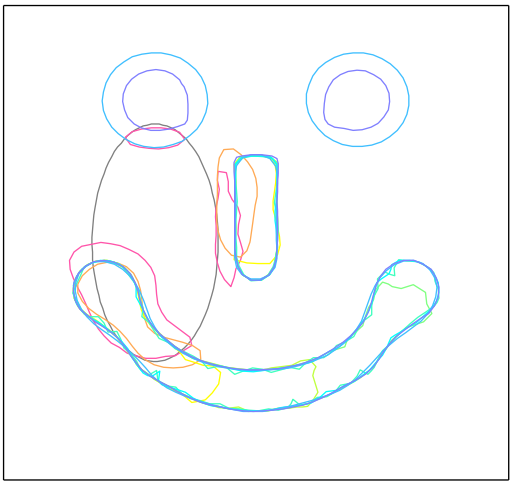


Shape Calculus Applied to State-Constrained Elliptic Optimal Control Problems

Dissertation

zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.)

der Fakultät für Mathematik, Physik und Informatik der Universität Bayreuth vorgelegt von

Dipl.-Math. Michael Frey

geboren am 11. Juli 1983 in Stuttgart

1. Gutachter: Prof. Dr. Hans Josef Pesch (Universität Bayreuth)
2. Gutachter: Prof. Dr. Fredi Tröltzsch (Technische Universität Berlin)
3. Gutachter: Prof. Dr. Eduardo Casas (Universidad de Cantabria)

Tag der Einreichung: 22.05.2012
Tag des Kolloquiums: 09.11.2012

Contents

Preface

Abstract

This thesis is devoted to the analysis of a very simple, pointwisely state-constrained optimal control problem of an elliptic partial differential equation. The transfer of an idea from the field of optimal control of ordinary differential equations, which proved fruitful with respect to both theoretical treatment and design of algorithms, is the starting point. On this, the state inequality constraint, which is regarded as an equation inside the active set, is differentiated in order to obtain a control law.

A geometrical splitting of the constraints is necessary to carry over this approach to the chosen model problem. The associated assertions are rigorously ensured. The subsequent derivation of a control law in the sense of the abovementioned idea yields an equivalent reformulation of the model problem. The active set appears as an independent and equal optimization variable in this new formulation. Thereby a new class of optimization problem is established, which forms a hybrid of optimal control and shape-/ topology optimization: set optimal control. This class is integrated into the very abstract framework of optimization on vector bundles; for that purpose some important notions from the field of calculus on manifolds are introduced and related with shape calculus.

First order necessary conditions of the set optimal control problem are derived by means of two different approaches: on the one hand a reduced approach via the elimination of the state variable, which uses a formulation as bilevel optimization problem, is pursued, and on the other hand a formal Lagrange principle is presented.

A comparison of the newly obtained optimality conditions with those known form literature yields relations between the Lagrange multipliers; in particular, it becomes apparent that the new approach involves higher regularity. The comparison is embedded to the theory of partial differential-algebraic equations, and it is shown that the new approach yields a reduction of the differential index.

Upon investigation of the gradient and the second covariant derivative of the objective functional different Newton- and trial algorithms are presented and discussed in detail. By means of a comparison with the well-established primal-dual active set method different benefits of the new approach become apparent. In particular, the new algorithms can be formulated in function space without any regularization. Some numerical tests illustrate that an efficient and competitive solution of state-constrained optimal control problems is achieved.

The whole work gives numerous references to different mathematical disciplines and encourages further investigations. All in all, it should be regarded as a first step towards a more comprehensive perspective on state-constrained optimal control of partial differential equations.

Zusammenfassung

Die vorliegende Arbeit befasst sich mit der Analyse eines sehr einfachen elliptischen Optimalsteuerungsproblems mit punktweisen Zustandsbeschränkungen. Ausgangspunkt ist die Übertragung einer Idee, die sich im Bereich der Optimalsteuerung gewöhnlicher Differenzialgleichungen sowohl bei theoretischer Behandlung als auch beim Entwurf von Lösungsalgorithmen als fruchtbar erwiesen hat. Hierzu wird die Zustandsbeschränkung in der aktiven Menge als Gleichung gesehen, aus der durch Differenziation ein Steuergesetz hergeleitet werden kann.

Um diese Herangehensweise auf das gewählte Modellproblem übertragen zu können, ist eine gebietsweise Aufspaltung der Nebenbedingung nötig, was durch den Beweis entsprechender Aussagen abgesichert wird. Die anschließende Herleitung eines Steuergesetzes im Sinne obengenannter Idee führt zu einer äquivalenten Umformulierung des Modellproblems. Die neue Formulierung beinhaltet die aktive Menge in natürlicher Art und Weise als eigenständige Optimierungsvariable, wodurch eine neuartige Klasse von Optimierungsproblemen begründet wird, die einen Hybrid aus Optimalsteuerung und Form-/Topologieoptimierung darstellt: Mengen-Optimalsteuerung. Diese Klasse wird eingebettet in einen sehr abstrakten Rahmen der Optimierung auf Vektorbündeln; hierzu werden insbesondere relevante Begriffe aus dem Bereich der Differenzialrechnung auf Mannigfaltigkeiten eingeführt und mit dem „Shape calculus“ in Beziehung gesetzt.

Auf zwei verschiedenen Wegen werden notwendige Optimalitätsbedingungen erster Ordnung für das Mengen-Optimalsteuerungsproblem hergeleitet: einerseits wird ein reduktionistischer Ansatz verfolgt, der die Zustandsvariable eliminiert und hier über eine Bilevelproblemformulierung führt, andererseits wird der Weg eines formalen Lagrangeprinzips präsentiert.

Ein Vergleich der neu erhaltenen Optimalitätsbedingungen mit denen aus der Literatur bekannten ermöglicht es Beziehungen zwischen Lagrangemultiplikatoren herzustellen; insbesondere wird klar, dass die neue Herangehensweise Regularitätsverbesserungen mit sich bringt. Der Vergleich der notwendigen Bedingungen wird eingebettet in die Theorie partiell differential-algebraischer Gleichungen und es wird nachgewiesen, dass man durch den neuen Ansatz eine Indexreduktion erhält.

Auf Basis der Untersuchung von Gradient und zweiter kovarianter Ableitung des Zielfunktionals werden verschiedene Newton- und Trialverfahren vorgestellt und eingehend untersucht. Durch einen Vergleich mit der etablierten primal-dualen aktiven Mengenstrategie werden verschiedene Vorzüge des neuen Ansatzes herausgearbeitet. Insbesondere sind die neuen Algorithmen ohne Regularisierung im Funktionenraum formulierbar. Verschiedene numerische Test zeigen, dass der neue hier verfolgte Ansatz die effiziente und konkurrenzfähige Lösung von zustandsbeschränkten Optimalsteuerungsproblemen ermöglicht.

Die gesamte Arbeit liefert zahlreiche Querbezüge zu anderen mathematischen Teilgebieten und regt an diese weiter zu verfolgen. Insgesamt ist sie als ein erster Schritt zu einer umfassenderen Betrachtung der zustandsbeschränkten Optimalsteuerung bei partiellen Differenzialgleichungen zu betrachten.

Structure of this work

The structure of the work is as follows: Chapter 2 is devoted to the presentation of the analytical approach to new necessary conditions of a very simple elliptic model problem which is introduced at the beginning. Subsequent to the introduction of the model problem, a very brief overview on preliminary work in the fields of optimal control of ordinary and partial differential equations and of shape optimization is given in Section 2.1. Starting from this basis, the original model problem undergoes a series of reformulations in Section 2.2, which yields a new type of optimization problem, called set optimal control problem. At this, two fundamental ideas of the whole approach become apparent, namely the geometrical splitting of the spacial domain into active and inactive sets, and the transformation of the state constraint into a control law. As a result of the geometrical splitting, the active set with respect to the state constraint becomes an optimization variable of its own. Hence, shape and topology calculus come on the scene in a natural way. The derivation of the control law, which is inspired by results from optimal control of ordinary differential equations, is directly connected to considerations of partial differential-algebraic equations, which are addressed in Section 2.7. The following two sections present two alternative ways on how to obtain first order necessary condition of the set optimal control problem. In particular, Section 2.3 uses a methodology based on a bilevel formulation and its reduction to a shape optimization problem, whereas Section 2.4 applies a formal Lagrange technique. Especially the first approach requires several subsequent steps, which are illustrated on page 10. It turns out, that the reformulation of the state constraint yields associated Lagrange multipliers, which are closely related to the well-known multipliers from previous work. By that means, the known specific inherent structure of the latter multiplier, to be a sum of a regular and a singular part, is reobtained. In view of efficient numerics, Section 2.5 is devoted to a brief second order analysis of the reduced objective functional of the shape optimization problem and the associated Lagrangian. The major result is that the second order derivative has a null at the optimum, which helps to understand some of the numerical findings of Chapter 4. Moreover, the algorithms of Chapter 3 require the identification of second order covariant derivatives of the shape functional and of the Lagrangian, respectively. In order to do so, Section 2.6 provides an abstract perspective on shape calculus. Hereunto, the calculus is imbedded to the more general framework of differential calculus on manifolds and vector bundles. This reasoning enables a very abstract point of view on shape optimization and optimal control, which provides valuable insight to the structure of the new class of set optimal control problems. Finally, Section 2.7 is devoted to a brief analysis of the new first order necessary conditions from the perspective of partial differential-algebraic equations. It is shown, that the new necessary conditions have a lower differentiation index than the well-known ones. This finding is related to the analog result from optimal control of ordinary differential equations.

Chapter 3 is devoted to the development of algorithms for solving the set optimal control problem, which was derived in Chapter 2. At first, descent algorithms on manifolds, in general, and on a specific set of feasible sets, in particular, are analyzed in detail in Section 3.1. In addition, it is shown that the optimum of the original model problem is no strict local minimum of the reduced shape functional. Consequently, gradient based algorithms are not applicable. Hence, some remarks on Newton's method on manifolds are presented in Section 3.2. Section 3.3 contains considerations how the new first order necessary conditions are accessible for numerical solution. At this, the perspectives from the reduced approach of Section 2.3, form the Lagrangian approach of Section 2.4 and of free boundary problems are used. This analysis yields different Newton type algorithms in Section 3.4. Moreover, some trial algorithms are presented there, which can be regarded as simplified Newton schemes. In order to get a better understanding of the benefits of the new algorithmic approach it is compared with the well-established primal-dual active set strategy in Section 3.5.

In order to get a first impression of the capability of the theoretical and algorithmic approach of the chapters 2 and 3, some basic numerical results are presented in Chapter 4. At first, Section 4.1 gives an overview on different aspects of the finite element discretization, which is applied. The focus is on the explanation of the problems that arise from the need of coping with different active sets during the iteration of the algorithms, such as updating the interface and mesh deformation. Finally, Section 4.2 contains different findings with respect to the numerical analysis of test examples. It turns out, that the new algorithmic approach, though not being globally convergent, features sufficient stability and indicates a mesh independent behavior. Moreover, it is shown, that certain types of changes of the topology of the active set can be attained automatically in the course of the iteration of the algorithms. A comparison with an

enhanced version of the primal-dual active set method reveals encouraging performance of the still quite basic new approach.

The different results of this work are summarized and placed within a broader context in Chapter 5. In particular, some selected open or undiscussed questions are seized.

Acknowledgements

I would like to take this opportunity to express my sincere gratitude to my supervisor Prof. Dr. Hans Josef Pesch for introducing me into the field of optimal control of partial differential equations. This thesis is essentially due to his continuous support, guidance and inspiration as well as to countless helpful discussions. His group at the University of Bayreuth is a creative and pleasant environment to work in.

The tight cooperation with my friends and colleges Dipl.-Math. Simon Bechmann and Dr. Armin Rund was characterized by deep felt esteem, intentness and the stubborn will to get to the bottom of mathematics. In this way, both of them have had a deep impact on the success of my research. Moreover, they greatly helped by proof-reading this thesis.

I would like to thank Prof. em. Dr. Christian G. Simader, Prof. Dr. Kurt Chudej, Dr. Julia Fischer and Dipl.-Math. Stefan Wendl for many helpful discussions and their inspiration.

I am grateful to Dr. Stephan Schmidt, who opened the field of shape optimization to me and who was on hand with help and advice in many discussions. In addition, Jun.-Prof. Dr. Winnifried Wollner and Dr. Anton Schiela helped to analyze specific questions of theory of partial differential equations, whereas Dr. Stefan Elsenhans and Dipl.-Math. Tim Kirschner introduced me to the fields of Lie groups and calculus on manifolds.

Finally, I must express my appreciation to my family and friends for their support, especially Salome for her love and patience.

This work has been supported by the German science foundation (DFG) in the context of the project "Restringierte Optimierungsprobleme mit partiellen Differentialgleichungen und Anwendungen auf Schweißprozesse".

Bayreuth, November 13, 2012

Michael Frey

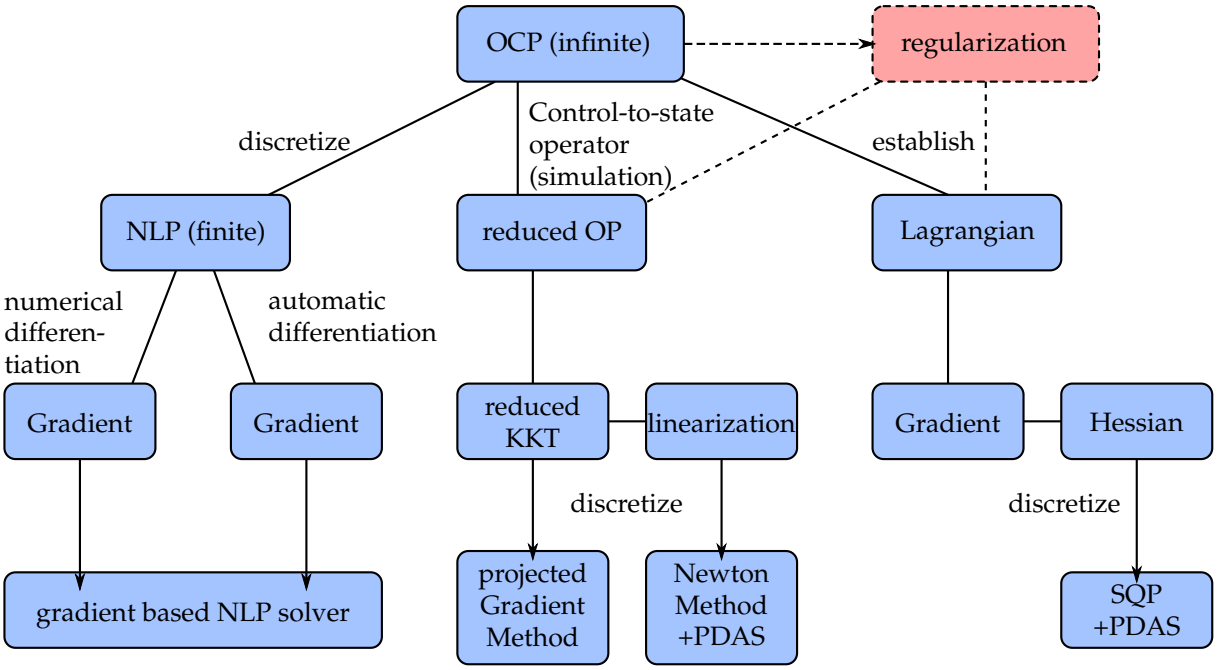
michael.frey[at]uni-bayreuth.de

CHAPTER 1

Introduction

Optimal control of partial differential equations (OC-PDE) has gained more and more attention in applied mathematics during the last three decades. On the one hand this discipline is appealing from a mathematical point of view, since many different branches meet there, and on the other hand this topic is interesting from a practical point of view, since many real-life problems in engineering (like cooling processes [35, 160], laser hardening [36, 57], laser welding [134, 135, 136, 137], control of fuel cells [34, 152, 32, 31, 153, 154, 33, 145] or crystal growth [127]), economics [47], biology [60] and many more can be modeled by that means. Though considerably progress is made, both theory and implementation of robust, efficient and easy to handle software packages are far from being complete. In particular, the treatment of state constraints, which are a natural part of almost any optimal control problem, constitute a striking challenge.

Based on the excellent overview of Herzog and Kunisch [81] different algorithmic approaches for solving *optimal control problems* (OCP) with PDE constraints can roughly be classified as follows.



- numeric differentiation is costly since the NLP is large scale and may yield poor approximation of gradient information
- automatic differentiation may be restricted to simple problems

+ high accuracy of gradient information (directly accessible)
○ requires efficient solver of the state equation (simulation)
- requires efficient solver of the adjoint equation

- consistent discretization has to be provided
+ linearized equations have to be solved in each iteration
- linear systems are larger than in reduced approach, since state

+ automatic differentiation ensures consistency in discretization + little software-user interaction required + highly sophisticated NLP solvers available	- consistency of the discretization of forward and adjoint solver has to be guaranteed - nonlinear state equation has to be solved iteratively in each iteration - KKT system might be not accessible for very complex OCPs	variable is an explicit optimization variable - enhanced simulation software of the state equation is typically not applicable - gradient of the Lagrangian might be not accessible for very complex OCPs

The left branch – often called "first discretize, then optimize" – is well-established nowadays in the field of *optimal control of ordinary differential equations* (OC-ODE) even for complex problems. In contrast, the other two branches – "first optimize, then discretize" – play a minor role there, since usability of corresponding software is more involved. Nonetheless, they are necessary, if very high accuracy of the solution is required, as for instance in problems of space travel. With respect to OC-PDE the situation changes considerably. The "fist discretize, then optimize" approach is confronted with two inherent difficulties: OCPs with partial differential equation yield large-scale nonlinear optimization problems (NLP) after discretization, such that even enhanced NLP solvers can be overcharged. Moreover, discretization of PDEs is not as straight forward as in the case of ODEs. Henceforth, a higher amount of software-user interaction is required so far. Consequently, the approach of "first optimize, then discretize" – with its two representative branches *black-box solvers* (middle) and *all-at-once solver* (right) – still is state of the art, and there is no evidence that this will change in the near future.

Pointwise state constraints play a crucial point in the treatment of OC-PDE and associated solvers. First order based projected gradient methods do not possess a natural extension to this situation, since the projection onto the feasible set cannot be performed easily there, since the set is characterized by means of the state, which is reduced within those methods. In addition, Newton differentiability of *first order necessary conditions* (NC), i. e. the *Karush-Kuhn-Tucker conditions* (KKT), is lost. Thus, higher order solvers cannot be applied (or suffer from mesh dependency), since they are based upon either linearization of the KKT system or differentiation of the gradient of the Lagrangian. A well-established and successful remedy is the application of a quadratic penalization of the state constraint, called *Moreau-Yosida regularization*. The price to pay is an extra loop in the algorithms. Hence, the numerical schemes contain (at least) three nested loops: the outer regularization loop, the Newton- or SQP-loop and the inner loop of the *primal-dual active set strategy* (PDAS). Basically the same holds true, when using *interior point methods* instead of SQP/PDAS. In contrast, the numerical schemes developed in this work come without regularization.

The content of this thesis emanated from the idea of construction new necessary conditions for state-constrained optimal control problems of partial differential equations. At this, the ideas of Bryson, Denham and Dreyfus [18] (*BDD approach*), which are situated in the field of OC-ODE, should serve as a blue print; so to speak of a bridge building between the two disciplines of OC-ODE and OC-PDE. This task is animated with two long-term goals, which have already been reached in the field of OC-ODE:

- gain an apriori insight into the structure of the active set, which is associated with the order of the state constraint, and
- construct efficient numerics upon the basis of the new necessary conditions, which exploit some inherent structure of the multipliers associated with the state constraint.

However, it has become apparent that developing the ideas of Bryson et al. in the world of OC-PDE is considerably more complex and requires results of several other mathematical disciplines, see Figure 1.1. This finding strongly influences the setup and the focus of this work. It is written from the perspective of OC-PDE; henceforth it is expected that the reader is familiar with theory and numerics of state constraint OC-PDE. Indeed, the reader needs not to be an expert in field of shape optimization, which enters the considerations in a very natural way. Unfortunately, brevity inhibits a satisfactory introduction to shape calculus and shape optimization and thus the reader is referred to literature as often as possible. It turns out, that the identification of the second covariant derivative in shape calculus (which is necessary for the algorithms of Chapter 3) requires a profound analysis of shape calculus, which is based upon infinite dimensional manifolds and vector bundles and not available in literature so far. It is not expected, that the reader is familiar with all notions used for it; hence, their definitions are included in this work. All in all, the presentation tends to have a bias on the discipline of shape calculus in order to built a bridge between shape calculus and shape optimization on the one hand and state-constrained optimal control of PDEs on the other hand. The disciplines are amalgamated in a hybrid problem formulation: the set optimal control problem.

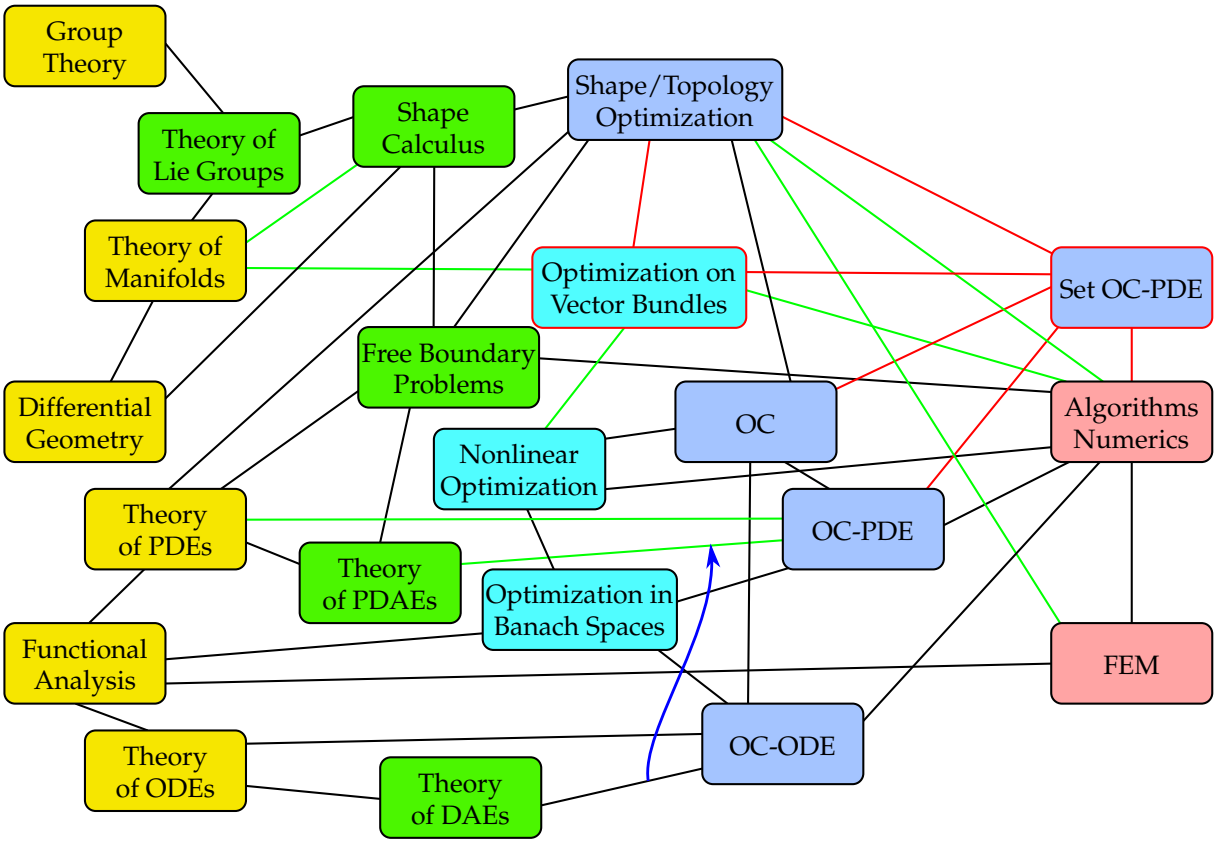


Figure 1.1: Mindmap of mathematical fields involved and their connections. At this, the coloring of the different fields displays their positioning in pure, applied and computational mathematics. Black links are used, whereas green ones are analyzed in-depth and/or partly extended. Red connections symbolize completely new ideas/results. The blue arrow illustrates the original goal of carrying some ideas from theory of OC-ODE over to OC-PDE.

The presentation of other important topics, such as the theory of *(partial) differential-algebraic equations* (P)DAE, are kept as short as possible. In this sense, the focus has been shifted from the analysis of the BDD approach towards a review of shape calculus and shape optimization in the context of state-constrained OC-PDE, which is the basis for any further research on the way to reach the abovementioned long-term goals. The first one still remains far from being reached, whereas some basic numerical results (for the probably simplest state constraint OC-PDE model problem) can be presented in Chapter 4.

It should be emphasized, that the analysis of the state-constrained OCP reveals, that shape and topology calculus/optimization play an equally important part. However, a profound investigation of the topology related part is beyond the scope of this work and now open for further research. Nonetheless, there are some minor tricks included in the numerical treatment, such that (some kind) of topology chances of the active set can be achieved.

Moreover, it is important to notice, that the investigation of the chosen model problem is only a first step towards a deep understanding of the fundamental ideas of this work. The depicted OCP is chosen to be linear quadratic (i. e. convex); hence, it is some sort of odd to construct algorithms which introduce a strongly nonlinear behavior by means of shape dependency. However, they are expected to be able to cope with fully nonlinear problems and even reveal their full performance there. Nonetheless, the simpler framework of the model OCP was chosen in order to keep theory as easy as possible, such that the whole reasoning starting from the reformulation of the OCP right up to the construction of algorithms can be exhibited here.

CHAPTER 2

Theory

This thesis is concerned throughout with the following *state-constrained elliptic optimal control problem* (OCP) of *tracking type*. Although this model problem is probably the most elementary state constraint representative of OC-PDE, it is possible to present and analyze the main ideas of this work.

Definition 1 (Model problem)**:**
Let $\Omega \subset \mathbb{R}^2$ be a bounded domain of class $C^{1,1}$ and let $\Gamma := \partial\Omega$ denote its boundary.[1] Let the *desired state* $y_d \in H^1(\Omega)$, the *control shift* $u_d \in H^2(\Omega)$, the *Tikhonov regularization parameter* $\lambda \in \mathbb{R}^+$, and let the state constraining functions $y_{\max}, y_{\min} \in H^4(\Omega)$, such that for all $x \in \Omega$ holds $y_{\min}(x) < y_{\max}(x)$.
The following state-constrained linear-quadratic elliptic optimal control problem is called *model problem*:
Find $(\bar{u}, \bar{y}) \in L^2(\Omega) \times H^1(\Omega)$ minimizing the tracking type *objective (functional)*

$$J(u,y) := \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\|u - u_d\|^2_{L^2(\Omega)} \tag{2.1a}$$

subject to the *state equation*

$$-\Delta y + y = u \quad \text{a. e. in } \Omega, \tag{2.1b}$$
$$\partial_n y = 0 \quad \text{a. e. on } \Gamma, \tag{2.1c}$$

and the pointwise *state constraints*

$$y - y_{\max} \leq 0 \quad \text{a. e. in } \Omega, \tag{2.1d}$$
$$y_{\min} - y \leq 0 \quad \text{a. e. in } \Omega, \tag{2.1e}$$

where the *state (variable)* y and the *control (variable)* u may vary in $H^1(\Omega)$ and $L^2(\Omega)$, respectively.

The regularity assumptions made for the different coefficient functions and the boundary regularity of the domain Ω are quite strong. They are required in order to achieve a fairly straight forward analysis, which is presented in the following. It is discussed, when these assumptions are needed. The control shift u_d has no practical meaning, but simplifies the construction of analytical test examples; see Paragraph 4.2.1.

2.1 Overview on preliminary work

Before starting the actual analysis, i. e. the derivation of new first order necessary conditions, this section is devoted to a very brief sketch of some preliminary work. For this purpose, Paragraph 2.1.1 contains results from OC-PDE, which are directly related to the model problem. In contrast, paragraphs 2.1.2 and 2.1.3 only list some literature from the fields of shape optimization and OC-ODE, which deals with related topics, since brevity inhibits a satisfying presentation of all relevant assertions.

[1] Further information on local characterization of sets are edited in [44, Chp. 2 Sec. 3–6]. In particular, a definition of sets of class $C^{1,1}$ can be found in [44, Chp. 2 Def. 3.1]. Moreover, its defining property is illustrated in the proof of Lemma 2.

2.1.1 Results in optimal control of PDEs

Since the model problem (2.1) is probably the simplest state-constrained optimal control problem of partial differential equations, it is well studied and a lot of literature concerning different details can be found, for instance, in [88]. The aim of this paragraph is to cite some selected results in order to provide the basis for the following treatment.

Proposition 1 (Unique solvability of the model problem)**:**
The model problem (2.1) is uniquely solvable; the optimum is denoted by $(\bar{u}, \bar{y}) \in L^2(\Omega) \times H^1(\Omega)$.

A proof of this well-known result can be found in e. g. [25], [159, Thm. 2.15] or [10, Satz 1.5b].

Remark (Higher regularity of the states)**:**
Due to the $C^{1,1}$-regularity of the boundary Γ each state of an admissible pair $(u, y) \in L^2(\Omega) \times H^1(\Omega)$ is even in $H^2(\Omega)$. Actually, the mapping (*control-to-state operator*) $L^2(\Omega) \to H^2(\Omega)$, $u \mapsto y$, where y is the unique solution of the state equation (2.1b), (2.1c) is a continuous isomorphism, cf., for instance, [69, Thm. 2.2.2.5 and Thm. 2.3.3.2].

Consequently, one can require that the state y is an element of $H^2(\Omega)$ without loss of generality. This consideration is of fundamental importance for the analysis for pointwisely state-constrained optimal control problems, as shown below.

Proposition 2 (First order necessary conditions; Casas)**:**
Assume that there exists a $\delta > 0$ such that $y_{\max}(x) - y_{\min}(x) \geq \delta$, $x \in \Omega$, and let $(\bar{u}, \bar{y})$ be the optimal solution of the model problem (2.1).

Then there are *Lagrange multipliers* $\mu^{\max} = \mu_\Omega^{\max} + \mu_\Gamma^{\max}$, $\mu^{\min} = \mu_\Omega^{\min} + \mu_\Gamma^{\min} \in C^0(\overline{\Omega})^*$ and an *adjoint state* $p^{trad} \in \bigcap_{s\in[1,2[} W^{1,s}(\Omega)$ such that the following first order necessary conditions (*Karush-Kuhn-Tucker conditions*; KKT) are fulfilled: The original constraints (2.2a)–(2.2d), the *adjoint equation* (2.2e), (2.2f), the *complementary slackness conditions* (2.2h), (2.2i) and the *sign conditions* (2.2j), (2.2k)

$$
\begin{aligned}
-\Delta\bar{y} + \bar{y} &= \bar{u} && \text{a. e. in } \Omega, && \text{(2.2a)}\\
\partial_{\boldsymbol{n}}\bar{y} &= 0 && \text{a. e. on } \Gamma, && \text{(2.2b)}\\
\bar{y} - y_{\max} &\leq 0 && \text{a. e. in } \Omega, && \text{(2.2c)}\\
y_{\min} - \bar{y} &\leq 0 && \text{a. e. in } \Omega, && \text{(2.2d)}\\
-\Delta p^{trad} + p^{trad} &= \bar{y} - y_d + \mu_\Omega^{\max} - \mu_\Omega^{\min} && \text{a. e. in } \Omega, && \text{(2.2e)}\\
\partial_{\boldsymbol{n}} p^{trad} &= \mu_\Gamma^{\max} - \mu_\Gamma^{\min} && \text{a. e. on } \Gamma, && \text{(2.2f)}\\
\lambda\,(\bar{u} - u_d) + p^{trad} &= 0 && \text{a. e. in } \Omega, && \text{(2.2g)}\\
\langle \mu^{\max}\,,\, \bar{y} - y_{\max}\rangle_{C^0(\overline{\Omega})^*, C^0(\overline{\Omega})} &= 0, && && \text{(2.2h)}\\
\langle \mu^{\min}\,,\, y_{\min} - \bar{y}\rangle_{C^0(\overline{\Omega})^*, C^0(\overline{\Omega})} &= 0, && && \text{(2.2i)}\\
\mu^{\max} &\geq 0, && && \text{(2.2j)}\\
\mu^{\min} &\geq 0. && && \text{(2.2k)}
\end{aligned}
$$

Proof. Due to Sobolev's embedding theorems, cf., for instance, [2, Thm. 4.12A], $H^2(\Omega)$ is continuously embedded in $C^0(\overline{\Omega})$. Define $\hat{y} := y_{\max} - \delta/2 \in H^4(\Omega) \subset H^2(\Omega)$. Due to the assumptions on the state constraints, the pair

$$(\hat{u}, \hat{y}) := (-\Delta\hat{y} + \hat{y}, \hat{y}) \in H^2(\Omega) \times C^0(\overline{\Omega})$$

is a *Slater point* of the optimal control problem (2.1). That is to say, $(\hat{u}, \hat{y})$ is an interior point of the admissible set $L^2(\Omega) \times \{y \in H^2(\Omega) \,|\, y_{\min} \leq y \leq y_{\max}\}$, where the topology of $C^0(\overline{\Omega})$ is used for the state associated component. The assertion follows now from, e. g., [25, Thm. 2] or [159, Thm. 6.5]. □

Remark:
Due to the theorem of Riesz-Radon the dual of $C^0(\overline{\Omega})$ can be identified with the space $\mathcal{M}(\overline{\Omega})$ of regular Borel measures on Ω, cf., for instance, [4, Thm. 4.22]. Consequently, the multipliers $\mu^{\max}$ and $\mu^{\min}$ can be identified with elements of $\mathcal{M}(\overline{\Omega})$ and thus do not necessarily possess a pointwise interpretation. Moreover, their decomposition into one part on Ω and a second part on the boundary Γ is just a splitting such that $\mu_\Omega^{\max}$ and $\mu_\Omega^{\min}$ have their support in Ω, whereas the support of $\mu_\Gamma^{\max}$ and $\mu_\Gamma^{\min}$ is localized within Γ.

The adjoint equation is only a symbolic notation for its weak formulation, cf. [25, 26, 3] and [159, Sec. 7.2.3].

This formulation strikes a nerve of the necessary conditions. The multiplier $\mu := \mu^{\max} - \mu^{\min}$ does not possess a pointwise interpretation in general. Moreover, the regularity can not be improved actually, since there are examples (e. g., cf. [128, 24], [91, Ex. 3]), where the optimal state hits the state constraint in isolated points, and consequently the multiplier is a *Dirac measure*, which is concentrated there. However, the Lagrange multipliers $\mu^{\max}$ and $\mu^{\min}$ reveal some intrinsic structure, provided there are some additional assumptions fulfilled for the active set, see Definition 3 and Assumption 1. In particular, they can be decomposed into a regular part on the interior of the active set and a singular part on the interface (that is the boundary of the active set). These results are due to Bergounioux and Kunisch [14].

