10^{-4}$	4
2	10^4	10^{-2}	$8.80{\cdot}10^{-4}$	$2.90{\cdot}10^{-5}$	$1.66{\cdot}10^{-4}$	$-3.53{\cdot}10^{-5}$	87
3	10^5	10^{-3}	$9.64{\cdot}10^{-4}$	$4.14{\cdot}10^{-6}$	$1.67{\cdot}10^{-5}$	$-3.37{\cdot}10^{-6}$	315
4	10^6	10^{-1}	$7.51{\cdot}10^{-6}$	$-3.68{\cdot}10^{-6}$	$-1.50{\cdot}10^{-5}$	$3.03{\cdot}10^{-6}$	13
5	10^6	10^{-2}	$9.68{\cdot}10^{-6}$	$2.83{\cdot}10^{-10}$	$6.62{\cdot}10^{-10}$	$-1.72{\cdot}10^{-9}$	124



(a) Example 6.1 **(b)** Example 6.2

Figure 6.2.: Flow around the final objects

Example 6.2. We consider now the setting of Example 6.1 in combination with geometric constraints. To be more precise, we impose an upper bound for the deformation of the body in outflow direction, i.e., the back end of the body has to lie in front of the threshold 0.265. We incorporate the geometric constraint via the Moreau-Yosida regularization technique into our overall objective. The progression of the combined Augmented Lagrangian Moreau-Yosida penalty method for Example 6.2 is presented in Table 6.2. The maximum point-wise violation of the geometric constraint is shown in the eighth column. The decrease of the quantity δ_{MY}^k is

not as fast as it was in Section 5.10. This might be explained by the fact that the optimization algorithm has to cope now with several competing constraints. The optimized body exhibits a drag of 1.04512679653. The final domain is depicted on the right hand side of Figure 6.2. To accommodate the geometric constraint, as well as the volume and center of mass constraints, the final object is wider and shorter than the result of Example 6.1. The differences can be nicely observed in Figure 6.3, where the objects are overlaid.

Table 6.2.: History of the Augmented Lagrangian method, Example 6.2

k	γ_1	β_k	$\left\|f'_\gamma(u_k)\right\|_{\mathcal{A}^{-1}}$	δ_V^k	$\delta_{c_1^M}^k$	$\delta_{c_2^M}^k$	δ_{MY}^k	# iter
1	10^4	10^0	$2.85{\cdot}10^{-5}$	$-1.74{\cdot}10^{-3}$	$2.13{\cdot}10^{-4}$	$-1.82{\cdot}10^{-4}$	0.00	7
2	10^4	10^{-1}	$8.13{\cdot}10^{-4}$	$-1.23{\cdot}10^{-3}$	$-2.32{\cdot}10^{-4}$	$5.53{\cdot}10^{-6}$	$1.61{\cdot}10^{-3}$	5
3	10^5	10^{-2}	$9.05{\cdot}10^{-5}$	$-1.55{\cdot}10^{-4}$	$-8.41{\cdot}10^{-5}$	$-1.40{\cdot}10^{-6}$	$5.33{\cdot}10^{-4}$	25
4	10^5	10^{-3}	$8.90{\cdot}10^{-4}$	$-6.21{\cdot}10^{-7}$	$-3.06{\cdot}10^{-6}$	$-2.86{\cdot}10^{-7}$	$3.46{\cdot}10^{-4}$	172
5	10^6	10^0	$9.15{\cdot}10^{-6}$	$2.12{\cdot}10^{-7}$	$1.65{\cdot}10^{-6}$	$2.54{\cdot}10^{-7}$	$7.75{\cdot}10^{-5}$	4
6	10^6	10^{-1}	$9.54{\cdot}10^{-6}$	$1.80{\cdot}10^{-7}$	$6.95{\cdot}10^{-7}$	$6.21{\cdot}10^{-10}$	$7.67{\cdot}10^{-5}$	8
7	10^7	10^{-2}	$7.78{\cdot}10^{-6}$	$2.08{\cdot}10^{-8}$	$7.63{\cdot}10^{-8}$	$7.37{\cdot}10^{-11}$	$1.46{\cdot}10^{-5}$	71
8	10^8	10^0	$7.40{\cdot}10^{-7}$	$-1.88{\cdot}10^{-8}$	$-6.91{\cdot}10^{-8}$	$-6.92{\cdot}10^{-11}$	$3.48{\cdot}10^{-6}$	5
9	10^8	10^{-1}	$4.25{\cdot}10^{-6}$	$1.02{\cdot}10^{-10}$	$3.75{\cdot}10^{-10}$	$2.06{\cdot}10^{-12}$	$3.43{\cdot}10^{-6}$	96

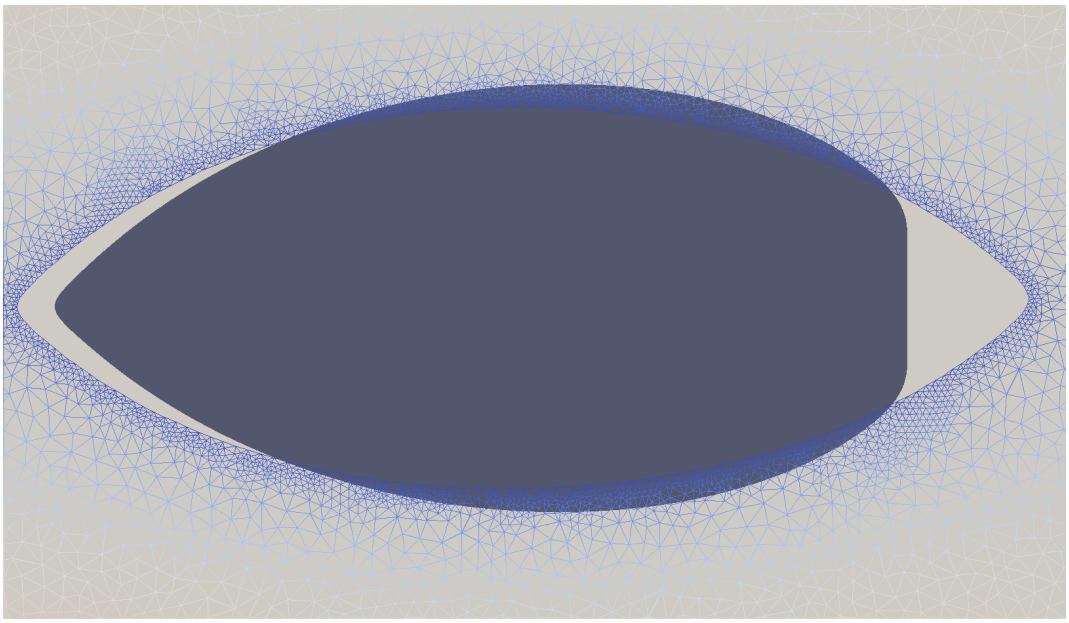


Figure 6.3.: Comparison of the final objects of Example 6.1 (blue grid) and Example 6.2 (grey surface)

As one can see, the tips of the final objects are still slightly rounded. This effect is very much influenced by our choice of the scalar product $\mathcal{A}$. The gradient with respect to such a smooth $\mathcal{A}$ does not develop a sharp kink, which would be necessary to obtain a pointed tip from our smooth initial domain. The presence of the regularization term If we choose

for one intermediate iteration of the Augmented Lagrangian method the H^1-scalar product a tip is formed and preserved by subsequent runs with a smoother scalar product. The final domains obtained with such a strategy is very similar to the ones of Examples 6.1 and 6.2, see Figure 6.4 for a comparison. The relative difference of the drag of the two objects is in the order of 10^{-5}. Let us note that a body with a sharp tip is more sensitive with respect to variations of the parameters of the system, in particular, with regard to changes in the attack angle, i.e., the inclination of the object with respect to the flow direction. In that sense the rounded tip is more robust with respect to uncertainties.

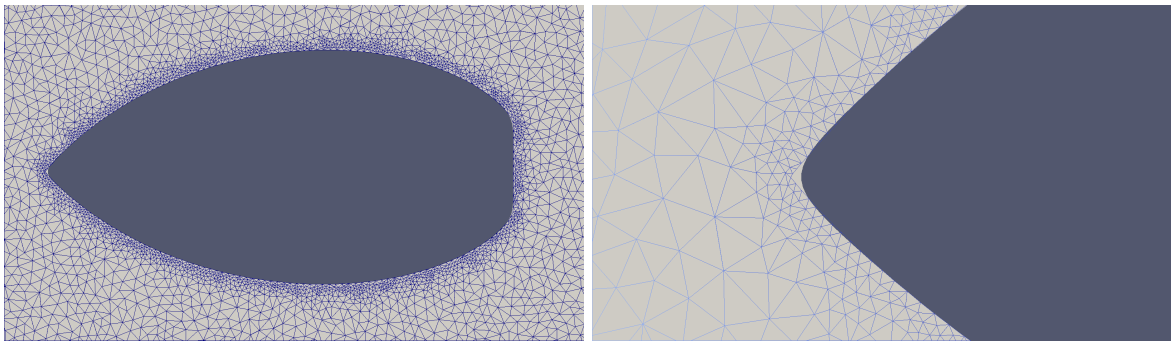


Figure 6.4.: Comparison with pointed object

Let us conclude this chapter by remarking that the proposed algorithm still needs fine tuning to be efficient. This is evident in the drastic increase of the required iterations per subproblem as we increase γ and decrease β. A careful balancing of the various parameters might significantly improve the overall performance. In particular it might be beneficial to revisit the heuristic updating scheme of [HK06a, HK06b] based on the value function, cf. Section 5.6. However, the question arises how one can incorporate a variable regularization term into this strategy. Furthermore, as already mentioned, the behavior of this updating scheme is again highly dependent on the choice of various parameters. Alternatively, one might consider more advanced strategies of handling the volume and center of mass constraints, e.g., SQP or interior point methods.

7. Shape optimization of a breakwater

This chapter is devoted to a shape optimization model problem which is motivated by a coastal engineering application. The objective is to reduce the resonance of a harbor due to long range ocean waves. Let us briefly describe this phenomenon. Harbors are designed to protect ships from incoming waves. More specifically one tries to place *breakwaters* in such a way that they absorb incoming waves, and shelter the *harbor basin*. Breakwaters are effective in absorbing waves with short wave periods (16 seconds or less according to [Xin09]), which covers the vast majority of ocean waves. However, this is no longer true for longer wave periods, whose periods may range from tens of seconds to several hours. These can be generated, for instance, by earthquakes or landslides, another example are tidal waves. 'In fact, it is possible that for certain semi-enclosed harbors the combined effect of wave diffraction [around the breakwaters], wave refraction [due to changing water depth] and multiple reflections from the boundaries can cause significant increase in the wave amplitude compared with the incident wave amplitude. This is commonly referred to as *harbor resonance*.' [Xin09, page 2]. We refer to [Rab09] and the references cited therein for a more detailed discussion of this topic. In practice, many publications employ the *mild slope equation* to model wave effects in coastal areas and harbor basins, cf., e.g., [FdSF04, LLL01, MH97, Xin09, XLR11]. It was first derived by Berkhoff in [Ber72], and can be written as

$$\nabla \cdot CC_g \nabla \varphi + k^2 CC_g \varphi = 0, \tag{7.1}$$

where φ is the horizontal variation in velocity potential, k is the wave number, ω is the wave frequency, $C = \omega/k$ is the wave celerity, $C_g = \frac{C}{2}\left(1 + \frac{2kh}{\sinh 2kh}\right)$ is the group velocity, and h is the water depth. Usually the mild slope equation is enriched with additional effects such as partial absorption boundaries, bottom friction, entrance loss, etc. In the references above several comparison studies with actual measured data were conducted, thus justifying the frequency domain approach.

Although we are motivated by the described application, a realistic treatment of harbor resonance is out of scope of this thesis, and we will significantly simplify the model in this chapter. We assume that the water depth is constant throughout the region of interest. Thus, the mild slope equation reduces to the well-known Helmholtz equation

$$\Delta \varphi + k^2 \varphi = 0. \tag{7.2}$$

There are various publications dealing with shape or topology optimization problems in combination with the Helmholtz equation, cf., e.g., [CK13, DJS07, SAM03] and the references therein. However, they are usually motivated by applications in acoustics. A lot of work is dedicated to inverse acoustic scattering problems, where the shape of some scatterer is to be determined from measurements of the far field. Other applications are, for example, finding the

optimal distribution of reflecting and nonreflecting materials in the walls of a room, or optimal design of sound barriers and wave guides. Although the resulting optimization problems are similar to the one considered here, none of those we found quite fits our setting. To the best of our knowledge a shape optimization problem involving the resonance of a harbor was, so far, only briefly considered in the thesis [BL98]. The author used the real-valued Helmholtz equation as state equation, an explicit discrete geometry description via the finite element mesh, and calculated the discrete shape derivative by differentiating with respect to nodal coordinates.

Some of the results presented in this chapter were already published in [KK15]. In parts, our presentation here follows the paper closely.

Assuming that the water depth is constant, we consider the complex valued Helmholtz equation as our model state equation. The objective is to minimize the average wave height in the harbor basin. We suppose that we are allowed to modify the shape of the breakwater which surrounds the harbor basin. The model problem naturally involves geometric constraints in the form of forbidden and contained regions. The harbor basin and the harbor approach should be part of the ocean, on the other hand the mainland should not be flooded. We strictly enforce these constraints by employing the projected descent method proposed in Section 2.13. The geometry will be described by the level set method, cf. Section 2.11.3, i.e., the domain Ω is given as the sub-zero level set of a function $\Phi\colon \mathcal{D} \to \mathbb{R}$, where $\mathcal{D} \subset \mathbb{R}^2$ is the holdall domain.