Proposition 3 (Enhancement of first order necessary conditions; Bergounioux and Kunisch)**:**
Let $(\bar{u}, \bar{y}) \in L^2(\Omega) \times H^1(\Omega)$ be the unique optimal solution of the model problem (2.1), let the adjoint state p^{trad} and the multipliers $\mu^{\max}$ and $\mu^{\min}$ be given by Proposition 2 and let Assumption 1 be fulfilled. Let $p_{\mathcal{I}}^{trad}$, $p_{\mathcal{A}_{\max}}^{trad}$ and $p_{\mathcal{A}_{\min}}^{trad}$ be the restrictions of p^{trad} on the inactive, respectively active sets (cf. Definition 3). Use the same notation for $\bar{y}$. Furthermore, for later use let (here $\gamma := \partial\mathcal{A}$, see Definition 3)

$$
\begin{aligned}
&\mu_{\mathcal{I}}^{\max} := \mu^{\max}|_{\mathcal{I}\cup\mathcal{A}_{\min}}, && \mu_{\mathring{\mathcal{A}}}^{\max} := \mu^{\max}|_{\mathring{\mathcal{A}}_{\max}}, && \mu_{\gamma}^{\max} := \mu^{\max}|_{\gamma_{\max}},\\
&\mu_{\mathcal{I}}^{\min} := \mu^{\min}|_{\mathcal{I}\cup\mathcal{A}_{\max}}, && \mu_{\mathring{\mathcal{A}}}^{\min} := \mu^{\min}|_{\mathring{\mathcal{A}}_{\min}}, && \mu_{\gamma}^{\min} := \mu^{\min}|_{\gamma_{\min}},\\
& && \mu_{\mathring{\mathcal{A}}} := \mu_{\mathring{\mathcal{A}}}^{\max} - \mu_{\mathring{\mathcal{A}}}^{\min}, && \mu_{\gamma} := \mu_{\gamma}^{\max} - \mu_{\gamma}^{\min},\\
&c_{\max} := \lambda(-\Delta^2 y_{\max} + 2\Delta y_{\max} - \Delta u_d + u_d) - y_{\max} + y_d, && && (2.3a)\\
&c_{\min} := \lambda(\quad \Delta^2 y_{\min} - 2\Delta y_{\min} + \Delta u_d - u_d) + y_{\min} - y_d, && && (2.3b)\\
&p_{\mathcal{A}}^{trad} := p^{trad}|_{\mathcal{A}}. && &&
\end{aligned}
$$

Then there holds:

1. In the active set everything is determined by the coefficient functions
$$
\begin{aligned}
p_{\mathcal{A}_{\max}}^{trad} &= \lambda(\Delta y_{\max} - y_{\max} + u_d) \in H^2(\mathring{\mathcal{A}}_{\max}), && (2.4a)\\
p_{\mathcal{A}_{\min}}^{trad} &= \lambda(\Delta y_{\min} - y_{\min} + u_d) \in H^2(\mathring{\mathcal{A}}_{\min}), && (2.4b)\\
\mu_{\mathring{\mathcal{A}}}^{\max} &= c_{\max} \geq 0 \text{ in } L^2(\mathring{\mathcal{A}}_{\max}), && (2.4c)\\
\mu_{\mathring{\mathcal{A}}}^{\min} &= c_{\min} \geq 0 \text{ in } L^2(\mathring{\mathcal{A}}_{\min}). && (2.4d)
\end{aligned}
$$
2. The adjoint state in the inactive set is given as the H^2-regular solution of
$$
\begin{aligned}
-\Delta p_{\mathcal{I}}^{trad} + p_{\mathcal{I}}^{trad} &= \bar{y}_{\mathcal{I}} - y_d \quad && \text{a. e. in } \mathcal{I}, && (2.4e)\\
\partial_n p_{\mathcal{I}}^{trad} &= 0 && \text{a. e. on } \Gamma, && (2.4f)\\
p_{\mathcal{I}}^{trad}|_{\gamma_{\max}} &= p_{\mathcal{A}_{\max}}^{trad}|_{\gamma_{\max}} && \text{a. e. on } \gamma_{\max}, && (2.4g)\\
p_{\mathcal{I}}^{trad}|_{\gamma_{\min}} &= p_{\mathcal{A}_{\min}}^{trad}|_{\gamma_{\min}} && \text{a. e. on } \gamma_{\min}. && (2.4h)
\end{aligned}
$$
3. In particular, $\mu_\Gamma^{\max} = \mu_\Gamma^{\min} = 0$, $\mu_{\mathcal{I}}^{\max} = 0$ and $\mu_{\mathcal{I}}^{\min} = 0$.

4. The interface parts of the multipliers ensue as the jump in the normal derivatives of the adjoint states and thus they are $H^{1/2}$-regular (cf. Lemma 1)

$$\mu_\gamma^{\max} = \partial_{\boldsymbol{n}}^{\mathcal{I}} p_{\mathcal{I}}^{trad} + \partial_{\boldsymbol{n}}^{\mathcal{A}} p_{\mathcal{A}_{\max}}^{trad} \quad \text{a. e. on } \gamma_{\max}, \tag{2.4i}$$

$$\mu_\gamma^{\min} = -\partial_{\boldsymbol{n}}^{\mathcal{I}} p_{\mathcal{I}}^{trad} - \partial_{\boldsymbol{n}}^{\mathcal{A}} p_{\mathcal{A}_{\min}}^{trad} \quad \text{a. e. on } \gamma_{\min}. \tag{2.4j}$$

Remark:
The just presented proposition says:

- The Lagrange multipliers $\mu_{\max}$ and $\mu_{\min}$ are concentrated on the active sets, which is basically a consequence of complementary slackness (2.2h), (2.2i) and the definition of the active sets. This result was already proven in [25, Sec. 8].
- Each of the multipliers can be decomposed into two parts. One of them, $\mu_\gamma^{\max}$ and $\mu_\gamma^{\min}$, respectively, is concentrated on the interface. If it is regarded as an object living on the interface, it is $H^{1/2}$-regular. But if one treats it as an object defined on the active set or even on $\overline{\Omega}$, it is not a function, but a measure in $\mathcal{M}(\overline{\Omega})$. Consequently, the assumptions made in Proposition 3 are weak enough in order to preserve the measure character of the multipliers. The other component, $\mu_{\mathcal{A}}^{\max}$ and $\mu_{\mathcal{A}}^{\min}$, respectively, is distributed in the active set and L^2-regular. Altogether one recognizes, that the measure nature appears only at the boundary of the active set.
- The regular, distributed part of the multipliers is prescribed by the choice of the coefficient functions. Consequently, the position of the active set in Ω is restricted apriori by means of the coefficient functions, provided that the optimal control problem is *strictly complementary*[2]: Those subsets of Ω in which the combination of coefficients (2.4c) and (2.4d) are negative cannot by parts of the active set. This insight can be used algorithmically, see the 2nd item of the discussion on page 109. Moreover, this fact should be minded when constructing test examples, cf. Paragraph 4.2.1.
- The adjoint state is a regular function locally. Its global regularity suffers from a kink at the interface between active and inactive set, which is induced by the singular component of the multiplier.
- The weak continuity of the adjoint state across the interfaces (2.4g), (2.4h) combined with the gradient equation (2.2g) reveals weak continuity of the optimal control across the interfaces

$$\bar{u}_{\mathcal{I}}|_{\gamma_{\max}} = \bar{u}_{\mathcal{A}_{\max}}|_{\gamma_{\max}} \tag{2.5a}$$

$$\bar{u}_{\mathcal{I}}|_{\gamma_{\min}} = \bar{u}_{\mathcal{A}_{\min}}|_{\gamma_{\min}}. \tag{2.5b}$$

2.1.2 Results in shape optimization

The model problem (2.1) is reformulated in Theorem 2, such that the active set (cf. Definition 3) becomes an optimization variable. Differentiation with respect to this variable requires the application of a suitable calculus, namely *shape-* and *topology calculus*. The foundation of modern shape calculus is close-knit with Céa, Gioan and Michel [27], Murat and Simon [129] and Zolésio [162], whereas the notion of a topology derivative goes back to Sokołowski and Żochowski, [101]. The latter field is left untouched within this thesis and consequently this paragraph focuses on the first one. Shape calculus and shape optimization has gained much attention during the last three decades and is consolidated in several textbooks, e. g. [150, 139, 151, 78, 115, 19, 44]. Especially the recent book of Delfour and Zolésio contains an extensive list of references, is an excellent starting point to get in touch with the theoretical basis of shape optimization and is the main reference of this work. Due to brevity a separate presentation of relevant results, such as the *Hadamard structure theorem* ([44, Chp. 9 Thm 3.6]), rules for differentiation of shape functionals ([44, Chp. 9 Thm. 4.2 and 4.3]) and the local shape derivative of elliptic boundary value problems (BVP) ([146, Sec. 3.4]) is abandoned here. However, detailed references are always given when results from those fields are applied.

[2] An optimization problem, or more precisely an inequality constraint of an optimization problem, is said to be strictly complementary, if the associated Lagrange multiplier is positive almost everywhere in the active set.

2.1.3 Results in optimal control of ODEs

As already indicated, an essential idea of this thesis is to transfer some ideas from optimal control of ordinary differential equations to OC-PDE. The Karush-Kuhn-Tucker conditions of Proposition 2 are analogously to the OC-ODE result of Jacobson, Lele and Speyer [102], often called *direct adjoining approach*, since the original state constraints are directly adjoint to the objective. However, even one decade before Bryson, Denham and Dreyfus published an alternative version of first order necessary [18] in 1963, where a reformulation of the state constraint is adjoint to the objective. This idea is often called *indirect adjoining approach*, but is referred to as *Bryson-Denham-Dreyfus-* or simply *BDD approach* here. The reformulation is based upon differentiation and is discussed in Paragraph 2.2.2 in more detail. Later on, Maurer succeeded in integrating both approaches into a more general framework in the habilitation [121]; later published in [122]. His investigations revealed that the multipliers associated with the different total time derivatives of the constraining function up to a certain order, by which the objective is augmented, become the more regular, the higher the order of the derivative is. An excellent survey on many more different contributions, which can be clustered roughly to the two approaches, is due to Hartl, Sethi and Vickson [75].

A second essential idea of this work is to use the active set of the state constraint as a separate and equal variable of the OCP. This is similar to introduce the starting and endpoints of active sets as optimization variables in context of OC-ODE, as it is done in *multiple shooting methods* and which are applied to solve *multipoint boundary value problems*; see [155, Sec. 7.3.5]. The combination of direct adjoining approach and multiple shooting methods proved to be a superior starting point for numerical treatment of complex OCPs; see, e. g. [20, 21, 123]

2.2 Reformulation into a set optimal control problem

After having commented on some preliminary work the actual involvement with content of this thesis starts now. This section is devoted to an specific reformulation of the original model problem (2.1). Hereto, two of the essential ideas of this thesis are applied, namely

- introducing the active set of the state constraint as a separate and equal variable of the optimal control problem and
- reformulating the state constraint in order to derive a control law.

It is important to notice, that the first idea can be realized without the second; however, they are presented in combination for brevity. Nonetheless, the procedure contains several steps such that it may be helpful to gain an overview of the whole reasoning by means of the illstration on page 10.

2.2.1 Geometrical Splitting

One essential idea, which is the basis for all the following, is a splitting of the state equation and the state constraint. The splitting is adapted to the geometrical partition of the domain Ω in the two parts of active and inactive set (cf. Definition 3) and should keep the original information. That is to say, the original constraints and the their split counterparts are to be equivalent.

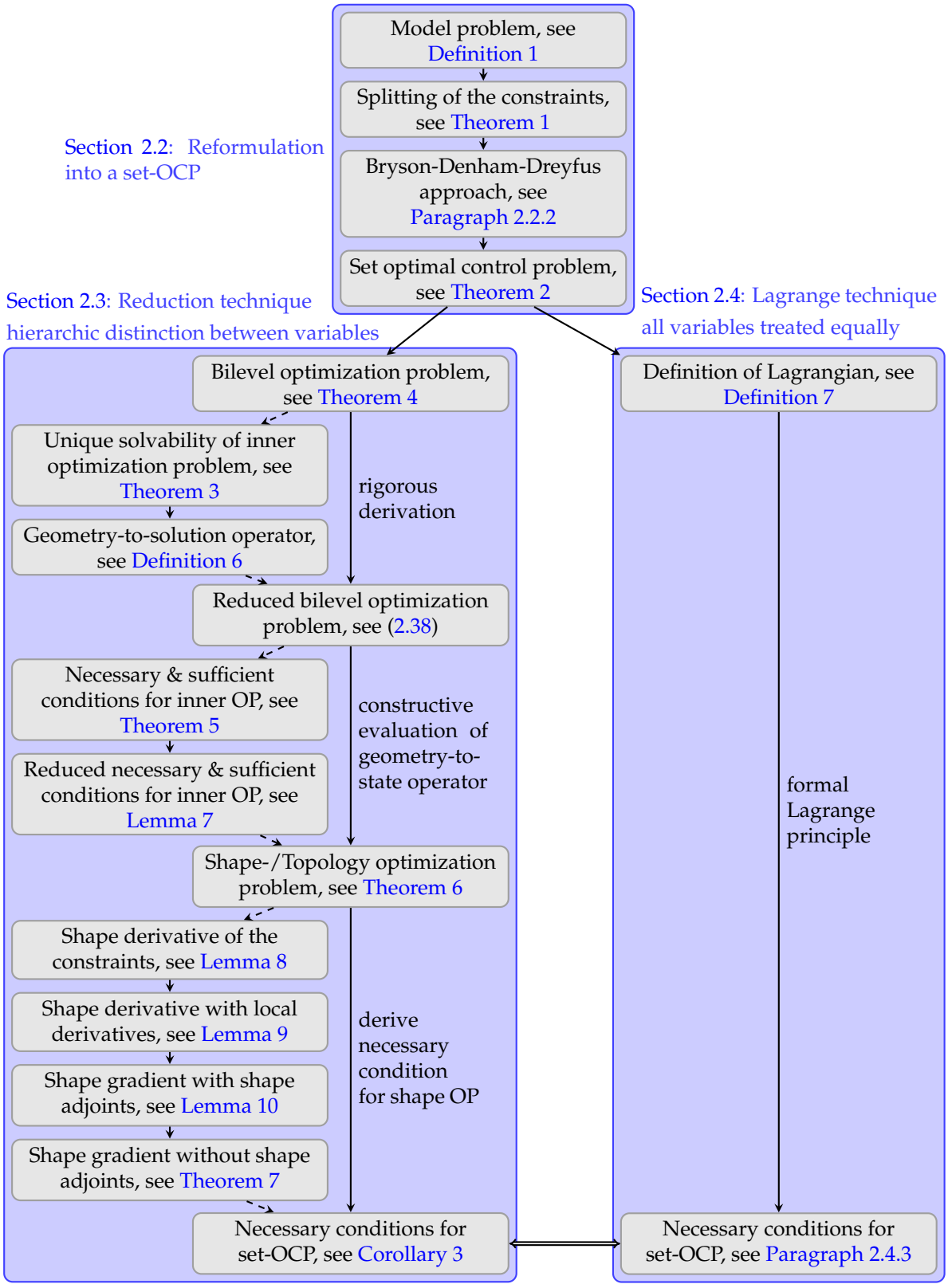
Proposition 4 states an equivalent reformulation of the state equation. In order to prove it, one requires some assertions on Sobolev spaces which are given by Definition 2, lemmas 1, 2 and an abstract version of *Green's formula*, which connects boundary value problems and their *variational formulation*; see Lemma 3.

Definition 2 (Trace operators)**:**
For $m, N \in \mathbb{N}$ let $G \subset \mathbb{R}^N$ be a bounded domain of class $C^{m-1,1}$ with boundary $\Gamma := \partial G$. Let $1 < p < \infty$ be given. Then:

1. The *trace operator*

$$\tau_G = \tau_G^1 : W^{m,p}(G) \to W^{m-\frac{1}{p},p}(\Gamma)$$

is defined as the extension of the trace operator for continuous functions.


Section 2.2: Reformulation into a set-OCP
Model problem, see Definition 1
Splitting of the constraints, see Theorem 1
Bryson-Denham-Dreyfus approach, see Paragraph 2.2.2
Set optimal control problem, see Theorem 2
Section 2.3: Reduction technique
hierarchic distinction between variables
Section 2.4: Lagrange technique
all variables treated equally
Bilevel optimization problem, see Theorem 4
Unique solvability of inner optimization problem, see Theorem 3
Geometry-to-solution operator, see Definition 6
rigorous derivation
Reduced bilevel optimization problem, see (2.38)
Necessary & sufficient conditions for inner OP, see Theorem 5
Reduced necessary & sufficient conditions for inner OP, see Lemma 7
constructive evaluation of geometry-to-state operator
Shape-/Topology optimization problem, see Theorem 6
Shape derivative of the constraints, see Lemma 8
Shape derivative with local derivatives, see Lemma 9
Shape gradient with shape adjoints, see Lemma 10
Shape gradient without shape adjoints, see Theorem 7
derive necessary condition for shape OP
Necessary conditions for set-OCP, see Corollary 3
Definition of Lagrangian, see Definition 7
formal Lagrange principle
Necessary conditions for set-OCP, see Paragraph 2.4.3

2. Additionally, let $\boldsymbol{n} = (n_1, \dots, n_N)^\top$ be a $C^{m-2,1}$-regular extension of the outer unit normal vector field of G, if $m > 1$, and let $\boldsymbol{n}$ be L^∞-regular, if $m = 1$ respectively. Then

$$\tau_G^m : W^{m,p}(G) \to \bigtimes_{i=0}^{m-1} W^{m-i-\frac{1}{p},p}(\Gamma)$$

$$f \mapsto \tau_G^m(f) := \left(\tau_G(f), \tau_G(\mathrm{D}(f)\,\boldsymbol{n}), \dots, \tau_G\Big(\sum_{i_1 \dots i_{m-1}=1}^{N} \mathrm{D}^{i_1 \dots i_m}(f) n_{i_1} \dots n_{i_{m-1}} \Big) \right)$$

is called the *trace operator of m-th order*.

3. To shorten the notation also define

$$\begin{aligned} f|_\Gamma &:= \tau_G(f) && \textit{Dirichlet trace (operator)} \text{ or } \textit{(Dirichlet-)trace} \\ \partial_{\boldsymbol{n}} f &:= \tau_G(\mathrm{D}(f)\,\boldsymbol{n}) && \textit{Neumann trace (operator)} \text{ or } \textit{normal derivative} \\ \partial_{\boldsymbol{nn}} f &:= \tau_G(\boldsymbol{n}^\top \mathrm{D}^2(f)\,\boldsymbol{n}) && \textit{binormal trace (operator)} \text{ or } \textit{binormal derivative}. \end{aligned}$$

Remark:
All components of Definition 2 are well-defined due to [69, p. 37, Thm. 1.5.1.2] and [151, Chp. 2.1].

Later on, the just defined trace operators are often applied to inner boundaries (*interfaces*) subdividing a set in two disjoint parts. In this context, it is important to distinguish between the trace operators acting on the same interface but related to either of the separated sets. In particular, this is relevant to the Neumann trace, since it uses the *outer* unit normal vector field $\boldsymbol{n}$. In this situation the notation

$$\partial_{\boldsymbol{n}}^G f := \tau_G(\mathrm{D}(f)\,\boldsymbol{n}_G)$$

is used to indicate that the outer unit normal vector field $\boldsymbol{n}_G$ of the set G is applied. Such kind of a notation is not necessary for the binormal derivative, due to the fact that the possible wrong choice of the unit normal vector field, i. e. the wrong sign, is compensated since it is used quadratic.

Lemma 1 (Properties of the trace operator)**:**
For $m, N \in \mathbb{N}$ let $G \subset \mathbb{R}^N$ be a bounded domain of class $C^{m-1,1}$ with boundary $\Gamma := \partial G$. Let $1 < p < \infty$ be given. Then the trace operator of m-th order

$$\tau_G^m : W^{m,p}(G) \to \bigtimes_{i=0}^{m-1} W^{m-i-\frac{1}{p},p}(\Gamma)$$

given by Definition 2 is *linear*, *continuous*, *onto* and possesses a continuous right inverse. This is an *extension operator* ω_G^m, such that

$$\tau_G^m \circ \omega_G^m = \mathrm{Id}_{\times_{i=0}^{m-1} W^{m-i-\frac{1}{p},p}(\Gamma)}.$$

A proof for Lemma 1 can by found in [69, Thm. 1.5.1.2].

The trace operators take up a central position in two respects: On the one hand they form the glue for Sobolev spaces on split domains (cf. Lemma 2) and on the other hand they are essential for the analysis of PDEs and boundary value problems (cf. Lemma 3).

Lemma 2 (Weak continuity in $W^{m,p}$)**:**
Let $m, N \in \mathbb{N}$, $1 < p < \infty$, $B \subset \mathbb{R}^N$ be a bounded domain, and let $G \subset\subset B$ be a compactly contained domain of class $C^{m-1,1}$ with complement $G^c := B \setminus \overline{G}$. Furthermore, let τ_G^m, $\tau_{G^c}^m$ denote the trace operators of m-th order, which were introduced in Definition 2, and let the map $f : B \to \mathbb{R}$ fulfill $f|_G \in W^{m,p}(G)$, $f|_{G^c} \in W^{m,p}(G^c)$. Then there holds

$$f \in W^{m,p}(B) \iff \tau_G^m(f|_G) = \tau_{G^c}^m(f|_{G^c})\,.$$

Proof. **1)** This preliminary part provides a localization of the boundary Γ such that it can be described as a graph in a local coordinate system; additionally, tangential and normal vectors in the local basis are given. The results are based on [4, p. 256–266] and will be used in the third part of the proof.
Since $G \subset \mathbb{R}^N$ is a bounded $C^{m-1,1}$-domain, there exist $r \in \mathbb{N}$ and $U_1, \dots, U_r \subset \mathbb{R}^N$ such that $U_1, \dots, U_r$ is an open cover of $\Gamma := \partial G$, and such that $\Gamma_q := \Gamma \cap U_q$ $(q = 1 \dots r)$ possesses a representation as a graph of a $C^{m-1,1}$-regular function, and such that G is above the graph locally. In particular, there exist domains $D^q \subset \mathbb{R}^{N-1}$, numbers $a^q > 0$, local coordinate systems $(e^{q,1}, \dots, e^{q,N})$ of $\mathbb{R}^N$ and $C^{m-1,1}$-functions $g^q : D^q \to \mathbb{R}$, such that for

$$\psi^q \; : \; D^q \times]-a^q; a^q[\to \mathbb{R}^N \,, \; \psi^q(y,h) := (y, g^q(y)+h) := \sum_{i=1}^{N-1} y^i e^{q,i} + (g^q(y)+h) e^{q,N}$$

there holds

$$\begin{aligned}
&\text{(i)} && \psi^q(D^q \times]-a^q; a^q[) = U_q \,, \\
&\text{(ii)} && \psi^q(D^q \times]0; a^q[) = U_q \cap G \,, \\
&\text{(iii)} && \psi^q(D^q \times 0) = \Gamma_q \,, \\
&\text{(iv)} && \psi^q \in C^{m-1,1}(D^q \times]-a^q; a^q[, \mathbb{R}^N) \,, \\
&\text{(v)} && \nabla \psi^q \in C^{m-2,1}(D^q \times]-a^q; a^q[, \mathbb{R}^{N \times N}) \,.
\end{aligned}$$

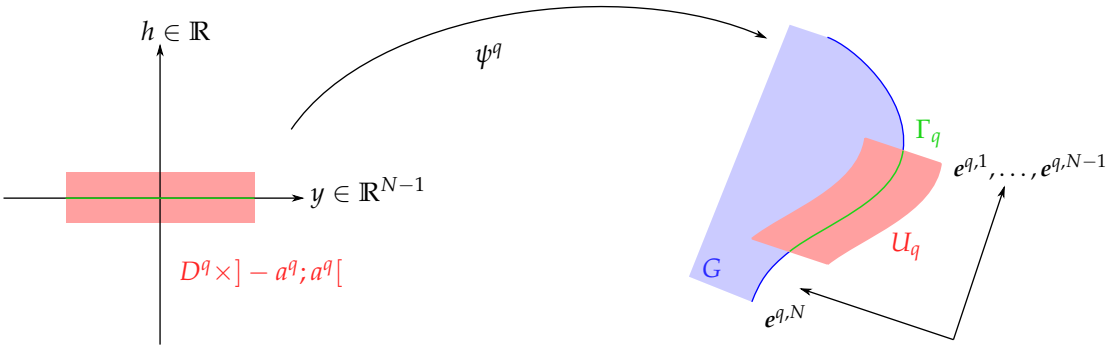


Figure 2.1: Illustration of ψ^q.

In addition, there exist bounded sets $U_0, U_{r+1} \subset B$, such that

$$\begin{aligned}
&U_0, \dots, U_r \text{ is an open cover of } G, \text{ where } U_0 \subset\subset G \text{ and} \\
&U_0, \dots, U_{r+1} \text{ is an open cover of } B.
\end{aligned}$$

Furthermore, there exists a *partition of unity* subordinated to the open cover of B, i. e.

$$\exists \Phi_q \in C_0^\infty(U_q),\, q = 0 \dots r+1 \text{ with } \begin{cases} \forall x_q \in U_q \; : \; \Phi_q(x_q) \in [0;1] \\ \forall x \in B \; : \; \sum_{q=0\dots r+1} \Phi_q(x) = 1 \,. \end{cases} \tag{2.6}$$

For all $q = 0 \dots r+1$ define the localizations of f

$$f_q := \Phi_q f \in W^{m,p}(U_q) \,.$$

Since ψ^q is a $C^{m-1,1}$ transformation, $f_q \circ \psi^q$ is measurable in particular. Fubini's theorem then yields the existence of zero sets $N^q \subset \mathbb{R}$, such that the functions

$$y \mapsto f_q \circ \psi^q(y,h) = \Phi_q f(y, g^q(y)+h)$$

are measurable and integrable for all $h \in]-a^q; a^q[\backslash N^q$. Then, the local part $\tau^1_{G,q}$ of the (Dirichlet-)trace operator τ^1_G on the set U_q can be defined as

$$f_q \mapsto \lim_{h \searrow 0} f_q \circ \psi^q(.,h) =: \tau^1_{G,q}(f_q).$$

As the last preliminary result define the tangential- and normal vectors

$$\boldsymbol{t}^{q,i}(x) := \partial^i \psi^q(y,h) = \boldsymbol{e}^{q,i} + \partial^i g^q(y)\boldsymbol{e}^{q,N}\,, \quad \forall i = 1\dots N-1 \tag{2.7}$$

$$\boldsymbol{n}^q(x) := \left(1+|\nabla g^q(y)|^2\right)^{-\frac{1}{2}} \left(\sum_{i=1}^{N-1} \partial^i g^q(y)\boldsymbol{e}^{q,i} - \boldsymbol{e}^{q,N}\right). \tag{2.8}$$

The local definitions of the trace operator and of the tangential and normal vectors are not dependent on the specific choice of the open cover and the local coordinate systems (cf. [4, p. 256–266]). Consequently, it is sufficient to prove the assertion of Lemma 1 locally and use the finite partition of unity for the globalizing step. Thus, the localization index q will be omitted in third the part of the proof, where the results of the present part are used.

2) This part is devoted to prove the if implication of the assertion.
Let $f \in W^{m,p}(B)$ be given. Since $W^{m,p}(B)\cap C^\infty(B)$ is dense in $W^{m,p}(B)$ (cf. [4, Satz 2.23]), there is a sequence $(f_n)_{n\in\mathbb{N}} \subset W^{m,p}(B)\cap C^\infty(B)$ with $f_n \to f$ in $W^{m,p}(B)$. Continuity of f_n yields $\tau_G^m(f_n|_G) = \tau_{G^c}^m(f_n|_{G^c})$ for all $n\in\mathbb{N}$. Furthermore, the continuity of the trace operators yields

$$\tau_G^m(f|_G) = \lim_{n\to\infty} \tau_G^m(f_n|_G) = \lim_{n\to\infty} \tau_{G^c}^m(f_n|_{G^c}) = \tau_{G^c}^m(f|_{G^c})\,,$$

where the limit is take in $\times_{i=0}^{m-1} W^{m-i-\frac{1}{p},p}(\Gamma)$.

3) This part is devoted to prove the only-if implication and uses mathematical induction with respect to $m\in\mathbb{N}$.
Let $f|_G$ and $f|_{G^c}$ be $W^{m,p}$-regular and let $\tau_G^m(f|_G) = \tau_{G^c}^m(f|_{G^c})$. Then $f\in L^p(B)$ and it remains to show that the composition of the partial derivatives of $f|_G$ and $f|_{G^c}$

$$\partial_{i_1}\dots\partial_{i_m} f(x) := \begin{cases} \partial_{i_1}\dots\partial_{i_m}(f|_G)(x)\,, & x\in G\,,\\ \partial_{i_1}\dots\partial_{i_m}(f|_{G^c})(x)\,, & x\in G^c\,, \end{cases} \quad i_1\dots i_m \in 1,\dots,N$$

defines partial derivatives of f. In the following, let $\phi\in C_0^\infty(B)$ be given arbitrarily.
$m=1$: Then there holds

$$\begin{aligned} -\int_B f\partial_i\phi &= -\int_G f|_G\partial_i\phi - \int_{G^c} f|_{G^c}\partial_i\phi \\ &= \int_G \partial_i f|_G\,\phi + \int_{G^c}\partial_i f|_{G^c}\,\phi - \int_\Gamma \underbrace{\left[\tau_G^1(f|_G) - \tau_{G^c}^1(f|_{G^c})\right]}_{=0} n_i\,\phi\,\mathrm{d}\sigma \\ &= \int_B \partial_i f\,\phi, \quad \forall i = 1\dots N\,. \end{aligned}$$

$m=2$: According to the case "$m=1$" $f\in W^{1,p}(B)$. Therefore, it holds for all $i,j=1\dots N$

$$\begin{aligned} (-1)^2\int_B f\partial_i\partial_j\,\phi &= -\int_B \partial_i f\partial_j\,\phi \\ &= \int_G \partial_j\partial_i f|_G\,\phi + \int_{G^c}\partial_j\partial_i f|_{G^c}\,\phi - \int_\Gamma \left[\tau_G^1(\partial_i f|_G) - \tau_{G^c}^1(\partial_i f|_{G^c})\right] n_j\,\phi\,\mathrm{d}\sigma\,, \end{aligned} \tag{2.9}$$

and it remains to show that $\tau_G^1(\partial_i f|_G) - \tau_{G^c}^1(\partial_i f|_{G^c}) = 0$. The basic idea to do so, is to express the partial derivatives in terms of tangential vectors $\boldsymbol{t}^k$ ($k=1\dots N-1$) and the outer unit normal vector $\boldsymbol{n}$, which were defined in (2.7) and (2.8). Consequently, define the coefficients ζ_j^s, which describe the transformation from the canonical basis $(\boldsymbol{e}_1,\dots,\boldsymbol{e}_N)$ to $(\boldsymbol{t}^1,\dots,\boldsymbol{t}^{N-1},\boldsymbol{n})$:

$$\boldsymbol{e}_j = \sum_{s=1}^{N-1}\zeta_j^s\boldsymbol{t}^s + \zeta_j^N\boldsymbol{n}\,.$$

This yields

$$\begin{aligned} \tau_G^1(\partial_i f|_G) &= \tau_G^1\left(\mathrm{D}(f|_G)\Big(\sum_{s=1}^{N-1}\zeta_i^s\boldsymbol{t}^s + \zeta_i^N\boldsymbol{n}\Big)\right) \\ &= \sum_{s=1}^{N-1}\zeta_i^s\underbrace{\tau_G^1\big(\mathrm{D}(f|_G)\boldsymbol{t}^s\big)}_{(1)} + \zeta_i^N\underbrace{\tau_G^1\big(\mathrm{D}(f|_G)\boldsymbol{n}\big)}_{(2)}\,. \end{aligned} \tag{2.10}$$

The term (2) equals to $\tau^2_G(f|_G)$ and consequently only the term (1) requires further investigation.

The partial derivative with respect to the s-th local basis vector $\boldsymbol{e}^s$ ($s = 1 \dots N-1$) is then given by

$$\begin{aligned}\partial^s f|_G(x) &:= \mathrm{D}(f|_G(x))\boldsymbol{e}^s \\ &= \mathrm{D}(f|_G \circ \psi(y,z))\boldsymbol{e}^s \\ &= (\mathrm{D}f|_G) \circ \psi(y,z)\, \partial^s \psi(y,s) \\ &= (\mathrm{D}f|_G)(x)\boldsymbol{t}^s(x)\,,\end{aligned}$$

which yields

$$\partial^s \tau^1_G(f|_G) = \tau^1_G(\mathrm{D}f|_G \boldsymbol{t}^s)\,. \tag{2.11}$$

The same arguments are valid for $\tau^1_{G^c}(\partial_i f|_{G^c})$ and consequently equation (2.9) can be reformulated by means of equations (2.10) and (2.11)

$$\begin{aligned}(-1)^2 \int_B f\partial_i\partial_j\phi &= \int_G \partial_j\partial_i f|_G\,\phi + \int_{G^c} \partial_j\partial_i f|_{G^c}\,\phi - \int_\Gamma \left[\tau^1_G(\partial_i f|_G) - \tau^1_{G^c}(\partial_i f|_{G^c})\right] n_j\,\phi\,\mathrm{d}\sigma \\ &= \int_B \partial_j\partial_i f\,\phi - \int_\Gamma \sum_{s=1}^{N-1} \zeta^s_i \underbrace{\partial^s \left[\tau^1_G(f|_G) - \tau^1_{G^c}(f|_{G^c})\right]}_{=0} n_j\,\phi\,\mathrm{d}\sigma \\ &\quad - \int_\Gamma \zeta^N_i \underbrace{\left[\tau^1_G(f|_G\boldsymbol{n}) - \tau^1_{G^c}(f|_{G^c}\boldsymbol{n})\right]}_{=\tau^2_G(f|_G) - \tau^2_{G^c}(f|_{G^c}) = 0} n_j\,\phi\,\mathrm{d}\sigma\,, \quad \forall i,j = 1\dots N\,.\end{aligned}$$

That is to say, $f \in W^{2,p}(B)$.
$m-1 \to m$: Assume that the only-if implication is valid for $m-1$, i. e. $f \in W^{m-1,p}(B)$. Consequently, for all $i_1, \dots, i_m = 1 \dots N$

$$(-1)^m \int_B f\partial_{i_1} \dots \partial_{i_m}\phi = \int_B \partial_{i_m} \dots \partial_{i_1} f\,\phi - \int_\Gamma \left[\tau^1_G(\partial_{i_m} \dots \partial_{i_1} f|_G) - \tau^1_{G^c}(\partial_{i_m} \dots \partial_{i_1} f|_{G^c})\right] n_{i_m}\,\phi\,\mathrm{d}\sigma\,. \tag{2.12}$$

By using the same arguments as in the "$m = 2$"-step, one obtains an expression in terms of $\boldsymbol{t}^s$ and $\boldsymbol{n}$:

$$\tau^1_G(\partial_{i_m} \dots \partial_{i_1} f|_G) = \tau^1_G\left(\underbrace{\mathrm{D}\dots\mathrm{D}}_{m \text{ times}}(f|_G) \prod_{l=1}^m \left(\sum_{s=1}^{N-1} \zeta^s_{i_l}\boldsymbol{t}^s + \zeta^N_{i_l}\boldsymbol{n}\right)\right).$$

For convenience define the abbreviations

$$\boldsymbol{T}^1_l := \sum_{s=1}^{N-1} \zeta^s_{i_l}\boldsymbol{t}^s\,, \qquad \boldsymbol{T}^2_l := \zeta^N_{i_l}\boldsymbol{n}\,,$$

and the product becomes

$$\prod_{l=1}^m (\boldsymbol{T}^1_l + \boldsymbol{T}^2_l) = \prod_{l=1}^m \boldsymbol{T}^1_l + \prod_{l=1}^m \boldsymbol{T}^2_l + \sum_{\substack{\alpha \in \{1,2\}^m \\ \alpha \notin \{(1,\dots,1),(2,\dots,2)\}}} \prod_{l=1}^m \boldsymbol{T}^{\alpha_l}_l\,.$$

Herein, the first summand contains tangential vectors only, the second summand only the normal vector, and the third one contains both tangential and normal vectors. This notice helps to structure the boundary integral in (2.12):

$$\begin{aligned}&\int_\Gamma \left[\tau^1_G(\partial_{i_m} \dots \partial_{i_1} f|_G) - \tau^1_{G^c}(\partial_{i_m} \dots \partial_{i_1} f|_{G^c})\right] n_{i_m}\phi\,\mathrm{d}\sigma \\ = &\int_\Gamma \left[\tau^1_G\left(\mathrm{D}\dots\mathrm{D}(f|_G)\prod_{l=1}^m \boldsymbol{T}^1_l\right) - \tau^1_{G^c}\left(\mathrm{D}\dots\mathrm{D}(f|_{G^c})\prod_{l=1}^m \boldsymbol{T}^1_l\right)\right]\mathrm{d}\sigma \\ &+ \int_\Gamma \left[\tau^1_G\left(\mathrm{D}\dots\mathrm{D}(f|_G)\prod_{l=1}^m \boldsymbol{T}^2_l\right) - \tau^1_{G^c}\left(\mathrm{D}\dots\mathrm{D}(f|_{G^c})\prod_{l=1}^m \boldsymbol{T}^2_l\right)\right]\mathrm{d}\sigma \\ &+ \int_\Gamma \sum_{\substack{\alpha \in \{1,2\}^m \\ \alpha \notin \{(1,\dots,1),(2,\dots,2)\}}} \left[\tau^1_G\left(\mathrm{D}\dots\mathrm{D}(f|_G)\prod_{l=1}^m \boldsymbol{T}^{\alpha_l}_l\right) - \tau^1_{G^c}\left(\mathrm{D}\dots\mathrm{D}(f|_{G^c})\prod_{l=1}^m \boldsymbol{T}^{\alpha_l}_l\right)\right]\mathrm{d}\sigma\end{aligned} \tag{2.13}$$

The first part can be treated like term (1) in equation (2.10). Again, denote the partial derivative with respect to the s-th local basis vector $\boldsymbol{e}^s$ with ∂^s ($s = 1 \dots N-1$), then one obtains

$$\begin{aligned}
\partial^{s_1} \dots \partial^{s_m} f|_G(x) &= \partial^{s_1} \dots \partial^{s_{m-1}} \Big((\mathrm{D}f|_G) \circ \psi(y,z) \big(\boldsymbol{e}^{s_m} + \partial_{s_m} g(y) \boldsymbol{e}^N\big) \Big) \\
&= \partial^{s_1} \dots \partial^{s_{m-2}} \Big((\mathrm{DD}f|_G) \circ \psi(y,z) \big(\boldsymbol{e}^{s_{m-1}} + \partial_{s_{m-1}} g(y) \boldsymbol{e}^N\big) \big(\boldsymbol{e}^{s_m} + \partial_{s_m} g(y) \boldsymbol{e}^N\big) \\
&\qquad + (\mathrm{D}f|_G) \circ \psi(y,z)\ \partial_{s_{m-1}} \partial_{s_m} g(y) \boldsymbol{e}^N \Big) \\
&\ \ \vdots \\
&= (\underbrace{\mathrm{D} \dots \mathrm{D}}_{m \text{ times}} f|_G)(x) \boldsymbol{t}^{s_1}(x) \dots \boldsymbol{t}^{s_m}(x) + \text{ additional terms}\,,
\end{aligned}$$

where the additional terms contain derivatives of $f|_G$ up to order $m-1$. This yields

$$\begin{aligned}
\tau_G^1 \left(\mathrm{D} \dots \mathrm{D}(f|_G) \prod_{l=1}^m \boldsymbol{T}_l^1 \right) &= \sum_{\alpha \in \{1 \dots N-1\}^m} \left(\prod_{l=1}^m \zeta_{i_l}^{\alpha_l} \right) \tau_G^1 \left(\mathrm{D} \dots \mathrm{D}(f|_G) \prod_{l=1}^m \boldsymbol{t}^{\alpha_l} \right) \\
&= \sum_{\alpha \in \{1 \dots N-1\}^m} \left(\prod_{l=1}^m \zeta_{i_l}^{\alpha_l} \right) \Big(\partial^{\alpha_1} \dots \partial^{\alpha_m} \tau_G^1 (f|_G) - \tau_G^1(\text{additional terms}) \Big)
\end{aligned} \tag{2.14}$$

Consequently, the first summand of equation (2.13) vanishes, since

- $\tau_G^1(f|_G) = \tau_{G^c}^1(f|_{G^c}) \quad \Longrightarrow \quad \partial^{\alpha_1} \dots \partial^{\alpha_m} \tau_G^1 (f|_G) = \partial^{\alpha_1} \dots \partial^{\alpha_m} \tau_{G^c}^1 (f|_{G^c})$,
- $f \in W^{m-1,p}(B)$ and the additional terms only contain derivatives up to order $m-1$ $\Longrightarrow \tau_G^1(\text{additional terms}) = \tau_{G^c}^1(\text{additional terms})$.