This chapter is organized as follows. We derive our model state equation on a bounded domain in Section 7.1. The shape optimization problem under consideration is described in Section 7.2. We sketch how the derivative of the reduced shape functional can be computed via the adjoint approach. Note that, in contrast to many publications, we use the level set method in combination with the volume expression of the shape derivative. It requires less regular finite element functions, and, in our experience, the volume expression is numerically more stable than the Hadamard form. This assessment is shared in the recent papers [HPS15, LS13]. In [LS13] the volume expression of the shape derivative and the level set method are also used. We present our optimization algorithm in Section 7.3, and discuss discretization and implementation aspects. Finally, we present the results of some numerical experiments in Section 7.4.

Let us fix some conventions and notation for this chapter. We define the usual bilinear L^2-scalar product for real-valued functions on some set $\Omega \subset \mathbb{R}^d$ as $(\cdot,\cdot)_{L^2(\Omega)}$, and the corresponding sesquilinear form as $(f,g)_{L^2_{\mathbb{C}}(\Omega)} := (f,\overline{g})_{L^2(\Omega)}$ for some complex valued functions f, g. Furthermore, we introduce the real-valued scalar product

$$(f,g)_{L^2_{\mathbb{R}}(\Omega)} := \operatorname{Re}(f,g)_{L^2_{\mathbb{C}}(\Omega)}.$$

The norm $\|\cdot\|_{L^2_{\mathbb{R}}(\Omega)}$ induced by this scalar product coincides with the norm induced by the sesquilinear form. Hence the elements of the space

$$L^2_{\mathbb{R}}(\Omega) := \{f\colon B \mapsto \mathbb{C} \mid \|f\|_{L^2_{\mathbb{R}}(\Omega)} < \infty\}$$

coincide with the elements of $\{f\colon \Omega \mapsto \mathbb{C} \mid \|f\|_{L^2_{\mathbb{C}}(\Omega)} < \infty\}$, but since we use the $(\cdot,\cdot)_{L^2_{\mathbb{R}}(\Omega)}$ scalar product we have a different Hilbert space structure. Other Hilbert spaces will be treated analogously (e.g., $H^1_{\mathbb{R}}(\Omega)$).

7.1. Description of the physical model

We consider the situation sketched in Figure 7.1. There is an isle bounded by the contour Γ_L, some breakwaters given by Γ_B and a surrounding ocean denoted by Ω^+. We are interested in the scattered wave u, induced by an incoming planar monochromatic wave $z(x) = \exp(ikw^T x)$ with incident direction $w \in \mathbb{R}^2$ and wave number $k > 0$. The total surface perturbation is then given by $y = u + z$, and satisfies

$$\begin{aligned} \Delta y + k^2 y &= 0 \quad \text{in } \Omega^+ \\ ay + \partial_n y &= 0 \quad \text{on } \Gamma_I := \Gamma_B \cup \Gamma_L. \end{aligned} \tag{7.3}$$

Here $a\colon \mathbb{R}^2 \to \mathbb{C}$ describes the absorption coefficient at the boundary. The boundaries Γ_L and Γ_B are assumed to be at least Lipschitz. Let us discuss appropriate boundary conditions for the far field. It is a standard assumption that the scattered wave satisfies the Sommerfeld radiation condition

$$iku - \partial_r u = o(r^{-\frac{1}{2}}), \quad \text{for } r \to \infty.$$

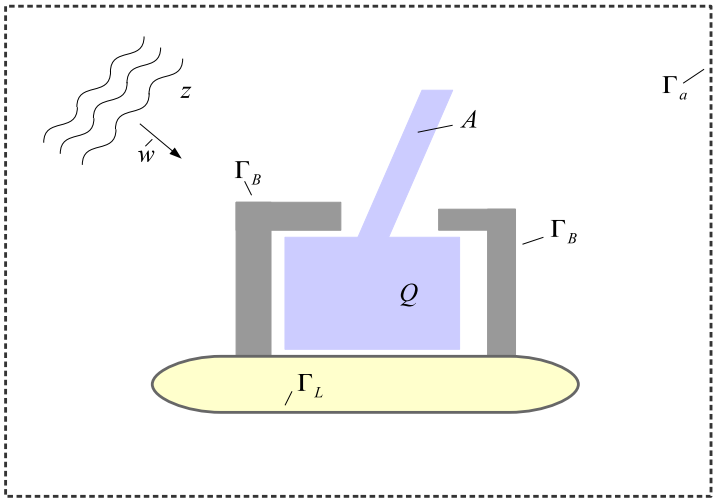


Figure 7.1.: The domain Ω

If one wants to study this problem in weak form on the unbounded domain Ω^+ one needs to introduce different weighted Sobolev spaces for the test and ansatz functions and include the Sommerfeld radiation condition in the ansatz space. See [Lei86, chapter 4] and [Ihl98, section 2.3] for more details. However, this approach leads to various difficulties in the numerical realization if a finite element discretization is employed. While it would be possible to handle the Helmholtz equation on an exterior domain with the *boundary element method*, cf., e.g., [Har08] and the references therein, an extension to the more realistic mild slope equation would be quite challenging. For this reason we focus on the finite element approach which can easily cope with such an extension of the model.

Alternatively one may decompose the domain Ω^+ disjointly into a bounded domain Ω and an unbounded domain Ω_a by introducing an artificial smooth boundary Γ_a such that

$$\Omega^+ = \Omega \cup \Gamma_a \cup \Omega_a.$$

The problem (7.3) is then equivalent to the following coupled problem (cf. [JN80])

$$\begin{aligned} \Delta y_- + k^2 y_- &= 0 && \text{in } \Omega \\ a y_- + \partial_n y_- &= 0 && \text{on } \Gamma_I \\ y_- &= y_+ && \text{on } \Gamma_a \\ \partial_n y_- &= \partial_n y_+ && \text{on } \Gamma_a \\ \Delta y_+ + k^2 y_+ &= 0 && \text{in } \Omega_a \\ iku_+ - \partial_r u_+ &= o(r^{-\frac{1}{2}}), && \text{for } r \to \infty. \end{aligned} \tag{7.4}$$

For a given u_- on Γ_a, recall $u = y - z$, one can solve the unbounded Dirichlet problem

$$\begin{aligned} \Delta u_+ + k^2 u_+ &= 0 && \text{in } \Omega_a \\ u_+ &= u_- && \text{on } \Gamma_a \\ iku_+ - \partial_r u_+ &= o(r^{-\frac{1}{2}}), && \text{for } r \to \infty, \end{aligned}$$

compare [Lei86]. Once the solution u_+ is available one can easily compute $\partial_n u_+ = \partial_n u_-$ on Γ_a. We denote the mapping $u_- \mapsto \partial_n u_-$ by G_e and observe that $G_e \in \mathcal{L}(H^{\frac{1}{2}}_{\mathbb{R}}(\Gamma_a), H^{-\frac{1}{2}}_{\mathbb{R}}(\Gamma_a))$. This operator is usually referred to as the Dirichlet-to-Neumann (DtN) operator. There exists an integral and a series representation of the non-local operator G_e, the integral variant is also called the Steklov-Poincaré operator. The Neumann condition $\partial_n y_- = \partial_n y_+$ on Γ_a can now be reformulated as

$$\partial_n y_- = \partial_n u_- + \partial_n z = G_e u_- + \partial_n z = G_e y_- - G_e z + \partial_n z.$$

Evaluating the exact solution operator G_e is expensive. For this reason it is replaced by some yet unspecified $G \in \mathcal{L}(H^{\frac{1}{2}}_{\mathbb{R}}(\Gamma_a), H^{-\frac{1}{2}}_{\mathbb{R}}(\Gamma_a))$, which might be some approximation of G_e. Hence we arrive at the bounded problem

$$\begin{aligned} \Delta y + k^2 y &= 0 && \text{in } \Omega \\ a y + \partial_n y &= 0 && \text{on } \Gamma_I \\ \partial_n y &= G y - G z + \partial_n z && \text{on } \Gamma_a, \end{aligned} \tag{7.5}$$

which is equivalent to (7.3) for the choice $G = G_e$. We simplify the problem by choosing G as the 0th-order approximation of G_e, cf. [Ihl98, section 3.]. Furthermore, we suppose that the the breakwater and isle have perfectly reflecting boundaries. This is summarized in the following assumption.

Assumption 7.1. *The operator $\mathcal{G} \in \mathcal{L}(H^{\frac{1}{2}}_{\mathbb{R}}(\Gamma_a), H^{-\frac{1}{2}}_{\mathbb{R}}(\Gamma_a))$ is given by*

$$\langle Gy, \varphi \rangle_{H^{-\frac{1}{2}}_{\mathbb{R}}(\Gamma_a), H^{\frac{1}{2}}_{\mathbb{R}}(\Gamma_a)} = (iky, \varphi)_{L^2_{\mathbb{R}}(\Gamma_a)}.$$

Furthermore, the absorption coefficient is set to $a \equiv 0$.

Remark 7.1. (i) The choice of G is a rather crude simplification, but our focus in this work is on methodology, and not so much on realistic modeling. The low-order approximation of G_e introduces artificial reflections at Γ_a. Since this makes the shape optimization problem presumably harder to solve, we accept that for the moment. For more evolved methods of treating the artificial boundary Γ_a and the operator G_e we refer to [Ihl98, chapter 3].

(ii) Setting the absorption coefficient to zero implies that there is no damping effect by absorption of energy at the reflecting boundary. We note again that this presumably makes the optimization problem harder, since small design changes might have large non-local effects due to wave interference. Furthermore, as already mentioned at the beginning of this chapter, long range ocean waves are, in fact, mostly reflected by breakwaters. Harbor oscillations and resonance due to long waves has been widely studied in the coastal engineering literature, see for example [Rab09] and the references therein.

The weak formulation of (7.5), with G and a chosen to satisfy Assumption 7.1, is given by

$$\begin{cases} \text{Find } y \in H^1_{\mathbb{R}}(\Omega): \\ b(y,\varphi) = f(\varphi), \quad \forall \varphi \in H^1_{\mathbb{R}}(\Omega), \end{cases} \tag{7.6}$$

where we define

$$\begin{aligned} b(y,\varphi) &:= (\nabla y, \nabla \varphi)_{L^2_{\mathbb{R}}(\Omega)} - k^2 (y,\varphi)_{L^2_{\mathbb{R}}(\Omega)} - (iky,\varphi)_{L^2_{\mathbb{R}}(\Gamma_a)}, \\ f(\varphi) &:= (\partial_n z - ikz, \varphi)_{L^2_{\mathbb{R}}(\Gamma_a)}. \end{aligned} \tag{7.7}$$

Results concerning the existence of a unique solution of (7.6) and its regularity are well known:

Theorem 7.2. *[KK15, Theorem 2.3] Let Ω be a Lipschitz domain. Then there exists a unique solution $y \in H^1_{\mathbb{R}}(\Omega)$ of (7.6) for any right-hand side $f \in H^1_{\mathbb{R}}(\Omega)^*$, and it holds $\|y\|_{H^1_{\mathbb{R}}(\Omega)} \leq c \|f\|_{H^1_{\mathbb{R}}(\Omega)^*}$ for some $c > 0$.*

Proof. We have the Gelfand-triple $H^1_{\mathbb{R}}(\Omega) \hookrightarrow L^2_{\mathbb{R}}(\Omega) \hookrightarrow H^1_{\mathbb{R}}(\Omega)^*$. Further, there exists a $C > 0$ such that $b(\cdot,\cdot) + C\,(\cdot,\cdot)_{L^2_{\mathbb{R}}(\Omega)}$ is $H^1_{\mathbb{R}}(\Omega)$-coercive. Hence the Fredholm alternative holds: Either there exists a unique solution of (7.6) for any $f \in H^1_{\mathbb{R}}(\Omega)$, or there exists a nontrivial solution $y_0 \neq 0$ of the homogenous problem

$$b(y,\varphi) = 0, \quad \forall \varphi \in H^1_{\mathbb{R}}(\Omega).$$

For our choice of the operator G the solution of the homogenous problem is unique, see [Ihl98, Theorem 3.2]. □

Theorem 7.3. *[KK15, Theorem 2.4] Let Ω be a C^2-domain. Then the solution $y \in H^1_{\mathbb{R}}(\Omega)$ of (7.6) has the additional regularity $y \in H^2_{\mathbb{R}}(\Omega)$.*

Proof. This follows from standard regularity results for elliptic equations, see for example [Hac92, Theorem 9.1.20]. □

7.2. Shape optimization problem

After deriving the model state equation (7.6) we are now ready to formulate the shape optimization problem under consideration in detail. As announced in the introduction, our objective is to minimize the average wave height in the harbor basin Q (compare Figure 7.1). Given a domain Ω and an associated solution y of (7.6) we define the cost functional as

$$J(y) = \frac{1}{2} \|y\|^2_{L^2_{\mathbb{R}}(Q)}.$$

Due to Theorem 7.2 there exists a design-to-state operator $\tilde{S}$ for Lipschitz domains $\Omega \subset \mathbb{R}^2$, and hence we consider as usual the shape functional

$$\mathfrak{j}(\Omega) := J(\tilde{S}(\Omega)).$$

Formulating the set of admissible domains requires some care. Enclosing the whole harbor basin by a breakwater is obviously not a feasible solution, so we introduce a harbor approach A and demand that $Q \cup A$ is always part of the ocean. Furthermore, we do not want to remove parts of the island (the inhabitants might complain). Finally, we do not want to change our model by modifying the artificial outer boundary Γ_a. Hence, given an initial layout Ω_0, and a suitable Banach space Θ, we define the admissible family of domains by

$$\mathcal{O}_{ad} := \{\Omega = \tau(\Omega_0) \mid \tau \in \mathcal{F}(\Theta),\ \tau(\Gamma_a) = \Gamma_a,\ (Q \cup A) \subset \Omega,\ \Omega \cap L = \emptyset\}.$$

Here $\Omega \in \mathcal{O}_{ad}$ represents the ocean. We are looking for a solution of

$$\min_{\Omega \in \mathcal{O}} \mathfrak{j}(\Omega) \quad \text{s.t. } \Omega \in \mathcal{O}_{ad}. \tag{7.8}$$

This problem fits exactly into the setting of Section 2.13, and we will apply the projected descent method proposed there. Note that we do not concern ourselves with the question whether a solution of the above shape optimization problem exists, we simply suppose that this is the case. In the next paragraph we briefly state the shape derivative of $\mathfrak{j}$, which we obtain by the function space parametrization approach described in Section 2.14. In Chapter 3 we presented the necessary procedure for a closely related state equation in detail.

Shape derivative

Suppose that Ω_0 is a Lipschitz domain whose boundary decomposes into

$$\partial\Omega_0 = \Gamma_a \cup \Gamma_L \cup \Gamma_B,$$

where Γ_a is the artificial boundary introduced in Section 7.1, Γ_L is the boundary of the island L and Γ_B is the boundary of the initial breakwater. Furthermore, we suppose that $\Omega_0 \in \mathcal{O}_{ad}$. We consider the space $\Theta = C^1(\mathbb{R}^d, \mathbb{R}^d)$, which ensures that every $\Omega \in \mathcal{O}_{ad}$ is again a Lipschitz domain. As in Section 2.13 we introduce the sets

$$\mathcal{V}_{feas} = \{V \in \Theta \mid V\ = 0 \text{ on } A \cup Q \cup L \cup \Gamma_a\},\ \text{and}\ \ \mathcal{V}_{ad} = \mathcal{H} \cap \mathcal{V}_{feas},$$

for some suitable Hilbert space $\mathcal{H}$ satisfying Assumption 2.11. Due to Proposition 2.99 it suffices to characterize $\langle \mathsf{j}'(\Omega), V\rangle_{\Theta^*,\Theta}$ for all $V \in \mathcal{V}_{ad}$. Hence, let $\Omega \in \mathcal{O}_{ad}$, and recall

$$A(U) = D\tau_U^{-1} D\tau_U^{-T} \det(D\tau_U)$$

from (3.2). For all $U \in \mathcal{V}_{ad}$ the transformed state equation operator is given by

$$\begin{aligned}\langle \varphi, E(U,y)\rangle_{\mathcal{Z}^*,\mathcal{Z}} := &\left(A(U)\nabla y, \nabla\varphi\right)_{L^2_{\mathbb{R}}(\Omega)} - k^2 \left(\det(D\tau_U) y, \varphi\right)_{L^2_{\mathbb{R}}(\Omega)} \\ &- \left(iky, \varphi\right)_{L^2_{\mathbb{R}}(\Gamma_a)} - \left(\partial_n z - ikz, \varphi\right)_{L^2_{\mathbb{R}}(\Gamma_a)},\end{aligned}$$

and the transformed objective by

$$J(U,y) = \frac{1}{2}\, \|y\|^2_{L^2_{\mathbb{R}}(Q)}.$$

We utilize now the adjoint approach, cf. Section A.1. Given a solution $y \in H^1_{\mathbb{R}}(\Omega)$ of the state equation (7.6), and a solution $p \in H^1_{\mathbb{R}}(\Omega)$ of the adjoint equation

$$b(\psi, p) = -\left(y, \psi\right)_{L^2_{\mathbb{R}}(Q)} \quad \forall \psi \in H^1_{\mathbb{R}}(\Omega), \tag{7.9}$$

the shape derivative of j in a direction $V \in \mathcal{V}_{ad}$ is determined by

$$\langle \mathsf{j}'(\Omega), V\rangle_{\Theta^*,\Theta} = \Big((\mathcal{I}\operatorname{div}(V) - DV - DV^T)\nabla y, \nabla p\Big)_{L^2_{\mathbb{R}}(\Omega)} - k^2 \left(\operatorname{div}(V) y, \varphi\right)_{\mathcal{L}^2_{\mathbb{R}}(\Omega)}, \tag{7.10}$$

where $\mathcal{I}$ denotes the identity matrix in $\mathbb{R}^2$.

Optimization strategy

As mentioned above, we employ the projected descent method proposed in Section 2.13. In our numerical experiments it performs quite well. Additionally, we experimented with some ideas which are not covered by our theory. We briefly mention two of them here, and describe the details of our implementation in Section 7.3.

A simple modification is inspired by the method of nonlinear conjugate gradients [NW06, Section 5.2], which applies a simple correction term to the search direction. In our experience this led to an improved performance of the descent method. The second idea is motivated by the standard projected gradient method. As mentioned in Section 2.13, projecting a domain Ω onto the admissible set of domains with respect to the metric $d_{\mathcal{F}}$ is challenging. However, under suitable assumptions, the projection with respect to the metric induced by the measure of the symmetric set difference or the Hausdorff metric is possible [Kra15b]. In fact, the projection amounts to removing forbidden regions from Ω and adding regions which should be contained in Ω. Unfortunately, this approach does not fit together with our shape sensitivity analysis. Hence, one is not guaranteed to achieve descent. However, since it is a quite obvious strategy which is easy to implement, we test it for comparison.

We describe the domains $\Omega \in \mathcal{O}_{ad}$ via the level set framework, see Section 2.11.3. In this context it is convenient to work with transformations obtained as flow maps of the vector fields

$V \in \mathcal{V}_{ad}$. In particular, given a descent direction $V \in \mathcal{V}_{ad}$, we employ the level set transport equation (2.37)

$$\partial_t \Psi + \nabla \Psi^T V = 0 \text{ with initial condition } \Psi(0, x) = b_\Omega(x),$$

to obtain a level set representation $\Psi(t)$ of the transformed domain $T_V(t)(\Omega)$.

Remark 7.4. Note that, while in theory, the solution of the transport equation satisfies

$$\Psi(t, x) = b_\Omega \circ T_V(t)^{-1}(x),$$

this is *no longer true* for the discrete approximate solution of the transport equation (2.37) obtained with a time stepping scheme. In particular, approximating the transport of the level set equation may lead to *topology changes*. However, for the shape optimization problem under consideration, this is actually a desired feature of the employed optimization strategy. Indeed, it is a priori not at all clear what topology a good breakwater should have. So while our optimization method in theory tries to find a solution of (7.8), we deliberately allow for 'accidental' topology changes of the domain.

We summarize the employed optimization strategy on the discrete level in Algorithm 7.1.

7.3. Optimization and discretization aspects

Let us briefly sketch Algorithm 7.1 before describing the necessary steps in detail.

Starting with a level set function Φ given on a regular grid, we extract a discretization of the domain Ω and the interface $\Gamma_I = \Gamma_B \cup \Gamma_L$. We solve the state and adjoint equations on the discretized domain using piecewise linear finite elements. Now we can compute the shape derivative of the reduced objective, and obtain a projected gradient representation from (2.40). Finally, the level set function is evolved according to (2.37) along the negative projected gradient for some time span Δt which we determine with a backtracking strategy. Since $V|_{\Gamma_a} = 0$ for all $V \in \mathcal{V}_{ad}$ we restrict our considerations to the bounded holdall domain $\mathcal{D} \subset \mathbb{R}^2$ with $\partial\mathcal{D} = \Gamma_a$.

Let us now describe the details of our algorithm. We approximate the interface Γ_I, given by the zero level set of Φ, with one or multiple polygonal curves. Between each pair of neighbouring points on the regular grid at which the level set function has different signs, there is an intersection point of the zero level set with the edges of the grid. We approximate this intersection point using an affine model for Φ along the edge. Connecting all these intersection points, we obtain the polygonal approximation $\Gamma_{I,h}$ to Γ_I, and thus the current domain Ω_h.

In the next step, the domains Ω_h and $\mathcal{D}$ are discretized with triangular meshes which resolve the polygonal boundary $\Gamma_{I,h}$. Furthermore, the mesh representing Ω_h consists of a subset of the triangles of the mesh for D. Cells of the rectangular grid for which all four vertices have the same sign of Φ are split along their diagonal into two triangles. Cells which are intersected by $\Gamma_{I,h}$ are split depending on how they are intersected. To avoid triangles that are too degenerate and cause numerical difficulties, we enforce a certain minimum ratio between

Algorithm 7.1: Projected descent method using the level set approach

Require: an initial level set function Φ^0 on a fixed grid and a scalar product $(\cdot,\cdot)_a$
1: set the iteration index to $k = 0$
2: **repeat**
3: approximate the zero level set of Φ^k by polygonal curves
4: construct a mesh of $\mathcal{D}$ and Ω^k which resolves the polygonal boundary
5: solve the state equation (7.6) on Ω^k to obtain y^k
6: evaluate $j(\Omega^k) = \frac{1}{2}\left\|y^k\right\|^2_{L^2(Q)}$
7: solve the adjoint equation (7.9) on Ω^k to obtain p^k
8: compute the derivative $\mathfrak{j}'(\Omega^k)$ by (7.10)
9: compute the negative projected gradient $U_k = P_a(-V_k)$ by (2.40)
10: determine the new level set function Φ^{k+1} via (7.12) such that the Armijo condition (7.13) is satisfied
11: **until** converged

edges of all mesh triangles. This strategy proved sufficient for our experiments. For a more advanced technique of transporting and resolving the zero level set we refer to [ADF14]. The mesh on Ω_h is used to solve the state and adjoint equations, and the mesh on $\mathcal{D}$ is used to solve (2.40) for the projected gradient. Furthermore, our mesh construction ensures that each point of the original regular grid is also a vertex of the triangle mesh, so that we can extract $P_a(-V_k)$ on each grid point to solve (2.37).

Let us briefly comment on the discretization of the state and adjoint equation. We employ the usual machinery of continuous, piecewise linear finite elements to discretize the scalar products $(f,g)_{L^2(\Omega_h)}$, $(\nabla f, \nabla g)_{L^2(\Omega_h)}$, and $(f,g)_{L^2(\Gamma_a)}$ for some real valued functions f, g. Recall the notation from section Section 7.1. Splitting every complex valued function f into $f = f_1 + if_2$, with $f_i : \Omega_h \to \mathbb{R}$, we find

$$(f,g)_{L^2_{\mathbb{R}}(\Omega_h)} = \mathrm{Re}(f_1 + if_2, g_1 + ig_2)_{L^2_{\mathbb{C}}(\Omega_h)} = (f_1,g_1)_{L^2(\Omega_h)} + (f_2,g_2)_{L^2(\Omega_h)}.$$

Analogously it holds that

$$(\nabla f, \nabla g)_{L^2_{\mathbb{R}}(\Omega_h)} = (\nabla f_1, \nabla g_1)_{L^2(\Omega_h)} + (\nabla f_2, \nabla g_2)_{L^2(\Omega_h)},$$

and

$$(if,g)_{L^2_{\mathbb{R}}(\Gamma_a)} = (f_1,g_2)_{L^2(\Gamma_a)} - (f_2,g_1)_{L^2(\Gamma_a)}. \tag{7.11}$$

Note that $z(x) = \exp(ikw^T x)$ which implies $\partial_n z - ikz = ik(w^T n - 1)z$. Hence we only need to implement the boundary expression (7.11). Combining these formulas with the usual mass and stiffness matrices, we can easily assemble the system matrix and right-hand side of the state equation (7.6), and adjoint equation (7.9).

Note that, in order to guarantee the minimum regularity of Θ for the velocity field, we would need to compute the gradient in $H^s(\mathcal{D},\mathbb{R}^2)$ with $s > 2$. In our numerical experiments we neglect to do so. We only use piecewise linear continuous finite elements to discretize the

ansatz and test spaces of (2.40). In the examples presented below the H^1-scalar product is used to determine the gradient and the projected gradient. We also experimented with different choices of the scalar product, cf. [KK15].

Once we have computed the projected gradient, we need to update our geometry. For this we solve

$$\partial_t \Psi(t) + U_k^T \nabla \Psi(t) = 0 \quad \forall t \in (0, \Delta t), \quad \Psi(0) = \Phi, \tag{7.12}$$

with $U_k = P_a(-V_k)$, and use $\Psi(\Delta t)$ as the new level set function in the next iteration. The time span Δt is determined by the Armijo rule, i.e.

$$\mathfrak{j}(T_{U_k}(\Delta t, \Omega_k)) \leq \mathfrak{j}(\Omega_k) + \Delta t \gamma \langle \mathfrak{j}'(\Omega_k), U_k \rangle_{\Theta^*, \Theta}. \tag{7.13}$$

As already mentioned, the discrete approximate solution of (7.12) may lead to topology changes. Since these are not allowed for in the shape derivative, this often causes an increase of the objective functional. Thus, enforcing monotonicity in the objective values may result in a stalling of the algorithm. Since such changes in the topology often lead to much better designs in the long run, we introduce a lower bound for Δt. If no better design is found by the Armijo rule we accept an increase in the objective functional and evolve the level set function for the minimum time span. Let us mention that, with proper care, nonmonotone linesearch methods have proven to be quite effective, and often outperform monotone linesearch methods. We refer to [GLL86, GLL89, Toi96, ZH04] for an introduction to nonmonotone optimization methods.