The second summand in equation (2.13) refers to the m-th component of τ_G^m

$$\tau_G^1 \left(\mathrm{D} \dots \mathrm{D}(f|_G) \prod_{l=1}^m \boldsymbol{T}_l^2 \right) = \left(\prod_{l=1}^m \zeta_{i_l}^N \right) \tau_G^1 \left(\mathrm{D} \dots \mathrm{D}(f|_G) \prod_{l=1}^m \boldsymbol{n} \right)$$

and consequently vanishes, too. The third summand in equation (2.13) basically consists of terms of the following type

$$\tau_G^1 \Big(\overbrace{\mathrm{D} \dots \mathrm{D}}^{k \text{ times}} \underbrace{\mathrm{D} \dots \mathrm{D}(f|_G) \prod \boldsymbol{n}}_{=:F|_G} \prod \boldsymbol{t}^{\alpha_l} \Big) - \tau_{G^c}^1 \Big(\mathrm{D} \dots \underbrace{\mathrm{D} \dots \mathrm{D}(f|_{G^c}) \prod \boldsymbol{n}}_{=:F|_{G^c}} \prod \boldsymbol{t}^{\alpha_l} \Big)\,,$$

where $F|_G$ and $F|_{G^c}$ are $W^{k,p}$-regular for a suitable $k \in \{2, \dots, m-1\}$ depending on the number of $\boldsymbol{n}$-factors in the considered term. Observing that

$$\begin{aligned}
\partial^s \Big(\mathrm{D}f|_G(x) \boldsymbol{n}(x) \Big) &= \mathrm{D} \Big((\mathrm{D}f|_G) \circ \psi(y,z)\, \boldsymbol{n}(y,z) \Big) \boldsymbol{e}^s \\
&= (\mathrm{DD}f|_G) \circ \psi(y,z) \mathrm{D}\psi(y,z) \boldsymbol{e}^s\, \boldsymbol{n}(y,z) + (\mathrm{D}f|_G) \circ \psi(y,z)\ (\mathrm{D}\boldsymbol{n}(y,z)) \boldsymbol{e}^s \\
&= (\mathrm{DD}f|_G)(x) \boldsymbol{t}^s(x) \boldsymbol{n}(x) + \text{ additional terms}\,,
\end{aligned}$$

where the additional terms contain derivatives of $f|_G$ up to order 1. This yields

$$\tau_G^1 \left(\mathrm{D} \dots \mathrm{D}(F|_G) \prod \boldsymbol{t}^{\alpha_l} \right) = \partial^{\alpha_1} \dots \partial_{\alpha_k} \tau_G^1(F|_G) - \tau_G^1(\text{additional terms})\,.$$

Consequently, one recognizes the structure of (2.14) in this type of terms. But since the number of differential operators D applied to F is $m-k < m$, it was already shown in inductive step $m-k-1 \to m-k$, that these type of terms vanish. Therefore, the third summand of equation (2.13) vanishes, which completes the proof of the inductive step and the whole proof.

□

Remark:
Lemma 2 provides sharp interface conditions which guarantee that a piecewise defined $W^{m,p}$-function globally exhibits the same regularity, and, vice versa, that a Sobolev function exhibits "weak continuity"

across sufficiently smooth interfaces. That is to say,

$$\int_{\partial G} \partial_{i_m} \dots \partial_{i_1} (f|_G)\, \phi \, \mathrm{d}\sigma = \int_{\partial G} \partial_{i_m} \dots \partial_{i_1} (f|_{G^c})\, \phi \, \mathrm{d}\sigma, \quad \forall \phi \in C_0^\infty(B), \quad i_1 \dots i_m \in \{1, \dots, N\}.$$

As already mentioned in the introducing text above Lemma 2 the second important property of trace operators is their application in the analysis of boundary value problems. The connection between boundary value problems and their corresponding *variational formulations* is based on *Green's formulae*, often called *integration by parts*.

Lemma 3 (Abstract Green's formula)**:**
Let V, H, T be Hilbert spaces, $\tau : V \to T$ be linear and continuous, and $a : V \times V \to \mathbb{R}$ be bilinear and continuous with the so called *trace properties*:

(i) τ maps V *onto* T (trace operator),
(ii) V is contained in H with a stronger topology,
(iii) $V_0 := \mathrm{kernel}(\tau)$ is *dense* in H.

H is referred to as the *pivot space* to V, since (ii) and (iii) imply the *Gelfand triples*

$$V_0 \subset H = H^* \subset V_0^*,$$
$$V \subset H = H^* \subset V^*.$$

Let $\Lambda : V \to V_0^*$ be the formal operator associated with the bilinear form a, i. e.

$$\langle \Lambda v \, , \, w \rangle_{V_0^*, V_0} = a(v, w), \quad \forall v \in V, \, w \in V_0 \, .$$

In addition, define the *domain Hilbert space*

$$V(\Lambda) := \{ v \in V \,|\, \Lambda v \in H \}, \text{ equipped with the norm } \|v\|_{V(\Lambda)} := (\|v\|_H^2 + \|\Delta v\|_H^2)^{\frac{1}{2}}. \tag{2.15}$$

Then there holds:
There exists a unique linear continuous operator $\delta : V(\Lambda) \to T^*$, such that the Green's formula holds

$$a(u, v) = (\Lambda u, v)_H + \langle \delta u \, , \, \tau v \rangle_{T^*, T} \quad \forall u \in V(\Lambda), v \in V, \tag{2.16}$$

where $(.,.)_H$ represents the *inner product* in H and $\langle . \, , \, . \rangle_{X^*, X}$ is the *duality pairing* of a Banach space X and its *dual (space)*.

This lemma and its proof can be found in [6, Thm. 6.2-1]. Additional information about *maximal domains of elliptic operators* (a closely related topic) can be found in [69, Sec. 1.5.3].

Remark:
A classical setting for the Green's formula of Lemma 3 is the following:

$$V = H^1(\Omega), \quad H = L^2(\Omega), \quad T = H^{\frac{1}{2}}(\Gamma), \quad a(u, v) := \int_\Omega \nabla u \cdot \nabla v + u \, v, \quad \tau = \tau_\Omega, \quad V_0 = H_0^1(\Omega),$$

where τ_Ω is the Dirichlet trace operator from Definition 2, and where Ω is of class $C^{1,1}$; see also [69, Rem. 1.5.3.5]. The formal operator associated with the bilinear form is

$$\Lambda = -\Delta + \mathrm{Id}.$$

It is well-known that (2.16) here is

$$\int_\Omega \nabla u \cdot \nabla v + u \, v = \int_\Omega -\Delta u \, v + u \, v + \langle \partial_{\boldsymbol{n}} u \, , \, \tau_\Omega v \rangle_{H^{-\frac{1}{2}}(\Gamma), H^{\frac{1}{2}}(\Gamma)}.$$

In other words, $\delta = \partial_{\boldsymbol{n}}$ is the Neumann trace operator. The idea how to prove this result is as follows: The Green's formula for (strongly) differentiable functions comes with the normal derivative $\partial_{\boldsymbol{n}}$. Since the operator δ is unique and the (strongly) differentiable functions are also weakly differentiable, δ is an extension of the classical normal derivative. In addition, it is compatible with the definition of the Neumann trace operator (cf. Definition 2) and therefore denoted by the same symbol.

From the perspective of functional analysis, the basis for a first step for the reformulation of the model problem (2.1) is provided now. Thus, the notion of *active* and *inactive set* is introduced and some requirements on their regularity are stated. Afterwards, an equivalent split reformulation of the state equation is presented by Proposition 4.

Definition 3 (Active set)**:**
The subsets of Ω in which the optimal state $\bar{y}$ hits the state constraints are called the *upper* and *lower active set*

$$\mathcal{A}_{\max} := \{x \in \overline{\Omega} \,|\, \bar{y}(x) = y_{\max}\}, \tag{2.17a}$$
$$\mathcal{A}_{\min} := \{x \in \overline{\Omega} \,|\, \bar{y}(x) = y_{\min}\}. \tag{2.17b}$$

Their boundaries are denoted by

$$\gamma_{\max} := \partial\mathcal{A}_{\max}, \tag{2.17c}$$
$$\gamma_{\min} := \partial\mathcal{A}_{\min}, \tag{2.17d}$$

and are called *upper* and *lower interface*. Their union and complement

$$\begin{aligned} \mathcal{A} &:= \mathcal{A}_{\max} \cup \mathcal{A}_{\min} \\ \mathcal{I} &:= \Omega \setminus \mathcal{A} \\ \gamma &:= \gamma_{\max} \cup \gamma_{\min} \end{aligned} \tag{2.17e}$$

are referred to as *(optimal) active set*, *(optimal) inactive set* and *(optimal) interface*.

Remark:
The active sets are closed due to $\bar{y} \in C^0(\overline{\Omega})$ and $y_{\max}, y_{\min} \in H^4(\Omega) \hookrightarrow C^0(\overline{\Omega})$ in $\mathbb{R}^2$, since they are the zero level set of $\bar{y} - y_{\max}$ and $y_{\min} - \bar{y}$, respectively.

In order to apply Lemma 2 and some subsequent results, there are some – unfortunately restrictive – assumptions to be made.

Assumption 1 (Regularity of the active sets)**:**
There is an $l \in \mathbb{N}$, such that the active set $\mathcal{A}$ fulfills

$$\mathcal{A} = \bigcup_{i=1}^{l} \mathcal{A}_i\,, \quad \overline{\mathring{\mathcal{A}}_i} = \mathcal{A}_i\,, \quad \mathcal{A} \cap \Gamma = \emptyset\,, \quad \mathcal{A}_i \cap \mathcal{A}_j = \emptyset\,,\ i \neq j\,,\ i,j \in \{1,\dots,l\}\,,$$

$\mathcal{A}_i$ has a $C^{1,1}$-boundary for each i.

At this, $\mathring{\mathcal{B}}$ denotes the interior of a set $\mathcal{B} \subset \mathbb{R}^2$ and $\overline{\mathcal{B}}$ its closure. Moreover, it is assumed that $\mathcal{A} \neq \emptyset$.

The geometrical consequences of Assumption 1 are illustrated in Figure 2.2.

Remark:
The assumptions on regularity of the active set are mainly due to technical reasons and require some explication.

- The active set is supposed to be non-empty to ensure a non-redundant formulation of the original model problem (2.1); otherwise the whole approach of this thesis is not possible and unnecessary. Hence, this assumption is natural and poses no true restriction of the general case.
- The assumption, that the active set shall be equal to the closure of its interior has two main implications.
  - Any lower dimensional connection component is forbidden. This is very restrictive, since it is known that the active set may consist of such kind of sets, such as isolated points and regular curves. To the best of the author's knowledge, there is no appropriate method, which

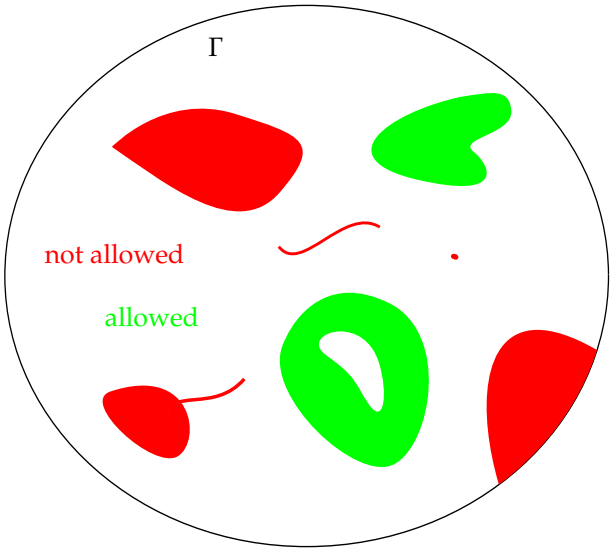


Figure 2.2: Illustration of allowed active sets.

is similar to the approach of this thesis, how to deal with such kind of sets. This is basically due to two different reasons. For one thing the derivation of a control law in the active set has to be adapted when the set has no interior. And for another thing – and this is much more fundamental – one has to apply a different kind of shape calculus, which can cope with sets of codimension greater than zero.

  - Sets with lower dimensional appendices are forbidden, too. This specific assumption does not seem to be very restrictive. It might be possible to prove that such kind of sets cannot occur in principle. However, this topic is beyond the scope of this thesis.

- The $C^{1,1}$-regularity of the boundaries enables a widespread application of shape calculus, which would not be possible with Lipschitzian boundaries. In this respect, confer the counterexamples of Adams, Aronszajn and Smith and of Murat and Simon which both are presented in [44, Chp. 2 Ex. 5.1, 5.2]. Moreover, the regularity ensures higher regularity of different entities on the boundaries (e. g. traces of distributed functions) and of extensions of such traces to the bulk of the domain.
- The active set shall consist of a finite number of connection components, which helps to avoid pathological situations. Moreover, this assumption ensures that the inactive set is of class $C^{1,1}$ as well. Otherwise, if the active set had infinitely many connection components, the inactive set would not be lying locally on one side of its boundary anymore. Hence, standard theory of elliptic boundary value problems can be applied.
- There are three major simplifications due to the fact that the active set may not intersect the outer boundary Γ.
  - Starting and endpoint of those parts of the boundary of the active set, which are subsets of Ω, would cause extra terms in shape calculus, see [151, Sec. 3.8].
  - If starting and endpoints of the boundary part in Ω have to be respected, theory of function spaces gets more involved, since for instance $H^{-1/2}(\gamma)$ is no longer the dual space of $H^{1/2}(\gamma)$, see [69, p. 57] and [117, Chp. 1 Thm. 11.7 and Rem. 12.1]. This type of problem occurs as well, when finite element discretization is used and the boundaries are approximated by polygons. Nevertheless, they are neglected in the numerical implementation and tests of the thesis (see Chapter 4).
  - If there is no intersection with the outer boundary Γ, the compactness of Γ yields that each connection component of the active set has a positive distance to it. Hence, there are no restrictions to variations of the active set, which considerably simplifies the analysis. Consequently, the active set turns out to be a critical shape of the reduced function $\mathcal{F}$ (Theorem 8) and there is no need for restriction to something similar like a "cone of admissible directions".

Since later on it will be referred to the assumptions frequently, it is useful to define the family of subsets of Ω which fulfill Assumption 1 and to fix some corresponding notation.

Definition 4 (Family of feasible sets)**:**
The family of feasible (active) sets is given by

$$\mathcal{O} := \{\mathcal{B} \subset \Omega \,|\, \mathcal{B} \text{ fulfills Assumption 1}\} \cup \emptyset.$$

Definition 5:
Let $\mathcal{B} \in \mathcal{O}$, where $\mathcal{O}$ is given by Definition 4. Then define the following symbols

$$\begin{aligned}
\mathcal{J} &:= \Omega \setminus \mathcal{B},\\
\beta &:= \partial\mathcal{B},\\
\boldsymbol{n}_{\mathcal{B}} &:= \text{ outer unit normal vector field of } \mathcal{B},\\
\boldsymbol{n}_{\mathcal{J}} &:= \text{ outer unit normal vector field of } \mathcal{J} \text{ restricted to } \beta,\\
\partial_{\boldsymbol{n}}^{\mathcal{B}}(.) &:= \tau_{\mathcal{B}}(\mathrm{D}(.)\boldsymbol{n}_{\mathcal{B}}),\\
\partial_{\boldsymbol{n}}^{\mathcal{J}}(.) &:= \tau_{\mathcal{J}}(\mathrm{D}(.)\boldsymbol{n}_{\mathcal{J}}).
\end{aligned}$$

Having the notation at hand, it is possible to introduce a split version of the state equation.

Proposition 4 (geometrical splitting of an elliptic boundary value problem[3])**:**
Let $\mathcal{B} \in \mathcal{O}$, where $\mathcal{O}$ is given by Definition 4 and use the notations of Definition 5. Furthermore, for $M \in \{\Omega, \mathring{\mathcal{B}}, \mathcal{J}\}$, define the domain Hilbert space $H^1(M,\Delta) := \{v \in H^1(M) \,|\, \Delta v \in L^2(M)\}$ of the operators $\Delta(.)$ and $-\Delta + \mathrm{Id}$, cf. (2.15) in Lemma 3.
Then the boundary value problems (for fixed u)

$$-\Delta y + y = u \quad \text{a. e. in } \Omega\,, \tag{2.18a}$$
$$\partial_{\boldsymbol{n}} y = 0 \quad \text{a. e. on } \Gamma\,, \tag{2.18b}$$
$$y \in H^1(\Omega,\Delta)\,, \tag{2.18c}$$
$$u \in L^2(\Omega) \tag{2.18d}$$

and

$$-\Delta y_{\mathcal{J}} + y_{\mathcal{J}} = u_{\mathcal{J}} \quad \text{a. e. in } \mathcal{J}\,, \tag{2.19a}$$
$$\partial_{\boldsymbol{n}} y_{\mathcal{J}} = 0 \quad \text{a. e. on } \Gamma\,, \tag{2.19b}$$
$$y_{\mathcal{J}}|_{\beta} - y_{\mathcal{B}}|_{\beta} = 0 \quad \text{a. e. on } \beta\,, \tag{2.19c}$$
$$y_{\mathcal{J}} \in H^1(\mathcal{J},\Delta)\,, \tag{2.19d}$$
$$u_{\mathcal{J}} \in L^2(\mathcal{J})\,, \tag{2.19e}$$
$$-\Delta y_{\mathcal{B}} + y_{\mathcal{B}} = u_{\mathcal{B}} \quad \text{a. e. in } \mathring{\mathcal{B}}\,, \tag{2.19f}$$
$$\partial_{\boldsymbol{n}}^{\mathcal{J}} y_{\mathcal{J}} + \partial_{\boldsymbol{n}}^{\mathcal{B}} y_{\mathcal{B}} = 0 \quad \text{a. e. on } \beta\,, \tag{2.19g}$$
$$y_{\mathcal{B}} \in H^1(\mathring{\mathcal{B}},\Delta)\,, \tag{2.19h}$$
$$u_{\mathcal{B}} \in L^2(\mathring{\mathcal{B}})\,, \tag{2.19i}$$

are equivalent in the following sense:
If $u_{\mathcal{J}} = u|_{\mathcal{J}}$ and $u_{\mathcal{B}} = u|_{\mathcal{B}}$, then the unique solutions y of (2.18) and the solutions $y_{\mathcal{B}}$ and $y_{\mathcal{J}}$ of (2.19) are connected by $y_{\mathcal{B}} = y|_{\mathcal{B}}$ and $y_{\mathcal{J}} = y|_{\mathcal{J}}$. In particular, (2.19) is uniquely solvable.

Proof. The proof is based on the idea to show that both (2.18) and (2.19) are equivalent to a variational formulation: Look for y satisfying

$$a_\Omega(y,\varphi) := \int_\Omega \nabla y \cdot \nabla\varphi + y\varphi = \int_\Omega u\varphi =: (u,\varphi)_{L^2(\Omega)}\,, \quad \forall \varphi \in H^1(\Omega) \tag{2.20a}$$
$$y \in H^1(\Omega)\,. \tag{2.20b}$$

[3] This result is similar to the discussion of a transmission problem and domain decomposition methods in [17, §I.4].

The bilinear form $a(y,\varphi)$ is known to be continuous and coercive on $H^1(\Omega)\times H^1(\Omega)$. Consequently, the theorem of Lax and Milgram guarantees existence and uniqueness of a solution y of (2.20).

1) (2.20) implies (2.19), which will be proven in this part.
Due to Lemma 2 the space $H^1(\Omega)$ can be identified with $W := \{(v_{\mathcal{J}}, v_{\mathcal{B}}) \in V \mid v_{\mathcal{J}}|_\beta = v_{\mathcal{B}}|_\beta\}$ and thus (2.20) is equivalent to look for $(y_{\mathcal{J}}, y_{\mathcal{B}}) \in W$ satisfying

$$a_\Omega(y,\varphi) = (u,\varphi)_H, \quad \forall \varphi := (\varphi_{\mathcal{J}}, \varphi_{\mathcal{B}}) \in W, \tag{2.21}$$

where $u = (u|_{\mathcal{J}}, u|_{\mathcal{B}}) \in H := L^2(\mathcal{J})\times L^2(\mathring{\mathcal{B}})$, cf. (2.19e) and (2.19i). In particular, there holds (2.19c), since $y \in H^1(\Omega) = W$. The next step is to apply the abstract Green's formula of Lemma 3. In order to check the assumptions, the following notations will be useful:

$$V := H^1(\mathcal{J})\times H^1(\mathring{\mathcal{B}})$$
$$T := H^{\frac12}(\partial\mathcal{J})\times H^{\frac12}(\partial\mathring{\mathcal{B}}) \cong H^{\frac12}(\Gamma)\times H^{\frac12}(\beta)\times H^{\frac12}(\beta)$$
$$\tau : V \to T, \quad (v_{\mathcal{J}}, v_{\mathcal{B}}) \mapsto (\tau_{\mathcal{J}}(v_{\mathcal{J}}), \tau_{\mathcal{B}}(v_{\mathcal{B}})) \equiv (v_{\mathcal{J}}|_\Gamma, v_{\mathcal{J}}|_\beta, v_{\mathcal{B}}|_\beta)$$
$$a : V\times V \to \mathbb{R}, \quad (v,w) \mapsto a_{\mathcal{J}}(v_{\mathcal{J}}, w_{\mathcal{J}}) + a_{\mathcal{B}}(v_{\mathcal{B}}, w_{\mathcal{B}}) := \int_{\mathcal{J}} \nabla v_{\mathcal{J}}\cdot\nabla w_{\mathcal{J}} + v_{\mathcal{J}} w_{\mathcal{J}} + \int_{\mathring{\mathcal{B}}} \nabla v_{\mathcal{B}}\cdot\nabla w_{\mathcal{B}} + v_{\mathcal{B}} w_{\mathcal{B}}$$
$$V_0 := H_0^1(\mathcal{J})\times H_0^1(\mathring{\mathcal{B}})$$
$$\Lambda = (-\Delta + \mathrm{Id}_{H^1(\mathcal{J})}, -\Delta + \mathrm{Id}_{H^1(\mathring{\mathcal{B}})}) : V \mapsto V_0^* = (H^{-1}(\mathcal{J}), H^{-1}(\mathring{\mathcal{B}})).$$

Then there holds

(i) τ is onto according to Lemma 1

(ii) $V \subset H$ according to the Sobolev embedding theorem and has a stronger topology

(iii) $C_0^\infty(\mathcal{J})\times C_0^\infty(\mathring{\mathcal{B}})$ is dense in H and V_0; consequently $V_0 \subset H$ is dense, too.

Since Λ is the formal operator associated with the continuous bilinear form a, there holds

$$a(y,\varphi) = \langle \Lambda y\ ,\ \varphi\rangle_{V_0{}^*, V_0}, \quad \forall \varphi \in V_0,$$
$$\xRightarrow[(2.21)]{V_0\subset W} \langle \Lambda y\ ,\ \varphi\rangle_{V_0{}^*, V_0} = (u,\varphi)_H, \quad \forall \varphi \in V_0,$$
$$\xRightarrow[\text{dense}]{V_0\subset H} \Lambda y = u \text{ in } H, \text{ i. e. } \Lambda y \in H.$$

Consequently, $y \in V(\Lambda) := \{v \in V \mid \Lambda v \in H\} = H^1(\mathcal{J},\Delta)\times H^1(\mathring{\mathcal{B}},\Delta)$; in other words (2.19a), (2.19d), (2.19f) and (2.19h) are fulfilled and the assumptions of Lemma 3, too. That is to say, there exists a unique operator

$$\delta = (\delta_\Gamma, \delta_\beta^{\mathcal{J}}, \delta_\beta^{\mathcal{B}}) : V(\Lambda) \to T^* \cong H^{-\frac12}(\Gamma)\times H^{-\frac12}(\beta)\times H^{-\frac12}(\beta),$$

and it holds

$$a(y,\varphi) = (\Lambda y,\varphi)_H + \langle \delta y\ ,\ \tau\varphi\rangle_{T^*,T}, \quad \forall \varphi \in V.$$

This equation is also fulfilled if φ only ranges in $W \subset V$ and a comparison with (2.21) yields

$$\begin{aligned}\langle \delta y\ ,\ \tau\varphi\rangle_{T^*,T} = 0, \quad \forall \varphi \in W, \quad \Leftrightarrow \qquad & \langle \delta_\Gamma y_{\mathcal{J}}\ ,\ \varphi_{\mathcal{J}}|_\Gamma\rangle_{H^{-\frac12}(\Gamma),H^{\frac12}(\Gamma)} \\ & + \langle \delta_\beta^{\mathcal{J}} y_{\mathcal{J}}\ ,\ \varphi_{\mathcal{J}}|_\beta\rangle_{H^{-\frac12}(\beta),H^{\frac12}(\beta)} \\ & + \langle \delta_\beta^{\mathcal{B}} y_{\mathcal{B}}\ ,\ \varphi_{\mathcal{B}}|_\beta\rangle_{H^{-\frac12}(\beta),H^{\frac12}(\beta)} = 0, \quad \forall (\varphi_{\mathcal{J}},\varphi_{\mathcal{B}}) \in W.\end{aligned}$$

Since $(\varphi_{\mathcal{J}},\varphi_{\mathcal{B}}) \in W$ one can make use of $\varphi_{\mathcal{J}}|_\beta = \varphi_{\mathcal{B}}|_\beta$ yielding

$$\langle \delta_\Gamma y_{\mathcal{J}}\ ,\ \varphi_{\mathcal{J}}|_\Gamma\rangle_{H^{-\frac12}(\Gamma),H^{\frac12}(\Gamma)} + \left\langle (\delta_\beta^{\mathcal{J}} y_{\mathcal{J}} + \delta_\beta^{\mathcal{B}} y_{\mathcal{B}})\ ,\ \varphi_{\mathcal{B}}|_\beta\right\rangle_{H^{-\frac12}(\beta),H^{\frac12}(\beta)} = 0, \quad \forall (\varphi_{\mathcal{J}},\varphi_{\mathcal{B}}) \in W.$$

Finally, the stepwise variation $\varphi \in H_0^1(\Omega) \subset H^1(\Omega) \cong W$ and $\varphi \in W$ reveals

$$\left\langle (\delta_\beta^{\mathcal{J}} y_{\mathcal{J}} + \delta_\beta^{\mathcal{B}} y_{\mathcal{B}})\ ,\ \varphi|_\beta\right\rangle_{H^{-\frac12}(\beta),H^{\frac12}(\beta)} = 0, \quad \forall \varphi \in H_0^1(\Omega)$$
$$\langle \delta_\Gamma y_{\mathcal{J}}\ ,\ \varphi|_\Gamma\rangle_{H^{-\frac12}(\Gamma),H^{\frac12}(\Gamma)} = 0, \quad \forall \varphi \in W.$$

Since the trace operator $(.)|_\Gamma : W \to H^{1/2}(\Gamma)$ is onto (cf. Lemma 1) and referring to the Remark on page 16

$$\partial_n y_\mathcal{J} = \delta_\Gamma = 0 \text{ in } H^{-\frac{1}{2}}(\Gamma), \text{ i. e. (2.19b)}.$$

The analog property of the trace operator $(.)|_\beta$ yields

$$\partial_n^\mathcal{J} y_\mathcal{J} + \partial_n^\mathcal{B} y_\mathcal{B} = \delta_\beta^\mathcal{J} y_\mathcal{J} + \delta_\beta^\mathcal{B} y_\mathcal{B} = 0 \text{ in } H^{-\frac{1}{2}}(\beta), \text{ i. e. (2.19g)}.$$

Altogether (2.20) implies (2.19).

2) This part is devoted to prove that (2.19) implies (2.20).
Let $\varphi \in H^1(\Omega)$ be arbitrary. Lemma 2 yields $\varphi_\mathcal{J} := \varphi|_\mathcal{J}$ and $\varphi_\mathcal{B} := \varphi|_\mathcal{B}$ are H^1-functions with $\varphi_\mathcal{J}|_\beta = \varphi_\mathcal{B}|_\beta$. Multiplying the PDEs (2.19a) and (2.19f) with $\varphi_\mathcal{J}$ and $\varphi_\mathcal{B}$ respectively, integration, and integration by parts results in

$$\begin{aligned} \int_\mathcal{J} \nabla y_\mathcal{J} \cdot \nabla \varphi_\mathcal{J} + y_\mathcal{J} \varphi_\mathcal{J} - \int_\Gamma \partial_n y_\mathcal{J} \varphi_\mathcal{J} - \int_\beta \partial_n^\mathcal{J} y_\mathcal{J} \varphi_\mathcal{J} &= \int_\mathcal{J} u_\mathcal{J} \varphi_\mathcal{J}, \\ \int_\mathcal{B} \nabla y_\mathcal{B} \cdot \nabla \varphi_\mathcal{B} + y_\mathcal{B} \varphi_\mathcal{B} \qquad\qquad - \int_\beta \partial_n^\mathcal{B} y_\mathcal{B} \varphi_\mathcal{B} &= \int_\mathcal{B} u_\mathcal{B} \varphi_\mathcal{B}. \end{aligned}$$

Addition of these equations, together with the conditions (2.19b), (2.19g) and $\varphi_\mathcal{J}|_\beta = \varphi_\mathcal{B}|_\beta$, yields (2.20).

3) The equivalence of (2.18) and (2.20) can be shown with the same arguments used in parts 1) and 2). □

Now that an equivalent split reformulation of the state equation is provided, the whole optimal control problem (2.1) can be divided. This is a first step towards introducing the active set as a separate and equal variable. But before stating this result, a technical assertion is presented for later use.

Lemma 4:
Let $\mathcal{O}$ be the family of feasible sets and let $\mathcal{B}_{\max}$, $\mathcal{B}_{\min} \in \mathcal{O}$ such that $\mathcal{B} := \mathcal{B}_{\max} \dot\cup \mathcal{B}_{\min} \in \mathcal{O}$.
Then there exists a function $y_{\min}^{\max} \in H^4(\Omega)$ such that

$$\begin{aligned} y_{\min}^{\max}(x) &= \begin{cases} y_{\max}(x), & x \text{ in a neighborhood } B_{\max} \text{ of } \mathcal{B}_{\max}, \\ y_{\min}(x), & x \text{ in a neighborhood } B_{\min} \text{ of } \mathcal{B}_{\min}, \end{cases} \\ \partial_n y_{\min}^{\max} &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{2.22}$$

The sets $B_{\max}$ and $B_{\min}$ are specified in the proof; also cf. Figure 2.3.

Remark:
Note, that the construction of $y_{\min}^{\max}$ depends on the particular choice of $\mathcal{B}$. However, the function can remain unchanged, if the boundary β of $\mathcal{B}$ is only slightly deformed in the sense of shape calculus; see Section 2.6.

The set dependency of $y_{\min}^{\max}$ is typically suppressed in the following, since the context tells to which set $\mathcal{B} \in \mathcal{O}$ the function has been constructed.

Proof. According to the choice $\mathcal{B} \in \mathcal{O}$, Assumption 1 and the assumption that $y_{\min} < y_{\max}$ (see Definition 1) there exists $\delta > 0$, such that

$$\operatorname{dist}(\Gamma, \beta) > \delta, \quad \operatorname{dist}(\beta_{\max}, \beta_{\min}) > \delta,$$

where $\beta_{\min} := \partial\mathcal{B}_{\min}$ and $\beta_{\max} := \partial\mathcal{B}_{\max}$ are in the style of notation (2.17c), (2.17d). Consequently, there exist open sets $B_{\max}$ and $B_{\min}$, which are compactly contained in Ω and which in turn contain the sets $\mathcal{B}_{\max}$ and respectively $\mathcal{B}_{\min}$ such that there holds

$$\operatorname{dist}(B_{\max}, B_{\min}) > \frac{\delta}{3}, \quad \operatorname{dist}(\mathcal{B}_{\max}, \partial B_{\max}) > \frac{\delta}{3}, \quad \operatorname{dist}(\mathcal{B}_{\min}, \partial B_{\min}) > \frac{\delta}{3}.$$

In addition, there is a set $J_{\min}^{\max} \subset \mathcal{J}$ with $\Omega = J_{\min}^{\max} \cup B_{\min} \cup B_{\max}$ and

$$\operatorname{dist}(\partial J_{\min}^{\max}, \beta_{\max}) > \frac{\delta}{6}, \quad \operatorname{dist}(\partial J_{\min}^{\max}, \beta_{\min}) > \frac{\delta}{6}, \quad \operatorname{dist}(\partial J_{\min}^{\max}, \partial B_{\max}) > \frac{\delta}{6}, \quad \operatorname{dist}(\partial J_{\min}^{\max}, \partial B_{\min}) > \frac{\delta}{6}.$$

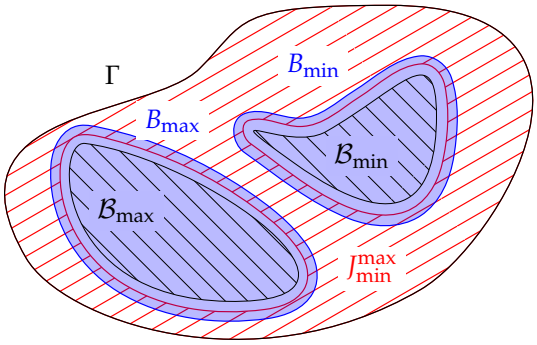


Figure 2.3: Illustration of $B_{\max}$, $B_{\min}$ and $J^{\max}_{\min}$

There are Φ^J, $\Phi^{\max}$ and $\Phi^{\min}$, a partition of unity subordinated to $J^{\max}_{\min}$, $B_{\max}$ and $B_{\min}$, comparable to (2.6). Finally, the function

$$y^{\max}_{\min} := \Phi^{\max} y_{\max} + \Phi^{\min} y_{\min}$$

fulfills all properties claimed, since $y_{\max}, y_{\min} \in H^4(\Omega)$ (cf. Definition 1). □

Theorem 1 (Split reformulation of the model problem)**:**
Let the active and inactive sets be given by Definition 3 and let $y^{\max}_{\min}$ be given by Lemma 4 and be constructed to the active set $\mathcal{A}$.