As was suggested in [LS13], we employ the local Lax-Friedrichs flux (cf. [OS91]), and an explicit Euler time stepping scheme to evolve Ψ. In our setting this leads to the level set function updates

$$\Psi_{ij}^{l+1} = \Psi_{ij}^l - \delta t H^{LLF}. \tag{7.14}$$

Here Ψ_{ij}^l is the value of the level set function in the node (x_i, y_j) of the regular grid at the l-th time step, and the local Lax-Friedrichs flux is given by

$$H^{LLF} = \frac{p^- + p^+}{2} U_1 + \frac{q^- + q^+}{2} U_2 - \frac{1}{2}(p^+ - p^-)|U_1| - \frac{1}{2}(q^+ - q^-)|U_2|,$$

where

$$p^- = \frac{\Psi_{ij}^l - \Psi_{i-1,j}^l}{\Delta x}, \qquad p^+ = \frac{\Psi_{i+1,j}^l - \Psi_{ij}^l}{\Delta x},$$
$$q^- = \frac{\Psi_{ij}^l - \Psi_{i,j-1}^l}{\Delta y}, \qquad q^+ = \frac{\Psi_{i,j+1}^l - \Psi_{ij}^l}{\Delta y}.$$

The time step size δt is chosen to satisfy the Courant-Friedrichs-Lewy condition which guarantees the stability of the explicit time stepping scheme. It is common to reinitialize the level set function every few steps for numerical stability if many time steps are necessary . We use the signed distance function of the current domain (as defined by Ψ^l), which we obtain via a fast marching algorithm [Set96]. In our numerical experiments the reinitialization was only rarely necessary.

Finally, we briefly describe the two heuristic extensions of our basic algorithm which were mentioned in Section 7.2. Our adaptation of the Fletcher- Reeves nonlinear conjugate gradients method (cf., e.g., [NW06, Section 5.2]) inserts the following vector field into (7.12)

$$U_k = P_a(-V_k) + \frac{\|P_a(-V_k)\|_a}{\|P_a(-V_{k-1})\|_a} U_{k-1}.$$

However, this modification of our original search direction is not necessarily a descent direction. As is often done, we restart the procedure every few steps by setting $U_{k-1} = 0$ and taking a step along the negative projected gradient. This ensures that the search direction is not dominated by old gradients. The other strategy mentioned above was to project the domain instead of the descent direction. In this case we first obtain a new level set function Φ_{aux}^{k+1} by solving (7.12) for some time Δt and $U_k = -V_k$. We then modify Φ_{aux}^{k+1} to be positive in forbidden regions and negative in regions which should be contained in Ω_{k+1}. Finally, we obtain the oriented distance function $b_{\Omega^{k+1}}$ associated with the corresponding domain by invoking the fast marching algorithm, and set $\Phi^{k+1} = b_{\Omega^{k+1}}$.

7.4. Numerical examples

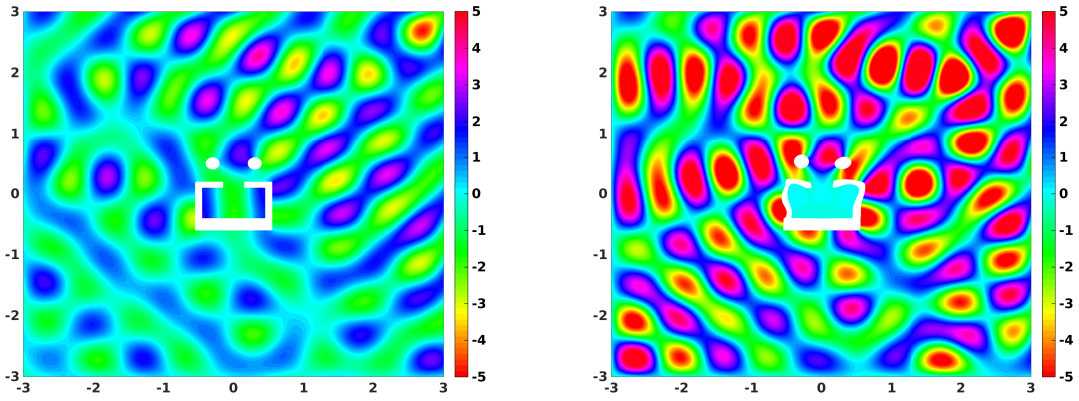
We implemented the proposed projected gradient method in `GNU Octave` [EBH09]. The routines for generating the geometry from the level set function, and assembling the finite element mesh are using the `Octave` package `level set` [Kra14] developed by Daniel Kraft. It also provides a method to compute the signed distance function using a fast marching algorithm. Note that optimization and discretization do not commute in our approach. Furthermore, we are only using a first-order method, hence convergence towards a critical point of the discrete problem is usually not to be expected in a reasonable number of steps. For this reason we simply terminate after 300 iterations.

The computations were carried out on a Linux cluster that was partially funded by the grant DFG INST 95/919-1 FUGG.

Example 7.1. In our first experiment the harbor basin Q corresponds to the rectangle $|x| \leq 0.4$, and $-0.3 \leq y \leq 0$. The harbor approach is given by $|x| \leq 0.1$, $y \geq 0$, and the isle by $|x| \leq 0.55$, $-0.6 \leq y \leq -0.4$. The computational effort is reduced by imposing that everything outside the box $-1 \leq x, y \leq 1$ is ocean. We study an incoming wave with wave number $k = 7$ and direction $w = (\sqrt{1/2}, -\sqrt{1/2})^T$, i.e., the wave front is advancing south-east from the upper left corner of the rectangle. The regular grid supporting the level set function consists of 501×501 nodes. The initial layout of the breakwater is depicted in Figure 7.2. The lower rectangle of the structure surrounding the harbor basin is the isle. Observe the wave resonance in the harbor basin.

We report the progress of the projected descent method in Figure 7.3. The value of the objective functional dropped in 300 iterations from $3.5{\cdot}10^{-1}$ to $1.7{\cdot}10^{-6}$. The H^1-norm of the projected gradient is $6.3{\cdot}10^{-4}$. Note that the objective functional is bounded from below by zero, hence the final domain seems to be close to optimal. Considering the various independent sources of discretization errors in our numerical scheme it seems plausible to say that we found

an approximation of the solution within the order of the discretization error. The final domain and the corresponding wave pattern are shown on the right side of Figure 7.2.



(a) Initial domain **(b)** Final domain

Figure 7.2.: Wave pattern for the initial and final domain of Example 7.1

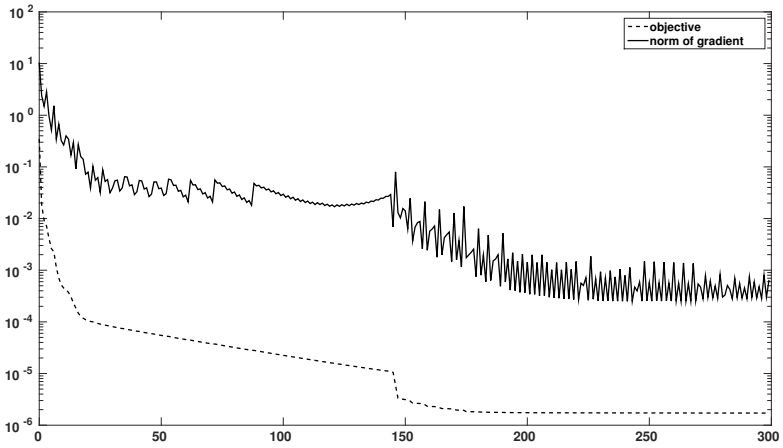
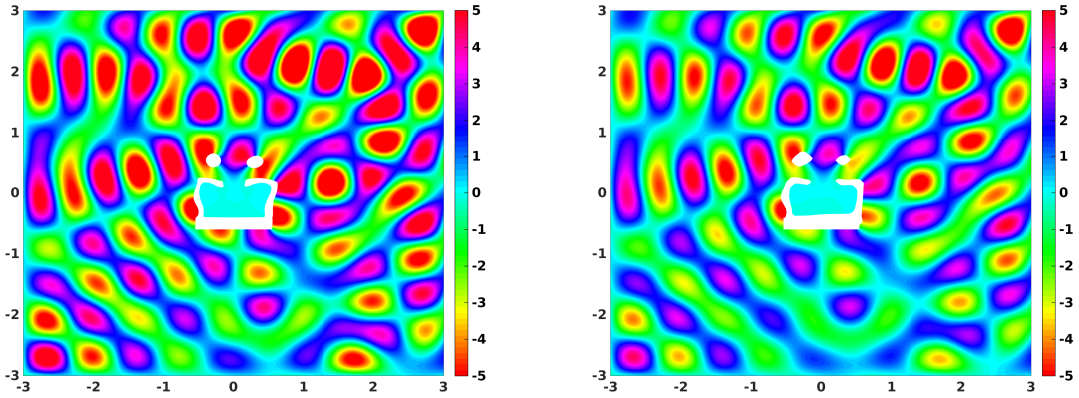


Figure 7.3.: History for Example 7.1

Example 7.2. We start with the same configuration as in Example 7.1, but employ the modification inspired by the nonlinear conjugate gradients method. The final domain is depicted on the left side of Figure 7.4. The difference to the final domain of Example 7.1 is quite small and essentially the same wave pattern can be observed. We report the progress of the algorithm in Figure 7.5. The final value of the objective $1.5{\cdot}10^{-6}$, and the respective H^1-norm of the projected gradient $5.6{\cdot}10^{-4}$ are quite similar to Example 7.1. However, comparing the history of the two methods shows that the modified algorithm took significantly fewer iterations till leveling out.



(a) Example 7.2 **(b)** Example 7.3

Figure 7.4.: Wave pattern for the final domains of Example 7.2 and Example 7.3

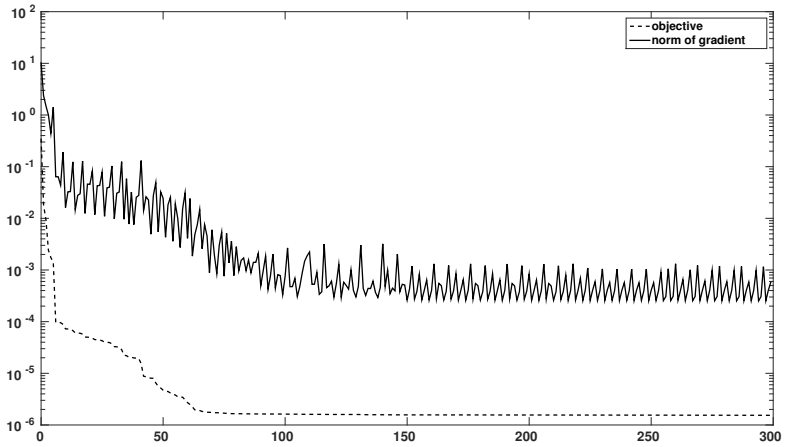


Figure 7.5.: History for Example 7.2

Example 7.3. We start with the same configuration as in Example 7.1, but this time we project the domains, respectively the level set functions, instead of the gradient. The final domain is depicted on the right side of Figure 7.4. There are notable difference to the other two examples. However, qualitatively the shape is similar, and also the wave pattern is resembles the other ones, albeit it exhibits a lower amplitude. We report the progress of the algorithm in Figure 7.6. The method seems to stall. The objective value after 300 iterations was $2.1{\cdot}10^{-4}$ and the norm of the gradient was $6.5{\cdot}10^{-2}$.

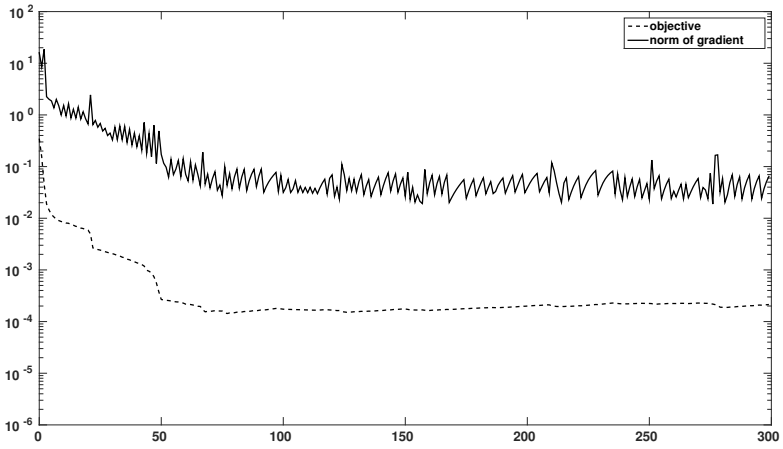


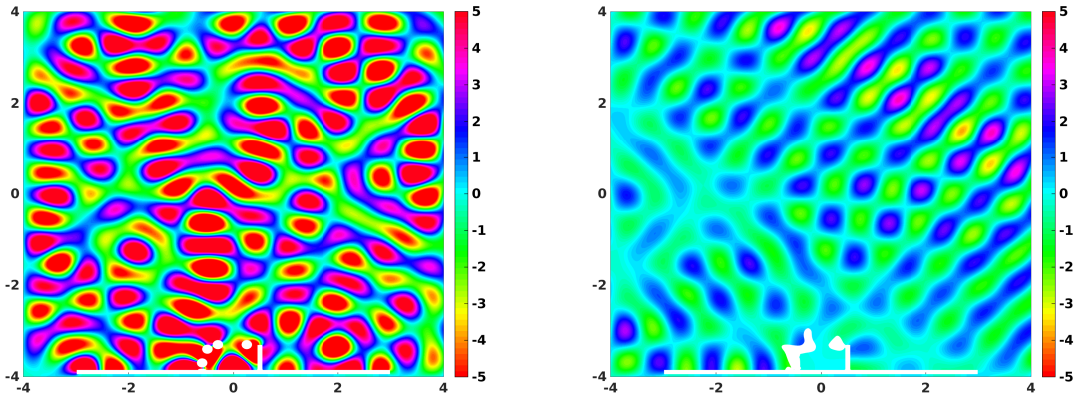
Figure 7.6.: History for Example 7.3

For the first setup of our experiment we can nicely observe resonance inside the harbor basin for an intuitive initial choice of the breakwater. A good configuration can be found by smoothly deforming the breakwater. In particular, no topology changes occur. However, the size of the isle compared to the harbor is very small. Furthermore, the deformed breakwater leads to significant wave interferences outside the harbor, which clearly dominate the incoming wave. In the next experiment we choose a much larger isle, and start with a setup which leads to significant wave interferences throughout the computational domain. They are caused by a large jetty on the right hand side of the harbor basin which we require to stay fixed. In particular, we want to demonstrate now the ability of the algorithm to handle topology changes. For this reason we seed the surrounding of the harbor basin with several small breakwaters.