Then the original model problem (2.1) and the following split reformulation

$$\begin{aligned} \text{minimize} \qquad \mathscr{J}(u_{\mathcal{I}}, u_{\mathcal{A}}, y_{\mathcal{I}}, y_{\mathcal{A}}) := \; & \frac{1}{2}\|y_{\mathcal{I}} - y_d\|^2_{L^2(\mathcal{I})} + \frac{1}{2}\|y_{\mathcal{A}} - y_d\|^2_{L^2(\mathring{\mathcal{A}})} \\ & + \frac{\lambda}{2}\|u_{\mathcal{I}} - u_d\|^2_{L^2(\mathcal{I})} + \frac{\lambda}{2}\|u_{\mathcal{A}} - u_d\|^2_{L^2(\mathring{\mathcal{A}})}, \end{aligned} \tag{2.23a}$$

subject to

$$\mathcal{A}, \mathcal{I}, \text{ and } \gamma \text{ given by Definition 3}, \tag{2.23b}$$

$$u_{\mathcal{I}} \in L^2(\mathcal{I}),\ y_{\mathcal{I}} \in H^1(\mathcal{I}, \Delta), \tag{2.23c}$$

$$u_{\mathcal{A}} \in L^2(\mathring{\mathcal{A}}),\ y_{\mathcal{A}} \in H^1(\mathring{\mathcal{A}}, \Delta), \tag{2.23d}$$

$$y_{\mathcal{A}} = y^{\max}_{\min} \quad \text{in } \mathcal{A}, \tag{2.23e}$$

$$y_{\min} < y_{\mathcal{I}} < y_{\max} \quad \text{in } \mathcal{I}, \tag{2.23f}$$

$$-\Delta y_{\mathcal{I}} + y_{\mathcal{I}} = u_{\mathcal{I}} \quad \text{a. e. in } \mathcal{I}, \tag{2.23g}$$

$$-\Delta y_{\mathcal{A}} + y_{\mathcal{A}} = u_{\mathcal{A}} \quad \text{a. e. in } \mathring{\mathcal{A}}, \tag{2.23h}$$

$$\partial_{\boldsymbol{n}} y_{\mathcal{I}} = 0 \quad \text{a. e. on } \Gamma, \tag{2.23i}$$

$$y_{\mathcal{I}}|_{\gamma} - y_{\mathcal{A}}|_{\gamma} = 0 \quad \text{a. e. on } \gamma, \tag{2.23j}$$

$$\partial^{\mathcal{I}}_{\boldsymbol{n}} y_{\mathcal{I}} + \partial^{\mathcal{A}}_{\boldsymbol{n}} y_{\mathcal{A}} = 0 \quad \text{a. e. on } \gamma, \tag{2.23k}$$

are equivalent in the following sense:

Let $(\bar{u}, \bar{y})$ be the optimal solution of (2.1) and let $(\bar{u}_{\mathcal{I}}, \bar{u}_{\mathcal{A}}, \bar{y}_{\mathcal{I}}, \bar{y}_{\mathcal{A}})$ be the optimal solution of (2.23), then

$$\bar{u}|_{\mathring{\mathcal{A}}} = \bar{u}_{\mathcal{A}}, \qquad \bar{y}|_{\mathring{\mathcal{A}}} = \bar{y}_{\mathcal{A}},$$
$$\bar{u}|_{\mathcal{I}} = \bar{u}_{\mathcal{I}}, \qquad \bar{y}|_{\mathcal{I}} = \bar{y}_{\mathcal{I}}.$$

Proof. Due to Proposition 4 the global state equation (2.1b), (2.1c) is equivalent to its split reformulation (2.23g)–(2.23k).

Let $(\bar{u}, \bar{y}) \in L^2(\Omega) \times H^1(\Omega)$ be the optimal solution of the model problem (2.1). Then the state is an element of $H^2(\Omega)$ actually (see the Remark on page 6); henceforth, it is an element of $H^1(\Omega, \Delta)$ as well. Consequently, $(\bar{u}|_{\mathcal{A}}, \bar{u}|_{\mathcal{I}}, \bar{y}|_{\mathcal{A}}, \bar{y}|_{\mathcal{I}})$ is a solution of the split reformulation of the state equation and fulfills the split state constraint (2.23e), (2.23f). That is to say, the tuple is feasible for the reformulation (2.23).

Vice versa, the concatenation of any admissible tuple of the reformulation is feasible for the original model problem, too. Moreover, such corresponding pairs of assembled and split tuples yield the same values of the objectives (2.1a) and (2.23a).

Thus, $\mathscr{J}$ is bounded from below by the optimal value of J. This lower bound is reached only for $(\bar{u}_{\mathcal{I}}, \bar{u}_{\mathcal{A}}, \bar{y}_{\mathcal{I}}, \bar{y}_{\mathcal{A}}) = (\bar{u}|_{\mathcal{I}}, \bar{u}|_{\mathcal{A}}, \bar{y}|_{\mathcal{I}}, \bar{y}|_{\mathcal{A}})$, since otherwise the original model problem were not uniquely solvable. □

2.2.2 Application of the Bryson-Denham-Dreyfus approach

As already indicated at the beginning of this section a second essential idea is a reformulation of the state constraint in order to reveal conditions which are set to the control variable implicitly. This is the topic for this paragraph. Since this approach is not mandatorily connected to the introduction of the active set as a separate variable, this paragraph can be skipped. However, due to brevity, the subsequent presentation relies on the reformulation of the state constraint developed here.

The approach is based on an idea of Bryson, Denham and Dreyfus [18], who dealt with optimal control problems of ordinary differential equations. Henceforth, it is called *Bryson-Denham-Dreyfus approach* (BDD approach). It is integrated in the wider context of index reduction of (partial) differential-algebraic equations in Section 2.7, since it basically tries to reveal additional information which are hidden in the algebraic constraint modelled by the state constraints.

From the perspective of optimal control the goal is to derive an equivalent form of the algebraic constraint, which reveals its impact on the control more explicitly than the original state constraint. In other words, the state constraint is transformed into a *control law*.

The heuristic procedure in order to reveal the control law is as follows:

1. Differentiate the algebraic equation formed by the state constraint in the active set. Use the differential operator of the state equation for this purpose.
2. Use the state equation to replace a derivative of the state whenever possible.
3. Stop, if the obtained equation reveals an explicit dependency on the control. Otherwise perform the first step.
4. Once, there is an equation obtained for the control – i. e. a control law – use the equations, which were produced by repeatedly performing steps 1 and 2 to replace the state with the derivatives of the constraining function.
5. Caused by (repeated) differentiation within the first step, there is some loss of information. Add a suitable boundary condition to compensate this loss, such that the original equation and its differentiated counterpart are equivalent.

The different steps applied to the split reformulation (2.23) read as follows:

1. Differentiation of the state constraint (2.23e), that is applying the operator $-\Delta$, yields
$$-\Delta y_{\mathcal{A}} = -\Delta y_{\min}^{\max} \quad \text{in } \mathring{\mathcal{A}}. \tag{2.24}$$
2. Comparison with the state equation in the active set (2.23h) exposes the control law
$$u_{\mathcal{A}} - y_{\mathcal{A}} = -\Delta y_{\min}^{\max} \quad \text{in } \mathring{\mathcal{A}}, \tag{2.25}$$
and an anew execution of step 1 is not necessary.
4. Since equation (2.25) still contains the state, use the original state constraint (2.23e) – for it being the zeroth differential order of itself – in order to replace the state
$$u_{\mathcal{A}} - y_{\min}^{\max} = -\Delta y_{\min}^{\max} \quad \text{in } \mathring{\mathcal{A}}. \tag{2.26}$$
5. If one replaces the state constraint (2.23e) by its newly obtained reformulation (2.26) within the constraints of problem (2.23), one suffers from a loss of sharpness of the constraints: Let $y_{\mathcal{I}}$, $y_{\mathcal{A}}$, $u_{\mathcal{I}}$ and $u_{\mathcal{A}}$ be a feasible choice for the constraints of problem (2.23) and suppose that only upper state constraints have to be fulfilled. Then define $z_{\mathcal{A}} \in H^1(\mathring{\mathcal{A}}, \Delta)$ as the solution of
$$\begin{cases} -\Delta z_{\mathcal{A}} + z_{\mathcal{A}} = 0 & \text{in } \mathring{\mathcal{A}}, \\ \partial_{\boldsymbol{n}}^{\mathcal{A}} z_{\mathcal{A}} = -1 & \text{on } \gamma. \end{cases}$$
The weak maximum principle yields $z_{\mathcal{A}} \not\equiv 0$, $z_{\mathcal{A}} \leq 0$. Due to elliptic regularity, $z_{\mathcal{A}}$ is in $H^2(\mathring{\mathcal{A}})$ actually and thus $z_{\mathcal{A}}|_\gamma \in H^{3/2}(\gamma)$, cf. Lemma 1. By means of the extension operator $\omega_{\mathcal{J}}^2$ of Lemma 1 there exists $z_{\mathcal{I}} \in H^2(\mathcal{J})$, such that
$$\begin{aligned} z_{\mathcal{I}}|_\Gamma &= 0 \text{ on } \Gamma, & z_{\mathcal{I}}|_\gamma &= z_{\mathcal{A}}|_\gamma \leq 0 \quad \text{on } \gamma, \\ \partial_{\boldsymbol{n}} z_{\mathcal{I}} &= 0 \text{ on } \Gamma, & \partial_{\boldsymbol{n}}^{\mathcal{I}} z_{\mathcal{I}} &= -\partial_{\boldsymbol{n}}^{\mathcal{A}} z_{\mathcal{A}} = 1 \quad \text{on } \gamma. \end{aligned}$$

Due to the boundary conditions, there is a choice with $z_{\mathcal{I}} \leq 0$ in $\mathcal{I}$. After all, the functions $y_{\mathcal{I}} + z_{\mathcal{I}}$, $y_{\mathcal{A}} + z_{\mathcal{A}}$, $u_{\mathcal{I}} - \Delta z_{\mathcal{I}} + z_{\mathcal{I}}$ and $u_{\mathcal{A}} + 0$ fulfill (2.26) and all of the constraints of problem (2.23), but (2.23e). Consequently, the state constraint (2.23e) cannot be replaced equivalently by the control law (2.26) within problem (2.23).

Therefore, one needs to provide a condition which is compatible with the constraints of (2.23) and which enables an equivalent replacement for the state constraint together with the control law. In view of the differential equation (2.24) a Dirichlet or Robin-type boundary condition would be adequate

$$\begin{aligned} y_{\mathcal{A}}|_\gamma &= y_{\min}^{\max}|_\gamma && \text{on } \gamma, \\ \partial_{\boldsymbol{n}}^{\mathcal{A}} y_{\mathcal{A}} + y_{\mathcal{A}} &= \partial_{\boldsymbol{n}}^{\mathcal{A}} y_{\min}^{\max} + y_{\min}^{\max} && \text{on } \gamma, \end{aligned}$$

but a Neumann boundary condition would not. However, in view of (2.26) a Neumann boundary condition would be suitable as well

$$\left.\begin{aligned} -\Delta y_{\mathcal{A}} + y_{\mathcal{A}} &= u_{\mathcal{A}} && \text{in } \mathring{\mathcal{A}} \\ -\Delta y_{\min}^{\max} + y_{\min}^{\max} &= u_{\mathcal{A}} && \text{in } \mathring{\mathcal{A}} \\ \partial_{\boldsymbol{n}}^{\mathcal{A}} y_{\mathcal{A}} &= \partial_{\boldsymbol{n}}^{\mathcal{A}} y_{\min}^{\max} && \text{on } \gamma \end{aligned}\right\} \quad \Rightarrow \quad y_{\mathcal{A}} = y_{\min}^{\max} \quad \text{in } \mathring{\mathcal{A}}.$$

The BDD approach has been applied to the reformulated model problem (2.23) within the preceding explanation, and it has been revealed that there are different options for the boundary condition. To be more precise there are three different types of reformulations for the state constraint

$$y_{\mathcal{A}} = y_{\min}^{\max} \quad \text{in } \mathcal{A} \quad \Leftrightarrow \quad \left\{\begin{aligned} -\Delta y_{\min}^{\max} + y_{\min}^{\max} &= u_{\mathcal{A}} && \text{in } \mathring{\mathcal{A}}, \\ y_{\min}^{\max}|_\gamma &= y_{\mathcal{A}}|_\gamma && \text{on } \gamma, \end{aligned}\right. \tag{2.27}$$

$$y_{\mathcal{A}} = y_{\min}^{\max} \quad \text{in } \mathcal{A} \quad \Leftrightarrow \quad \left\{\begin{aligned} -\Delta y_{\min}^{\max} + y_{\min}^{\max} &= u_{\mathcal{A}} && \text{in } \mathring{\mathcal{A}}, \\ \partial_{\boldsymbol{n}}^{\mathcal{A}} y_{\min}^{\max} &= \partial_{\boldsymbol{n}}^{\mathcal{A}} y_{\mathcal{A}} && \text{on } \gamma, \end{aligned}\right. \tag{2.28}$$

$$y_{\mathcal{A}} = y_{\min}^{\max} \quad \text{in } \mathcal{A} \quad \Leftrightarrow \quad \left\{\begin{aligned} -\Delta y_{\min}^{\max} + y_{\min}^{\max} &= u_{\mathcal{A}} && \text{in } \mathring{\mathcal{A}}, \\ \partial_{\boldsymbol{n}}^{\mathcal{A}} y_{\min}^{\max} + y_{\min}^{\max} &= \partial_{\boldsymbol{n}}^{\mathcal{A}} y_{\mathcal{A}} + y_{\mathcal{A}} && \text{on } \gamma. \end{aligned}\right. \tag{2.29}$$

Each of these different BDD reformulations of the state-constrained can be used within all of the following considerations. They yield different type of optimality systems; in particular, the interface condition of the adjoint equations differs, see Appendix A. Due to brevity, the remainder of this thesis is focused on the first choice (2.27) only.

Obviously, the above reasoning essentially relies on the regularity of the active set induced by Assumption 1. In particular, it would not be possible to recover the original constraint after differentiation, if the active set had lower dimensional connection components or lower dimensional appendices. Additionally, there are new ideas required, if the control of the considered OCP only acts on the boundary Γ, since differentiation of the state constraint cannot directly reveal a control law. It is beyond the scope of this thesis to extend the approach to such problems, however the concept of a virtual distributed control developed by Krumbiegel and Rösch [110] might be suitable. Moreover, it should be noted, that it is not always expedient to apply the differential operator of the state equation to the constraint, for instance, in the case of gradient constraints. It might be helpful to rewrite the state equation to a system of first order differential equations, but even then there are situations (e. g. constrained L^2-norm of the gradient) where additional ideas are required. Since this thesis is to be understood as a first step to investigate index reduction of partial differential-algebraic equations in the context of optimal control of partial differential equations, theses questions are left unattained.

2.2.3 Resulting set optimal control problem

Until now, the active set $\mathcal{A}$ and the inactive set $\mathcal{I}$ were prescribed by Definition 3. From the point of view of solving the state-constrained model problem via the splitting technique of Paragraph 2.2.1 and the Bryson-Denham-Dreyfus approach of Paragraph 2.2.2 these definitions are unsatisfactory, since they are implicit: without having solved the optimal control problem, one does not know $\mathcal{A}$ and $\mathcal{I}$; without knowledge of these sets, one is not able to state the reformulated problem, which is to be solved.

This tautology can be tackled by regarding the active set as unknown and treating it as optimization variable. Hence, there are three different steps of reformulation:

- geometrical splitting of the constraints,
- application of the BDD approach (derivation of a control law) and
- introduction of the active set as an optimization variable.

Theorem 2 (Reformulation as set optimal control problem)**:**
Let the family of admissible sets $\mathcal{O}$ be given by Definition 4 and use the notations of Definition 5. Then there holds:

The original model problem (2.1) and the *set optimal control problem* (set-OCP)

$$\text{minimize} \qquad \mathfrak{J}(\mathcal{B}; u_{\mathcal{J}}, u_{\mathcal{B}}, y_{\mathcal{J}}, y_{\mathcal{B}}) := \frac{1}{2}\|y_{\mathcal{J}} - y_d\|^2_{L^2(\mathcal{J})} + \frac{1}{2}\|y_{\mathcal{B}} - y_d\|^2_{L^2(\mathring{\mathcal{B}})} + \frac{\lambda}{2}\|u_{\mathcal{J}} - u_d\|^2_{L^2(\mathcal{J})} + \frac{\lambda}{2}\|u_{\mathcal{B}} - u_d\|^2_{L^2(\mathring{\mathcal{B}})} \tag{2.30a}$$

subject to

$$\mathcal{B} \in \mathcal{O}, \tag{2.30b}$$

$$u_{\mathcal{J}} \in L^2(\mathcal{J}),\ y_{\mathcal{J}} \in H^1(\mathcal{J}, \Delta), \tag{2.30c}$$

$$u_{\mathcal{B}} \in L^2(\mathring{\mathcal{B}}),\ y_{\mathcal{B}} \in H^1(\mathring{\mathcal{B}}, \Delta), \tag{2.30d}$$

$$-\Delta y^{\max}_{\min} + y^{\max}_{\min} = u_{\mathcal{B}} \quad \text{in } \mathring{\mathcal{B}}, \tag{2.30e}$$

$$y^{\max}_{\min}|_{\beta} = y_{\mathcal{B}}|_{\beta} \quad \text{on } \beta, \tag{2.30f}$$

$$y_{\min} < y_{\mathcal{J}} < y_{\max} \quad \text{in } \mathcal{J}, \tag{2.30g}$$

$$-\Delta y_{\mathcal{J}} + y_{\mathcal{J}} = u_{\mathcal{J}} \quad \text{a. e. in } \mathcal{J}, \tag{2.30h}$$

$$-\Delta y_{\mathcal{B}} + y_{\mathcal{B}} = u_{\mathcal{B}} \quad \text{a. e. in } \mathring{\mathcal{B}}, \tag{2.30i}$$

$$\partial_{\boldsymbol{n}} y_{\mathcal{J}} = 0 \quad \text{a. e. on } \Gamma, \tag{2.30j}$$

$$y_{\mathcal{J}}|_{\beta} - y_{\mathcal{B}}|_{\beta} = 0 \quad \text{a. e. on } \beta, \tag{2.30k}$$

$$\partial^{\mathcal{J}}_{\boldsymbol{n}} y_{\mathcal{J}} + \partial^{\mathcal{B}}_{\boldsymbol{n}} y_{\mathcal{B}} = 0 \quad \text{a. e. on } \beta, \tag{2.30l}$$

are equivalent in the following sense:

Let $(\bar{u}, \bar{y})$ be the optimal solution of (2.1) and let $(\bar{\mathcal{B}}; \bar{u}_{\mathcal{J}}, \bar{u}_{\mathcal{B}}, \bar{y}_{\mathcal{J}}, \bar{y}_{\mathcal{B}})$ be the optimal solution of (2.30), then

$$\mathcal{A} = \bar{\mathcal{B}}, \qquad \bar{u}|_{\mathcal{A}} = \bar{u}_{\mathcal{B}}, \qquad \bar{y}|_{\mathcal{A}} = \bar{y}_{\mathcal{B}},$$
$$\bar{u}|_{\mathcal{I}} = \bar{u}_{\mathcal{J}}, \qquad \bar{y}|_{\mathcal{I}} = \bar{y}_{\mathcal{J}}.$$

In particular, (2.30) is uniquely solvable.

Proof. It will be shown that the unique optimal solution to the original model problem is optimal for the reformulation, as well.

Thereto, let $(\bar{u}, \bar{y})$ be the optimal solution to (2.1), and let $\mathcal{A}$ be given by Definition 3. By defining $\bar{\mathcal{B}} := \mathcal{A}$, $\mathcal{J} := \Omega \setminus \mathcal{A}$, $\bar{u}_{\mathcal{J}} := \bar{u}|_{\mathcal{J}}$, $\bar{u}_{\mathcal{B}} := \bar{u}|_{\mathcal{B}}$, $\bar{y}_{\mathcal{J}} := \bar{y}|_{\mathcal{J}}$, and $\bar{y}_{\mathcal{B}} := \bar{y}|_{\mathcal{B}}$, one obtains a feasible point for (2.30): The specific choice of $\bar{\mathcal{B}}$ and $\mathcal{J}$ guarantees (2.30b) and (2.30g), Proposition 4 ensures feasibility for (2.30h)–(2.30l), whereas the construction of the BDD ansatz, cf. (2.27), ensures (2.30e) as well as and (2.30f). Now suppose this point is not optimal, in other words there exists an admissible tuple $(\hat{\mathcal{B}}; \hat{u}_{\mathcal{J}}, \hat{u}_{\mathcal{B}}, \hat{y}_{\mathcal{J}}, \hat{y}_{\mathcal{B}})$ with

$$\mathfrak{J}(\hat{\mathcal{B}}; \hat{u}_{\mathcal{J}}, \hat{u}_{\mathcal{B}}, \hat{y}_{\mathcal{J}}, \hat{y}_{\mathcal{B}}) < \mathfrak{J}(\bar{\mathcal{B}}; \bar{u}_{\mathcal{J}}, \bar{u}_{\mathcal{B}}, \bar{y}_{\mathcal{J}}, \bar{y}_{\mathcal{B}}).$$

Regarding the equivalence between the state equation and its split version (see Proposition 4), it can easily be verified that $(\hat{u}, \hat{y})$ defined by

$$\hat{u} = \begin{cases} \hat{u}_{\mathcal{J}} & \text{in } \hat{\mathcal{J}}, \\ \hat{u}_{\mathcal{B}} & \text{in } \hat{\mathcal{B}}, \end{cases} \qquad \text{and} \qquad \hat{y} = \begin{cases} \hat{y}_{\mathcal{J}} & \text{in } \hat{\mathcal{J}}, \\ \hat{y}_{\mathcal{B}} & \text{in } \hat{\mathcal{B}}, \end{cases}$$

is a feasible point for the original problem (2.1). This yields

$$J(\hat{u}, \hat{y}) = \mathfrak{J}(\hat{\mathcal{B}}; \hat{u}_{\mathcal{J}}, \hat{u}_{\mathcal{B}}, \hat{y}_{\mathcal{J}}, \hat{y}_{\mathcal{B}}) < \mathfrak{J}(\bar{\mathcal{B}}; \bar{u}_{\mathcal{J}}, \bar{u}_{\mathcal{B}}, \bar{y}_{\mathcal{J}}, \bar{y}_{\mathcal{B}}) = J(\bar{u}, \bar{y}),$$

in contradiction to optimality of $(\bar{u}, \bar{y})$. Uniqueness of the solution of the reformulation is shown with the same arguments as above together with the uniqueness of the solution of the original model problem. □

Remark:
To the best of the author's knowledge, problems of this type have not been studied in literature so far and the notion of a set optimal control problem is introduced. It can be seen as a generalization of multipoint boundary value problems from the world of ODEs (see Paragraph 2.1.3), and thus the split BVP may also be called *multiset boundary value problem*.

The family $\mathcal{O}$ of all admissible sets $\mathcal{B}$ (cf. Definition 4) does not possess a vector space structure, cf. Section 2.6 and the 13th item on page 68 in particular. Consequently, problem (2.30) does not fit in the usual framework of nonlinear optimization in Banach spaces. It is a hybrid problem which possesses Banach space variables and a set variable, whose topology and shape has to be determined. Hence, it forms an integration of shape/topology optimization on the one hand, and optimal control of PDEs on the other hand.

It becomes apparent in Paragraph 2.6.3 that a set optimal control problem should be regarded as an optimization problem on a vector bundle on $\mathcal{O}$ and that some details of notation should be adapted to this fact. In particular, if the function space variables and the set variable are regarded as equal, it is important to ensure, that the former can be chosen without having specified the latter. However, the chosen notation seems to be natural and hence it is retained for convenience.

2.2.4 Role of the strict inequality constraint

Before proceeding to the derivation of first order necessary optimality conditions in sections 2.3 and 2.4 it is indicated to have a closer look at the set-OCP (2.30). In particular, it is necessary to investigate the strict inequality constraint (2.30g), i. e. $y_{\min} < y_{\mathcal{J}} < y_{\max}$, in more detail. The aim of this paragraph is to rise the awareness, that strict inequality constraints within infinite dimensional optimization are a topic of its own and that there seems to be a fluent passage from non-strict to strict inequality constraints.

Strict inequality constraints are non-standard in nonlinear optimization, since they lead to nonclosed admissible sets typically. Hence, they may yield unsolvable optimization problems, since existence of a minimizer is not guaranteed. Unsolvability is not an issue in the present context, since existence of a minimizer is ensured by Theorem 4. However, there is no standard procedure available how to treat such kind of constraints. Nonetheless, strict inequality constraints naturally emerge in the context of shape optimization and free boundary value problems, where the optimal set sometimes may be characterized by the domain of the positive part of the optimal state; cf. the weak formulation of free boundary problems in [163] and the references therein. It should be interesting to pursue the approach presented there, however this reaches beyond the scope of this thesis. The analysis of strict inequality constraints remains restricted to the considered specific situation.

A nearby idea is to rewrite it as non-strict inequality $y_{\min} \leq y_{\mathcal{J}} \leq y_{\max}$ and regard it as not active. This is a very natural point of view, since this is its original interpretation. At a first glance inactive constraints should have no impact on the derivation of necessary conditions – they do not seem to restrict the set of admissible directions of variation (cf. tangent and linearizing cone, [103, 124]) – because there is a positive distance between the optimal solution and the part of the boundary of the admissible set, which corresponds to the inactive constraint. Unfortunately, this point of view is only true, if there are finitely many inactive constraints present only. In the specific situation the constraint $y_{\min} < y_{\mathcal{J}} < y_{\max}$ can be interpreted as infinitely many pointwise, inactive constraints

$$y_{\min}(x) < y_{\mathcal{J}}(x) < y_{\max}(x), \quad \forall x \in \mathcal{J},$$

since elliptic regularity ensures $y_{\mathcal{J}} \in H^2(\mathcal{J}) \subset C^0(\overline{\mathcal{J}})$. Each of those pointwise constraints is inactive and there is an (individual) positive distance to the active situation

$$\forall x \in \mathcal{J} \quad \exists \delta(x) > 0 : \quad y_{\max} - y_{\mathcal{J}}(x) \geq \delta(x) \quad \text{and} \quad y_{\mathcal{J}}(x) - y_{\min} \geq \delta(x).$$

But on the other hand one has

$$y_{\mathcal{J}}|_{\beta_{\max}} = y_{\max}|_{\beta_{\max}} \quad \text{and} \quad y_{\mathcal{J}}|_{\beta_{\min}} = y_{\min}|_{\beta_{\min}},$$

which yields that there is no such distance $\delta > 0$ that works for all $x \in \mathcal{J}$ simultaneously. Consequently, the constraint as a whole should be seen as "quasi active" in the sense that the optimum does not lie in the interior of the admissible set.[4]

From this perspective there is little hope to adequately deal with the strict inequality constraint. However, there is some structure which is exploited in the following.

Assumption 2 (Gamma is strictly inactive)**:**
There is a $\delta > 0$ such that the optimal state $\bar{y}_{\mathcal{I}}$ in the (optimal) inactive set $\mathcal{I}$ fulfills

$$y_{\max}|_\Gamma - \bar{y}_{\mathcal{I}}|_\Gamma \geq \delta \quad \text{and} \quad \bar{y}_{\mathcal{I}}|_\Gamma - y_{\min}|_\Gamma \geq \delta.$$

Lemma 5:
Let the assumptions 1 and 2 be fulfilled and let $\bar{y}$ be the composed optimal state of the set optimal control problem of Theorem 2. Then

$$\forall M_{\max} \subset (\mathcal{A}_{\min} \cup \mathcal{I} \cup \Gamma) \text{ with } \operatorname{dist}(M_{\max}, \gamma_{\max}) > 0 \text{ and } M_{\max} \text{ is closed}$$

there holds

$$\exists\, \delta > 0 : \quad y_{\max}|_{M_{\max}} - \bar{y}|_{M_{\max}} \geq \delta.$$

The analogous assertion holds true for the lower constraints $y_{\min}$.

Proof. Assume there is such a set M for which there is no $\delta > 0$ as claimed. Since M is compact and since $\bar{y}$ and $y_{\max}$ are continuous, it follows that there is an $x \in M$ with $\bar{y}(x) = y_{\max}(x)$. Furthermore, $\bar{y}$ respects $y_{\min} < \bar{y}_{\mathcal{I}} < y_{\max}$ in $\mathcal{I}$ and $\bar{y} = y_{\min} < y_{\max}$ in $\mathcal{A}_{\min}$; consequently, one has $x \in M \setminus (\mathcal{I} \cup \mathcal{A}_{\min}) = M \cap \Gamma$. This is a contradiction to Assumption 2. □

The assertion of Lemma 5 is illustrated in Figure 2.4.

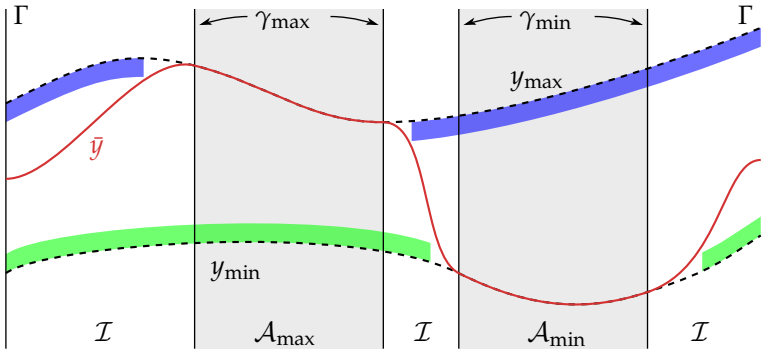


Figure 2.4: Illustration of the sets $M_{\max}$ and $M_{\min}$.

Lemma 6 (Second reformulation as set optimal control problem)**:**
Let the family of admissible sets $\mathcal{O}$ be given by Definition 4, use the notations of Definition 5, and denote the two components of the "candidate active set" $\mathcal{B}$ corresponding to the upper and the lower state constraint with $\mathcal{B}_{\max}$ and $\mathcal{B}_{\min}$, in the style of $\mathcal{A}_{\max}$ and $\mathcal{A}_{\min}$. Then there holds:

The original model problem (2.1) and the set optimal control problem (where in comparison to (2.30) the interface condition (2.30f) is replaced by the weaker conditions (2.31g) and (2.31h))

$$\begin{aligned} \text{minimize} \qquad \mathfrak{J}(\mathcal{B}; u_{\mathcal{J}}, u_{\mathcal{B}}, y_{\mathcal{J}}, y_{\mathcal{B}}) := \; & \frac{1}{2}\|y_{\mathcal{J}} - y_d\|^2_{L^2(\mathcal{J})} + \frac{1}{2}\|y_{\mathcal{B}} - y_d\|^2_{L^2(\mathring{\mathcal{B}})} \\ & + \frac{\lambda}{2}\|u_{\mathcal{J}} - u_d\|^2_{L^2(\mathcal{J})} + \frac{\lambda}{2}\|u_{\mathcal{B}} - u_d\|^2_{L^2(\mathring{\mathcal{B}})} \end{aligned} \tag{2.31a}$$

[4] Note, that the cones defined by $\bar{y} \leq y_{\max}$ and $y_{\min} \leq \bar{y}$ indeed have interior points, since the space $C^0(\overline{\Omega})$ is used here.

subject to

$$\mathcal{B} \in \mathcal{O}, \tag{2.31b}$$
$$u_{\mathcal{J}} \in L^2(\mathcal{J}),\ y_{\mathcal{J}} \in H^1(\mathcal{J},\Delta), \tag{2.31c}$$
$$u_{\mathcal{B}} \in L^2(\mathring{\mathcal{B}}),\ y_{\mathcal{B}} \in H^1(\mathring{\mathcal{B}},\Delta), \tag{2.31d}$$
$$-\Delta y_{\min}^{\max} + y_{\min}^{\max} = u_{\mathcal{B}} \quad \text{in } \mathring{\mathcal{B}}, \tag{2.31e}$$
$$y_{\min} < y_{\mathcal{J}} < y_{\max} \quad \text{in } \mathcal{J}, \tag{2.31f}$$
$$y_{\mathcal{B}}|_{\mathcal{B}_{\min}} < y_{\max}|_{\mathcal{B}_{\min}} \text{ in } \mathcal{B}_{\min}, \tag{2.31g}$$
$$y_{\min}|_{\mathcal{B}_{\max}} < y_{\mathcal{B}}|_{\mathcal{B}_{\max}} \text{ in } \mathcal{B}_{\max}, \tag{2.31h}$$
$$-\Delta y_{\mathcal{J}} + y_{\mathcal{J}} = u_{\mathcal{J}} \quad \text{in } \mathcal{J}, \tag{2.31i}$$
$$-\Delta y_{\mathcal{B}} + y_{\mathcal{B}} = u_{\mathcal{B}} \quad \text{in } \mathring{\mathcal{B}}, \tag{2.31j}$$
$$\partial_{\boldsymbol{n}} y_{\mathcal{J}} = 0 \quad \text{on } \Gamma, \tag{2.31k}$$
$$y_{\mathcal{J}}|_{\beta} - y_{\mathcal{B}}|_{\beta} = 0 \quad \text{on } \beta, \tag{2.31l}$$
$$\partial_{\boldsymbol{n}}^{\mathcal{J}} y_{\mathcal{J}} + \partial_{\boldsymbol{n}}^{\mathcal{B}} y_{\mathcal{B}} = 0 \quad \text{on } \beta, \tag{2.31m}$$

are equivalent in the following sense:

Let $(\bar{u},\bar{y})$ be the optimal solution of (2.1) and let $(\bar{\mathcal{B}};\bar{u}_{\mathcal{J}},\bar{u}_{\mathcal{B}},\bar{y}_{\mathcal{J}},\bar{y}_{\mathcal{B}})$ be the optimal solution of (2.31), then

$$\mathcal{A} = \bar{\mathcal{B}}, \qquad \bar{u}|_{\mathring{\mathcal{A}}} = \bar{u}_{\mathcal{B}}, \qquad \bar{y}|_{\mathring{\mathcal{A}}} = \bar{y}_{\mathcal{B}},$$
$$\bar{u}|_{\mathcal{I}} = \bar{u}_{\mathcal{J}}, \qquad \bar{y}|_{\mathcal{I}} = \bar{y}_{\mathcal{J}}.$$

In particular (2.31) is uniquely solvable.

Proof. **1)** According to Proposition 4, the minimizer $(\bar{u},\bar{y}) \in L^2(\Omega)\times H^2(\Omega)$ of the original model problem is feasible for the state equations of the reformulation, if one chooses $\mathcal{B} := \mathcal{A}$ and $\mathcal{J} := \mathcal{I}$. Additionally, the BDD reformulation of the state constraint (2.31e) is fulfilled. Consequently, $(\bar{u}|_{\mathcal{J}},\bar{u}|_{\mathcal{B}},\bar{y}|_{\mathcal{J}},\bar{y}|_{\mathcal{B}})$ is a feasible point for the reformulation and

$$\min \mathfrak{J}(\mathring{\mathcal{B}};u_{\mathcal{J}},u_{\mathcal{B}},y_{\mathcal{J}},y_{\mathcal{B}}) \le \mathfrak{J}(\mathring{\mathcal{A}};\bar{u}|_{\mathcal{I}},\bar{u}|_{\mathcal{A}},\bar{y}|_{\mathcal{I}},\bar{y}|_{\mathcal{A}}) = J(\bar{u},\bar{y}).$$

2) Assume there exists an admissible point $(\hat{\mathcal{B}};\hat{u}_{\mathcal{J}},\hat{u}_{\mathcal{B}},\hat{y}_{\mathcal{J}},\hat{y}_{\mathcal{B}})$ of the set set optimal control problem with

$$\mathfrak{J}(\hat{\mathcal{B}};\hat{u}_{\mathcal{J}},\hat{u}_{\mathcal{B}},\hat{y}_{\mathcal{J}},\hat{y}_{\mathcal{B}}) < \mathfrak{J}(\mathring{\mathcal{A}};\bar{u}|_{\mathcal{I}},\bar{u}|_{\mathcal{A}},\bar{y}|_{\mathcal{I}},\bar{y}|_{\mathcal{A}}).$$

Again using Proposition 4 it can easily be verified that $(\hat{u},\hat{y})$ defined by

$$\hat{u} = \begin{cases} \hat{u}_{\mathcal{J}} & \text{in } \hat{\mathcal{J}}, \\ \hat{u}_{\mathcal{B}} & \text{in } \hat{\mathcal{B}}, \end{cases} \qquad \text{and} \qquad \hat{y} = \begin{cases} \hat{y}_{\mathcal{J}} & \text{in } \hat{\mathcal{J}}, \\ \hat{y}_{\mathcal{B}} & \text{in } \hat{\mathcal{B}}, \end{cases}$$

fulfills the state equation of the original model problem. Suppose that there exists $\overline{x} \in \beta_{\max}$ such that $\hat{y}(\overline{x}) > y_{\max}(\overline{x})$ (note that due to elliptic regularity $\hat{y} \in H^2(\Omega) \subset C^0(\overline{\Omega})$). Continuity of $\hat{y}$ and $y_{\max}$ yields existence of $\delta > 0$ with

$$\forall x \in B_\delta(\overline{x}): \quad \hat{y}(x) > y_{\max}(x).$$

Consequently, $\hat{y}_{\mathcal{J}} > y_{\max}$ in $B_\delta(\overline{x}) \cap \hat{\mathcal{J}} \neq \emptyset$, which contradicts the feasibility of $\hat{y}_{\mathcal{J}}$, cf. (2.31f). That is, $\hat{y} \le y_{\max}$ on $\beta_{\max}$. Furthermore, $z := \hat{y}_{\mathcal{B}} - y_{\max}$ fulfills

$$\begin{aligned} -\Delta z + z &= 0, \quad \text{in } \hat{\mathcal{B}}_{\max}, \\ z &\le 0, \quad \text{on } \beta_{\max}. \end{aligned}$$

The weak maximum principle yields $z \le 0$ in $\hat{\mathcal{B}}_{\max}$, i. e. $\hat{y}_{\mathcal{B}} \le y_{\max}$ everywhere in $\hat{\mathcal{B}}_{\max}$. Consequently, $\hat{y} \le y_{\max}$ in Ω. Analogously one obtains $\hat{y}_{\mathcal{B}} \ge y_{\min}$ everywhere in Ω.

All in all, $(\hat{u}_{\mathcal{J}},\hat{u}_{\mathcal{B}},\hat{y}_{\mathcal{J}},\hat{y}_{\mathcal{B}})$ is admissible for the original model problem, which yields

$$\mathfrak{J}(\mathring{\mathcal{A}};\bar{u}|_{\mathcal{I}},\bar{u}|_{\mathcal{A}},\bar{y}|_{\mathcal{I}},\bar{y}|_{\mathcal{A}}) = J(\bar{u},\bar{y}) = \min J(u,y) \le \mathfrak{J}(\hat{\mathcal{B}};\hat{u}_{\mathcal{J}},\hat{u}_{\mathcal{B}},\hat{y}_{\mathcal{J}},\hat{y}_{\mathcal{B}}),$$

in contradiction to the assumption.

3) Uniqueness of the solution of the reformulation is shown with the same arguments as above together with the uniqueness of the solution of the original model problem. □

Remarks (on the strict inequality constraint and optimality)**:**
In view of the lemmas 5 and 6 the BDD interface condition

$$y_{\mathcal{B}} = y_{\min}^{\max} \quad \text{on } \beta$$

can be regarded as the "active part" of the inequality constraint $y_{\min} < y_{\mathcal{J}} < y_{\max}$ in $\mathcal{J}$ and compensates its influence on admissibly near the optimum. In other words, if the interface condition is used as constraint – this is formulation (2.30) –, one can expect, that *the strict inequality has no local impact on optimality*. That is to say, it does not restrict the cone of feasible directions in which the directional derivative of the objective must vanish at the optimal configuration.

One can expect, in particular, that the first order necessary conditions obtained for the *set optimal control problem without strict inequality* actually are compatible with the strict inequality constraint. Indeed, this expectation holds true as it is shown by Theorem 8. In this respect, see also the Remarks on the strict inequality constraint on page 49.

Nonetheless, the strict inequality constraint may not be abandoned, since it has a *global effect on optimal solutions*. This circumstance is illustrated in Figure 2.5.

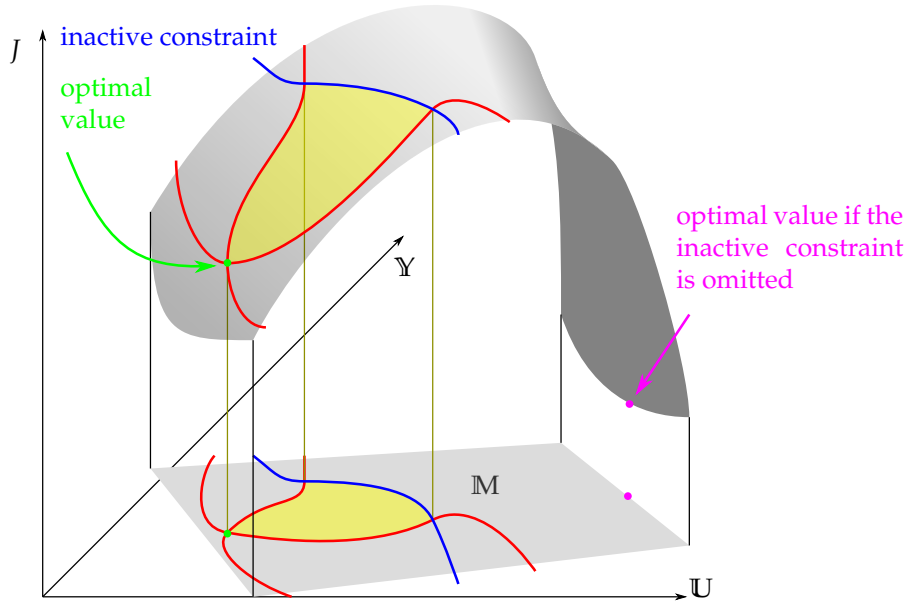


Figure 2.5: Illustration of the global effect of an inactive constraint.

In addition, if it is not abandoned completely, but relaxed to the non-strict case

$$y_{\min} \leq y_{\mathcal{J}} \leq y_{\max}$$

and not treated as mandatorily inactive, unique solvability of the set optimal control problems (2.30) and respectively of (2.31) is lost. The interpretation of $\mathcal{B}$ and $\mathcal{J}$ as active, respectively inactive set does not hold any more and for example the choice $\mathcal{B} = \emptyset$, $\mathcal{J} = \Omega$, $\bar{y}_{\mathcal{J}} := \bar{y}$ and $\bar{u}_{\mathcal{J}} := \bar{u}$ is feasible and yields the optimal value. Hence, arbitrarily small perturbations of (2.30g) (the relaxation from the strict to the non-strict case is the smallest perturbation one might think of) seem to have an impact on solvability of the optimization problem. This finding is illustrated in Figure 2.6.

Remark:
In view of the deeper analysis in Section 2.4, especially from the perspective of the reduced formulation (2.38), where the strict inequality constraint is used as a constraint for the set variable, one is mandatorily faced with this global effect on optimal solutions. Consequently, the constraint is necessary in general, but can be omitted while deriving first order necessary conditions for it has no local impact on optimality.

Unfortunately, the question whether the original unique optimum can be a cluster point of such additional critical points mentioned above is left open here. In view of the reduced bilevel optimization problem (2.38), this would mean, that there are sets $\mathcal{B}_n \in \mathcal{O}$ such that for each $n \in \mathbb{N}$ the relaxed constraint $y_{\min} \leq \bar{y}_{\mathcal{J}} \leq y_{\max}$ is fulfilled and $\mathcal{B}_n$ converges to the optimal active set $\mathcal{A}$ (with respect to a reasonable

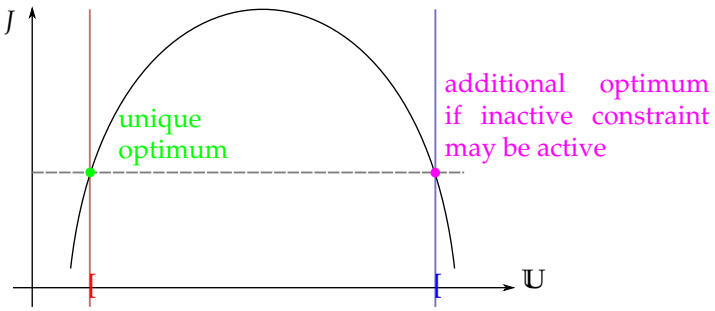


Figure 2.6: Illustration of a "quasi active" constraint.

topology on $\mathcal{O}$), while $\mathcal{F}(\mathcal{B}_n) = \mathcal{F}(\mathcal{A})$. In other words, it is not clear, whether the original unique optimum is an interior point of the set $\{\mathcal{B} \in \mathcal{O} \,|\, y_{\min} \leq \bar{y}_{\mathcal{J}} \leq y_{\max}\}$ with respect to a reasonable topology on $\mathcal{O}$. Therefore, the role of the strict inequality constraint is closely related to the question whether the optimal active set $\mathcal{A}$ is an isolated critical point of the shape functional $\mathcal{F}$; cf. Paragraph 2.5.2.

2.3 First order analysis via reduction technique

After having found an equivalent reformulation of the original model problem (2.1), in which the active set arises as separate and equal variable, this section and Section 2.4 are attended to derive first order necessary conditions. Whereas an informal approach via the Lagrange principle is used in Section 2.4, a more rigorous reasoning by means of the formulation of a bilevel optimization problem is applied here. The latter procedure requires a sequence of steps as illustrated on page 10 and roughly follows the general recipe for deriving first order necessary conditions form Paragraph 2.3.2. In view of this, it is gainful to envision the abstract framework of optimal control (cf. also [19, Chp. 3]).

2.3.1 Abstract framework of optimal control

Let $\mathbb{U}$ and $\mathbb{Y}$ be two sets, whose intrinsic structure is known sufficiently well. These are, for instance, Banach spaces, topological spaces or Riemannian manifolds. In addition, let $J : \mathbb{U} \times \mathbb{Y} \to \mathbb{R}$ be a functional, which has to be minimized. One fundamental ingredient of optimal control is a distinction between the *control* $u \in \mathbb{U}$ and the *state* $y \in \mathbb{Y}$, which is due the so called *control-to-state operator* $S : \mathbb{U} \to \mathbb{Y}$. It is motivated by the common situation in applications, that a controlable input u to a system evokes an unique answer y.[5] Since not every input might be realizable and not every output might be desirable, the minimization takes place on a set $\mathbb{M} \subset \mathbb{U} \times \mathbb{Y}$.

$$\text{minimize } J(u,y) \quad \text{subject to} \quad \begin{cases} (u,y) \in \mathbb{M} \subset \mathbb{U} \times \mathbb{Y} \\ y = S(u) \end{cases}$$

The control-to-state operator plays a decisive role in the analysis of optimal control problems. Due to it, the state can be regarded as dependent on the control, and hence the whole optimization problem can be reduced:

$$\text{minimize} \quad f(u) := J(u, S(u)) \quad \text{subject to} \quad (u, S(u)) \in \mathbb{M}.$$

The derivation of first order necessary conditions is substantially connected with differentiability of the operator S now.

With respect to shape and topology optimization, the control is a domain, whose optimal design has to be found. The state is typically given as solution of a boundary value problem defined on this domain. Hence, the "control-to-state operator" is the evaluation of a BVP for a given domain and the appropriate framework for the analysis is the *shape-* or *topology calculus*, respectively.

In the context of the set optimal control problem (2.30), the abstract control is composed of the function control variables $u_{\mathcal{J}}$ and $u_{\mathcal{B}}$ and of the set variable $\mathcal{B}$. Consequently, the family $\mathcal{O}$ rather defines the

[5] The reasonable possibility that S is a set valued operator and that the optimization singles out one specific answer is disregarded here. Such kind of problems are investigated in [59].

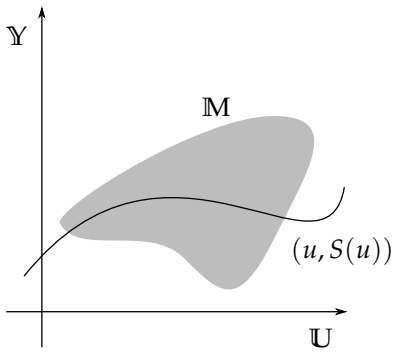


Figure 2.7: The abstract setting of optimal control.

set of admissible sets (comparable to $u_{\mathcal{J}} \in L^2(\mathcal{J})$), than being an explicit constraint (cf. (2.30b)). The corresponding control-to-state operator maps the controls $(u_{\mathcal{J}}, u_{\mathcal{B}}, \mathcal{B})$ to the states $(y_{\mathcal{J}}, y_{\mathcal{B}})$ here.

However, it is possible to focus on another structure of the set-OCP, which is more similar to shape- and topology optimization. Regard the set $\mathcal{B}$ as the control and the minimizing tuple $(\bar{u}_{\mathcal{J}}, \bar{u}_{\mathcal{B}}, \bar{y}_{\mathcal{J}}, \bar{y}_{\mathcal{B}})$ as corresponding state. Thus, the map which assigns $\mathcal{B}$ to the tuple (called geometry-to-solution operator in the following) plays the role of a control-to-state operator then. This point of view will be used for the remainder of this section. In particular, it has to be ensured, that the mentioned operator is well-defined and differentiable. But before proceeding to the analysis of the specific problem under consideration, it might be helpful to illustrate the basic steps in an abstract but compact way.

2.3.2 General recipe for deriving first order necessary conditions

An approach, which is often applied to derive first order necessary conditions, is the analysis via reduction, i. e. via introduction of a control-to-state operator. The derivation of first order necessary conditions is along the lines of the following informal recipe then. It should be noted, that this approach may have a major drawback. In particular, there are optimization problems, which are differentiable, but whose control-to-state operator is not.[6] In such a situation, this approach cannot be applied, but it may be possible to use a Lagrange principle, see Section 2.4 and [100].

1. Identify a control-to-state operator S.
   a) Prove its well-definedness.
   b) Formulate the reduced optimization problem
   $$\text{minimize} \quad f(u) := J(u, S(u)) \quad \text{subject to} \quad (u, S(u)) \in \mathbb{M}.$$
2. Derive the derivative of the reduced objective f:
   a) Prove differentiability of J and S.
   b) Use the chain rule
   $$\big((\mathrm{D}f)(u)\big)(v) = \big((\mathrm{D}J)(u, S(u))\big)(v) = \langle (\partial_u J)(u, S(u)) \,,\, v\rangle + \big\langle (\partial_y J)(u, S(u)) \,,\, (\mathrm{D}S)(u)v\big\rangle.$$
3. Identify the Hadamard form of the derivative of the reduced objective J, or even the gradient in the case of Hilbert spaces:
   a) Introduce adjoint variables.
   b) Transform $\big((\mathrm{D}f)(u)\big)(v)$ into *Hadamard form* $\langle (\mathrm{D}f)(u) \,,\, v\rangle$ (see below).
   c) Identify the gradient $\langle (\mathrm{D}f)(u) \,,\, v\rangle = ((\nabla f)(u) \,,\, v)$.

[6] Regard the following two-dimensional optimization problem due to S. Bechmann. Minimize the objective $J(x_1, x_2) = x_2$ subject to $x_1^2 - x_2^2 = 0$ and $x_2 \leq 0$. The obvious unique solution is $\bar{x} = (0, 0)$. Moreover, there exist Lagrange multipliers, since the tangent cone and the cone of descent directions do not intersect, cf. [61, Prop. 2.2.1]. However, if the equality constraint is eliminated, one obtains the reduced optimization problem: minimize $f(x_1) = |x_1|$, which obviously not differentiable.
In the context of shape optimization there is an example given in [100].

4. Establish first order necessary conditions $\langle (\mathrm{D}f)(\bar{u}) \, , \, v \rangle \geq 0, \quad \forall v \in \mathbb{K}(\bar{u})$.
   At this, the set $\mathbb{K}(\bar{u})$ is a suitable set of allowed directions of variation, which locally approximates the projection of the intersection of $\mathbb{M}$ with the graph of S onto the set $\mathbb{U}$. This can be for instance the tangential cone, or if additional nonlinear equality constraints are present the *derived cone*; see [82, Chp. 4 Sec. 2]. In the case of $\mathbb{U}$ being a manifold, $\mathbb{K}(\bar{u})$ can be a suitable part of the tangent space $T_{\bar{u}}\mathbb{U}$; cf. Definition 8.

In order to get a better understanding of the Hadamard form, the steps 2 and 3 are reviewed in more detail. Thereto, the common situation is used, where the constraint only implicitly defines the control-to-state operator S. However, specific difficulties, such as differentiability can only be obtained for some subsets or with respect to some directions, are not addressed here for simplicity.

$$\text{minimize } J(u,y) \quad \text{subject to} \quad \begin{cases} (u,y) \in \mathbb{M} \subset \mathbb{U} \times \mathbb{Y} \\ T(u,y) = 0 \text{ in } \mathbb{Z} \end{cases} \tag{2.32}$$

Assume J and T are sufficiently smooth and there exists an implicit function (locally around the optimal control $\bar{u}$) $S : \mathbb{U} \to \mathbb{Y}, \; u \mapsto y = S(u)$ with $0 = T(u, S(u))$. This yields

$$\begin{aligned} 0 = (\mathrm{D}T)(u,S(u)) &= (\partial_u T)(u,S(u)) + (\partial_y T)(u,S(u)) \circ (\mathrm{D}S)(u) \\ \Leftrightarrow \quad (\mathrm{D}S)(u) &= -(\partial_y T)^{-1}(u,S(u)) \circ (\partial_u T)(u,S(u)). \end{aligned} \tag{2.33}$$

And consequently, there holds[7] for all $v \in T_{\bar{u}}\mathbb{U}$

$$\Big((\mathrm{D}J)(\bar{u},S(\bar{u}))\Big)(v) = \Big\langle (\partial_u J)(\bar{u},S(\bar{u})) \, , \, v \Big\rangle_{T^*_{\bar{u}}\mathbb{U},T_{\bar{u}}\mathbb{U}} + \Big\langle (\partial_y J)(\bar{u},S(\bar{u})) \, , \, \Big((\mathrm{D}S)(\bar{u})\Big)(v) \Big\rangle_{T^*_{S(\bar{u})}\mathbb{Y},T_{S(\bar{u})}\mathbb{Y}}.$$

This representation of the derivative is not in *Hadamard form*, since it is not obvious, what its representative in $T^*_{\bar{u}}\mathbb{U}$ should be

$$\Big((\mathrm{D}J)(\bar{u},S(\bar{u}))\Big)(v) = \Big\langle (\mathrm{D}J)(\bar{u},S(\bar{u})) \, , \, v \Big\rangle_{T^*_{\bar{u}}\mathbb{U},T_{\bar{u}}\mathbb{U}}.$$

By the use of (2.33) the derivative reads

$$\begin{aligned} \Big((\mathrm{D}J)(\bar{u},S(\bar{u}))\Big)(v) = &\Big\langle (\partial_u J)(\bar{u},S(\bar{u})) \, , \, v \Big\rangle_{T^*_{\bar{u}}\mathbb{U},T_{\bar{u}}\mathbb{U}} \\ &+ \Big\langle (\partial_y J)(\bar{u},S(\bar{u})) \, , \, -(\partial_y T)^{-1}(\bar{u},S(\bar{u})) \circ \Big((\partial_u T)(\bar{u},S(\bar{u}))\Big)(v) \Big\rangle_{T^*_{S(\bar{u})}\mathbb{Y},T_{S(\bar{u})}\mathbb{Y}}. \end{aligned}$$

Adjoining the operator in the second duality pairing and introducing the *adjoint state* $p \in \mathbb{Z}^*$ yields

$$\begin{aligned} \Big((\mathrm{D}J)(\bar{u},S(\bar{u}))\Big)(v) &= \Big\langle (\partial_u J)(\bar{u},S(\bar{u})) \, , \, v \Big\rangle_{T^*_{\bar{u}}\mathbb{U},T_{\bar{u}}\mathbb{U}} \\ &\quad + \Big\langle (\partial_u T)^*(\bar{u},S(\bar{u})) \circ \underbrace{\Big(-(\partial_y T)^{-1}(\bar{u},S(\bar{u}))\Big)^* \Big((\partial_y J)(\bar{u},S(\bar{u}))\Big)}_{=:p} \, , \, v \Big\rangle_{T^*_{\bar{u}}\mathbb{U},T_{\bar{u}}\mathbb{U}} \\ &= \Big\langle (\partial_u J)(\bar{u},S(\bar{u})) + \Big((\partial_u T)^*(\bar{u},S(\bar{u}))\Big)(p) \, , \, v \Big\rangle_{T^*_{\bar{u}}\mathbb{U},T_{\bar{u}}\mathbb{U}}. \end{aligned} \tag{2.34}$$

Consequently, the Hadamard form is obtained. Moreover, if $T_{\bar{u}}\mathbb{U}$ is a Hilbert space one can proceed and identify the ($\mathbb{U}$-)gradient[8] of the reduced objective f

$$\langle \mathrm{D}f(\bar{u}) \, , \, v \rangle_{T^*_{\bar{u}}\mathbb{U},T_{\bar{u}}\mathbb{U}} = (\nabla f(u) \, , \, v)_{T_{\bar{u}}\mathbb{U}}.$$

That is

$$(\nabla f)(\bar{u}) = (\nabla J)(\bar{u},S(\bar{u})) = (\nabla_u J)(\bar{u},S(\bar{u})) + \Big((\nabla_u T)^*(\bar{u},S(\bar{u}))\Big)(p) \in T_{\bar{u}}\mathbb{U}. \tag{2.35}$$

[7] The notions of the tangent and cotangent spaces (cf. Definition 8 and Definition 10) are used, since they are needed if the spaces have no linear structure. If such a structure is present, however, the co/tangent spaces coincide with the space itself and its dual, respectively. Due to the subscript of the tangent space, there should be no confusion with the constraining operator T.

[8] If there is a *Gelfand triple* $T_{\bar{u}}\mathbb{U} \subset H = H^* \subset T^*_{\bar{u}}\mathbb{U}$ with pivot Hilbert space H it is possible to identify an H-gradient, too

$$(\nabla_{\mathbb{U}} f(\bar{u}) \, , \, v)_{T_{\bar{u}}\mathbb{U}} = (\nabla_H f(\bar{u}) \, , \, v)_H.$$

This topic is closely related to *Sobolev gradients*, cf. the Remark to Theorem 7.

Hence, the introduction of adjoint variables can be regarded as typical step in the derivation of the Hadamard form of the derivative of the reduced objective f.

Although the gradient of the reduced objective for the model problem (2.1) is well-known, it may be instructive to transfer the abstract derivation of the gradient to the specific situation.

It is well-known (cf. Proposition 4), that for each $u \in L^2(\Omega)$, there exists a unique $y \in H^1(\Omega)$ with

$$a(y,\phi) := \int_{\mathcal{J}} \nabla y \cdot \nabla \phi + y\,\phi = \int_{\mathcal{J}} u\,\phi =: F(u), \quad \forall \phi \in H^1(\Omega).$$

The variational form can also be written as an operator equation

$$\Lambda(y) = F(u),$$

where the operator $\Lambda : H^1(\Omega) \to H^1(\Omega)^*$ is associated with the bilinear form a. That is to say, the constraint operator is given by

$$T : L^2(\Omega) \times H^1(\Omega) \to H^1(\Omega)^*, \quad (u,y) \mapsto \Lambda(y) - F(u).$$

Moreover, since for every $u \in L^2(\Omega)$ there exists a unique $y \in H^1(\Omega)$ with $T(u,y) = 0$, there is a control-to-state operator

$$S : L^2(\Omega) \to H^1(\Omega), \quad u \mapsto y = (\Lambda^{-1} \circ F)(u).$$

Hence, by means of (2.34) and the objective functional $J(u,y) = \frac{1}{2}\|E(y - y_d)\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\|u - u_d\|^2_{L^2(\Omega)}$ one obtains

$$\begin{aligned} p &= -\Big((\partial_y T)^{-1}(u, S(u))\Big)^* (\partial_y J)(u, S(u)) \\ &= -(\Lambda^{-1})^* E(S(u) - y_d) \in H^1(\Omega)^{**} \cong H^1(\Omega), \end{aligned}$$

where $E : H^1(\Omega) \to L^2(\Omega)$ is the embedding operator. In other words, the adjoint state fulfills the boundary value problem[9]

$$\Lambda^* p = -E(S(u) - y_d) \quad \Longleftrightarrow \quad \begin{cases} -\Delta p + p = -(y - y_d) & \text{in } \Omega, \\ \partial_{\boldsymbol{n}} p = 0 & \text{on } \Gamma. \end{cases}$$

Finally, the gradient of the reduced objective can be identified with

$$(\nabla f)(u) = (\nabla_u J)(u, S(u)) + (\partial_u T)^*(u, S(u))p = \lambda(u - u_d) - F^* p = \lambda(u - u_d) - p.$$

These results are in total agreement with well-known approach of deriving the adjoint state of the state-unconstrained analog of the model problem (cf. [159, Sec. 2.8]), except for the different sign of p, which is due to fact that the adjoint state usually is defined as the negative Lagrange multiplier ("Russian minus").

2.3.3 Reformulation into a bilevel optimization problem

The aim of this paragraph is a first analysis of the set optimal control problem (2.30) with respect to its inherent structure. Since the set optimal control problem shelters the two types of optimization problems – shape/topology optimization on the one hand, and optimal control on the other hand – it is obvious to look for an bilevel formulation, which separates these two parts. Thus, examine the *bilevel optimization problem* (BiOP) (2.36)–(2.37)

$$\text{outer optimization problem (oOP)} \begin{cases} \text{minimize} & \mathfrak{J}(\mathcal{B}; \bar{u}_{\mathcal{J}}, \bar{u}_{\mathcal{B}}, \bar{y}_{\mathcal{J}}, \bar{y}_{\mathcal{B}}), & \text{(2.36a)} \\ \text{subject to} & \begin{cases} \mathcal{B} \in \mathcal{O}, & \text{(2.36b)} \\ y_{\min} < \bar{y}_{\mathcal{J}} < y_{\max} \quad \text{in } \mathcal{J}, & \text{(2.36c)} \end{cases} \end{cases}$$

where $(\bar{u}_{\mathcal{J}}, \bar{u}_{\mathcal{B}}, \bar{y}_{\mathcal{J}}, \bar{y}_{\mathcal{B}})$ is the solution to the inner optimization problem, which is parametrized by $\mathcal{B}$:

[9] Due to symmetry of the bilinear form a, its associated operator Λ is self adjoint.

inner optimization problem (iOP)

$$
\begin{aligned}
&\text{minimize} \quad \mathfrak{J}(\mathcal{B}; u_{\mathcal{J}}, u_{\mathcal{B}}, y_{\mathcal{J}}, y_{\mathcal{B}}) := \frac{1}{2}\|y_{\mathcal{J}} - y_d\|^2_{L^2(\mathcal{J})} + \frac{1}{2}\|y_{\mathcal{B}} - y_d\|^2_{L^2(\mathring{\mathcal{B}})} \\
&\qquad\qquad + \frac{\lambda}{2}\|u_{\mathcal{J}} - u_d\|^2_{L^2(\mathcal{J})} + \frac{\lambda}{2}\|u_{\mathcal{B}} - u_d\|^2_{L^2(\mathring{\mathcal{B}})} \qquad (2.37a)\\
&\text{subject to}\\
&u_{\mathcal{J}} \in L^2(\mathcal{J}),\ y_{\mathcal{J}} \in H^1(\mathcal{J}, \Delta), \quad (2.37b) \qquad -\Delta y_{\mathcal{J}} + y_{\mathcal{J}} = u_{\mathcal{J}} \quad \text{in } \mathcal{J}, \quad (2.37f)\\
&u_{\mathcal{B}} \in L^2(\mathring{\mathcal{B}}),\ y_{\mathcal{B}} \in H^1(\mathring{\mathcal{B}}, \Delta), \quad (2.37c) \qquad -\Delta y_{\mathcal{B}} + y_{\mathcal{B}} = u_{\mathcal{B}} \quad \text{in } \mathring{\mathcal{B}}, \quad (2.37g)\\
&-\Delta y^{\max}_{\min} + y^{\max}_{\min} = u_{\mathcal{B}} \quad \text{in } \mathring{\mathcal{B}}, \quad (2.37d) \qquad \partial_{\boldsymbol{n}} y_{\mathcal{J}} = 0 \quad \text{on } \Gamma, \quad (2.37h)\\
&y^{\max}_{\min}|_{\beta} = y_{\mathcal{B}}|_{\beta} \quad \text{on } \beta, \quad (2.37e) \qquad y_{\mathcal{J}}|_{\beta} - y_{\mathcal{B}}|_{\beta} = 0 \quad \text{on } \beta, \quad (2.37i)\\
&\qquad\qquad \partial^{\mathcal{J}}_{\boldsymbol{n}} y_{\mathcal{J}} + \partial^{\mathcal{B}}_{\boldsymbol{n}} y_{\mathcal{B}} = 0 \quad \text{on } \beta. \quad (2.37j)
\end{aligned}
$$

Provided that the set parametrized inner optimization problem is uniquely solvable for each set $\mathcal{B} \in \mathcal{O}$, this BiOP fits in the framework of Paragraph 2.3.1. The deeper analysis will be the scope of paragraphs 2.3.4–2.3.7.

2.3.4 Geometry-to-solution operator

At first, the most important question concerning the bilevel problem is, if it is well-defined. This question is twofold:

- Is there a minimizer of the inner optimization problem (2.37) for any $\mathcal{B} \in \mathcal{O}$?
- Is the outer optimization problem (2.36) well-defined? In particular, how can the strict inequality constraint (2.36c) be interpreted?

Since the second question has already been discussed in Paragraph 2.2.4, it remains to answer the first one. Indeed, the inner optimization problem is well-defined:

Theorem 3 (Unique solvability of the inner optimization problem)**:**
Let the family of admissible sets $\mathcal{O}$ be given by Definition 4 and use the notations of Definition 5.
Then the (parametrized) optimization problem (2.37) has a unique solution $(\bar{u}_{\mathcal{J}}, \bar{u}_{\mathcal{B}}, \bar{y}_{\mathcal{J}}, \bar{y}_{\mathcal{B}})$ for any $\mathcal{B} \in \mathcal{O}$.

Proof. Let $\Omega = \mathcal{B} \dot\cup \mathcal{J}$ be a splitting with $\mathcal{B} \in \mathcal{O}$. Due to constraints (2.37d), (2.37e), and (2.37g), the variables $u_{\mathcal{B}} = -\Delta y^{\max}_{\min} + y^{\max}_{\min}$ and $y_{\mathcal{B}} = y^{\max}_{\min}$ are fixed on the set $\mathcal{B}$. Consequently, it has to be optimized in the set $\mathcal{J}$ only.

With the solution operator

$$
S : L^2(\mathcal{J}) \to H^1(\mathcal{J}, \Delta), \quad u_{\mathcal{J}} \mapsto y_{\mathcal{J}}, \text{ which is the solution to } \begin{cases} -\Delta y_{\mathcal{J}} + y_{\mathcal{J}} = u_{\mathcal{J}} & \text{a. e. in } \mathcal{J},\\ \partial_{\boldsymbol{n}} y_{\mathcal{J}} = 0 & \text{a. e. on } \Gamma,\\ y_{\mathcal{J}}|_{\beta} = y^{\max}_{\min}|_{\beta} & \text{a. e. on } \beta. \end{cases}
$$

at hand, the inner optimization problem can be reduced to the set $\mathcal{J}$ equivalently:

$$
\begin{aligned}
&\text{minimize} \quad \mathfrak{J}(\mathcal{B}; u_{\mathcal{J}}, -\Delta y^{\max}_{\min} + y^{\max}_{\min}, S(u_{\mathcal{J}}), y^{\max}_{\min})\\
&\text{subject to} \quad u_{\mathcal{J}} \in U := \{u_{\mathcal{J}} \in L^2(\mathcal{J}) \,|\, \partial^{\mathcal{J}}_{\boldsymbol{n}} S(u_{\mathcal{J}}) = -\partial^{\mathcal{B}}_{\boldsymbol{n}} y_{\max} \text{ on } \beta\}.
\end{aligned}
$$

The solution operator S is continuous from $L^2(\mathcal{J})$ to $H^2(\mathcal{J})$ (cf. [69, Thm. 2.3.3.2]), and since $H^1(\mathcal{J}, \Delta) \hookrightarrow H^2(\mathcal{J})$, S is continuous from $L^2(\mathcal{J})$ to $H^1(\mathcal{J}, \Delta)$, too. In addition, $\partial^{\mathcal{J}}_{\boldsymbol{n}}$ is continuous as well (cf. Lemma 1), and the admissible set U is a closed, convex and nonempty subset of $L^2(\mathcal{J})$. The last property is a consequence of the extension operator $\omega^2_{\mathcal{J}}$: let $z \in H^2(\mathcal{J})$ given by $z := \omega^2_{\mathcal{J}}(0, y^{\max}_{\min}|_{\beta}, 0, \partial^{\mathcal{J}}_{\boldsymbol{n}} y^{\max}_{\min})$, then $-\Delta z + z \in U$. Consequently, the reduced problem has a unique optimal solution (cf. [159, Thm. 2.16]). □

Unique solvability of the bilevel optimization problem is an easy consequence of theorems 2 and 3.

Theorem 4 (Reformulation as bilevel optimization problem)**:**
Let the family of admissible sets $\mathcal{O}$ be given by Definition 4 and use the notations of Definition 5.
Then the bilevel optimization problem (2.36), (2.37) and the set optimal control problem (2.30) are equivalent. In particular, the bilevel optimization problem is uniquely solvable.

Proof. Let $(\mathcal{A}, \bar{u}_\mathcal{I}, \bar{u}_\mathcal{A}, \bar{y}_\mathcal{I}, \bar{y}_\mathcal{A})$ be the unique solution of the set optimal control problem according to Theorem 2. In a first step, it will be shown, that this tuple is feasible for the bilevel optimization problem:
Theorem 1 says, that $(\bar{u}_\mathcal{I}, \bar{u}_\mathcal{A}, \bar{y}_\mathcal{I}, \bar{y}_\mathcal{A})$ is the unique solution to (2.23). The state constraint (2.23e) can be replaced by the equivalent BDD ansatz (2.27) and since the constraint $y_{\min} < \bar{y}_\mathcal{I} < y_{\max}$ is not active, (2.23f) can be omitted. By means of these reformulations, (2.23) becomes the inner optimization problem (2.37) for the set $\mathcal{A}$. Consequently, $(\bar{u}_\mathcal{I}, \bar{u}_\mathcal{A}, \bar{y}_\mathcal{I}, \bar{y}_\mathcal{A})$ is an optimal solution of the inner optimization problem (2.37) for the set $\mathcal{A}$, too, and unique due to Theorem 3. Additionally, $(\mathcal{A}, \bar{u}_\mathcal{I}, \bar{u}_\mathcal{A}, \bar{y}_\mathcal{I}, \bar{y}_\mathcal{A})$ is admissible for the outer optimization problem (2.36).
It remains to show, that it is optimal and the unique optimum. Both properties can be proven by the contradiction ideas form the proof of Theorem 2. □

In view of the abstract framework of optimal control from Paragraph 2.3.1, the existence and uniqueness results of the theorems 3 and 4 transcend their assertion: They ensure the existence of a control-to-state operator, which should more accurately be called *geometry-to-solution operator* in this specific context. Beyond this, the theorems enable a formulation as reduced optimization problem.

Definition 6 (Geometry-to-solution operator)**:**
Let the family of admissible sets $\mathcal{O}$ be given by Definition 4. Then the map

$$G : \mathcal{O} \to L^2(\mathcal{J}) \times L^2(\mathring{\mathcal{B}}) \times H^1(\mathcal{J}, \Delta) \times H^1(\mathring{\mathcal{B}}, \Delta), \quad \mathcal{B} \mapsto (\bar{u}_\mathcal{J}, \bar{u}_\mathcal{B}, \bar{y}_\mathcal{J}, \bar{y}_\mathcal{B}),$$

where the image is the solution of the inner optimization problem (2.37), is called the *geometry-to-solution operator* for the set optimal control problem (2.30) and the bilevel optimization problem (2.36), (2.37), respectively.[10]

With help of this definition, one can reduce the bilevel optimization problem (2.36), (2.37), and the set optimal control problem (2.30) as well:

$$\begin{aligned} &\text{minimize} && \mathcal{F}(\mathcal{B}) := \mathfrak{J}(\mathcal{B}; G(\mathcal{B})), && \text{(2.38a)}\\ &\text{subject to} && \begin{cases} \mathcal{B} \in \mathcal{O}, & \text{(2.38b)}\\ y_{\min} < G_3(\mathcal{B}) < y_{\max} \quad \text{in } \mathcal{J}. & \text{(2.38c)} \end{cases} \end{aligned}$$

At this, G_3 denotes the third component of the geometry-to-solution operator.

If one steps back to the general recipe of how to derive first order necessary conditions of Paragraph 2.3.2, one recognizes, that the first step is successfully completed.
In order to tackle the second step of the recipe, it is necessary to analyze the differentiability of the reduced functional $\mathcal{F}$. Since the reduced problem (2.38) suppresses the OC-PDE character of the set optimal control problem, but maintains the shape/topology-optimization character, it is nearby to check whether $\mathcal{F}$ can be differentiated by means of the corresponding calculus. However, the bilevel character is only suppressed via the geometry-to-solution operator. Shape/topological differentiability of G is not obvious and requires further investigation. Conceptually this means, that one needs to prove shape/topological differentiability of a set parametrized minimization problem, which is the goal of the following paragraph. Another approach to tackle such type of problems can be found in [44, Chp. 10].

[10] In view of the results of Section 2.6 the geometry-to-solution operator G maps from $\mathcal{O}$ to a vector bundle E on $\mathcal{O}$. Hence, to be precise, one should write $G : \mathcal{O} \to E$, where the vector bundle E is given by (2.86), since only the specific choice of $\mathcal{B} \in \mathcal{O}$ determines the image space $L^2(\mathcal{J}) \times L^2(\mathring{\mathcal{B}}) \times H^1(\mathcal{J}, \Delta) \times H^1(\mathring{\mathcal{B}}, \Delta)$.