Example 7.4. The holdall domain $\mathcal{D}$ is given by the square $-4 \leq x, y \leq 4$. The initial layout of the breakwater is depicted in Figure 7.2, where the long rectangle at the bottom is the isle. The harbor basin Q corresponds to the rectangle $|x| \leq 0.4$, and $-3.8 \leq y \leq -3.5$. The harbor approach is given by $|x| \leq 0.1$, $y \geq -3.5$, the isle by $|x| \leq 3$, $-3.95 \leq y \leq -3.85$, and the fixed jetty by $0.45 \leq x \leq 0.55$, $-3.85 \leq y \leq -3.3$. The computational effort is reduced by imposing that everything outside the box $|x| \leq 1, -4 \leq y \leq -2.5$ is ocean. We study again the incoming wave with wave number $k = 7$ and direction $w = (\sqrt{1/2}, -\sqrt{1/2})^T$. The regular grid supporting the level set function consists of 601×601 points.

We report the progress of the projected descent method in Figure 7.8. The value of the objective functional dropped in 300 iterations from 9.3 to $3.3{\cdot}10^{-6}$. The H^1-norm of the

projected gradient is $2.5{\cdot}10^{-4}$. It seems that we found again an approximation of the solution within numerical accuracy. The final domain and the corresponding wave pattern are shown on the right side of Figure 7.7. As one can see the separate small breakwaters have merged into two large ones. Furthermore, one can see the benefit of our nonmonotone linesearch method. During some of those topology changes the objective increased steeply. However, in the long run, this helped find a much better configuration. Note that for this example, in the end, the wave pattern from the incident wave seems to dominate in a large part of the domain.



(a) Initial domain **(b)** Final domain

Figure 7.7.: Wave pattern for the initial and final domain of Example 7.4

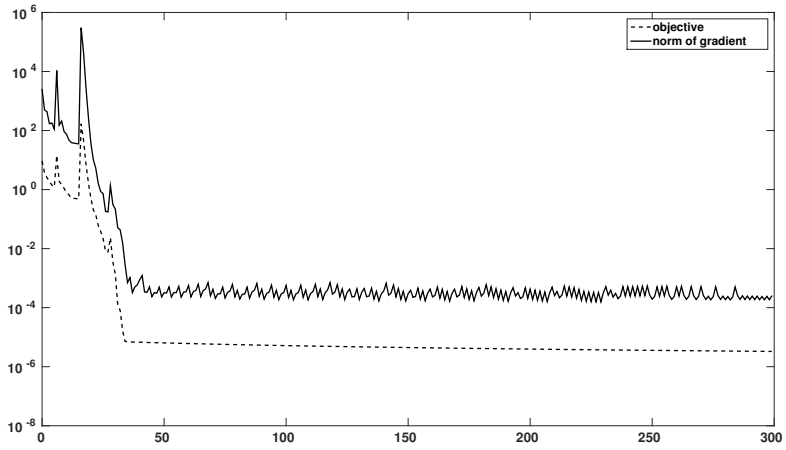
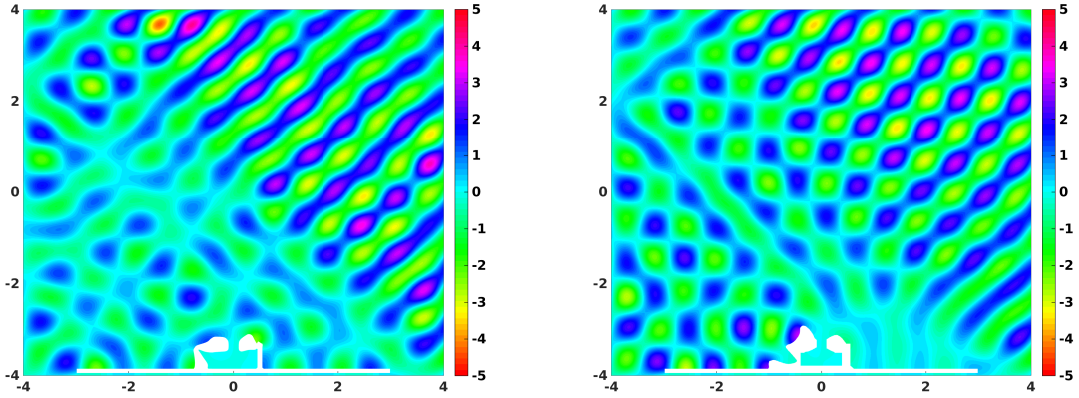


Figure 7.8.: History for Example 7.4

Example 7.5. We start with the same configuration as in Example 7.4, but employ the modification inspired by the nonlinear conjugate gradients method. The final domain is depicted on the left side of Figure 7.9. It differs markedly from the final domain of Example 7.1. Again the wave pattern from the incident wave seems to dominate in a large part of the domain.

We report the progress of the algorithm in Figure 7.10. The final value of the objective functional is $9{\cdot}10^{-8}$, and the respective H^1-norm of the projected gradient is $8{\cdot}10^{-6}$. Note again the fast decrease of both quantities. However, considering our discretization scheme, these values seem to be artificially small.



(a) Example 7.5 **(b)** Example 7.6

Figure 7.9.: Wave pattern for the final domains of Example 7.5 and Example 7.6

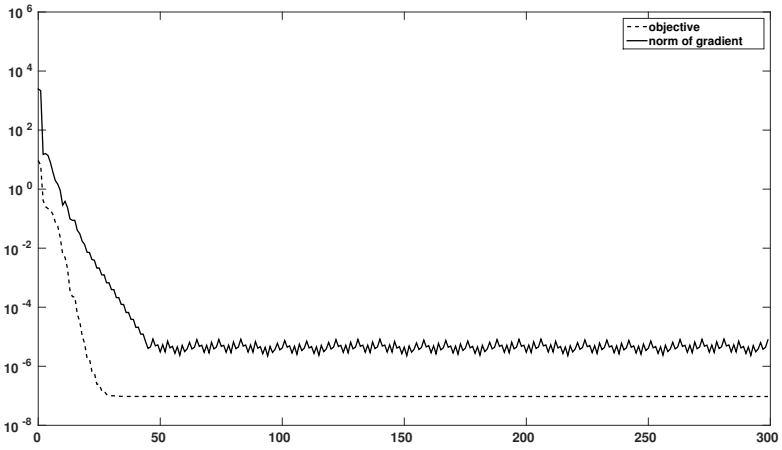


Figure 7.10.: History for Example 7.5

Example 7.6. We start with the same configuration as in Example 7.4, but this time we project the domains, respectively the level set functions, instead of the gradient. The final domain is depicted on the right side of Figure 7.9. There are notable difference to the other two examples. While these mostly keep some distance to the harbor basin and approach, the projected domain method generates a breakwater which touches those two regions to a much greater extend. Furthermore, a closer look at the outline of the breakwater (not depicted here) shows that it is rougher than the counterparts from Example 7.4 and Example 7.5.

The same observation can be made for the first experiment. A reason for this could be the reinitialization of the level set function in each iteration. We report the progress of the algorithm in Figure 7.11. It shows the danger of our quite basic linesearch strategy. Only after over a 100 ascent iterations the algorithm manages to connect two breakwaters and afterwards rapidly finds a good configuration. The objective function value after 300 iterations is $4{\cdot}10^{-8}$ and the norm of the gradient is $2.4{\cdot}10^{-4}$.

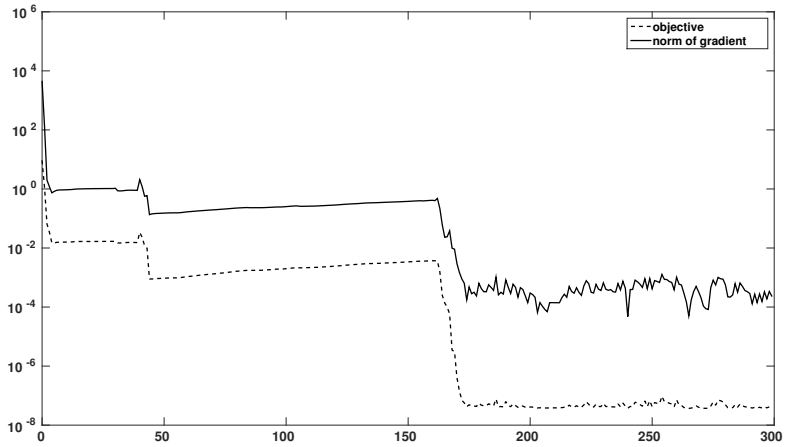


Figure 7.11.: History for Example 7.6

Although our comparison is far from being exhaustive, the modified projected descent method seems to be the most promising of the three presented alternatives. Moreover, in our examples it appears that the projected descent method is able to find approximately *global* solutions of the shape optimization problem. Of course, there are many other possible strategies which could be explored. In particular, it would be very interesting to see how projected (quasi-) Newton methods perform in our example. Furthermore, other strategies for handling the geometric constraints might be investigated.

8. Conclusion and further perspectives

In this thesis, we considered shape optimization problems with a special focus on constraints described by partial differential equations, as well as point-wise geometric constraints. We developed several suitable optimization algorithms, and analyzed their convergence properties in function space. In particular, we addressed Newton-type methods, which offer the potential of fast local convergence. The developed methods were applied to several model problems and substantiated by numerical tests.

Shape optimization bears a close resemblance to optimization on manifolds. We provided a new perspective on the connection between these different fields of optimization, in particular regarding the concepts of second order derivatives and retractions. Based on these insights we discussed several related algorithms for shape optimization. A well established approach supposes that the initial domain is already close to a solution of the shape optimization problem. In that case, the shape optimization problem can be reformulated on a fixed reference domain, and a nonlinear optimization problem in a Banach space setting is obtained. In particular, standard results and techniques can be applied, and second-order methods are readily available. In most of this thesis we considered this setting, more precisely we described the admissible family of domains via perturbations of a reference boundary. We derived an approximation of the Hessian via its operator symbol for a class of elliptic model problems motivated by potential flow pressure matching. The approximation can be used either instead of the Hessian in a Newton-type strategy, or as a preconditioner for the true Hessian. Our numerical experiments indicate that it has great potential in both roles. Inspired by the analogies to optimization on manifolds we recently developed optimization methods which explore the whole admissible family of domains. While the application of at least first order methods of this type is quite common in shape optimization, the available literature regarding the convergence analysis of these algorithms is very scarce. Inspired by linesearch methods along retractions, we developed a framework for a globally convergent linesearch descent method. To the best of our knowledge such a rigorous and general analysis has not been presented before. We further extended the available theory by discussing generalized Newton methods in this framework. Note however, that this approach is subject of current research, and has so far not been compared to more established methods in shape optimization.

Besides the analysis of second order methods, the second focus of this thesis are point-wise geometric constraints. These restrict the admissible shapes to be located inside or outside some given regions. We considered two quite different approaches to handle such situations. In the special case where the free part of the boundary is required to be located inside some convex set, the situation resembles an optimal control problem with control constraints. However, due to the smoothness of the transformations, similar problems as in state constrained optimal control arise. We extended the theory of Moreau-Yosida path following in the field of optimal control and show its applicability to shape optimization. Note, that our results are also applicable in

a wider context. For instance, they fit into the setting of [BU15], where full-waveform seismic inversion with additional constraints on the parameters is discussed. We successfully applied the developed method to shape optimization problems in fluid dynamics, and considered potential flow as well as Stokes flow. We would like to point out that second order methods are, so far, only rarely applied in shape optimization, especially in combination with point-wise geometric constraints. More general geometric constraints in the form of some regions which should be contained, or which should not be contained in the optimal domain, were also considered. We strictly enforced these constraints by projecting the search directions onto a suitable admissible set. This strategy was applied to the problem of minimizing the resonance of a harbor basin by modifying the shape of the breakwaters. Usage of the level set method allows for topological changes during the optimization. Note that such a shape optimization problem has, so far, not been studied in the literature.

Throughout this thesis we pointed out possible extensions and open questions. Let us summarize some of these. Our analysis is based on the group of transformations $\mathcal{F}(\Theta)$ from (2.3) and the associated group of images $\mathcal{O}_\Theta(\Omega_0)$ of a set Ω_0. However, several different transformations may result in the same set $\Omega \in \mathcal{O}_\Theta(\Omega_0)$. An isomorphism between transformations and associated sets can be obtained by studying the quotient group $\mathcal{F}(\Theta)/\mathcal{G}(\Omega_0)$, cf. Section 2.2.1. It would be very interesting to properly study the tangent space of this quotient group and extend our analysis of Chapter 2 to this setting. In particular this might facilitate the analysis of Newton's method, cf. Section 2.9. This section poses several open questions for the presented approach which are all related to the fact that the shape Hessian with respect to $\mathcal{F}(\Theta)$ has necessarily a nontrivial kernel. In this context an extension of the developed algorithms to Quasi-Newton methods seems particularly promising. An alternative might be to employ approximations of the Hessian via its operator symbol. Our results for the potential flow pressure matching are very promising, and motivate the application of our strategy for the derivation of the symbol to other applications. Another issue which has to be addressed are suitable globalization strategies and transition to fast local convergence. A combination of our globally convergent descent method with second order methods would be the most obvious choice, but trust-region or filter based globalization methods are also attractive candidates.