2.3.5 Necessary conditions for the inner optimization problem

As already mentioned, the geometry-to-solution operator G hides the OC-PDE character of the bilevel problem. Applying G to a set $\mathcal{B} \in \mathcal{O}$ means solving the inner optimization problem (2.37) for $\mathcal{B}$. The aim of the present paragraph is to provide a constructive way to evaluate G. Since the set parametrized inner optimization problem is strictly convex for any choice of $\mathcal{B} \in \mathcal{O}$, solving it is equivalent to solve its first order necessary (and sufficient) conditions. Moreover, since (iOP) is an optimization problem in Banach spaces, it is possible to use the sophisticated tools of Karush-Kuhn-Tucker theory to prove existence of Lagrange multipliers.

Theorem 5 (Existence of Lagrange multipliers for the inner optimization problem)**:**
Let the family of admissible sets $\mathcal{O}$ be given by Definition 4, let $\mathcal{B} \in \mathcal{O}$ be arbitrarily chosen, and let $(\bar{u}_{\mathcal{J}}, \bar{u}_{\mathcal{B}}, \bar{y}_{\mathcal{J}}, \bar{y}_{\mathcal{B}})$ be the optimal solution to the inner optimization problem (2.37) for the fixed parameter $\mathcal{B}$. Then there exist multipliers $\bar{q}_{\mathcal{B}} \in H^2(\mathring{\mathcal{B}})$, $\bar{\sigma}_{\mathcal{J}} \in H^{-3/2}(\beta)$, and functions $\bar{p}_{\mathcal{J}} \in L^2(\mathcal{J}, \Delta)$, $\bar{p}_{\mathcal{B}} \in H^2(\mathring{\mathcal{B}})$, such that there holds (this is the dual part of *Karush-Kuhn-Tucker (KKT) conditions*)[11]

$$-\Delta \bar{p}_{\mathcal{J}} + \bar{p}_{\mathcal{J}} = \bar{y}_{\mathcal{J}} - y_d \quad \text{a. e. in } \mathcal{J}, \tag{2.39a}$$

$$\partial_{\boldsymbol{n}} \bar{p}_{\mathcal{J}} 0 \quad \text{a. e. on } \Gamma, \tag{2.39b}$$

$$\partial_{\boldsymbol{n}}^{\mathcal{J}} \bar{p}_{\mathcal{J}} = \bar{\sigma}_{\mathcal{J}} \quad \text{a. e. on } \beta, \tag{2.39c}$$

$$\lambda\,(\bar{u}_{\mathcal{J}} - u_d) = -\bar{p}_{\mathcal{J}} \quad \text{a. e. in } \mathcal{J}, \tag{2.39d}$$

$$-\Delta \bar{p}_{\mathcal{B}} + \bar{p}_{\mathcal{B}} = \bar{y}_{\mathcal{B}} - y_d \quad \text{a. e. in } \mathring{\mathcal{B}}, \tag{2.39e}$$

$$\bar{p}_{\mathcal{B}}|_{\beta} = 0 \quad \text{a. e. on } \beta, \tag{2.39f}$$

$$\lambda\,(\bar{u}_{\mathcal{B}} - u_d) = -\bar{p}_{\mathcal{B}} - \bar{q}_{\mathcal{B}} \quad \text{a. e. in } \mathring{\mathcal{B}}. \tag{2.39g}$$

Remarks:

1. Note, that due to the low regularity of $\bar{\sigma}_{\mathcal{J}}$, the equations defining $\bar{p}_{\mathcal{J}}$ make sense for $\bar{p}_{\mathcal{J}} \in L^2(\mathcal{J}, \Delta)$, and for Γ, β of class $C^{1,1}$ only. In particular, this result is one cause of the Assumption 1 on the regularity of the boundaries.
2. Further properties of the Hilbert space $L^2(\mathring{\mathcal{A}}, \Delta)$ (its definition is given by (2.15)) can be found in [6, Sec. 7.1], [69, Sec. 1.5.3] and [64]. Especially it is mentioned there ([6, Thm. 7.1-2]), that
$$(-\Delta + \mathrm{Id}, \partial_{\boldsymbol{n}}, \partial_{\boldsymbol{n}}^{\mathcal{J}}) : L^2(\mathcal{J}, \Delta) \to L^2(\mathcal{J}) \times H^{-\frac{3}{2}}(\Gamma) \times H^{-\frac{3}{2}}(\beta), \quad \phi \mapsto (-\Delta\phi + \phi,\ \partial_{\boldsymbol{n}}\phi,\ \partial_{\boldsymbol{n}}^{\mathcal{J}}\phi)$$
is an isomorphism. This is used in the following proof actually.
3. Obviously one gains regularity of the multiplier $\bar{q}_{\mathcal{B}}$ by means of the BDD ansatz, since the corresponding distributed Lagrange multiplier $\mu_{\mathring{\mathcal{A}}}$ is in $L^2(\mathring{\mathcal{A}})$ only, cf. Proposition 3. Besides index reduction (cf. Paragraph 2.7.3) this is an essential motivation for the BDD approach.
4. However, there is no improvement of regularity of $\bar{\sigma}_{\mathcal{B}}$ vs. μ_γ. Even worse, the new interface multiplier is less regular. But the comparison is flawed, since μ_γ lives on the optimal interface, whereas $\bar{\sigma}_{\mathcal{B}}$ does not (have to do it) and since it is known, that characterizing properties of the optimum may yield higher regularity. In particular, Corollary 2 shows that regularity of primal and dual variables is higher at the optimum at least. Moreover, the BDD ansatz applied here (i. e. (2.27)) does not contain a differentiation step of the boundary condition, such that improvement of the regularity of the corresponding multiplier cannot be expected as contrasted with the BDD approach (2.28), see Appendix A.
5. Low regluarity of $\bar{\sigma}_{\mathcal{J}}$ and $\bar{p}_{\mathcal{J}}$ is a very crucial point in the following considerations. For one thing, it is responsible for different fruitless efforts to prove shape differentiability of the KKT system, cf. Lemma 8 and Appendix C. For another thing, low regularity may cause issues in the numerical treatment, cf. the 3rd item on page 101.
6. While Bergounioux and Kunisch had to search for a suitable decomposition of the Lagrange multiplier associated with the state constraint into a distributed regular part $\mu_{\mathring{\mathcal{A}}}$ and a singular interface part μ_γ (cf. Proposition 3), the BDD reformulation (2.27) of the state constraint entails the introduction of two multipliers $\bar{q}_{\mathcal{B}}$ and $\bar{\sigma}_{\mathcal{J}}$. Thus, an outcome of the BDD approach is a quite natural decomposition of the multiplier.

[11] The dual variables are denoted with a bar $\bar{\cdot}$ here in order to distinguish them from another set of dual variables, which are introduced in Theorem 9 in Appendix B. Moreover, the indices $\mathcal{B}$ and $\mathcal{J}$ are used in order to mark, that the multipliers are associated with the set parametrized inner optimization problem.

7. It would be desirable to formulate an adjoint equation, whose form is closer to the state equation of (iOP). In other words, the adjoint states $\bar{p}_{\mathcal{J}}$ and $\bar{p}_{\mathcal{B}}$ should be directly connected via interface conditions. This topic is discussed in Appendix B.

Proof. The proof consists of three parts. The first one provides existence of $\bar{q}_{\mathcal{B}}$ and $\bar{\sigma}_{\mathcal{J}}$ as Lagrange multipliers to a reduced problem. The second part shows, that the functions $\bar{p}_{\mathcal{J}}$ and $\bar{p}_{\mathcal{B}}$ are well-defined and that the relations (2.39) hold. Finally, higher regularity of $\bar{q}_{\mathcal{B}}$ is shown.

1) The coupled system (2.37d)–(2.37j) can be written equivalently (this is without changing the feasible set) as

$$-\Delta y_{\mathcal{J}} + y_{\mathcal{J}} = u_{\mathcal{J}} \quad \text{in } \mathcal{J}, \tag{2.40a}$$
$$\partial_{\boldsymbol{n}} y_{\mathcal{J}} = 0 \quad \text{on } \Gamma, \tag{2.40b}$$
$$\partial_{\boldsymbol{n}}^{\mathcal{J}} y_{\mathcal{J}} = \partial_{\boldsymbol{n}}^{\mathcal{J}} y_{\min}^{\max} \quad \text{on } \beta, \tag{2.40c}$$
$$y_{\mathcal{J}}|_{\beta} = y_{\min}^{\max}|_{\beta} \quad \text{on } \beta, \tag{2.40d}$$
$$-\Delta y_{\mathcal{B}} + y_{\mathcal{B}} = u_{\mathcal{B}} \quad \text{in } \mathring{\mathcal{B}}, \tag{2.40e}$$
$$y_{\min}^{\max}|_{\beta} = y_{\mathcal{B}}|_{\beta} \quad \text{on } \beta, \tag{2.40f}$$
$$-\Delta y_{\min}^{\max} + y_{\min}^{\max} = u_{\mathcal{B}} \quad \text{in } \mathring{\mathcal{B}}. \tag{2.40g}$$

In particular, the constraints are separated in one block which acts on $\mathcal{J}$ and one which acts on $\mathring{\mathcal{B}}$ now. Consequently, the minimization on the two sets are independent of each other.

Let $\mathcal{B} \in \mathcal{O}$ be arbitrarily chosen, but fix. Consider the linear control-to-state operators $S_{\mathcal{J}}$ and $S_{\mathcal{B}}$ of the split system (2.40)

$$S_{\mathcal{J}} : L^2(\mathcal{J}) \to H^2(\mathcal{J}), \quad u_{\mathcal{J}} \mapsto y_{\mathcal{J}}, \text{ where } y_{\mathcal{J}} \text{ is the solution to (2.40a)–(2.40c)},$$
$$S_{\mathcal{B}} : L^2(\mathring{\mathcal{B}}) \to H^2(\mathring{\mathcal{B}}), \quad u_{\mathcal{B}} \mapsto y_{\mathcal{B}}, \text{ where } y_{\mathcal{B}} \text{ is the solution to (2.40e)–(2.40f)}.$$

$S_{\mathcal{J}}$ and $S_{\mathcal{B}}$ are known to be continuous (cf. [69, Thm. 2.3.3.2] or [6, Thm. 7.1-2]). With use of the Dirichlet trace operators on the interface (cf. Definition 2 and Lemma 1)

$$\tau_{\mathcal{J}} : H^2(\mathcal{J}) \to H^{\frac{3}{2}}(\beta),$$
$$\tau_{\mathcal{B}} : H^2(\mathring{\mathcal{B}}) \to H^{\frac{3}{2}}(\beta),$$

the inner optimization problem (2.37) can be reduced to

$$\begin{aligned} &\text{minimize} && f(u_{\mathcal{J}}, u_{\mathcal{A}}) := \mathfrak{J}(\mathcal{B}; u_{\mathcal{J}}, u_{\mathcal{B}}, S_{\mathcal{J}}(u_{\mathcal{J}}), S_{\mathcal{B}}(u_{\mathcal{B}})) \\ &\text{subject to} && T(u_{\mathcal{J}}, u_{\mathcal{B}}) := \begin{pmatrix} \tau_{\mathcal{J}} S_{\mathcal{J}}(u_{\mathcal{J}}) - \tau_{\mathcal{J}} y_{\min}^{\max} \\ \Delta y_{\min}^{\max} - y_{\min}^{\max} + u_{\mathcal{B}} \end{pmatrix} = 0, \end{aligned} \tag{2.41}$$

where $T : L^2(\mathcal{J}) \times L^2(\mathring{\mathcal{B}}) \to H^{3/2}(\beta) \times L^2(\mathring{\mathcal{B}})$ collects the remaining constraints (2.40d), (2.40g). This reduced problem fits in the usual framework of nonlinear optimization in Banach spaces.

In order to prove existence of multipliers, one has to show that a constraint qualification is valid. In the current context, the Zowe-Kurcyusz constraint qualification (cf. [164] and [159, p. 330]) is suitable, and its validity for the operator T in $(\bar{u}_{\mathcal{J}}, \bar{u}_{\mathcal{B}})$ will be proven next.

Note, that T is continuously Fréchet differentiable, since $S_{\mathcal{J}}$ is continuous and affine. Thus, for each arbitrary $z_1 \in H^{3/2}(\beta)$ and $z_2 \in L^2(\mathring{\mathcal{B}})$ one has to find $(h_{\mathcal{J}}, h_{\mathcal{B}}) \in L^2(\mathcal{J}) \times L^2(\mathring{\mathcal{B}})$ such that

$$(\mathrm{D}T(\bar{u}_{\mathcal{J}}, \bar{u}_{\mathcal{B}}))(h_{\mathcal{J}}, h_{\mathcal{B}}) = \begin{pmatrix} \tau_{\mathcal{J}} S_{\mathcal{J}}^0(h_{\mathcal{J}}) \\ h_{\mathcal{B}} \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix},$$

where $S_{\mathcal{J}}^0$ is the linear part of the affine operator $S_{\mathcal{J}}$. This is the solution operator to

$$\begin{aligned} -\Delta v + v &= h_{\mathcal{J}} && \text{in } \mathcal{J}, \\ \partial_{\boldsymbol{n}} v &= 0 && \text{on } \Gamma, \\ \partial_{\boldsymbol{n}}^{\mathcal{J}} v &= 0 && \text{on } \beta. \end{aligned}$$

As a start, this defines $h_{\mathcal{B}} := z_2$. According to Lemma 1, there exists $v \in H^2(\mathcal{J})$ such that

$$\partial_{\boldsymbol{n}} v = 0 \quad \text{on } \Gamma, \qquad \partial_{\boldsymbol{n}}^{\mathcal{J}} v = 0 \quad \text{on } \beta, \qquad \tau_{\mathcal{J}} v = z_1 \quad \text{on } \beta.$$

Consequently, $h_{\mathcal{J}} := -\Delta v + v \in L^2(\mathcal{J})$ is well-defined and $\tau_{\mathcal{J}} S_{\mathcal{J}}^0(h_{\mathcal{J}}) = z_1$. Hence, the Zowe-Kurcyusz constraint qualification is fulfilled and there exist Lagrange multipliers $\bar{\sigma}_{\mathcal{J}} \in H^{3/2}(\beta)^* = H^{-3/2}(\beta)$ and $q \in L^2(\mathring{\mathcal{B}})$.

2) In addition, $(\bar{u}_{\mathcal{J}}, \bar{u}_{\mathcal{B}}, \bar{q}_{\mathcal{B}}, \bar{\sigma}_{\mathcal{J}})$ is a saddle point of the Lagrange function (cf. [159, Thm. 6.3])

$$\mathcal{L} : L^2(\mathcal{J}) \times L^2(\mathring{\mathcal{B}}) \times L^2(\mathring{\mathcal{B}}) \times H^{-\frac{3}{2}}(\beta) \to \mathbb{R},$$
$$\mathcal{L}(u_{\mathcal{J}}, u_{\mathcal{B}}, q_{\mathcal{B}}, \sigma_{\mathcal{J}}) := f(u_{\mathcal{J}}, u_{\mathcal{B}}) + \int_{\mathring{\mathcal{B}}} q_{\mathcal{B}} \, (u_{\mathcal{B}} + \Delta y_{\min}^{\max} - y_{\min}^{\max}) + \left\langle \sigma_{\mathcal{J}} \, , \, \tau_{\mathcal{J}} S_{\mathcal{J}}(u_{\mathcal{J}}) - \tau_{\mathcal{J}} y_{\min}^{\max} \right\rangle_{H^{-\frac{3}{2}}(\beta), H^{\frac{3}{2}}(\beta)}.$$

Since the constraints (2.41) solely consist of equations, one has (analogously to $S_{\mathcal{J}}^0$, $S_{\mathcal{B}}^0$ denotes the linear part of the affine operator $S_{\mathcal{B}}$ here)

$$0 = \partial_{u_{\mathcal{B}}} \mathcal{L}(\bar{u}_{\mathcal{J}}, \bar{u}_{\mathcal{B}}, \bar{q}_{\mathcal{B}}, \bar{\sigma}_{\mathcal{J}}) \, h = \int_{\mathring{\mathcal{B}}} (S_{\mathcal{B}}(\bar{u}_{\mathcal{B}}) - y_d) \, S_{\mathcal{B}}^0(h) + \lambda(\bar{u}_{\mathcal{B}} - u_d) \, h + \bar{q}_{\mathcal{B}} \, h, \quad \forall h \in L^2(\mathring{\mathcal{B}}).$$

Now let $\bar{p}_{\mathcal{B}} \in H^2(\mathcal{J}) \cap H_0^1(\mathring{\mathcal{B}})$ be introduced as the solution to (2.39e), (2.39f). Using integration by parts one can proceed, that for all $h \in L^2(\mathring{\mathcal{B}})$ there holds

$$\begin{aligned} 0 &= \int_{\mathring{\mathcal{B}}} (-\Delta \bar{p}_{\mathcal{B}} + \bar{p}_{\mathcal{B}}) \, S_{\mathcal{B}}^0(h) + \lambda(\bar{u}_{\mathcal{B}} - u_d) \, h + \bar{q}_{\mathcal{B}} \, h \\ &= \int_{\mathring{\mathcal{B}}} \underbrace{(-\Delta S_{\mathcal{B}}^0(h) + S_{\mathcal{B}}^0(h))}_{=h} \, p + \lambda(\bar{u}_{\mathcal{B}} - u_d) \, h + \bar{q}_{\mathcal{B}} \, h + \int_{\beta} -\partial_{\boldsymbol{n}}^{\mathcal{B}} \bar{p}_{\mathcal{B}} \underbrace{\tau_{\mathcal{B}} S_{\mathcal{B}}^0(h)}_{=0} + \underbrace{\tau_{\mathcal{B}} \bar{p}_{\mathcal{B}}}_{=0} \partial_{\boldsymbol{n}}^{\mathcal{B}} S_{\mathcal{B}}^0(h). \end{aligned}$$

Hence, one obtains (2.39g). In addition, the saddle point property, together with $\bar{p}_{\mathcal{J}} \in L^2(\mathcal{J}, \Delta)$, which is well-defined by (2.39a)–(2.39c) (cf. the 2nd item of the Remarks above this proof), and a suitable Green's formula (cf. [69, Thm. 1.5.3.6] in the special case without corners S_j) yields

$$\begin{aligned} 0 &= \partial_{u_{\mathcal{J}}} \mathcal{L}(\bar{u}_{\mathcal{J}}, \bar{u}_{\mathcal{B}}, \bar{q}_{\mathcal{B}}, \bar{\sigma}_{\mathcal{J}}) \, h \\ &= \int_{\mathcal{J}} (S_{\mathcal{J}}(\bar{u}_{\mathcal{J}}) - y_d) \, S_{\mathcal{J}}^0(h) + \lambda(\bar{u}_{\mathcal{J}} - u_d) \, h + \left\langle \bar{\sigma}_{\mathcal{J}} \, , \, \tau_{\mathcal{J}} S_{\mathcal{J}}^0(h) \right\rangle_{H^{-\frac{3}{2}}(\beta), H^{\frac{3}{2}}(\beta)} \\ &= \int_{\mathcal{J}} (-\Delta \bar{p}_{\mathcal{J}} + \bar{p}_{\mathcal{J}}) \, S_{\mathcal{J}}^0(h) + \lambda(\bar{u}_{\mathcal{J}} - u_d) \, h + \left\langle \bar{\sigma}_{\mathcal{J}} \, , \, \tau_{\mathcal{J}} S_{\mathcal{J}}^0(h) \right\rangle_{H^{-\frac{3}{2}}(\beta), H^{\frac{3}{2}}(\beta)} \\ &= \int_{\mathcal{J}} \underbrace{(-\Delta S_{\mathcal{J}}^0(h) + S_{\mathcal{J}}^0(h))}_{=h} \, \bar{p}_{\mathcal{J}} + \lambda(\bar{u}_{\mathcal{J}} - u_d) \, h \\ &\quad - \Big\langle \underbrace{\partial_{\boldsymbol{n}} \bar{p}_{\mathcal{J}}}_{=0} \, , \, S_{\mathcal{J}}^0(h)|_{\Gamma} \Big\rangle_{H^{-\frac{3}{2}}(\Gamma), H^{\frac{3}{2}}(\Gamma)} + \Big\langle \underbrace{\bar{\sigma}_{\mathcal{J}} - \partial_{\boldsymbol{n}}^{\mathcal{J}} \bar{p}_{\mathcal{J}}}_{=0} \, , \, \tau_{\mathcal{J}} S_{\mathcal{J}}^0(h) \Big\rangle_{H^{-\frac{3}{2}}(\beta), H^{\frac{3}{2}}(\beta)} \\ &\quad + \Big\langle \bar{p}_{\mathcal{J}}|_{\Gamma} \, , \, \underbrace{\partial_{\boldsymbol{n}} S_{\mathcal{J}}^0(h)}_{=0} \Big\rangle_{H^{-\frac{1}{2}}(\Gamma), H^{\frac{1}{2}}(\Gamma)} + \Big\langle \tau_{\mathcal{J}} \bar{p}_{\mathcal{J}} \, , \, \underbrace{\partial_{\boldsymbol{n}}^{\mathcal{J}} S_{\mathcal{J}}^0(h)}_{=0} \Big\rangle_{H^{-\frac{1}{2}}(\beta), H^{\frac{1}{2}}(\beta)}. \end{aligned}$$

Finally, one obtains (2.39d). This completes the derivation of the claimed conditions.

3) Equation (2.39g) guarantees higher regularity of $\bar{q}_{\mathcal{B}}$ indeed

$$\bar{q}_{\mathcal{B}} = -\lambda(\bar{u}_{\mathcal{B}} - u_d) - \bar{p}_{\mathcal{B}} \in H^2(\mathring{\mathcal{B}}).$$ □

In view of strict convexity of the inner optimization problem, the KKT conditions are also sufficient. Consequently, the goal of constructively evaluate the geometry-to-solution operator G is reached, if one is able to solve the first order necessary conditions. This topic is investigated next. Due to Theorem 5 the full first order necessary and sufficient conditions for the inner optimization problem iOP for $\mathcal{B} \in \mathcal{O}$ are

$$\begin{aligned}
-\Delta \bar{y}_{\mathcal{J}} + \bar{y}_{\mathcal{J}} &= \bar{u}_{\mathcal{J}} && \text{in } \mathcal{J}, && (2.42a) \\
-\Delta \bar{y}_{\mathcal{B}} + \bar{y}_{\mathcal{B}} &= \bar{u}_{\mathcal{B}} && \text{in } \mathring{\mathcal{B}}, && (2.42b) \\
\partial_{\boldsymbol{n}} \bar{y}_{\mathcal{J}} &= 0 && \text{on } \Gamma, && (2.42c) \\
\bar{y}_{\mathcal{J}}|_{\beta} &= y_{\min}^{\max}|_{\beta} && \text{on } \beta, && (2.42d) \\
\partial_{\boldsymbol{n}}^{\mathcal{J}} \bar{y}_{\mathcal{J}} &= \partial_{\boldsymbol{n}}^{\mathcal{J}} y_{\min}^{\max} && \text{on } \beta, && (2.42e) \\
-\Delta y_{\min}^{\max} + y_{\min}^{\max} &= \bar{u}_{\mathcal{B}} && \text{in } \mathring{\mathcal{B}}, && (2.42f) \\
y_{\min}^{\max}|_{\beta} &= \bar{y}_{\mathcal{B}}|_{\beta} && \text{on } \beta, && (2.42g)
\end{aligned}$$

$$\begin{aligned}
-\Delta \bar{p}_{\mathcal{J}} + \bar{p}_{\mathcal{J}} &= \bar{y}_{\mathcal{J}} - y_d && \text{in } \mathcal{J}, && (2.42h) \\
-\Delta \bar{p}_{\mathcal{B}} + \bar{p}_{\mathcal{B}} &= \bar{y}_{\mathcal{B}} - y_d && \text{in } \mathring{\mathcal{B}}, && (2.42i) \\
\partial_{\boldsymbol{n}} \bar{p}_{\mathcal{J}} &= 0 && \text{on } \Gamma, && (2.42j) \\
\bar{p}_{\mathcal{B}}|_{\beta} &= 0 && \text{on } \beta, && (2.42k) \\
\partial_{\boldsymbol{n}}^{\mathcal{J}} \bar{p}_{\mathcal{J}} &= \bar{\sigma}_{\mathcal{J}} && \text{on } \beta, && (2.42l) \\
\lambda \, (\bar{u}_{\mathcal{J}} - u_d) + \bar{p}_{\mathcal{J}} &= 0 && \text{in } \mathcal{J}, && (2.42m) \\
\lambda \, (\bar{u}_{\mathcal{B}} - u_d) + \bar{p}_{\mathcal{B}} + \bar{q}_{\mathcal{B}} &= 0 && \text{in } \mathring{\mathcal{B}}, && (2.42n)
\end{aligned}$$

where the functions involved possess the following regularities

$$\bar{u}_{\mathcal{J}} \in L^2(\mathcal{J},\Delta), \quad (2.43a) \qquad \bar{p}_{\mathcal{J}} \in L^2(\mathcal{J},\Delta), \quad (2.43e)$$
$$\bar{u}_{\mathcal{B}} \in L^2(\mathring{\mathcal{B}},\Delta), \quad (2.43b) \qquad \bar{p}_{\mathcal{B}} \in H^2(\mathring{\mathcal{B}}), \quad (2.43f)$$
$$\bar{y}_{\mathcal{J}} \in H^2(\mathcal{J}), \quad (2.43c) \qquad \bar{q}_{\mathcal{B}} \in H^2(\mathring{\mathcal{B}}), \quad (2.43g)$$
$$\bar{y}_{\mathcal{B}} \in H^2(\mathring{\mathcal{B}}), \quad (2.43d) \qquad \bar{\sigma}_{\mathcal{J}} \in H^{-\frac{3}{2}}(\beta). \quad (2.43h)$$

Fortunately this optimality system can be reduced considerably. However, this property is only due to the very simple structure of the original model problem (2.1), such that the assertion of Lemma 7 is not representative for the presented approach.

Lemma 7 (Reduced optimality system of the inner optimization problem)**:**
The optimality system (2.42) can be reduced to

$$-\Delta\bar{y}_{\mathcal{J}} + \bar{y}_{\mathcal{J}} + \frac{1}{\lambda}\bar{p}_{\mathcal{J}} = u_d \quad \text{in } \mathcal{J}, \quad (2.44a) \qquad -\Delta\bar{p}_{\mathcal{J}} + \bar{p}_{\mathcal{J}} - \bar{y}_{\mathcal{J}} = -y_d \text{ in } \mathcal{J}, \quad (2.44e)$$
$$\partial_{\boldsymbol{n}}\bar{y}_{\mathcal{J}} = 0 \quad \text{on } \Gamma, \quad (2.44b) \qquad \partial_{\boldsymbol{n}}\bar{p}_{\mathcal{J}} = 0 \quad \text{on } \Gamma, \quad (2.44f)$$
$$\bar{y}_{\mathcal{J}}|_\beta = y_{\min}^{\max}|_\beta \text{ on } \beta, \quad (2.44c)$$
$$\partial_{\boldsymbol{n}}^{\mathcal{J}}\bar{y}_{\mathcal{J}} = \partial_{\boldsymbol{n}}^{\mathcal{B}} y_{\min}^{\max} \text{ on } \beta. \quad (2.44d)$$

This means, that instead of solving (2.42) all at once, one can solve it step by step:

1. assign $\bar{y}_{\mathcal{B}} = y_{\min}^{\max}$ in $\mathring{\mathcal{B}}$
2. solve (2.44)
3. assign $\begin{cases} \bar{u}_{\mathcal{J}} = -\frac{1}{\lambda}\bar{p}_{\mathcal{J}} + u_d & \text{in } \mathcal{J} \\ \bar{u}_{\mathcal{B}} = -\Delta y_{\min}^{\max} + y_{\min}^{\max} & \text{in } \mathring{\mathcal{B}} \end{cases}$
4. solve $\begin{cases} -\Delta\bar{p}_{\mathcal{B}} + \bar{p}_{\mathcal{B}} = y_{\min}^{\max} - y_d & \text{in } \mathring{\mathcal{B}} \\ \bar{p}_{\mathcal{B}}|_\beta = 0 & \text{on } \beta \end{cases}$
5. assign $\begin{cases} \bar{q}_{\mathcal{B}} = \lambda(\bar{u}_{\mathcal{B}} - u_d) - \bar{p}_{\mathcal{B}} & \text{in } \mathring{\mathcal{B}} \\ \bar{\sigma}_{\mathcal{J}} = \partial_{\boldsymbol{n}}^{\mathcal{J}}\bar{p}_{\mathcal{J}} & \text{on } \beta. \end{cases}$

Remark:
On the first glance, it may be confusing, that the lemma states on the one hand, that the optimality system can be reduced to (2.44), and on the other hand, that there are four additional solving steps. Usually, one is not really interested in solving the optimality system itself, but find the optimal solution of iOP. In other words, one aims at solving the optimality system in the primal variables. This is done in the first three steps, indeed. But since the assignments in steps one and three are very cheap in the current context, solving the inner optimization problem means solving the $(\bar{y}_{\mathcal{J}}, \bar{p}_{\mathcal{J}})$-system (2.44), here. This fact is reflected in Theorem 6.

Proof. First of all, it will be shown, that the optimality system (2.42) can be solved by executing the five steps of solving and assigning. Afterwards, the attention is concentrated on the question, whether the solving steps are well-defined.

1) The BDD reformulation of the state constraint (2.42f), (2.42g), together with the state equation (2.42b), ensures $\bar{y}_{\mathcal{B}} = y_{\min}^{\max}$; cf. (2.27). Consequently, $\bar{y}_{\mathcal{B}}$ can be replaced by $y_{\min}^{\max}$ within (2.42). Plugging the gradient equation (2.42m) into the state equation (2.42a) yields the reduced system (2.44). As soon as (2.44) is solved, the right hand sides of the assignment in step three are known. After having solved the adjoint boundary value problem of step four, the remaining part of solving (2.42) means executing assignments of the fifth step.

2) In contrast to the boundary value problem of step four, it is not obvious if the reduced system (2.44) is uniquely solvable, since the distribution of boundary condition is unusual.

Strict convexity of the inner optimization problem (2.37) ensures unique solvability (cf. Theorem 3) as well as sufficiency of its first order necessary conditions (2.42). Assume the BVP (2.44) were not uniquely solvable. Hence, in virtue of the first assignment in step three, each of its solutions would produce an extra optimal solution of the inner optimization problem. This contradiction completes the proof. □

At this point, the goal to provide a constructive way to evaluate the geometry-to-solution operator G is attained. Instead of solving the inner optimization problem (2.37), one can solve its first order necessary and sufficient conditions (2.42). Consequently, the inner optimization problem can be replaced by this conditions within the bilevel optimization problem (2.36), (2.37). Thereby – in view of the explanation at the end of Paragraph 2.3.4 – one has already gained a lot: It is not necessary to prove shape/topological differentiability of a set parametrized minimization problem any more; now one has to prove shape/topological differentiability of the optimality system, which is a partial differential algebraic equation (PDAE).

Beyond this, it is even not necessary to solve the whole optimality system. It is sufficient to solve its reduced form of Lemma 7, i. e. solve BVP (2.44). And as a result, shape/topological differentiability is only needed for this boundary value problem. This is much closer to the standard problem formulation of shape/topology optimization.

That is to say, this paragraph helps to prove differentiability of the geometry-to-solution operator G, and therefore is an essential ingredient to tackle the second step of the general recipe of Paragraph 2.3.2.

2.3.6 Analysis of the outer optimization problem

By means of the preliminary work of Paragraph 2.3.5 it is possible to derive the derivative of the reduced objective $\mathcal{F}$ of the reduced bilevel optimization problem (2.38). This is step two of the general recipe of Paragraph 2.3.2 within the analysis of the set optimal control problem (2.30).

By means of the geometry-to-solution operator G (cf. Paragraph 2.3.4), it is possible to reduce the bilevel optimization problem BiOP to (2.38). The detailed analysis of the inner optimization problem (cf. Paragraph 2.3.5) yields necessary and sufficient conditions, which enable an easy evaluation of the geometry-to-solution operator. Altogether the set optimal control problem (2.30) is equivalent to a strongly reduced shape/topology optimization problem.

Theorem 6 (Set optimal control problem as shape/topology optimization problem)**:**
The set optimal control problem (2.30) is equivalent to the shape/topology optimization problem

$$\text{minimize} \quad \mathcal{F}(\mathcal{B}) := \frac{1}{2}\|\bar{y}_{\mathcal{J}} - y_d\|^2_{L^2(\mathcal{J})} + \frac{1}{2}\|y^{\max}_{\min} - y_d\|^2_{L^2(\mathring{\mathcal{B}})} + \frac{1}{2\lambda}\|\bar{p}_{\mathcal{J}}\|^2_{L^2(\mathcal{J})} + \frac{1}{2\lambda}\|p^{\max}_{\min}\|^2_{L^2(\mathring{\mathcal{B}})} \tag{2.45a}$$

subject to

$$\mathcal{B} \in \mathcal{O}, \tag{2.45b}$$

$$y_{\min} < \bar{y}_{\mathcal{J}} < y_{\max} \quad \text{in } \mathcal{J}, \tag{2.45c}$$

$$-\Delta\bar{y}_{\mathcal{J}} + \bar{y}_{\mathcal{J}} + \frac{1}{\lambda}\bar{p}_{\mathcal{J}} = u_d \quad \text{in } \mathcal{J}, \tag{2.45d}$$

$$\partial_n \bar{y}_{\mathcal{J}} = 0 \quad \text{on } \Gamma, \tag{2.45e}$$

$$\bar{y}_{\mathcal{J}}|_\beta = y^{\max}_{\min}|_\beta \quad \text{on } \beta, \tag{2.45f}$$

$$\partial^{\mathcal{J}}_n \bar{y}_{\mathcal{J}} = \partial^{\mathcal{B}}_n y^{\max}_{\min} \quad \text{on } \beta, \tag{2.45g}$$

$$-\Delta\bar{p}_{\mathcal{J}} + \bar{p}_{\mathcal{J}} - \bar{y}_{\mathcal{J}} = -y_d \quad \text{in } \mathcal{J}, \tag{2.45h}$$

$$\partial_n \bar{p}_{\mathcal{J}} = 0 \quad \text{on } \Gamma, \tag{2.45i}$$

$$\bar{p}_{\mathcal{J}} \in L^2(\mathcal{J}, \Delta),\ \bar{y}_{\mathcal{J}} \in H^2(\mathcal{J}), \tag{2.45j}$$

in the following sense:
Let $(\mathcal{A}; \bar{u}_{\mathcal{I}}, \bar{u}_{\mathcal{A}}, \bar{y}_{\mathcal{I}}, \bar{y}_{\mathcal{A}})$ be the optimal solution of (2.30) and let $\bar{\mathcal{B}}$ be the optimal solution of (2.45), then

$$\mathcal{A} = \bar{\mathcal{B}}, \qquad \bar{u}_{\mathcal{A}} = -\frac{1}{\lambda}p^{\max}_{\min} + u_d, \qquad \bar{y}_{\mathcal{A}} = y^{\max}_{\min},$$

$$\bar{u}_{\mathcal{I}} = -\frac{1}{\lambda}\bar{p}_{\mathcal{J}} + u_d, \qquad \bar{y}_{\mathcal{I}} = \bar{y}_{\mathcal{J}}.$$

In particular, (2.45) is uniquely solvable. At this, the coefficient function $p_{\min}^{\max} \in H^2(\Omega)$ is constructed the same way as $y_{\min}^{\max}$ in Lemma 4, but such that

$$p_{\min}^{\max}(x) = \begin{cases} \lambda\,(\Delta y_{\max}(x) - y_{\max}(x) + u_d(x)), & x \text{ in a neighborhood } B_{\max} \text{ of } \mathcal{B}_{\max}, \\ \lambda\,(\Delta y_{\min}(x) - y_{\min}(x) + u_d(x)), & x \text{ in a neighborhood } B_{\min} \text{ of } \mathcal{B}_{\min}, \end{cases} \tag{2.46}$$
$$\partial_{\boldsymbol{n}} p_{\min}^{\max} = 0 \quad \text{on } \Gamma.$$

Remarks:
Although the set optimal control problem (2.30) and the shape/topology optimization problem look similar, there is an essential discrepancy: The boundary value problem (2.45d)–(2.45i) is uniquely solvable for any given $\mathcal{B} \in \mathcal{O}$, whereas (2.30h)–(2.30l) is not. Consequently, the set optimal control problem requires optimization with respect to the function space variables, whereas its reduced counterpart does not.

The strict inequality constraint (2.45c) plays the role of a constraint here, which influences the admissibility of the geometrical splitting of $\Omega = \mathcal{B} \dot\cup \mathcal{J}$. That is to say, the constraint is a state constraint in shape/topology optimization. Moreover, in view of the discussion of Paragraph 2.2.4, it is expected that is has no effect on first order necessary conditions. This actually turns out to be true in Paragraph 2.3.7.

Proof. The set-OCP (2.30) is equivalent to the bilevel optimization problem (2.36), (2.37) according to Theorem 4. By means of strict convexity of the inner optimization problem (2.37) – confer the proof of Theorem 3 – its first order necessary conditions (2.42) are sufficient, too. Hence, the inner optimization problem can equivalently be replaced by its optimality system within the bilevel optimization problem. However, one is only interested in the optimal primal variables $\bar u_{\mathcal{J}}, \bar u_{\mathcal{B}}, \bar y_{\mathcal{J}}, \bar y_{\mathcal{B}}$ and not in the dual ones. Consequently, it is sufficient to execute the first three solving steps of Lemma 7:

1. assign $\bar y_{\mathcal{B}} = y_{\min}^{\max}$,
2. solve (2.44), i. e. solve (2.45d)–(2.45i)
3. assign $\bar u_{\mathcal{J}} = -\frac{1}{\lambda}\bar p_{\mathcal{J}} + u_d$ and $\bar u_{\mathcal{B}} = -\Delta y_{\min}^{\max} + y_{\min}^{\max} = -\frac{1}{\lambda} p_{\min}^{\max} + u_d$.

Plugging these results into the objective $\mathfrak{J}$ in (2.30a) yields (2.45a)

$$\mathcal{F}(\mathcal{B}) := \mathfrak{J}(\mathcal{B}; \bar u_{\mathcal{J}}, \bar u_{\mathcal{B}}, \bar y_{\mathcal{J}}, \bar y_{\mathcal{B}}) = \mathfrak{J}(\mathcal{B}; -\frac{1}{\lambda}\bar p_{\mathcal{J}} + u_d, -\frac{1}{\lambda} p_{\min}^{\max} + u_d, \bar y_{\mathcal{J}}, y_{\min}^{\max}).$$

All in all, one has reached the reduced reformulation (2.45). □

As already mentioned, the scope of this paragraph shall be to execute the second step of the general recipe of Paragraph 2.3.2. The first part therein is to prove differentiability of the control-to-state operator S, which means to prove shape differentiability of the equality constraints in the present context. However, differentiability of the constraints could not yet be proven (see Appendix C for a more detailed discussion), and has to be assumed here.

Lemma 8 (Shape derivative of the constraints)**:**
Let the family of admissible sets $\mathcal{O}$ be given by Definition 4. Let $\mathcal{B} \in \mathcal{O}$ be given, such that the solution $(\bar y_{\mathcal{J}}, \bar p_{\mathcal{J}})$ of (2.45d)–(2.45j) lies in $H^2(\mathcal{J}) \times H^1(\mathcal{J})$.[12] Moreover, let $p_{\min}^{\max}$ be defined as in (2.46). Additionally, assume that the boundary value problem (2.45d)–(2.45j) is shape differentiable.

Then for each

$$V \in \mathcal{V} := \{W \in C^{1,1}(\Omega, \mathbb{R}^2) \,|\, W \cdot \boldsymbol{n} = 0 \text{ on } \Gamma\} \tag{2.47}$$

the (local[13]) shape derivatives $y'_{\mathcal{J}}[V] \in H^2(\mathcal{J})$ and $p'_{\mathcal{J}}[V] \in L^2(\mathcal{J}, \Delta)$ are given as the unique solution of the boundary value problem

$$-\Delta y'_{\mathcal{J}}[V] + y'_{\mathcal{J}}[V] = -\frac{1}{\lambda} p'_{\mathcal{J}}[V] \quad \text{in } \mathcal{J}, \tag{2.48a}$$
$$\partial_{\boldsymbol{n}} y'_{\mathcal{J}}[V] = 0 \quad \text{on } \Gamma, \tag{2.48b}$$
$$y'_{\mathcal{J}}[V]|_{\beta} = 0 \quad \text{on } \beta, \tag{2.48c}$$
$$\partial_{\boldsymbol{n}}^{\mathcal{J}} y'_{\mathcal{J}}[V] = V \cdot \boldsymbol{n}_{\mathcal{J}} \frac{1}{\lambda}(p_{\min}^{\max}|_{\beta} - \bar p_{\mathcal{J}}|_{\beta}) \text{ on } \beta, \tag{2.48d}$$
$$-\Delta p'_{\mathcal{J}}[V] + p'_{\mathcal{J}}[V] = y'_{\mathcal{J}}[V] \quad \text{in } \mathcal{J}, \tag{2.48e}$$
$$\partial_{\boldsymbol{n}} p'_{\mathcal{J}}[V] = 0 \quad \text{on } \Gamma. \tag{2.48f}$$

Remark:

1. The definition of the space of velocity fields $\mathcal{V}$ is advisedly chosen:
   - it ensures that the *holdall* Ω remains unchanged under the action of V;
   - $C^{1,1}$ regularity of the transported candidate active set $\mathcal{B}_t$ is preserved; see Paragraph 2.6.1;
   - the regularity assumptions (V) (cf. [44, Chp. 4 Eq. (5.5)]), required for the definition Hadamard differentiability (cf. [44, Chp. 9, Def. 3.1]), which is the basis for the definition of shape differentiability (cf. [44, Chp. 9, Def. 3.4]), are fulfilled. In particular, attend to [44, Chp. 4, Rem. 5.2 and the introduction to Sec. 5.2].
2. As already mentioned in the 5th item of the Remarks on page 36 the low regularity of $\bar{p}_{\mathcal{J}}$ is crucial. In particular, it has not yet been possible to prove Lemma 8 without the additional regularity assumption at $\bar{p}_{\mathcal{J}}$. From the perspective of necessary conditions of the set-OCP (2.30) the assumption is without problems, since it is fulfilled at the optimum. However, from an algorithmic stand point, the assumption made may be a true restriction, since the optimality system of iOP and its local shape derivative system (2.48) have to be solved at non-optimal configurations as well; see Algorithm 1.
3. The notation $(.)'[V]$ is used here for the local shape derivative. The explicit usage of the velocity field V indicates, that this object is a semiderivative, and hence requires a "direction".

Proof. The proof consists of two parts. Firstly, it is shown, that the shape derivatives $y'[V]$ and $p'_{\mathcal{J}}[V]$ are solutions to the coupled BVP (2.48). Afterwards unique solvability in of the system is provided.

1) Since each component – except the Neumann interface condition – of the BVP (2.45d)–(2.45j) is pretty much standard, the reader is referred to the rules for shape differentiation of boundary value problems [147, Lem. 14, Lem. 15] or [151, Prop. 3.1, Prop 3.3]. For convenience the derivation of the non-standard Neumann boundary condition (2.45g) is given here. Its special character is, that although the function $y_{\min}^{\max}$ does not depend on the choice of $\mathcal{B}$ locally (cf. the Remark to Lemma 4), its normal derivative $\partial_{\boldsymbol{n}}^{\mathcal{J}} y_{\min}^{\max}$ does, indeed.

Before the derivation of the shape derivative can be addressed, it is useful to notice the following finding: In contrast to the Neumann trace operator $\partial_{\boldsymbol{n}}^{\mathcal{J}}(.)$, the *tangential gradient* $\nabla_\beta(.)$ and the *Laplace-Beltrami operator* $\Delta_\beta(.)$ are directly acting on the submanifold $\beta \subset \mathbb{R}^2$. That is, they act on the image space of the Dirichlet trace operator $\tau_\beta(.) = (.)|_\beta$. Consequently, there holds

$$\phi \in H^2(\mathcal{J}) \text{ with } \phi|_\beta \equiv 0 \quad \text{on } \beta \quad \Rightarrow \quad \nabla_\beta \phi \equiv 0 \quad \text{and} \quad \Delta_\beta \phi \equiv 0,$$

whereas

$$\phi \in H^2(\mathcal{J}) \text{ with } \phi|_\beta \equiv 0 \quad \text{on } \beta \quad \not\Rightarrow \quad \partial_{\boldsymbol{n}}^{\mathcal{J}} \phi = 0.$$

Transferred to the Dirichlet boundary condition (2.45f), this yields

$$\nabla_\beta(\bar{y}_{\mathcal{J}} - y_{\min}^{\max}) \equiv 0 \quad \text{and} \quad \Delta_\beta(\bar{y}_{\mathcal{J}} - y_{\min}^{\max}) \equiv 0. \tag{2.49}$$

According to [147, Lem. 15], which provides the derivative of Neumann boundary conditions, and with use of the notation $\partial_{\boldsymbol{nn}}$ for the binormal derivative (cf. Definition 2), there holds

$$\begin{aligned}
\partial_{\boldsymbol{n}}^{\mathcal{J}}(y'_{\mathcal{J}}[V] - \underbrace{y_{\min}^{\max\prime}[V]}_{=0}) &= -V\cdot\boldsymbol{n}_{\mathcal{J}}\,\partial_{\boldsymbol{nn}}(\bar{y}_{\mathcal{J}} - y_{\min}^{\max}) + \underbrace{\nabla_\beta(\bar{y}_{\mathcal{J}} - y_{\min}^{\max})}_{=0}\cdot\nabla_\beta(V\cdot\boldsymbol{n}_{\mathcal{J}}) \\
&= V\cdot\boldsymbol{n}_{\mathcal{J}}\big(\Delta(y_{\min}^{\max} - \bar{y}_{\mathcal{J}})|_\beta - \underbrace{\Delta_\beta(y_{\min}^{\max} - \bar{y}_{\mathcal{J}})}_{=0} - \underbrace{\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\min}^{\max} - \bar{y}_{\mathcal{J}})}_{=0}\kappa_{\mathcal{J}}\big) \\
&= V\cdot\boldsymbol{n}_{\mathcal{J}}\Big(\Delta y_{\min}^{\max} - \underbrace{\bar{y}_{\mathcal{J}}}_{=y_{\min}^{\max}} + u_d - \frac{1}{\lambda}\bar{p}_{\mathcal{J}}\Big)|_\beta \\
&= V\cdot\boldsymbol{n}_{\mathcal{J}}\frac{1}{\lambda}(p_{\min}^{\max}|_\beta - \bar{p}_{\mathcal{J}}|_\beta),
\end{aligned} \tag{2.50}$$

[12] Actually, this condition is fulfilled for the active set $\mathcal{B} = \mathcal{A}$ at least, see Corollary 2. Note in addition, that higher regularity at the optimum can be proven without knowledge of shape differentiability, since it only relies on weak continuity of the optimal control across the optimal interface γ.

[13] A detailed background to this notion can be found in Paragraph 2.4.2.

where (2.45d), (2.49) and the identity [151, Prop. 2.68]

$$\partial_{nn}(.) = \Delta(.)|_\beta - \Delta_\beta(.) - \partial_{\boldsymbol{n}}^{\mathcal{J}}(.)\,\kappa_{\mathcal{J}}. \tag{2.51}$$

are applied. Here $\kappa_{\mathcal{J}}$ denotes the *mean curvature* of β, where β is interpreted as boundary of $\mathcal{J}$; cf. [44, p. 74][14].

2) Unique solvability of the BVP is ensured by the following reasoning. Regard the auxiliary strictly convex optimization problem

$$\begin{aligned} &\text{minimize} \quad f(u') := \int_{\mathcal{J}} \frac{1}{2} S(u')^2 + \frac{\lambda}{2}(u')^2 \\ &\text{subject to} \quad u' \in U := \{u' \in L^2(\mathcal{J}) \,|\, \partial_{\boldsymbol{n}}^{\mathcal{J}} S(u') = \frac{1}{\lambda} V\cdot \boldsymbol{n}_{\mathcal{J}} (p_{\min}^{\max} - \bar{p}_{\mathcal{J}}) \text{ on } \beta\}, \end{aligned}$$

where $S : L^2(\mathcal{J}) \to H^2(\mathcal{J})$ is the solution operator of the boundary value problem

$$\begin{aligned} -\Delta y' + y' &= u' \quad \text{in } \mathcal{J}, \\ \partial_{\boldsymbol{n}} y' &= 0 \quad \text{on } \Gamma, \\ y' &= 0 \quad \text{on } \beta. \end{aligned}$$

Since $V \in \mathcal{V}$, and since the normal vector field $\boldsymbol{n}_{\mathcal{J}}$ is Lipschitzian (cf. Definition 2), their scalar product $V\cdot\boldsymbol{n}_{\mathcal{J}}$ is Lipschitz continuous, too. In consequence of [69, Thm. 1.4.1.1] and of $\bar{p}_{\mathcal{J}} \in H^1(\mathcal{J})$, there holds

$$\frac{1}{\lambda} V\cdot\boldsymbol{n}_{\mathcal{J}} (p_{\min}^{\max} - \bar{p}_{\mathcal{J}}) \in H^1(\Omega).$$

Furthermore, the right hand side of the inhomogeneous Neumann boundary condition on β within the definition of U is an element of $H^{1/2}(\beta)$. Lemma 1 ensures U to be nonempty now.

The auxiliary optimization problem is uniquely solvable, which is obtained with the same proof as that of Theorem 3. Furthermore, with the same reasoning as in the proofs of Theorem 5 and Lemma 7, one recognizes that the BVP (2.48) can be interpreted as the reduced first order necessary and sufficient conditions of the auxiliary problem. Unique solvability of (2.48) is a consequence of unique solvability of the auxiliary problem now. □

Lemma 8 offers the opportunity to derive the shape derivative of the reduced functional $\mathcal{F}$.

Lemma 9 (Shape differentiability of $\mathcal{F}$)**:**
Let the family of admissible sets $\mathcal{O}$ be given by Definition 4 and let $\mathcal{B} \in \mathcal{O}$ such that the assumption of Lemma 8 are fulfilled. Furthermore, let $V \in \mathcal{V}$ – see (2.47) – be arbitrarily chosen.

Then the *shape semiderivative* of the reduced objective $\mathcal{F}$ – see (2.45a) – in the direction V is given by

$$\begin{aligned} d\mathcal{F}(\mathcal{B};V) = &\int_{\mathcal{J}} (\bar{y}_{\mathcal{J}} - y_d)\, y'_{\mathcal{J}}[V] + \int_\beta \frac{1}{2}(\bar{y}_{\mathcal{J}} - y_d)^2\, V\cdot\boldsymbol{n}_{\mathcal{J}} + \int_\beta \frac{1}{2}(y_{\min}^{\max} - y_d)^2\, V\cdot\boldsymbol{n}_{\mathcal{B}} \\ &+ \int_{\mathcal{J}} \frac{1}{\lambda}\bar{p}_{\mathcal{J}}\, p'_{\mathcal{J}}[V] + \int_\beta \frac{1}{2\lambda}\bar{p}_{\mathcal{J}}^2\, V\cdot\boldsymbol{n}_{\mathcal{J}} + \int_\beta \frac{1}{2\lambda} p_{\min}^{\max 2}\, V\cdot\boldsymbol{n}_{\mathcal{B}}\,, \end{aligned} \tag{2.52}$$

where $y'_{\mathcal{J}}[V] \in H^2(\mathcal{J})$ and $p'_{\mathcal{J}}[V] \in L^2(\mathcal{J},\Delta)$ are the unique solutions of the BVP (2.48).

Proof. Due to the rules of shape calculus (cf. [151, Eq. (2.168)]), the first summand can be differentiated. Since y_d is not dependent on the shape $\mathcal{J}$, since $y'_{\mathcal{J}}[V] \in H^2(\mathcal{J})$ is well-defined (see Lemma 8) and since $V \in \mathcal{V}$, one has

$$\begin{aligned} d\left(\frac{1}{2}\|\bar{y}_{\mathcal{J}} - y_d\|^2_{L^2(\mathcal{J})}; V\right) &= \int_{\mathcal{J}} \left(\frac{1}{2}(\bar{y}_{\mathcal{J}} - y_d)^2\right)'[V] + \int_{\partial\mathcal{J}} \frac{1}{2}(\bar{y}_{\mathcal{J}} - y_d)^2\, V\cdot\boldsymbol{n} \\ &= \int_{\mathcal{J}} (\bar{y}_{\mathcal{J}} - y_d)\,(y'_{\mathcal{J}}[V] - \underbrace{y'_d[V]}_{=0}) \\ &\quad + \int_\beta \frac{1}{2}(\bar{y}_{\mathcal{J}} - y_d)^2\, V\cdot\boldsymbol{n}_{\mathcal{J}} + \int_\Gamma \frac{1}{2}(\bar{y}_{\mathcal{J}} - y_d)^2 \underbrace{V\cdot\boldsymbol{n}}_{=0}. \end{aligned}$$

[14] Note, that the sign of the mean curvature depends on the choice of the orientation of the normal vector field of the boundary. Hence, since $\boldsymbol{n}_{\mathcal{B}} = -\boldsymbol{n}_{\mathcal{J}}$, there holds $\kappa_{\mathcal{B}} = -\kappa_{\mathcal{J}}$.

The second summand of $\mathcal{F}$ yields

$$d\left(\frac{1}{2}\|y_{\min}^{\max}-y_d\|^2_{L^2(\mathcal{J})};V\right) = \int_{\mathcal{B}}\left(\frac{1}{2}(y_{\min}^{\max}-y_d)^2\right)'[V] + \int_{\partial\mathcal{B}}\frac{1}{2}(y_{\min}^{\max}-y_d)^2\,V\cdot\boldsymbol{n}$$
$$= \int_{\mathcal{B}}(y_{\min}^{\max}-y_d)\,(\underbrace{y_{\min}^{\max\prime}[V]}_{=0}-\underbrace{y_d'[V]}_{=0}) + \int_{\beta}\frac{1}{2}(y_{\min}^{\max}-y_d)^2\,V\cdot\boldsymbol{n}_{\mathcal{B}},$$

since neither y_d nor $y_{\min}^{\max}$ is dependent on the shape $\mathcal{B}$ (at least locally; cf. the Remark of Lemma 4). The analog results for the remaining two terms of the sum lead to (2.52). □

In view of the general recipe to derive first order necessary conditions in Paragraph 2.3.2, a closer look at the representation (2.52) of the shape semiderivative reveals, that the Hadamard form (cf. [44, Chp. 9 Thm. 3.6, Chp. 9 Cor. 1]) – i. e. a gradient representation – has not yet been obtained. Generally speaking, it is necessary to identify the adjoint operator of the derivative of the geometry-to-solution operator, such that adjoint states can be derived; see (2.34).

Lemma 10 (L^1-shape gradient of $\mathcal{F}$)**:**
Let the family of admissible sets $\mathcal{O}$ be given by Definition 4, let $\mathcal{B}\in\mathcal{O}$ be chosen such that the assumptions of Lemma 8 are fulfilled and let $p_{\min}^{\max}$ be defined as in (2.46). Furthermore, let the *shape adjoint states* $Y_{\mathcal{J}}$ and $P_{\mathcal{J}}$ be the unique solution to the *shape adjoint equation*

$$-\Delta Y_{\mathcal{J}} + Y_{\mathcal{J}} + \frac{1}{\lambda}P_{\mathcal{J}} = \frac{1}{\lambda}\bar{p}_{\mathcal{J}} \quad \text{in } \mathcal{J}, \tag{2.53a}$$
$$\partial_{\boldsymbol{n}} Y_{\mathcal{J}} = 0 \quad \text{on } \Gamma, \tag{2.53b}$$
$$Y_{\mathcal{J}}|_{\beta} = 0 \quad \text{on } \beta, \tag{2.53c}$$
$$\partial_{\boldsymbol{n}}^{\mathcal{J}} Y_{\mathcal{J}} = 0 \quad \text{on } \beta, \tag{2.53d}$$
$$-\Delta P_{\mathcal{J}} + P_{\mathcal{J}} - Y_{\mathcal{J}} = \bar{y}_{\mathcal{J}} - y_d \quad \text{in } \mathcal{J}, \tag{2.53e}$$
$$\partial_{\boldsymbol{n}} P_{\mathcal{J}} = 0 \quad \text{on } \Gamma, \tag{2.53f}$$
$$Y_{\mathcal{J}}\in H^2(\mathcal{J}),\ P_{\mathcal{J}}\in L^2(\mathcal{J},\Delta). \tag{2.53g}$$

Then the shape semiderivative of the reduced objective $\mathcal{F}$ – see (2.45a) – evaluated at the set $\mathcal{B}$ in the direction $V\in\mathcal{V}$ can be expressed as

$$d\mathcal{F}(\mathcal{B};V) = \int_{\beta}\left(\frac{1}{\lambda}(p_{\min}^{\max}-\bar{p}_{\mathcal{J}})P_{\mathcal{J}} - \frac{1}{2\lambda}(p_{\min}^{\max 2}-\bar{p}_{\mathcal{J}}^2)\right)V\cdot\boldsymbol{n}_{\mathcal{J}}.$$

Thus, the (L^1-) *shape gradient* can be identified with

$$\nabla\mathcal{F}(\mathcal{B}) = \frac{1}{\lambda}(p_{\min}^{\max}|_{\beta}-\bar{p}_{\mathcal{J}}|_{\beta})P_{\mathcal{J}}|_{\beta} - \frac{1}{2\lambda}(p_{\min}^{\max 2}|_{\beta}-\bar{p}_{\mathcal{J}}^2|_{\beta}) \in L^1(\beta). \tag{2.54}$$

In other words, one has

$$d\mathcal{F}(\mathcal{B};V) = \left\langle\nabla\mathcal{F}(\mathcal{B})\ ,\ V|_{\beta}\cdot\boldsymbol{n}_{\mathcal{J}}\right\rangle_{C^{0,1}(\beta)^*,C^{0,1}(\beta)} = \int_{\beta}\nabla\mathcal{F}(\mathcal{B})V\cdot\boldsymbol{n}_{\mathcal{J}}.$$

In respect of higher regularity of the shape gradient confer the Remark to Theorem 7.

Remark:
In defiance of the original usage of the notion of the shape gradient [44, Chp. 9 Def. 3.4 and Thm. 3.6], the associated but strictly speaking distinguished scalar distribution (2.54) (cf. [44, Chp. 9 Cor. 1]) is called shape gradient in the following. See also 15th item of the discussion on page 77.

Proof. **1)** This part concerns the unique solvability of the shape adjoint system (2.53).
Since it has the same form as (2.48), the assertion follows along the lines of the second part of the proof of Lemma 8, in which one uses the auxiliary strictly convex optimization problem

$$\begin{aligned}\text{minimize}\quad & f(U) := \int_{\mathcal{J}}\frac{1}{2}(S(U)-(y_d+\bar{y}_{\mathcal{J}}))^2 + \frac{\lambda}{2}(U-\frac{1}{\lambda}\bar{p}_{\mathcal{J}})^2\\ \text{subject to}\quad & U\in\{U\in L^2(\mathcal{J})\,|\,\partial_{\boldsymbol{n}}^{\mathcal{J}}S(U)=0 \text{ on } \beta\},\end{aligned}$$

where $S : L^2(\mathcal{J}) \to H^2(\mathcal{J})$ is the solution operator for the boundary value problem

$$\begin{aligned} -\Delta Y + Y &= U && \text{in } \mathcal{J}, \\ \partial_{\boldsymbol{n}} Y &= 0 && \text{on } \Gamma, \\ Y &= 0 && \text{on } \beta. \end{aligned}$$

2) Let $V \in \mathcal{V}$ be arbitrary, but fixed. One recognizes that the shape semiderivative of the reduced objective can be transformed into Hadamard form by means of the shape adjoint system (2.53) and the local shape derivative BVP (2.48)

$$\begin{aligned} d\mathcal{F}(\mathcal{B}; V) &= \int_{\mathcal{J}} (\bar{y}_{\mathcal{J}} - y_d)\, y'_{\mathcal{J}}[V] + \int_\beta \frac{1}{2}(\bar{y}_{\mathcal{J}} - y_d)^2\, V \cdot \boldsymbol{n}_{\mathcal{J}} + \int_\beta \frac{1}{2}(y_{\min}^{\max} - y_d)^2\, V \cdot \underbrace{\boldsymbol{n}_{\mathcal{B}}}_{=-\boldsymbol{n}_{\mathcal{J}}} \\ &\quad + \int_{\mathcal{J}} \frac{1}{\lambda} \bar{p}_{\mathcal{J}}\, p'_{\mathcal{J}}[V] + \int_\beta \frac{1}{2\lambda} \bar{p}_{\mathcal{J}}^2\, V \cdot \boldsymbol{n}_{\mathcal{J}} + \int_\beta \frac{1}{2\lambda} p_{\min}^{\max 2}\, V \cdot \underbrace{\boldsymbol{n}_{\mathcal{B}}}_{=-\boldsymbol{n}_{\mathcal{J}}} \\ &= \int_{\mathcal{J}} (-\Delta P_{\mathcal{J}} + P_{\mathcal{J}} - Y_{\mathcal{J}})\, y'_{\mathcal{J}}[V] + \int_{\mathcal{J}} (-\Delta Y_{\mathcal{J}} + Y_{\mathcal{J}} + \frac{1}{\lambda} P_{\mathcal{J}})\, p'_{\mathcal{J}}[V] \\ &\quad + \int_\beta \Big(\underbrace{\frac{1}{2}(\bar{y}_{\mathcal{J}} - y_d)^2 - \frac{1}{2}(y_{\min}^{\max} - y_d)^2}_{=0} + \frac{1}{2\lambda}(\bar{p}_{\mathcal{J}}^2 - p_{\min}^{\max 2}) \Big) V \cdot \boldsymbol{n}_{\mathcal{J}} \\ &= \int_{\mathcal{J}} \Big(\underbrace{-\Delta y'_{\mathcal{J}}[V] + y'_{\mathcal{J}}[V] + \frac{1}{\lambda} p'_{\mathcal{J}}[V]}_{=0} \Big) P_{\mathcal{J}} + \Big(\underbrace{-\Delta p'_{\mathcal{J}}[V] + p'_{\mathcal{J}}[V] - y'_{\mathcal{J}}[V]}_{=0} \Big) Y_{\mathcal{J}} \\ &\quad + \int_\Gamma \underbrace{-\partial_{\boldsymbol{n}} P_{\mathcal{J}}}_{=0}\, y'_{\mathcal{J}}[V] + P_{\mathcal{J}} \underbrace{\partial_{\boldsymbol{n}} y'_{\mathcal{J}}[V]}_{=0} \underbrace{-\partial_{\boldsymbol{n}} Y_{\mathcal{J}}}_{=0}\, p'_{\mathcal{J}}[V] + Y_{\mathcal{J}} \underbrace{\partial_{\boldsymbol{n}} p'_{\mathcal{J}}[V]}_{=0} \\ &\quad + \int_\beta -\partial_{\boldsymbol{n}}^{\mathcal{J}} P_{\mathcal{J}} \underbrace{y'_{\mathcal{J}}[V]}_{=0} + P_{\mathcal{J}} \underbrace{\partial_{\boldsymbol{n}}^{\mathcal{J}} y'_{\mathcal{J}}[V]}_{=V\cdot\boldsymbol{n}_{\mathcal{J}} \frac{1}{\lambda}(p_{\min}^{\max} - \bar{p}_{\mathcal{J}})} \overbrace{-\partial_{\boldsymbol{n}}^{\mathcal{J}} Y_{\mathcal{J}}}^{=0}\, p'_{\mathcal{J}}[V] + \underbrace{Y_{\mathcal{J}}}_{=0} \partial_{\boldsymbol{n}}^{\mathcal{J}} p'_{\mathcal{J}}[V] + \frac{1}{2\lambda} \left(\bar{p}_{\mathcal{J}}^2 - p_{\min}^{\max 2} \right) V \cdot \boldsymbol{n}_{\mathcal{J}} \\ &= \int_\beta \Big(\frac{1}{\lambda}(p_{\min}^{\max} - \bar{p}_{\mathcal{J}}) P_{\mathcal{J}} - \frac{1}{2\lambda}(p_{\min}^{\max 2} - \bar{p}_{\mathcal{J}}^2) \Big) V \cdot \boldsymbol{n}_{\mathcal{J}}. \end{aligned}$$

Herefrom, the shape gradient (2.54) can be identified by means of the fundamental lemma of calculus of variations. □

Remark (Constructive heuristic to derive the shape adjoint system)**:**
The shape adjoint boundary value problem (2.53) can be obtained constructively by means of the following heuristic:

- Multiply the homogeneous PDEs (2.48a) and (2.48e), which define the shape derivatives $y'_{\mathcal{J}}[V]$ and $p'_{\mathcal{J}}[V]$, by $P_{\mathcal{J}}$ and respectively $Y_{\mathcal{J}}$, integrate and add the two terms to $d\mathcal{F}(\mathcal{B}; V)$ (2.52).
- Perform an integration by parts and define $Y_{\mathcal{J}}$ and $P_{\mathcal{J}}$, such that all contributions of $y'_{\mathcal{J}}[V]$ and $p'_{\mathcal{J}}[V]$ vanish.

This yields

$$\begin{aligned} d\mathcal{F}(\mathcal{B}; V) &= \int_{\mathcal{J}} (\bar{y}_{\mathcal{J}} - y_d)\, y'_{\mathcal{J}}[V] + \int_\beta \frac{1}{2}(\bar{y}_{\mathcal{J}} - y_d)^2\, V \cdot \boldsymbol{n}_{\mathcal{J}} + \int_\beta \frac{1}{2}(y_{\min}^{\max} - y_d)^2\, V \cdot \boldsymbol{n}_{\mathcal{B}} \\ &\quad + \int_{\mathcal{J}} \frac{1}{\lambda} \bar{p}_{\mathcal{J}}\, p'_{\mathcal{J}}[V] + \int_\beta \frac{1}{2\lambda} \bar{p}_{\mathcal{J}}^2\, V \cdot \boldsymbol{n}_{\mathcal{J}} + \int_\beta \frac{1}{2\lambda} p_{\min}^{\max 2}\, V \cdot \boldsymbol{n}_{\mathcal{B}} \\ &\quad - \int_{\mathcal{J}} (-\Delta y'_{\mathcal{J}}[V] + y'_{\mathcal{J}}[V] + \frac{1}{\lambda} p'_{\mathcal{J}}[V]) P_{\mathcal{J}} - \int_{\mathcal{J}} (-\Delta p'_{\mathcal{J}}[V] + p'_{\mathcal{J}}[V] - y'_{\mathcal{J}}[V]) Y_{\mathcal{J}} \\ &= \int_{\mathcal{J}} \underbrace{(\bar{y}_{\mathcal{J}} - y_d + \Delta P_{\mathcal{J}} - P_{\mathcal{J}} + Y_{\mathcal{J}})}_{\Rightarrow \text{(2.53e)}} y'_{\mathcal{J}}[V] + \int_{\mathcal{J}} \underbrace{(\frac{1}{\lambda} \bar{p}_{\mathcal{J}} + \Delta Y_{\mathcal{J}} - Y_{\mathcal{J}} - \frac{1}{\lambda} P_{\mathcal{J}})}_{\Rightarrow \text{(2.53a)}} p'_{\mathcal{J}}[V] \end{aligned}$$

$$
\begin{aligned}
&+ \int_\Gamma \underbrace{\partial_{\boldsymbol{n}} y'_{\mathcal{J}}[V]}_{=0} P_{\mathcal{J}} - y'_{\mathcal{J}}[V] \underbrace{\partial_{\boldsymbol{n}} P_{\mathcal{J}}}_{\Rightarrow (2.53f)} + \underbrace{\partial_{\boldsymbol{n}} p'_{\mathcal{J}}[V]}_{=0} Y_{\mathcal{J}} - p'_{\mathcal{J}}[V] \underbrace{\partial_{\boldsymbol{n}} Y_{\mathcal{J}}}_{\Rightarrow (2.53b)} \\
&+ \int_\beta \underbrace{\partial^{\mathcal{J}}_{\boldsymbol{n}} y'_{\mathcal{J}}[V]}_{=V\cdot \boldsymbol{n}_{\mathcal{J}} \frac{1}{\lambda}(\bar{p}_{\mathcal{J}} - p^{\max}_{\min})} P_{\mathcal{J}} - \underbrace{y'_{\mathcal{J}}[V]}_{=0} \partial^{\mathcal{J}}_{\boldsymbol{n}} P_{\mathcal{J}} + \partial^{\mathcal{J}}_{\boldsymbol{n}} p'_{\mathcal{J}}[V] \underbrace{Y_{\mathcal{J}}}_{\Rightarrow (2.53c)} - p'_{\mathcal{J}}[V] \underbrace{\partial^{\mathcal{J}}_{\boldsymbol{n}} Y_{\mathcal{J}}}_{\Rightarrow (2.53d)} \\
&+ \int_\beta \frac{1}{2}\Big((\bar{y}_{\mathcal{J}} - y_d)^2 - (y^{\max}_{\min} - y_d)^2 + \frac{1}{\lambda}\bar{p}^2_{\mathcal{J}} - \frac{1}{\lambda} p^{\max 2}_{\min} \Big) V \cdot \boldsymbol{n}_{\mathcal{J}}.
\end{aligned}
$$

The Hadamard form is obtained in Lemma 10 at the price of solving the shape adjoint boundary value problem (2.53). This drawback can be overcome.

Theorem 7 (Shape gradient of $\mathcal{F}$ without shape adjoints)**:**
Let the family of admissible sets $\mathcal{O}$ be given by Definition 4 and let $\mathcal{B} \in \mathcal{O}$ be chosen such that the assumptions of Lemma 8 are fulfilled.

Then the shape gradient of the reduced objective $\mathcal{F}$ – see (2.45a) – evaluated at the set $\mathcal{B}$ has a representation as

$$
\nabla \mathcal{F}(\mathcal{B}) = -\frac{1}{2\lambda}(p^{\max}_{\min}|_\beta - \bar{p}_{\mathcal{J}}|_\beta)^2 \in L^1(\beta), \tag{2.55}
$$

where $p^{\max}_{\min}$ is defined in (2.46) and $\bar{p}_{\mathcal{J}}$ is given by (2.45d)–(2.45j). In particular, the shape gradient comes without shape adjoint variables.

Proof. Lemma 10 ensures, that the shape adjoint system (2.53) is uniquely solvable. A closer look reveals, that the unique solution to the shape adjoint system is given by $P_{\mathcal{J}} = \bar{p}_{\mathcal{J}}$ and $Y_{\mathcal{J}} = 0$. Thus, one finally obtains (2.55)

$$
\begin{aligned}
\nabla \mathcal{F}(\mathcal{B}) &= \Big(\frac{1}{\lambda}(p^{\max}_{\min}|_\beta - \bar{p}_{\mathcal{J}}|_\beta) P_{\mathcal{J}}|_\beta - \frac{1}{2\lambda}(p^{\max 2}_{\min}|_\beta - \bar{p}^2_{\mathcal{J}}|_\beta) \Big) \\
&= \Big(\frac{1}{\lambda}(p^{\max}_{\min}|_\beta - \bar{p}_{\mathcal{J}}|_\beta) \bar{p}_{\mathcal{J}}|_\beta - \frac{1}{2\lambda}(p^{\max 2}_{\min}|_\beta - \bar{p}^2_{\mathcal{J}}|_\beta) \Big) \\
&= -\frac{1}{2\lambda}\Big(p^{\max 2}_{\min}|_\beta - 2 p^{\max}_{\min}|_\beta \, \bar{p}_{\mathcal{J}}|_\beta + \bar{p}^2_{\mathcal{J}}|_\beta \Big).
\end{aligned}
$$

□

Remark:

- If it is possible to ensure that $\bar{p}_{\mathcal{J}}|_\beta \in H^{3/2}(\beta)$ for a given $\mathcal{J}$, the regularity of the shape gradient improves to $H^{3/2}(\beta)$. This is due to [69, Thm. 1.4.4.2], which states, that $H^2(\mathcal{J})$ is an algebra, i. e. the product of two H^2 functions are H^2 regular, as well. And this fact can be carried over to the trace spaces. Indeed, Corollary 2 ensures, that $\bar{p}_{\mathcal{I}} \in H^2(\mathcal{I})$ and thus the shape gradient is $H^{3/2}$-regular at least the optimum.
- As a result, one can identify a so called *Sobolev gradient* of the reduced objective $\mathcal{F}$ (see [130], [146, Sec. 5.3]). Hereunto, let $\nabla \mathcal{F}(\mathcal{B}) \in L^2(\beta)$ and consider the variational problem

$$
\begin{aligned}
\int_\beta \nabla_\beta(\nabla_S \mathcal{F}(\mathcal{B})) \cdot \nabla_\beta \varphi + \nabla_S \mathcal{F}(\mathcal{B})\, \varphi &= \int_\beta \nabla \mathcal{F}(\mathcal{B})\, \varphi \qquad \forall \varphi \in H^1(\beta), \\
\Leftrightarrow \quad (\nabla_S \mathcal{F}(\mathcal{B})\,,\, \varphi)_{H^1(\beta)} &= (\nabla \mathcal{F}(\mathcal{B})\,,\, \varphi)_{L^2(\beta)} \quad \forall \varphi \in H^1(\beta),
\end{aligned}
$$

  which is associated with the *surface PDE* problem of the Laplace-Beltrami operator

$$
-\Delta_\beta(\nabla_S \mathcal{F}(\mathcal{B})) + \nabla_S \mathcal{F}(\mathcal{B}) = \nabla \mathcal{F}(\mathcal{B}) \quad \text{a. e. on } \beta.
$$

  This surface PDE is known to be uniquely solvable, cf. for instance [49].
- Sobolev gradients have the appealing property that the operator which maps $\nabla \mathcal{F}$ to $\nabla_S \mathcal{F}$ may be used for preconditioning in steepest descent algorithms, cf. for instance [146, 56].