Regarding the topic of point-wise geometric constraints, it would be interesting to study the convergence properties of (inexact) Moreau-Yosida path following in more detail. In particular, the topic of second order necessary and sufficient conditions deserves further attention. Another possible direction of research would be the integration of second order information into the inner product used in our projected descent method in the spirit of variable metric methods. Furthermore, a closer inspection of the idea of projecting domains with respect to a suitable metric is warranted. Of course, also other methods for handling geometric constraints can be conceived. In particular, it would be interesting to see wether one could carry the ideas of sequential quadratic programming or interior point methods over to the setting of shape optimization. Similarly, an extension to efficient Lagrange-Newton strategies for PDE-constrained shape optimization would be beneficial. In this thesis we only considered state equations in which the state enters linearly. We believe that the reduced approach offers more advantages for such equations. However, in the case of nonlinear state equations, the effort of evaluating the reduced objective increases dramatically, and a Lagrange-Newton method might be better suited. The connection to optimization problems on manifolds, or rather to optimization over vector bundles, may point the way to suitable algorithms.

Acknowledgments

I gratefully acknowledge the funding received towards my Ph.D. from the German Research Foundation (DFG) through the International Research Training Group (IGDK) 1754 Munich – Graz 'Optimization and Numerical Analysis for Partial Differential Equations with Nonsmooth Structures', which is co-funded by the Austrian Science Fund (FWF).

This work would not have been possible without the support of a lot of people.

First of all, I would like to express my gratitude to my supervisor Prof. Michael Ulbrich. He introduced me to the fascinating field of optimization in general and shape optimization in particular. His ideas as well as valuable critical questions and remarks helped me find my way into science. Many of the topics in this thesis would not have been addressed without his guidance.

I would also like to thank Prof. Wolfgang Ring, who acted as my second supervisor. He hosted me for my two stays in Graz, and opened a new perspective on shape optimization for me. His friendly support and enthusiasm encouraged me. I had the pleasure to attend two times the wondeful 'Chemnitzer Seminar zur Optimalsteuerung', where I got to know Prof. Roland Herzog. I am grateful for our interesting discussions, and that he agreed to review this thesis.

Furthermore, I would like to thank all my friends and colleagues at the chair of mathematical optimization and the chair of optimal control for the enjoyable time we spent together both at university and outside. You always had an open door for my concerns and made me feel at home. Thanks to Alana, Andre, Andreas, Anne-Céline, Bernhard, Christian, Daniel, Dennis, Dominik, Florian, Florian, Ira, Konstantin, Lucas, Lukas, Martin, Moritz, Philipp, Sebastian, and Sebastian. A special thanks to Konstantin for countless fruitful discussions.

I had the great privilege to be also a member of the IGDK, which was a wonderful experience for me. The many possibilities offered through the program enriched both my academic as well as my personal live. The exchange with students from various other work groups and universities broadened my horizon and helped me see things in a different light. I am thinking especially of our annual summer workshops, but also the graduate seminar, the many compact courses, and similar events. In particular, the collaboration with Daniel Kraft was inspiring. Furthermore, the generous support of the IGDK enabled me to visit various scientific conferences and workshops, both in Germany and abroad, for which I am very grateful. Aside from academics, the people in the IGDK always created a great atmosphere. A special thanks to Armin, Bao, Behzad, Daniel, David, Felix, Jelena, Max, and Philip for the enjoyable time we spent together in Graz.

Finally, I would like to thank my closest friends and family. Without you it would not be worth it. Thank you.

A. Appendix

A.1. The adjoint approach

In this section we show how to efficiently compute the derivatives of the reduced objective

$$\mathfrak{j} : \mathcal{V} \to \mathbb{R}, \quad \mathfrak{j}(U) = J(U, S(U)).$$

This can be achieved by the adjoint approach, which is for example described in [HPUU09, section 1.6]. We only state the necessary equations for the convenience of the reader.

Let us briefly fix the setting. $\mathcal{V}, \mathcal{Y}, \mathcal{Z}$ are Banach spaces, the objective is given by

$$J : \mathcal{V} \times \mathcal{Y} \to \mathbb{R},$$

the state equation by

$$E(U, y) = 0, \text{ where } E : \mathcal{V} \times \mathcal{Y} \to \mathcal{Z}^*,$$

and the design-to state-operator $S : \mathcal{V} \to \mathcal{Y}$ satisfies $E(U, S(U)) = 0$ for all $U \in \mathcal{V}$. We assume that J, E, and S are smooth enough such that all required derivatives exist. Recall the short notation

$$p^* E(U, y) = \langle E(U, y), p \rangle_{\mathcal{Z}^*, \mathcal{Z}}.$$

For a given U we consider the solution $y = S(U) \in \mathcal{Y}$ of the state equation

$$E(U, y) = 0 \quad \text{in } \mathcal{Z}^*,$$

and the solution $p = p(U) \in \mathcal{Z}$ of the adjoint equation

$$p^* E_y(U, y) = 0 \quad \text{in } \mathcal{Y}^*. \tag{A.1}$$

The derivative of the reduced objective is then given by

$$\mathfrak{j}'(U) = J_U(U, y) + p^* E_U(U, y) \in \mathcal{V}^*.$$

If we want to solve the Newton equation with an iterative method like conjugate gradients we need to evaluate $\mathfrak{j}''(U)V \in \mathcal{V}^*$. For this, we further require the solution $z_V \in \mathcal{Y}$ of the linearized state equation

$$E_y(U, y) z_V = -E_U(U, y) V \quad \text{in } \mathcal{Z}^*,$$

and the solution $q_V \in \mathcal{Z}$ of the linearized adjoint equation

$$\begin{aligned} E_y(U,y)q_V = &- J_{yy}(U,y)z_V - J_{Uy}(U,y)V \\ &- p^* E_{yy}(U,y)z_V - pE_{Uy}(U,y)V \text{ in } \mathcal{Y}^*. \end{aligned}$$

Now it holds

$$\begin{aligned} \mathfrak{j}''(U)V = &J_{UU}(U,y)V + J_{yU}(U,y)z_V + q_V^* E_U(U,y) \\ &+ p^* E_{UU}(U,y)V + p^* E_{yU}(U,y)z_V \quad \in \mathcal{V}^*. \end{aligned}$$

A.2. The adjoint approach for the extension operator

We demonstrate in this section how one can compute the derivatives of $j(u) = \mathfrak{j}_\Omega(T(u))$, when T is given as the solution operator of the equation

$$\begin{aligned} KU &= 0 \quad \text{in } \Omega \\ U &= u \quad \text{on } \partial\Omega. \end{aligned} \tag{A.2}$$

Note that this problem fits the setting of Section A.1, hence the procedure is already described there. Nevertheless, we provide the necessary steps here in detail for completeness.

Let us specify (A.2) in more detail. We consider

$$\begin{aligned} KU_0 &= -KFu \quad \text{in } \mathcal{V}_0, \\ U &= U_0 + Fu \quad \text{in } \mathcal{V}, \end{aligned}$$

where $\mathcal{V}$ is a Banach space, $\mathcal{V}_0$ its subspace with zero trace on $\partial\Omega$, and $F : \mathcal{U} \to \mathcal{V}$ is some suitable smooth linear extension operator.

Remark A.1. Imagine for example the linear elasticity equation, $d = 2$, $\mathcal{U} = H^2(\partial\Omega, \mathbb{R}^2)$, and $\mathcal{V} = H^1(\Omega, \mathbb{R}^2)$. In Section 2.11.2 we discussed conditions which ensure then that $T(u)$ has enough regularity to serve as a domain displacement. Introducing the operator F is a standard technique. It can be assumed to be smooth and equal to zero away from the boundary.

We assume that we can find a solution U for every boundary displacement $u \in \mathcal{U}$, and denote the solution operator by $T : U \to \mathcal{V}$. Given a smooth functional $\mathfrak{j} : \mathcal{V} \to \mathbb{R}$ we now consider

$$j : \mathcal{U} \to \mathbb{R}, \quad j(u) = \mathfrak{j}(T(u)).$$

The associated Lagrangian is given by

$$\begin{aligned} L : \mathcal{U} \times \mathcal{V} \times \mathcal{V}_0 \times \mathcal{V}^* \times \mathcal{V}_0^* \to \mathbb{R}, \\ L(u, U, U_0, p_1, p_2) = \mathfrak{j}(U) + \langle p_1, U - U_0 - Fu \rangle_{\mathcal{V}^*,\mathcal{V}} + \langle p_2, KU_0 + KFu \rangle_{\mathcal{V}_0^*,\mathcal{V}_0}. \end{aligned}$$

We obtain

$$j'(u) = -F^* p_1 + F^* K^* p_2 \in \mathcal{U}^*,$$

where $F^* : \mathcal{V}^* \to \mathcal{U}^*$ denotes the dual of F, and p_1, p_2 solve the adjoint equations

$$\begin{aligned} p_1 &= -\mathrm{j}'(T(u)) \quad \text{in } \mathcal{V}^* \\ K^* p_2 &= p_1 \quad \text{in } \mathcal{V}_0^*. \end{aligned}$$

Analogously we can evaluate the second derivative $j''(u)v$ with the adjoint approach. We introduce $(z_1, z_2) \in \mathcal{V} \times \mathcal{V}_0$, and $(q_1, q_2) \in \mathcal{V}^*, \mathcal{V}_0^*$ which solve the linearized state equations

$$\begin{aligned} Kz_2 &= -KFv \quad \text{in } \mathcal{V}_0, \\ z_1 &= z_2 + Fv \quad \text{in } \mathcal{V}, \end{aligned}$$

as well as the adjoint equations for the Hessian

$$\begin{aligned} q_1 &= -\mathrm{j}''(T(u))z_1 \quad \text{in } \mathcal{V}^* \\ K^* q_2 &= q_1 \quad \text{in } \mathcal{V}_0^*. \end{aligned}$$

One obtains

$$j''(u)v = -F^* q_1 + F^* K^* q_2 \in \mathcal{U}^*.$$

List of Figures

List of Tables

List of Algorithms

Bibliography

[ABV13] P. Antonietti, A. Borzì, and M. Verani. Multigrid Shape Optimization Governed by Elliptic PDEs. *SIAM Journal on Control and Optimization*, 51(2):1417–1440, 2013.

[AC08] L. Ambrosio and G. Crippa. Existence, uniqueness, stability and differentiability properties of the flow associated to weakly differentiable vector fields. In *Transport Equations and Multi-D Hyperbolic Conservation Laws*, volume 5 of *Lecture Notes of the Unione Matematica Italiana*, pages 3–57. Springer Berlin Heidelberg, 2008.

[ADF14] G. Allaire, C. Dapogny, and P. Frey. Shape optimization with a level set based mesh evolution method. Technical report, Université Pierre et Marie Curie - Paris VII, 2014.

[AF03] R. A. Adams and J. J. F. Fourier. *Sobolev Spaces*. Pure and Applied Mathematics. Elsevier/Academic Press, 2nd edition, 2003.

[AH01] G. Allaire and A. Henrot. On some recent advances in shape optimization. *Comptes Rendus de l'Académie des Sciences - Series {IIB} - Mechanics*, 329(5):383 – 396, 2001.

[AJT02] G. Allaire, F. Jouve, and A.-M. Toader. A level-set method for shape optimization. *Comptes Rendus Mathematique*, 334(12):1125 – 1130, 2002.

[AJT04] G. Allaire, F. Jouve, and A.-M. Toader. Structural optimization using sensitivity analysis and a level-set method. *J. Comput. Phys.*, 194(1):363–393, 2004.

[All02] G. Allaire. *Shape Optimization by the Homogenization Method*, volume 146 of *Applied Mathematical Sciences*. Springer New York, 2002.

[All07] G. Allaire. *Conception optimale de structures*, volume 58 of *Mathématiques & Applications*. Springer Berlin Heidelberg, 2007.

[AMS08] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[Ang83] F. Angrand. Optimum design for potential flows. *International Journal for Numerical Methods in Fluids*, 3:265–282, 1983.

[Ari95] E. Arian. Analysis of the Hessian for aeroelastic optimization. Technical report, Institute for Computer Applications in Science and Engineering (ICASE), 1995.

[AT95] E. Arian and S. Ta'asan. Shape Optimization in One Shot. In *In Optimal Design and Control*, pages 8–9, 1995.

[AT96] E. Arian and S. Ta'asan. Analysis of the Hessian for aerodynamic optimization: inviscid flow. Technical report, Institute for Computer Applications in Science and Engineering (ICASE), 1996.

[AV99] E. Arian and V. N. Vatsa. A Preconditioning Method for Shape Optimization Governed by the Euler Equations. *International Journal of Computational Fluid Dynamics*, 12(1):17, 27 1999.

[BBM14] M. Bauer, M. Bruveris, and P.W. Michor. Overview of the geometries of shape spaces and diffeomorphism groups. *Journal of Mathematical Imaging and Vision*, 50(1-2):60–97, 2014.

[BC11] H. Bauschke and P. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer New York, 2011.

[BDL09] J. Bolte, A. Daniilidis, and Adrian Lewis. Tame functions are semismooth. *Mathematical Programming*, 117:5–19, 2009.

[Ben95] M. Bendsœ. *Optimization of Structural Topology, Shape, and Material*. Springer Berlin Heidelberg, 1995.

[Ber72] J.C.W. Berkhoff. Computation of combined refraction-diffraction. In *Proceedings 13th Coastal Engineering Conference, Vancouver*, pages 471–490. American Society of Civil Engineers, 1972.

[BFCLS97] J. Bello, E. Fernández-Cara, J. Lemoine, and J. Simon. The Differentiability of the Drag with Respect to the Variations of a Lipschitz Domain in a Navier–Stokes Flow. *SIAM Journal on Control and Optimization*, 35(2):626–640, 1997.

[BGHFSS14] L. Blank, H. Garcke, M. Hassan Farshbaf-Shaker, and V. Styles. Relating phase field and sharp interface approaches to structural topology optimization. *ESAIM: Control, Optimisation and Calculus of Variations*, 20:1025–1058, 10 2014.

[BL98] F. J. Baron Lopez. *Quelque problèmes d'optimisation de formes en électromagnétisme et méchanique de fluides*. PhD thesis, Paris VI, Grenoble, 1998.

[BLUU09] Ch. Brandenburg, F. Lindemann, M. Ulbrich, and S. Ulbrich. A Continuous Adjoint Approach to Shape Optimization for Navier-Stokes Flow. In K. Kunisch, J. Sprekels, G. Leugering, and F. Tröltzsch, editors, *Optimal Control of Coupled Systems of Partial Differential Equations*, volume 158 of *International Series of Numerical Mathematics*, pages 35–56. Birkhäuser Basel, 2009.

[BLUU11] Ch. Brandenburg, F. Lindemann, M. Ulbrich, and S. Ulbrich. Advanced Numerical Methods for PDE Constrained Optimization with Application to Optimal Design in Navier Stokes Flow. In G. Leugering, S. Engell, A. Griewank, M. Hinze, R. Rannacher, V. Schulz, M. Ulbrich, and S. Ulbrich, editors, *Constrained Optimization and Optimal Control for Partial Differential Equations*, volume 160 of *International Series of Numerical Mathematics*, pages 257–275. Birkhäuser Basel, 2011.

[BM14] M. Braack and P.B. Mucha. Directional Do-Nothing Condition for the Navier-Stokes Equations. *Journal of Computational Mathematics*, 32(5):507–521, 2014.

[BO05] M. Burger and S. Osher. A survey on level set methods for inverse problems and optimal design. *European Journal of Applied Mathematics*, 16(2):263–301, 4 2005.

[BR15] L. Blank and C. Rupprecht. An extension of the projected gradient method to a Banach space setting with application in structural topology optimization. Technical report, ArXiv e-prints, 2015.

[Bra11] Ch. Brandenburg. *Adjoint-based adaptive multilevel shape optimization based on goal-oriented error estimators for the instationary Navier-Stokes equations*. PhD thesis, TU Darmstadt, 2011.

[BU15] C. Böhm and M. Ulbrich. A Semismooth Newton-CG Method for Constrained Parameter Identification in Seismic Tomography. *SIAM Journal on Scientific Computing*, 2015.

[Bur04] Martin Burger. Levenberg–marquardt level set methods for inverse obstacle problems. *Inverse Problems*, 20(1):259, 2004.

[But93] R. Butt. Optimal shape design for a nozzle problem. *The ANZIAM Journal*, 35:71–86, 1993.

[CGT00] Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint. *Trust Region Methods*. Society for Industrial and Applied Mathematics, 2000.

[CIL92] M. G. Crandall, H. Ishii, and P. L. Lions. User's guide to viscosity solutions of second order partial differential equations. *Bulletin of the American Mathematical Society*, 27(1):1–67, 1992.

[CK08] I.M. Cohen and P.K. Kundu. *Fluid Mechanics*. Elsevier/Academic Press, 4 edition, 2008.

[CK13] David Colton and Rainer Kress. *Inverse Acoustic and Electromagnetic Scattering Theory*, volume 93 of *Applied Mathematical Sciences*. Springer New York, 2013.

[Cla83] F. H. Clarke. *Optimization and Nonsmooth Analysis*. John Wiley, 1983.

[Cla90] F. Clarke. *Optimization and Nonsmooth Analysis*. Society for Industrial and Applied Mathematics, 1990.

[CT02] E. Casas and F. Tröltzsch. Error estimates for the finite-element approximation of a semilinear elliptic control problem. *Control and Cybernetics*, 31:695–712, 2002.

[CT12] E. Casas and F. Tröltzsch. A general theorem on error estimates with application to a quasilinear elliptic optimal control problem. *Computational Optimization and Applications*, 53(1):173–206, 2012.

[CT15] E. Casas and F. Tröltzsch. Second Order Optimality Conditions and Their Role in PDE Control. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 117(1):3–44, 2015.

[CW05] P. Combettes and V. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[DJS07] Maria Bayard Dühring, Jakob Søndergaard Jensen, and Ole Sigmund. Acoustic design by topology optimization. *Journal of Sound and Vibration*, 317(3-5):557–575, 2007.

[DM74] J. Dennis and J. Moré. A characterization of superlinear convergence and its application to quasi–Newton methods. *Math. Comp.*, 28:549–560, 1974.

[DW12] P. Deuflhard and M. Weiser. *Adaptive Numerical Solution of PDEs*. De Gruyter, 2012.

[DZ92] M.C Delfour and J.P Zolésio. Structure of shape derivatives for nonsmooth domains. *Journal of Functional Analysis*, 104(1):1 – 33, 1992.

[DZ11] M. C. Delfour and J.-P. Zolésio. *Shapes and Geometries*. Society for Industrial and Applied Mathematics, Philadelphia, second edition, 2011.

[EBH09] John W. Eaton, David Bateman, and Sören Hauberg. *GNU Octave version 3.0.1. manual: a high-level interactive language for numerical computations*. CreateSpace Independant Publishing Platform, 2009. ISBN 1441413006.

[EH12] K. Eppler and H. Harbrecht. Shape optimization for free boundary problems – analysis and numerics. In G. Leugering, S. Engell, A. Griewank, M. Hinze, R. Rannacher, V. Schulz, M. Ulbrich, and S. Ulbrich, editors, *Constrained Optimization and Optimal Control for Partial Differential Equations*, volume 160 of *International Series of Numerical Mathematics*, pages 277–288. Springer Basel, 2012.

[EHM08] K. Eppler, H. Harbrecht, and M. Mommer. A new fictitious domain method in shape optimization. *Computational Optimization and Applications*, 40(2):281–298, 2008.

[EHS07] K. Eppler, H. Harbrecht, and R. Schneider. On convergence in elliptic shape optimization. *SIAM Journal on Control and Optimization*, 46(1):61–83, 2007.

[Epp00] K. Eppler. Second derivatives and sufficient optimality conditions for shape functionals. *Control and Cybernetics*, 29(2):458–512, 2000.

[ESSI09] K. Eppler, S. Schmidt, V. Schulz, and C. Ilic. Preconditioning the Pressure Tracking in Fluid Dynamics by Shape Hessian Information. *Journal of Optimization Theory and Applications*, 141(3):513–531, 2009.

[FdSF04] J.L.M. Fernandes, M.A. Vaz dos Santos, and C.J.E.M. Fortes. An element-by-element mild-slope model for wave propagation studies. In *ICS 2004 (Proceedings)*, Journal of Coastal Research, pages 1869–1874, 2004.

[FO03] R. Fedkiw and S. Osher. *Level Set Methods and Dynamic Implicit Surfaces*, volume 153 of *Applied Mathematical Sciences*. Springer New York, 2003.

[FP03] F. Facchinei and J.-S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2003.

[Fre12] M. Frey. *Shape Calculus Applied to State-Constrained Elliptic Optimal Control Problems*. PhD thesis, Universität Bayreuth, 2012.

[Gal11] G.P. Galdi. *An Introduction to the Mathematical Theory of the Navier-Stokes Equations*. Springer Monographs in Mathematics. Springer New York, 2011.

[GHS14] A. Günnel, R. Herzog, and E. Sachs. A Note on Preconditioners and Scalar Products in Krylov Subspace Methods for Self-Adjoint Problems in Hilbert Space. *Electronic Transactions on Numerical Analysis*, 41:13–20, 2014.

[Gig06] Y. Giga. *Surface Evolution Equations: a level set approach*. Monographs in mathematics. Birkhäuser Basel, 2006.

[GLL86] L. Grippo, F. Lampariello, and S. Lucidi. A Nonmonotone Line Search Technique for Newton's Method. *SIAM Journal on Numerical Analysis*, 23(4):707–716, 1986.

[GLL89] L. Grippo, F. Lampariello, and S. Lucidi. A truncated Newton method with nonmonotone line search for unconstrained optimization. *Journal of Optimization Theory and Applications*, 60(3):401–419, 1989.

[GR87] V. Girault and P.-A. Raviart. *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms.* Springer Series in Computational Mathematics. Springer Berlin Heidelberg, 1987.

[Gri85] P. Grisvard. *Elliptic problems in nonsmooth domains*. Pitman Publishing Inc., Marshfield, Massachusetts, 1985.

[Gri89] P. Grisvard. Singularités en elasticité. *Archive for Rational Mechanics and Analysis*, 107(2):157–180, 1989.

[Gri92] P. Grisvard. *Singularities in Boundary Value Problems*. Masson, Paris, 1992.

[Gri06] R. Griesse. Lipschitz stability of solutions to some state-constrained elliptic optimal control problems. *Zeitschrift für Analysis und ihre Anwendungen*, 25(4), 2006.

[Hac92] W. Hackbusch. *Elliptic differential equations: theory and numerical treatment*, volume 18. Springer series in computational mathematics, 1992.

[Har08] H. Harbrecht. Analytical and numerical methods in shape optimization. *Mathematical Methods in the Applied Sciences*, 31(18):2095–2114, 2008.

[Hay54] R.M. Hayes. Iterative methods of solving linear problems on Hilbert space. In *Contributions to the solution of systems of linear equations and the determination of eigenvalues*, number 39 in National Bureau of Standards Applied Mathematics Series, pages 71–103. U. S. Government Printing Office, 1954.

[HC08] S.-H. Ha and S. Cho. Level set based topological shape optimization of geometrically nonlinear structures using unstructured mesh. *Computers & Structures*, 86(13–14):1447 – 1455, 2008.

[Hin05] M. Hintermüller. Fast level-set based algorithms using shape and topological sensitivity information. *Control and Cybernetics*, 34(1):305–324, 2005.

[Hin07] M. Hintermüller. A combined shape Newton and topology optimization technique in real time image segmentation. In *Real-Time PDE-Constrained Optimization*, volume 3, pages 253–274. SIAM, 2007.

[HK01] Michael Hinze and Karl Kunisch. Second order methods for optimal control of time-dependent fluid flow. *SIAM Journal on Control and Optimization*, 40(3):925–946, 2001.

[HK06a] M. Hintermüller and K. Kunisch. Feasible and noninterior path following in constrained minimization with low multiplier regularity. *SIAM Journal on Control and Optimization*, 45(4):1198–1221, 2006.

[HK06b] M. Hintermüller and K. Kunisch. Path-following methods for a class of constrained minimization problems in function space. *SIAM Journal on Optimization*, 17(1):159–187, 2006.

[HLA08] R. H. W. Hoppe, C. Linsenmann, and H. Antil. Adaptive Path Following Primal Dual Interior Point Methods for Shape Optimization of Linear and Nonlinear Stokes Flow Problems. In Ivan Lirkov, Svetozar Margenov, and Jerzy Waśniewski, editors, *Large-Scale Scientific Computing*, volume 4818 of *Lecture Notes in Computer Science*, pages 259–266. Springer Berlin Heidelberg, 2008.

[HM03] J. Haslinger and R. A. E. Mäkinen. *Introduction to Shape Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, 2003.

[HP05] A. Henrot and M. Pierre. *Variation et optimisation de formes. Une analyse géométrique*, volume 48 of *Mathématiques & Applications*. Springer Berlin Heidelberg, 2005.

[HPS15] R. Hiptmair, A. Paganini, and S. Sargheini. Comparison of approximate shape gradients. *BIT Numerical Mathematics*, 55(2):459–485, 2015.

[HPUU09] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints*. Springer Netherlands, 2009.

[HR04] M. Hintermüller and W. Ring. A second order shape optimization approach for image segmentation. *SIAM Journal on Applied Mathematics*, 64(2):442–467, 2004.

[HR12] M. Hinze and A. Rösch. Discretization of optimal control problems. In G. Leugering, S. Engell, A. Griewank, M. Hinze, R. Rannacher, V. H. Schulz, M. Ulbrich, and S. Ulbrich, editors, *Constrained Optimization and Optimal Control for Partial Differential Equations*, volume 160 of *International Series of Numerical Mathematics*, pages 391–430. Springer Basel, 2012.

[HRT96] J. Heywood, R. Rannacher, and S. Turek. Artificial boundaries and flux and pressure conditions for the incompressible Navier–Stokes equations. *International Journal for Numerical Methods in Fluids*, 22(5):325–352, January 1996.

[HS52] M. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49:409–436, 1952.

[HS15] R. Herzog and E. Sachs. Superlinear Convergence of Krylov Subspace Methods for Self-Adjoint Problems in Hilbert Space. *SIAM Journal on Numerical Analysis*, (to appear) 2015.

[HSW14] M. Hintermüller, A. Schiela, and W. Wollner. The Length of the Primal-Dual Path in Moreau–Yosida-Based Path-Following Methods for State Constrained Optimal Control. *SIAM Journal on Optimization*, 24(1):108–126, 2014.

[HT10] M. Hinze and F. Tröltzsch. Discrete concepts versus error analysis in PDE-constrained optimization. *GAMM-Mitteilungen*, 33(2):148–162, 2010.

[HWH03] Michael A. Heroux, James M. Willenbring, and Robert Heaphy. Trilinos Developers Guide. Technical Report SAND2003-1898, Sandia National Laboratories, 2003.

[Ihl98] F. Ihlenburg. *Finite Element Analysis of Acoustic Scattering*. Springer New York, 1998.

[IK03] K. Ito and K. Kunisch. Semi-smooth newton methods for state-constrained optimal control problems. *Systems & Control Letters*, 50(3):221 – 228, 2003.

[JN80] C. Johnson and J. C. Nedelec. On the coupling of boundary integral and finite element methods. *Mathematics of Computation*, 35(152):1063–1079, 1980.

[Kin15] B. Kiniger. A transformation approach in shape optimization: Existence and regularity results. *Numerical Functional Analysis and Optimization*, 2015.

[Kir14] A. Kirchner. *Adaptive regularization and discretization for nonlinear inverse problems with PDEs*. PhD thesis, Technische Universität München, 2014.

[KK15] M. Keuthen and D. Kraft. Shape Optimization of a Breakwater. *Inverse Problems in Science and Engineering*, (to appear) 2015.

[Kli11] W. Klingenberg. *Riemannian Geometry*. de Gruyter Studies in Mathematics. De Gruyter, Berlin Boston, 2011.

[KP91] J. Katz and A. Plotkin. *Low speed aerodynamics: From wing theory to panel methods*. McGraw-Hill, New York, 1991.

[KR09] K. Krumbiegel and A. Rösch. A virtual control concept for state constrained optimal control problems. *Computational Optimization and Applications*, 43(2):213–233, 2009.

[Kra14] D. Kraft. The "level-set" package for GNU Octave. Octave Forge, 2014.

[Kra15a] D. Kraft. A Hopf-Lax Formula for the Time Evolution of the Level-Set Equation and a New Approach to Shape Sensitivity Analysis. Technical Report IGDK-2015-18, University of Graz, 2015.

[Kra15b] D. Kraft. *A Level-Set Framework for Shape Optimisation.* PhD thesis, Karl-Franzens-Universität Graz, Graz, 2015.

[Kru14] F. Kruse. *Interior point methods for optimal control problems with state constraints.* PhD thesis, Technische Universität München, 2014.

[KU15] M. Keuthen and M. Ulbrich. Moreau-Yosida Regularization in Shape Optimization with Geometric Constraints. *Computational Optimization and Applications*, 62(1):181–216, 2015.

[KV13] Bernhard Kiniger and Boris Vexler. A priori error estimates for finite element discretizations of a shape optimization problem. *ESAIM: Mathematical Modelling and Numerical Analysis*, 47:1733–1763, 2013.

[Lau00] M. Laumen. Newton's method for a class of optimal shape design problems. *SIAM Journal on Optimization*, 10(2):503–533, 2000.

[LBvBW12] Kevin Long, Paul T. Boggs, and Bart G. van Bloemen Waanders. Sundance: High-level software for pde-constrained optimization. *Sci. Program.*, 20(3):293–310, July 2012.

[Lei86] R. Leis. *Initial Boundary Value Problems in Mathematical Physics.* J. Wiley & Teubner Verlag, Stuttgart, 1986.

[Lin12] F. Lindemann. *Theoretical and Numerical Aspects of Shape Optimization with Navier-Stokes Flows.* PhD thesis, Technische Universität München, München, 2012.

[LLL01] Jiin-Jen Lee, Ching-Piau Lai, and Yigong Li. Application of computer modeling for harbor resonance studies of Long Beach & Los Angeles harbor basins. *Coastal Engineering Proceedings*, 1(26), 2001.

[LS13] A. Laurain and K. Sturm. Domain expression of the shape derivative and application to electrical impedance tomography. Technical Report 1863, Weierstrass Institute for Applied Analysis and Stochastics, 2013.

[MH97] J.P.-Y. Maa and H.-H. Hwung. A wave transformation model for harbor planning. In *Proceedings, Waves '97, Virginia Beach*, volume 1, pages 256–270, 1997.

[Mic72] Anna Maria Micheletti. Metrica per famiglie di domini limitati e proprietà generiche degli autovalori. *Annali della Scuola Normale Superiore di Pisa - Classe di Scienze*, 26(3):683–694, 1972.

[Mic14] G. Michailidis. *Manufacturing Constraints and Multi-Phase Shape and Topology Optimization via a Level-Set Method.* PhD thesis, École Ploytechnique, 2014.

[Mic15] P.W. Michor. Manifolds of mappings and shapes. Technical report, Universität Wien, 2015.

[MM06] P.W. Michor and D.B. Mumford. Riemannian geometries on spaces of plane curves. *Journal of the European Mathematical Society*, 8(1):1–48, 2006.

[MP01] B. Mohammadi and O. Pironneau. *Applied Shape Optimization for Fluids*. Oxford University Press, Oxford, 2001.

[MRT06] C. Meyer, A. Rösch, and F. Tröltzsch. Optimal Control of PDEs with Regularized Pointwise State Constraints. *Comput. Optim. Appl.*, 33(2-3):209–228, March 2006.

[MS76] F. Murat and J. Simon. Sur le contrôl par un domaine géométrique. Technical report, Universite P. et M. Curie (Paris IV), 1976.

[MT99] K. Malanowski and F. Tröltzsch. Lipschitz stability of solutions to parametric optimal control for parabolic equations. *Journal for Analysis and its Applications (ZAA)*, 18:469–489, 1999.

[MY09] C. Meyer and I. Yousept. Regularization of state-constrained elliptic optimal control problems with nonlocal radiation interface conditions. *Computational Optimization and Applications*, 44(2):183–212, 2009.

[Neč12] J. Nečas. *Direct Methods in the Theory of Elliptic Equations*. Springer Berlin Heidelberg, 2012.

[NR95] A. Novruzi and J. R. Roche. Second order derivatives, Newton method, application to shape optimization. Rapport de Recherche 2555, Institut National de Recherche en Informatique et en Automatique, 1995.

[NT08] I. Neitzel and F. Tröltzsch. On convergence of regularization methods for nonlinear parabolic optimal control problems with control and state constraints. *Control and Cybernetics*, 37(4):1013–1043, 2008.

[NW06] J. Nocedal and S. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2006.

[NZP04] M. Nemec, D. W. Zingg, and T. H. Pulliam. Multipoint and multiobjective aerodynamic shape optimization. *AIAA Journal*, 42(6):1057–1065, 2004.

[OS88] S. Osher and J.A. Sethian. Fronts Propagating with Curvature-Dependent Speed: Algorithms Based on Hamilton-Jacobi Formulations. *Journal of Computational Physics*, 79:12–49, 1988.

[OS91] S. Osher and C. Shu. High-Order Essentially Nonoscillatory Schemes for Hamilton–Jacobi Equations. *SIAM Journal on Numerical Analysis*, 28(4):907–922, 1991.

[OS01] S. Osher and F. Santosa. Level Set Methods for Optimization Problems Involving Geometry and Constraints: I. Frequencies of a Two-Density Inhomogeneous Drum. *Journal of Computational Physics*, 171(1):272 – 288, 2001.

[Per04] P.-O. Persson. *Mesh Generation for Implicit Geometries*. PhD thesis, Massachusetts Institute of Technology, Cambridge (Massachusetts), 2004.

[Pir73] O. Pironneau. On optimum profiles in Stokes flow. *Journal of Fluid Mechanics*, 59:117–128, 6 1973.

[Pir82] O. Pironneau. Optimal shape design for elliptic systems. In R.F. Drenick and F. Kozin, editors, *System Modeling and Optimization*, volume 38 of *Lecture Notes in Control and Information Sciences*, pages 42–66. Springer Berlin Heidelberg, 1982.

[Pir84] O. Pironneau. *Optimal Shape Design For Elliptic Systems*. Springer Series in Computational Physics. Springer Berlin Heidelberg, 1984.

[PTW08] U. Prüfert, F. Tröltzsch, and M. Weiser. The convergence of an interior point method for an elliptic control problem with mixed control-state constraints. *Comput. Optim. Appl.*, 39(2):183–218, March 2008.

[Rab09] A.B. Rabinovich. Seiches and harbor oscillations. In Y.C. Kim, editor, *Handbook of Coastal and Ocean Engineering*, pages 93–236. World Scientific Publishing Company Singapore, 2009.

[RW12] W. Ring and B. Wirth. Optimization Methods on Riemannian Manifolds and Their Application to Shape Space. *SIAM Journal on Optimization*, 22(2):596–627, 2012.

[SAM03] B. Samet, S. Amstutz, and M. Masmoudi. The Topological Asymptotic for the Helmholtz Equation. *SIAM Journal on Control and Optimization*, 42(5):1523–1544, 2003.

[Sch69] J.T. Schwartz. *Nonlinear functional analysis*. Gordon and Breach Science, 1969.

[Sch08] A. Schiela. A Simplified Approach to Semismooth Newton Methods in Function Space. *SIAM Journal on Optimization*, 19(3):1417–1432, 2008.

[Sch09] A. Schiela. Barrier methods for optimal control problems with state constraints. *SIAM Journal on Optimization*, 20(2):1002–1031, 2009.

[Sch10] S. Schmidt. *Efficient Large Scale Aerodynamic Design Based on Shape Calculus*. PhD thesis, University of Trier, Germany, 2010.

[Sch14] Volker Schulz. A Riemannian View on Shape Optimization. *Foundations of Computational Mathematics*, 14(3):483–501, 2014.

[Set96] J. A. Sethian. A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences*, 93(4):1591–1595, 1996.

[Set99] J. Sethian. *Level Set Methods and Fast Marching Methods : Evolving Interfaces in Computational Geometry,Fluid Mechanics, Computer Vision, and Materials Science*. Cambridge University Press, Cambridge, 2nd edition, 1999.

[SG11] A. Schiela and A. Günther. An interior point algorithm with inexact step computation in function space for state constrained optimal control. *Numerische Mathematik*, 119(2):373–407, 2011.

[Sha13] Alexander Shapiro. Differentiability properties of metric projections onto convex sets. Technical report, Optimization Online, 2013.

[Sim91] J. Simon. Domain variation for drag in Stokes flow. In X. Li and J. Yong, editors, *Control Theory of Distributed Parameter Systems and Applications*, volume 159 of *Lecture Notes in Control and Information Sciences*, pages 28–42. Springer Berlin Heidelberg, 1991.

[SSW14] V. H. Schulz, M. Siebenborn, and K. Welker. Towards a Lagrange-Newton approach for PDE constrained shape optimization. *ArXiv e-prints*, May 2014.

[STD+96] M. Schäfer, S. Turek, F. Durst, E. Krause, and R. Rannacher. Benchmark computations of laminar flow around a cylinder. In E. Hirschel, editor, *Flow Simulation with High-Performance Computers II*, volume 48. Vieweg+Teubner Verlag, Wiesbaden, 1996.

[Ste83] Trond Steihaug. The conjugate gradient method and trust regions in large scale optimization. *SIAM Journal on Numerical Analysis*, 20(3):626–637, 1983.

[SW00] J.A. Sethian and Andreas Wiegmann. Structural boundary design via level set and immersed interface methods. *Journal of Computational Physics*, 163(2):489 – 528, 2000.

[SZ92] J. Sokolowski and J.-P. Zolésio. *Introduction to Shape Optimization*. Series in Computational Mathematic. Springer Berlin Heidelberg, 1992.

[SZ07] Joachim Schöberl and Walter Zulehner. Symmetric Indefinite Preconditioners for Saddle Point Problems with Applications to PDE-Constrained Optimization Problems. *SIAM Journal on Matrix Analysis and Applications*, 29(3):752–773, 2007.

[Tem77] Roger Temam. *Navier-Stokes Equations: Theory and Numerical Analysis*. North Holland, 1977.

[THM08] J.I. Toivanen, J. Haslinger, and R.A.E. Mäkinen. Shape optimization of systems governed by Bernoulli free boundary problems. *Computer Methods in Applied Mechanics and Engineering*, 197(45–48):3803 – 3815, 2008.

[TM15] Inc. The MathWorks. *MATLAB R2012a*. The MathWorks, Inc., Natick, Massachusetts, United States, 2015.

[Toi96] Philippe L. Toint. An assessment of nonmonotone linesearch techniques for unconstrained optimization. *SIAM Journal on Scientific Computing*, 17(3):725–739, 1996.

[Trö05] F. Tröltzsch. Regular lagrange multipliers for control problems with mixed pointwise control-state constraints. *SIAM Journal on Optimization*, 15(2):616–634, 2005.

[Ulb11] M. Ulbrich. *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. MOS-SIAM series on optimization, 2011.

[vDMLvK13] N.P. van Dijk, K. Maute, M. Langelaar, and F. van Keulen. Level-set methods for structural topology optimization: a review. *Structural and Multidisciplinary Optimization*, 48(3):437–472, 2013.

[WW14] T. Wick and W. Wollner. On the differentiability of stationary fluid-structure interaction problems with respect to the problem data. Technical Report 25, Universität Hamburg, 2014.

[WWG03] Michael Yu Wang, Xiaoming Wang, and Dongming Guo. A level set method for structural topology optimization. *Computer Methods in Applied Mechanics and Engineering*, 192(1–2):227 – 246, 2003.

[Xin09] Xiuying Xing. *Computer modeling for wave oscillation problems in harbors and coastal regions*. PhD thesis, University of Southern California, Los Angeles, 2009.

[XLR11] Xiuying Xing, Jiin-Jen Lee, and Fredric Raichlen. Harbor resonance: a comparison of field measurements to numerical results. *Coastal Engineering Proceedings*, 1(32), 2011.

[XSLW12] Qi Xia, Tielin Shi, Shiyuan Liu, and Michael Yu Wang. A level set solution to the stress-based structural shape and topology optimization. *Computers & Structures*, 90–91(0):55 – 64, 2012.

[You10] L. Younes. *Shapes and Diffeomorphisms*, volume 171 of *Applied Mathematical Sciences*. Springer Berlin Heidelberg, 2010.

[ZH04] Hongchao Zhang and William W. Hager. A nonmonotone line search technique and its application to unconstrained optimization. *SIAM Journal on Optimization*, 14(4):1043–1056, 2004.

[Zol73] Jean-Paul Zolesio. *Sur la localisation d'un domaine*. PhD thesis, Université de Nice, 1973. Thèse de 3e cycle Mathématiques Nice 1973.

[Zol79] Jean-Paul Zolesio. *Identification de domaines par déformations*. PhD thesis, Université de Nice, 1979. Thèse d'État Mathématiques Nice 1979.