The assertion of Theorem 7 is significantly affected by the observance that the shape adjoint variables are either zero or already given by the adjoint state of the inner optimization problem. This is a special case of a more general result concerning bilevel optimization problems with the following structure:

$$\begin{aligned} &\text{minimize} && J(b;u_b,y_b) \\ &\text{subject to} && b \in \mathbb{M} \subset \mathbb{B} \\ & && (u_b,y_b) := \operatorname*{arg\,min}_{(u,y)\in\mathbb{U}\times\mathbb{Y}} J(b;u,y) \text{ subject to } T(b;u,y)=0 \text{ in } \mathbb{Z}. \end{aligned} \tag{2.56}$$

It is assumed here, that there exists operators

$$\begin{aligned} &S_b : \mathbb{U} \to \mathbb{Y}, \quad u \mapsto y = S_b(u) \text{ with } T(b;u,S_b(u)) = 0, \quad \forall b \in \mathbb{B}, \\ &G : \mathbb{B} \to \mathbb{U}\times\mathbb{Y}, \quad b \mapsto (u_b,y_b), \end{aligned}$$

and that all operators are sufficiently smooth for the following analysis. In particular, the parametrized inner optimization problem is uniquely solvable for each parameter $b \in \mathbb{B}$. In other words, the bilevel optimization problem can equivalently be formulated as

$$\begin{aligned} &\text{minimize} && J(b;G(b)) \\ &\text{subject to} && b \in \mathbb{M} \subset \mathbb{B}. \end{aligned}$$

Moreover, it is assumed, that the first order necessary conditions for the parametrized inner optimization problem are also sufficient. Then the effect of the operator G is the same as solving the optimality system. In view of Paragraph 2.3.2, the optimality system can be written as

$$\mathcal{T}(b;u,y,p) = 0,$$

where the operator $\mathcal{T}$ is

$$\mathcal{T} : \mathbb{B}\times\mathbb{U}\times\mathbb{Y}\times\mathbb{Z}^* \to \mathbb{Z}\times\mathbb{Y}^*\times\mathbb{U}^*, \quad (b;u,y,p) \mapsto \begin{pmatrix} T(b;u,y) \\ ((\partial_y T)(b;u,y))^* p + \nabla_y J(b;u,y) \\ ((\partial_u T)(b;u,y))^* p + \nabla_u J(b;u,y) \end{pmatrix}.$$

Unique solvability of the optimality system for arbitrarily chosen $b \in \mathbb{B}$ induces

$$\mathcal{G} : \mathbb{B} \to \mathbb{U}\times\mathbb{Y}\times\mathbb{Z}^*, \quad b \mapsto (u_b,y_b,p_b) = (G(b),p_b).$$

All in all, the bilevel optimization problem can equivalently be replaced by

$$\begin{aligned} &\text{minimize} && J(b;u_b,y_b) \\ &\text{subject to} && b \in \mathbb{M} \subset \mathbb{B}, \\ & && \mathcal{T}(b;u_b,y_b,p_b) = 0. \end{aligned}$$

Referring to the definition (2.34) of the adjoint state p, one can introduce another adjoint $P = (P^u, P^y, P^p)$ in $\mathbb{Z}^*\times\mathbb{Y}^{**}\times\mathbb{U}^{**}$

$$\Big(\partial_{u,y,p}\mathcal{T}(b,\mathcal{G}(b))\Big)^* P = -\nabla_{u,y,p} J(b,\mathcal{G}(b)),$$

where the objective J is only formally dependent on p. This yields

$$\begin{pmatrix} (\partial_u T(b;u_b,y_b))^* & (\partial_u(\partial_y T(b;u_b,y_b))^* p + \partial_u\nabla_y J(b;u_b,y_b))^* & (\partial_u(\partial_u T(b;u_b,y_b))^* p + \partial_u\nabla_u J(b;u_b,y_b))^* \\ (\partial_y T(b;u_b,y_b))^* & (\partial_y(\partial_y T(b;u_b,y_b))^* p + \partial_y\nabla_y J(b;u_b,y_b))^* & (\partial_y(\partial_u T(b;u_b,y_b))^* p + \partial_y\nabla_u J(b;u_b,y_b))^* \\ 0 & \partial_y T(b;u_b,y_b) & \partial_u T(b;u_b,y_b) \end{pmatrix} \cdot$$

$$\cdot \begin{pmatrix} P^u \\ P^y \\ P^p \end{pmatrix} = \begin{pmatrix} -\nabla_u J(b;u_b,y_b) \\ -\nabla_y J(b;u_b,y_b) \\ 0 \end{pmatrix}.$$

By setting

$$P^u = p_b, \quad P^y = 0, \quad P^p = 0.$$

this equation system reduces to

$$\begin{aligned} (\partial_u T(b;u_b,y_b))^* p &= -\nabla_u J(b;u_b,y_b), \\ (\partial_y T(b;u_b,y_b))^* p &= -\nabla_y J(b;u_b,y_b) \end{aligned}$$

which is part of $\mathcal{T}(b;u_b,y_b,p_b) = 0$. Consequently, the choice for P is a solution.

Remark:
One recognizes, that the bilevel structure of set optimal control problem – confer (2.30) and (2.36), (2.37) – fits into the more general framework (2.56). Additionally, the reduction of the optimality system of the inner optimization problem (cf. Lemma 7) has no impact on the results on principle. The reduction only is for convenience in order to avoid large systems. With it, the whole procedure of paragraphs 2.3.4–2.3.6 is not restricted to the special case under consideration. In particular, one might think of state-constrained optimal control problems with multiple controls and/or states, where the reduction step of Lemma 7 is not applicable.

2.3.7 New necessary conditions

The aim of this paragraph is the derivation of necessary conditions for the outer optimization problem (2.36), which parallels the fourth step of the general recipe of Paragraph 2.3.2.

Theorem 8 (Necessary conditions for the outer optimization problem oOP)**:**
Let $(\mathcal{A}, \bar{u}_{\mathcal{I}}, \bar{u}_{\mathcal{A}}, \bar{y}_{\mathcal{I}}, \bar{y}_{\mathcal{A}})$ be the unique solution of the bilevel optimization problem (2.36), (2.37), $\gamma = \partial\mathcal{A}$ and let $p_{\min}^{\max}$ be given by (2.46).
Then the shape gradient of the reduced objective (2.45a) has a null

$$\nabla\mathcal{F}(\mathcal{A}) = -\frac{1}{2\lambda}(p_{\min}^{\max}|_\gamma - \bar{p}_{\mathcal{I}}|_\gamma)^2 = 0.$$

In particular, there holds

$$p_{\min}^{\max}|_\gamma = \bar{p}_{\mathcal{I}}|_\gamma \quad \text{on } \gamma. \tag{2.57}$$

Proof. According to Theorem 6, the bilevel optimization problem (2.36), (2.37) is equivalent to the reduced shape/topology optimization problem (2.45), which can be written compactly as (2.38)

$$\begin{aligned} &\text{minimize} \quad \mathcal{F}(\mathcal{B}) := \mathfrak{J}(\mathcal{B}; G(\mathcal{B})), \\ &\text{subject to} \quad \begin{cases} \mathcal{B} \in \mathcal{O}, \\ y_{\min} < G_3(\mathcal{B}) < y_{\max} \quad \text{in } \mathcal{J}. \end{cases} \end{aligned}$$

Now omit the strict inequality constraint for a short while and look for necessary conditions of the relaxed optimization problem. In view of the discussion in Paragraph 2.2.4, this approach is not unreasonable, since only an inactive constraint is omitted. Again only concentrating on the shape optimization aspect, a necessary condition of the unconstrained, relaxed problem obviously is

$$0 = \nabla\mathcal{F}(\mathcal{A}) = -\frac{1}{2\lambda}(p_{\min}^{\max}|_\gamma - \bar{p}_{\mathcal{I}}|_\gamma)^2.$$

The definition of $p_{\min}^{\max} = \lambda(\Delta y_{\min}^{\max} - y_{\min}^{\max} + u_d)$ (cf. (2.46)) and the optimality conditions $\bar{u}_{\mathcal{A}} = -\Delta y_{\min}^{\max} + y_{\min}^{\max}$ and $-\bar{p}_{\mathcal{I}} = \lambda(\bar{u}_{\mathcal{I}} - u_d)$ (cf. step 3 in Lemma 7) yield

$$\begin{aligned} 0 &= -\frac{1}{2\lambda}(p_{\min}^{\max}|_\gamma - \bar{p}_{\mathcal{I}}|_\gamma)^2 \\ &= -\frac{1}{2\lambda}(\lambda(\Delta y_{\min}^{\max} - y_{\min}^{\max} + u_d)|_\gamma + \lambda(\bar{u}_{\mathcal{I}} - u_d)|_\gamma)^2 \\ &= -\frac{\lambda}{2}((u_d - \bar{u}_{\mathcal{A}})|_\gamma + (\bar{u}_{\mathcal{I}} - u_d)|_\gamma)^2 \\ &= -\frac{\lambda}{2}(\bar{u}_{\mathcal{I}}|_\gamma - \bar{u}_{\mathcal{A}}|_\gamma)^2, \end{aligned}$$

where the last step is due to H^1-regularity of u_d and Lemma 2 (mind the twofold meaning of $(.)|_\gamma$, as trace with respect to $\mathcal{I}$ and $\mathcal{A}$, respectively). In other words, the optimal control is weakly continuous across the optimal interface

$$\bar{u}_{\mathcal{I}}|_\gamma = \bar{u}_{\mathcal{A}}|_\gamma, \quad \text{on } \gamma.$$

A comparison with the necessary conditions of Proposition 3 shows, that this condition is a necessary condition for the original state-constrained optimal control problem (2.1): The Dirichlet traces of the adjoint state p^{trad} coincide on the interface, cf. (2.4g) and (2.4h). With use of the gradient equation (2.2g) this matching is transferred to the optimal control, see (2.5).

The original model problem is equivalent to the considered outer shape optimization problem and hence – following the steps in reversed order – shape gradient equals zero indeed is a necessary optimality condition. □

As an easy consequence of Theorem 8 one has

Corollary 1 (Local shape derivatives at the optimum)**:**
Let $(\mathcal{A}, \bar{u}_{\mathcal{I}}, \bar{u}_{\mathcal{A}}, \bar{y}_{\mathcal{I}}, \bar{y}_{\mathcal{A}})$ be the unique solution of the bilevel optimization problem (2.36), (2.37), let $\gamma = \partial\mathcal{A}$, let $p_{\min}^{\max}$ be given by (2.46), and let $V \in \mathcal{V}$ be arbitrarily chosen.
Then the local shape derivatives $y'_{\mathcal{I}}[V]$ and $p'_{\mathcal{I}}[V]$ defined by (2.48) in Lemma 8 vanish

$$y'_{\mathcal{I}}[V] \equiv 0, \qquad p'_{\mathcal{I}}[V] \equiv 0.$$

Proof. According to Theorem 8, there holds $p_{\min}^{\max}|_\gamma = \bar{p}_{\mathcal{I}}|_\gamma$ on γ. Hence, the BVP (2.48) is homogeneous and its unique solution is $y'_{\mathcal{I}}[V] = p'_{\mathcal{I}}[V] \equiv 0$. □

Remarks (on the strict inequality constraint)**:**
Since $y'_{\mathcal{I}}[V] \equiv 0$ for all $V \in \mathcal{V}$, the equation $y_{\mathcal{J}}|_\beta = y_{\min}^{\max}|_\beta$ holds true for $\mathcal{J}$ "near" $\mathcal{I}$ up to first order in perturbations V of $\mathcal{I}$. Near means here, that $\mathcal{J} = \mathcal{I}_t := T_t(V)(\mathcal{I})$ with $t > 0$ sufficiently small, where the transformation $T_t(V)$ is defined in the 2nd item of the discussion on page 72.

In view of the reasoning of Paragraph 2.2.4, in particular the first of the Remarks to Lemma 6 on page 28, it turns out to be sufficient to use the constraint $y_\beta = y_{\min}^{\max}$ on β and not the whole inequality constraint $y_{\min} < y_{\mathcal{J}} < y_{\max}$ in $\mathcal{J}$ in order to derive first order necessary conditions for the set optimal control problem (2.30). That is to say, the interpretation of the interface condition as the "active part" of the inequality constraints, seems to hold true. This result justifies to derive necessary conditions in the given approach while omitting the strict inequality constraint.

In other words, the strict inequality constraint seems to have no impact on the admissible directions of variation. Supposing that this holds true, indeed, directly yields

$$\begin{aligned} & 0 = d\mathcal{F}(\mathcal{A}; V) = (\nabla\mathcal{F}(\mathcal{A})\ ,\ V \cdot \boldsymbol{n}_{\mathcal{I}})_{L^2(\mathcal{J})}, \quad \forall V \in \mathcal{V}, \\ \Leftrightarrow \quad & 0 = \nabla\mathcal{F}(\mathcal{A}). \end{aligned}$$

In particular, the assertion of Theorem 8 would follow without referring to results of [14], cf. Proposition 3, which are not embedded in the approach presented here.

Corollary 2 (Higher regularity at the optimum)**:**
Let $(\mathcal{A}, \bar{u}_{\mathcal{I}}, \bar{u}_{\mathcal{A}}, \bar{y}_{\mathcal{I}}, \bar{y}_{\mathcal{A}})$ be the unique solution of the bilevel optimization problem (2.36), (2.37).
Then the adjoint states, the multipliers provided by Theorem 5, and the optimal controls feature higher regularity

$$\bar{u}_{\mathcal{I}} \in H^2(\mathcal{I}), \tag{2.58a}$$
$$\bar{p}_{\mathcal{I}} \in H^2(\mathcal{I}), \tag{2.58b}$$
$$\bar{\sigma}_{\mathcal{I}} \in H^{\frac{1}{2}}(\gamma), \tag{2.58c}$$
$$\bar{u}_{\mathcal{A}} \in H^2(\mathring{\mathcal{A}}), \tag{2.58d}$$
$$\bar{p}_{\mathcal{A}} \in H^2(\mathring{\mathcal{A}}), \tag{2.58e}$$
$$\bar{q}_{\mathcal{A}} \in H^2(\mathring{\mathcal{A}}). \tag{2.58f}$$

Proof. According to the defining equations (2.39a), (2.39b) and Theorem 8 the adjoint state $\bar{p}_{\mathcal{I}}$ fulfills

$$\begin{aligned} -\Delta\bar{p}_{\mathcal{I}} + \bar{p}_{\mathcal{I}} &= \bar{y}_{\mathcal{I}} - y_d && \text{a. e. in } \mathcal{I}, \\ \partial_{\boldsymbol{n}}\bar{p}_{\mathcal{I}} &= 0 && \text{a. e. on } \Gamma, \\ \bar{p}_{\mathcal{I}}|_\gamma &= p_{\min}^{\max}|_\gamma && \text{a. e. on } \gamma. \end{aligned}$$

This BVP is uniquely solvable in $H^2(\mathcal{J})$, since $\bar{y}_\mathcal{I} - y_d \in H^2(\mathcal{J})$ and since $p_{\min}^{\max} \in H^2(\mathring{\mathcal{B}})$ (cf. (2.46)), which yields $p_{\min}^{\max}|_\beta \in H^{3/2}(\gamma)$ (cf. Lemma 1). This is (2.58b). Hence, one obtains (2.58c) via the properties of the Neumann trace operator (cf. Lemma 1) and $\bar{\sigma}_\mathcal{I} = \partial_{\boldsymbol{n}}^{\mathcal{I}} \bar{p}_\mathcal{I} \in H^{1/2}(\gamma)$. Furthermore, the gradient equation (2.39d) yields (2.58a)

$$\bar{u}_\mathcal{I} = -\frac{1}{\lambda}\bar{p}_\mathcal{I} + u_d \in H^2(\mathcal{J}).$$

Due to the control law (2.42f) and $y_{\min}^{\max} \in H^4(\mathring{\mathcal{A}})$ one obtains H^2-regularity of $\bar{u}_\mathcal{A}$, see (2.58d). Moreover, the adjoint state in the active set solves the Dirichlet BVP (2.39e), (2.39f)

$$\begin{aligned} -\Delta \bar{p}_\mathcal{A} + \bar{p}_\mathcal{A} &= \bar{y}_\mathcal{A} - y_d && \text{a. e. in } \mathring{\mathcal{A}}, \\ \bar{p}_\mathcal{A}|_\gamma &= 0 && \text{a. e. on } \gamma. \end{aligned}$$

Consequently, $\bar{p}_\mathcal{A} \in H^2(\mathring{\mathcal{A}})$, which proves (2.58e). Finally the gradient equation (2.39g) yields (2.58f)

$$\bar{q}_\mathcal{A} = -\lambda(\bar{u}_\mathcal{A} - u_d) - \bar{p}_\mathcal{A} \in H^2(\mathring{\mathcal{B}}). \qquad \square$$

At this point the derivation of the first order necessary conditions of the bilevel optimization problem and its equivalent set optimal control problem is completed. The entire optimality system is repeated for convenience. At this, three different but equivalent formulations, which express that the shape gradient must vanish are given: prescribed inhomogeneous Dirichlet trace of $\bar{p}_\mathcal{I}$ on γ, weak continuity across the interface of the optimal control, and prescribed inhomogeneous Dirichlet trace of the multiplier $\bar{q}_\mathcal{A}$ on γ. The first two conditions are known from the proof of Theorem 8, whereas the last one is a consequence of the first two and the gradient equations (2.59n) and (2.59o).

Corollary 3 (Full first order necessary conditions of the set optimal control problem)**:**
Let $\mathcal{A} \in \mathcal{O}$ and $(\bar{u}_\mathcal{I}, \bar{u}_\mathcal{A}, \bar{y}_\mathcal{I}, \bar{y}_\mathcal{A}) \in H^2(\mathcal{I}) \times H^2(\mathring{\mathcal{A}}) \times H^2(\mathcal{I}) \times H^2(\mathring{\mathcal{A}})$ be the unique solution of the set optimal control problem (2.30).
Then there holds

$$\begin{aligned}
-\Delta \bar{y}_\mathcal{I} + \bar{y}_\mathcal{I} &= \bar{u}_\mathcal{I} && \text{in } \mathcal{I}, && \text{(2.59a)} \\
-\Delta \bar{y}_\mathcal{A} + \bar{y}_\mathcal{A} &= \bar{u}_\mathcal{A} && \text{in } \mathring{\mathcal{A}}, && \text{(2.59b)} \\
\partial_{\boldsymbol{n}} \bar{y}_\mathcal{I} &= 0 && \text{on } \Gamma, && \text{(2.59c)} \\
\bar{y}_\mathcal{I}|_\gamma &= y_{\min}^{\max}|_\gamma && \text{on } \gamma, && \text{(2.59d)} \\
\partial_{\boldsymbol{n}}^{\mathcal{I}} \bar{y}_\mathcal{I} &= \partial_{\boldsymbol{n}}^{\mathcal{I}} y_{\min}^{\max} && \text{on } \gamma, && \text{(2.59e)}
\end{aligned}$$

$$\begin{aligned}
-\Delta \bar{p}_\mathcal{I} + \bar{p}_\mathcal{I} &= \bar{y}_\mathcal{I} - y_d && \text{in } \mathcal{I}, && \text{(2.59i)} \\
-\Delta \bar{p}_\mathcal{A} + \bar{p}_\mathcal{A} &= \bar{y}_\mathcal{A} - y_d && \text{in } \mathring{\mathcal{A}}, && \text{(2.59j)} \\
\partial_{\boldsymbol{n}} \bar{p}_\mathcal{I} &= 0 && \text{on } \Gamma, && \text{(2.59k)} \\
\bar{p}_\mathcal{A}|_\gamma &= 0 && \text{on } \gamma, && \text{(2.59l)} \\
\partial_{\boldsymbol{n}}^{\mathcal{I}} \bar{p}_\mathcal{I} &= \bar{\sigma}_\mathcal{I} && \text{on } \gamma, && \text{(2.59m)}
\end{aligned}$$

$$\begin{aligned}
-\Delta y_{\min}^{\max} + y_{\min}^{\max} &= \bar{u}_\mathcal{A} && \text{in } \mathring{\mathcal{A}}, && \text{(2.59f)} \\
y_{\min}^{\max}|_\gamma &= \bar{y}_\mathcal{A}|_\gamma && \text{on } \gamma, && \text{(2.59g)} \\
y_{\min} < \bar{y}_\mathcal{I} &< y_{\max} && \text{in } \mathcal{I}, && \text{(2.59h)}
\end{aligned}$$

$$\begin{aligned}
\lambda\,(\bar{u}_\mathcal{I} - u_d) + \bar{p}_\mathcal{I} &= 0 && \text{in } \mathcal{I}, && \text{(2.59n)} \\
\lambda\,(\bar{u}_\mathcal{A} - u_d) + \bar{p}_\mathcal{A} + \bar{q}_\mathcal{A} &= 0 && \text{in } \mathring{\mathcal{A}}, && \text{(2.59o)}
\end{aligned}$$

$$\begin{cases}
\bar{p}_\mathcal{I}|_\gamma - p_{\min}^{\max}|_\gamma = 0 & \text{on } \gamma, \quad \text{(2.59p)} \\
\bar{u}_\mathcal{I}|_\gamma - \bar{u}_\mathcal{A}|_\gamma = 0 & \text{on } \gamma, \quad \text{(2.59q)} \\
\bar{q}_\mathcal{A}|_\gamma - p_{\min}^{\max}|_\gamma = 0 & \text{on } \gamma, \quad \text{(2.59r)}
\end{cases}$$

where the last three equations are equivalent formulations for the condition that the shape gradient must vanish at the optimal configuration.

This optimality system deserves closer attention. Its connection with the necessary condition of [14], cf. Proposition 3, is of major interest.

Proposition 5 (Connection between common and new necessary conditions)**:**
Let $\mathcal{A}$ be the (optimal) active set and let $\bar{p}_\mathcal{I}$, $\bar{p}_\mathcal{A}$, $\bar{q}_\mathcal{A}$ and $\bar{\sigma}_\mathcal{I}$ be the multipliers of the optimality system (2.59) and let $p_\mathcal{I}^{trad}$, $p_\mathcal{A}^{trad}$, $\mu^{\max}$ and μ_γ be given by Proposition 3.

Then there holds

$$\bar{p}_{\mathcal{I}} = p_{\mathcal{I}}^{trad} \qquad \text{in } \mathcal{I}, \tag{2.60a}$$

$$\bar{p}_{\mathcal{A}} + \bar{q}_{\mathcal{A}} = p_{\mathcal{A}}^{trad} \qquad \text{in } \mathcal{A}, \tag{2.60b}$$

$$-\Delta\bar{q}_{\mathcal{A}} + \bar{q}_{\mathcal{A}} = \mu_{\mathring{\mathcal{A}}}^{\max} \qquad \text{a. e. in } \mathring{\mathcal{A}}_{\max}, \tag{2.60c}$$

$$-\Delta\bar{q}_{\mathcal{A}} + \bar{q}_{\mathcal{A}} = -\mu_{\mathring{\mathcal{A}}}^{\min} \qquad \text{a. e. in } \mathring{\mathcal{A}}_{\min}, \tag{2.60d}$$

$$\bar{\sigma}_{\mathcal{I}} + \partial_{\boldsymbol{n}}^{\mathcal{A}}(\bar{q}_{\mathcal{A}} + \bar{p}_{\mathcal{A}}) = \mu_{\gamma} \qquad \text{a. e. on } \gamma. \tag{2.60e}$$

The proof is included in the proof of Corollary 6.

Remarks (Concluding results)**:**

1. The Bryson-Denham-Dreyfus approach yields an additive decomposition of the adjoint state $p_{\mathcal{A}}^{trad}$ into $\bar{p}_{\mathcal{A}}$ and $\bar{q}_{\mathcal{A}}$, whereas everything remains unchanged in the inactive set.
2. The relationship between the multipliers $\mu_{\mathring{\mathcal{A}}}$ and $\bar{q}_{\mathcal{A}}$ is directly linked with the BDD reformulation of the state constraint. Instead of the original state constraint, a differentiated counterpart – the control law (2.28) – was used, and, in the end, an analog differential equation holds true for the corresponding multipliers (cf. (2.60c), (2.60d)). In particular, one recognizes that the differential operations made within the BDD ansatz in order to derive the control law (primal regime) finds expression in an inverse manner on the dual regime and consequently yield higher regularity for the multiplier.
3. The new multipliers $\bar{q}_{\mathcal{A}}$ and $\bar{\sigma}_{\mathcal{I}}$ do not feature any sign conditions. On the one hand this fact is not surprising, since they correspond to equality constraints and as such do not have a fixed sign. On the other hand the multipliers $\mu_{\mathring{\mathcal{A}}}$ and μ_{γ} are known to be nonnegative and one might wonder, why this property is not mirrored the new ones, especially if one bares in mind their tight connection shown in Proposition 5. This topic is examined in Corollary 6 in the Appendix B.

2.4 First order analysis via formal Lagrange technique

The first order analysis of the set optimal control problem (2.30) via reduction technique in Section 2.3 means taking a mathematically rigorous, but sophisticated path. The formal Lagrange technique is known to be a powerful tool in order to derive the first order necessary conditions in a possibly unjustified, but easy to handle and constructive manner. In general one cannot expect to gain a deeper insight in the function spaces in which the multipliers can be found, but its probably most salient property is, that the derivation of necessary conditions only requires partial derivatives.

A revision of Section 2.3 reveals (see page 10) that many subsequent steps are required in order to derive the total shape derivative of the reduced objective $\mathcal{F}$ (cf. (2.45a)). Most of them can be bypassed within the Lagrange technique, which yields two benefits. The obvious one is – as already mentioned – that much effort is saved, if the local shape derivative of the constraints has not to be derived. The second, more profound benefit emerges with a review of the approach of Section 2.3:

- The derivation of the shape derivative of the reduced objective $\mathcal{F}$ enforces the introduction of the local shape derivatives $y_{\mathcal{J}}'[V]$ and $p_{\mathcal{J}}'[V]$, cf. Lemma 9.
- One can get rid of the local shape derivatives by introducing the shape adjoint variables $Y_{\mathcal{J}}$ and $P_{\mathcal{J}}$, cf. Lemma 10.
- Finally it turns out, that the shape derivative, respectively the shape gradient can be expressed without referring to the shape adjoint variables, cf. Theorem 7.

In the light of these experiences, it seems reasonable to wonder if this approach is really suitable for the considered class of problems. It is easily conceivable that there are problems, where $\mathcal{F}$ is shape

differentiable, but one of the mentioned steps is not applicable. An indication that such situations do occur indeed can be found in a paper of Ito, Kunisch and Peichl [100]. A technique to compute the shape derivative of a shape optimization problem is proposed therein, which copes without local shape derivatives; cf. also [77, 76]. Such ideas to use the Lagrange formalism are also known from the field of topology optimization; cf. [58, 132, 133].

It is beyond the scope of this thesis, to develop and prove such a technique in the more general framework of set optimal control problems, but these hints shall motivate, that raison d'être of the Lagrange technique reaches beyond the first-mentioned benefit.

For convenience, the formal Lagrange principle is worked out in the notation of the abstract framework of optimal control, which was given in Paragraph 2.3.2. Thereto, define the Lagrangian with respect to the optimization problem (2.32)

$$L(u,y,p) := J(u,y) + \langle p \,,\, T(u,y)\rangle_{\mathbb{Z}^*,\mathbb{Z}}.$$

Then its partial derivatives are

$$\begin{aligned}
\Big((\partial_u L)(u,y,p)\Big)(v) &= \Big\langle (\partial_u J)(u,y) \,,\, v \Big\rangle_{\mathbb{U}^*,\mathbb{U}} + \Big\langle p \,,\, \Big((\partial_u T)(u,y)\Big)(v) \Big\rangle_{\mathbb{Z}^*,\mathbb{Z}} \\
&= \Big\langle (\partial_u J)(u,y) + \Big((\partial_u T)^*(u,y)\Big)(p) \,,\, v \Big\rangle_{\mathbb{U}^*,\mathbb{U}}, \\
\Big((\partial_y L)(u,y,p)\Big)(z) &= \Big\langle (\partial_y J)(u,y) \,,\, z \Big\rangle_{\mathbb{Y}^*,\mathbb{Y}} + \Big\langle p \,,\, \Big((\partial_y T)(u,y)\Big)(z) \Big\rangle_{\mathbb{Z}^*,\mathbb{Z}} \\
&= \Big\langle (\partial_y J)(u,y) + \Big((\partial_y T)^*(u,y)\Big)(p) \,,\, z \Big\rangle_{\mathbb{U}^*,\mathbb{U}}, \\
\Big((\partial_p L)(u,y,p)\Big)(s) &= \Big\langle s \,,\, T(u,y) \Big\rangle_{\mathbb{Z}^*,\mathbb{Z}}.
\end{aligned}$$

A comparison with equation (2.35) reveals, that for an admissible pair $(u,y) \in \mathbb{M}$, i. e.

$$T(u,y) = 0 \quad \Leftrightarrow \quad (\partial_p L)(u,y,p) = 0,$$

there holds

$$(\mathrm{D}f)(u) = (\partial_u L)(u,y,p),$$

if the adjoint state p is chosen equally in both approaches as

$$p := -\Big((\partial_y T)^{-1}(u,y)\Big)^* \Big((\partial_y J)(u,y)\Big),$$

which is equivalent to

$$(\partial_y L)(u,y,p) = 0.$$

In other words, the Lagrange technique enables the derivation of the gradient of the reduced objective f without differentiating the implicitly defined control-to-state operator S, cf. (2.33).

This scheme is applied to the set optimal control problem (2.30) within the remaining part of this section.

2.4.1 Lagrangian

The Lagrangian for the set optimal control problem (2.30) is defined as usual, by augmenting the objective with duality products of multipliers and constraints. At this, the strict inequality constraint $y_{\min} < y_{\mathcal{J}} < y_{\max}$ in $\mathcal{J}$ (i. e. (2.30g)) is disregarded, which is motivated by the considerations of Paragraph 2.2.4, and by the success in deriving first order necessary conditions in Paragraph 2.3.7.

To simplify the notation, use the following abbreviations for the duality pairings for the remaining part of this paragraph.

$$\begin{aligned}
\langle .\,,\,. \rangle_M &:= \langle .\,,\,. \rangle_{H^{-\frac{1}{2}}(M),H^{\frac{1}{2}}(M)}, \quad \text{for } M \in \{\Gamma, \beta, \gamma\}, \\
\langle\!\langle .\,,\,. \rangle\!\rangle_M &:= \langle .\,,\,. \rangle_{H^{-\frac{3}{2}}(M),H^{\frac{3}{2}}(M)}, \quad \text{for } M \in \{\Gamma, \beta, \gamma\}.
\end{aligned}$$

Definition 7 (Lagrangian)**:**
The Lagrangian of the set optimal control problem

$$\text{minimize} \qquad \mathfrak{J}(\mathcal{B}; u_{\mathcal{J}}, u_{\mathcal{B}}, y_{\mathcal{J}}, y_{\mathcal{B}}) := \frac{1}{2}\|y_{\mathcal{J}} - y_d\|^2_{L^2(\mathcal{J})} + \frac{1}{2}\|y_{\mathcal{B}} - y_d\|^2_{L^2(\mathring{\mathcal{B}})} + \frac{\lambda}{2}\|u_{\mathcal{J}} - u_d\|^2_{L^2(\mathcal{J})} + \frac{\lambda}{2}\|u_{\mathcal{B}} - u_d\|^2_{L^2(\mathring{\mathcal{B}})} \tag{2.61a}$$

subject to

$$\mathcal{B} \in \mathcal{O}, \tag{2.61b}$$
$$u_{\mathcal{J}} \in L^2(\mathcal{J}),\ y_{\mathcal{J}} \in H^1(\mathcal{J}, \Delta), \tag{2.61c}$$
$$u_{\mathcal{B}} \in L^2(\mathring{\mathcal{B}}),\ y_{\mathcal{B}} \in H^1(\mathring{\mathcal{B}}, \Delta), \tag{2.61d}$$
$$-\Delta y^{\max}_{\min} + y^{\max}_{\min} = u_{\mathcal{B}} \quad \text{in } \mathring{\mathcal{B}}, \tag{2.61e}$$
$$y^{\max}_{\min}|_\beta = y_{\mathcal{B}}|_\beta \quad \text{on } \beta, \tag{2.61f}$$
$$y_{\min} < y_{\mathcal{J}} < y_{\max} \quad \text{in } \mathcal{J}, \tag{2.61g}$$
$$-\Delta y_{\mathcal{J}} + y_{\mathcal{J}} = u_{\mathcal{J}} \quad \text{in } \mathcal{J}, \tag{2.61h}$$
$$-\Delta y_{\mathcal{B}} + y_{\mathcal{B}} = u_{\mathcal{B}} \quad \text{in } \mathring{\mathcal{B}}, \tag{2.61i}$$
$$\partial_{\boldsymbol{n}} y_{\mathcal{J}} = 0 \quad \text{on } \Gamma, \tag{2.61j}$$
$$y_{\mathcal{J}}|_\beta = y^{\max}_{\min}|_\beta \quad \text{on } \beta, \tag{2.61k}$$
$$\partial^{\mathcal{J}}_{\boldsymbol{n}} y_{\mathcal{J}} = \partial^{\mathcal{J}}_{\boldsymbol{n}} y^{\max}_{\min} \quad \text{on } \beta, \tag{2.61l}$$

which is an equivalent reformulation[15] of the original set optimal control problem (2.30), is defined as

$$\mathcal{L} : \mathcal{O} \times L^2(\mathcal{J}) \times L^2(\mathring{\mathcal{B}}) \times H^2(\mathcal{J}) \times H^2(\mathring{\mathcal{B}}) \times L^2(\mathcal{J}, \Delta) \times L^2(\mathring{\mathcal{B}}, \Delta) \times L^2(\mathring{\mathcal{B}}) \times H^{-\frac{3}{2}}(\beta) \times H^{-\frac{3}{2}}(\beta) \to \mathbb{R} \tag{2.62}$$

$$\begin{aligned}\mathcal{L}(\mathcal{B}; u_{\mathcal{J}}, u_{\mathcal{B}}, y_{\mathcal{J}}, y_{\mathcal{B}};\ p_{\mathcal{J}}, p_{\mathcal{B}}, q_{\mathcal{B}}, \sigma_{\mathcal{J}}, \sigma_{\mathcal{B}}) &:= \mathfrak{J}(\mathcal{B}; u_{\mathcal{J}}, u_{\mathcal{B}}, y_{\mathcal{J}}, y_{\mathcal{B}}) \\ &\quad - \int_{\mathcal{J}} (-\Delta y_{\mathcal{J}} + y_{\mathcal{J}} - u_{\mathcal{J}})\, p_{\mathcal{J}} - \langle p_{\mathcal{J}}\ ,\ \partial_{\boldsymbol{n}} y_{\mathcal{J}} \rangle_\Gamma - \left\langle p_{\mathcal{J}}\ ,\ \partial^{\mathcal{J}}_{\boldsymbol{n}} y_{\mathcal{J}} - \partial^{\mathcal{J}}_{\boldsymbol{n}} y^{\max}_{\min} \right\rangle_\beta \\ &\quad - \int_{\mathring{\mathcal{B}}} (-\Delta y_{\mathcal{B}} + y_{\mathcal{B}} - u_{\mathcal{B}})\, p_{\mathcal{B}} \\ &\quad + \int_{\mathring{\mathcal{B}}} (\Delta y^{\max}_{\min} - y^{\max}_{\min} + u_{\mathcal{B}})\, q_{\mathcal{B}} + \langle\!\langle \sigma_{\mathcal{B}}\ ,\ y_{\mathcal{B}} - y^{\max}_{\min} \rangle\!\rangle_\beta + \langle\!\langle \sigma_{\mathcal{J}}\ ,\ y_{\mathcal{J}} - y^{\max}_{\min} \rangle\!\rangle_\beta .\end{aligned} \tag{2.63}$$

2.4.2 Partial shape derivatives

As already mentioned, one advantage of the formal Lagrange technique is, that all variables are treated as independent. Hence, there is no need for applying the chain rule, and in this sense the formalism only relies on partial derivatives. Thus, the notion of a partial shape derivatives is required in the present context.

By way of illustration, let f be a shape functional (whose domain actually is a vector bundle and should not be written as Cartesian product)

$$f : \mathcal{O} \times \mathbb{X} \to \mathbb{R}, \quad (\mathcal{B}, x) \mapsto f(\mathcal{B}, x),$$

which is assumed to be shape differentiable at $\mathcal{A} \in \mathcal{O}$. In addition, let $x(\mathcal{B}) \in \mathbb{X}$ be uniquely determined by the choice of the set $\mathcal{B}$ and be shape differentiable at $\mathcal{A}$. Then the chain rule[16] reveals, that the *(total) shape derivative* consists of two separate parts

- the *partial shape derivative*, which describes the explicit dependency of f on the shape variable $\mathcal{B}$ and
- the partial derivative with respect to the function variable x composed with the *local shape derivative* $x'(\mathcal{A})[V]$, which describes the implicit dependency of f on the shape variable, which is caused by the shape dependent variable x

$$\underbrace{df(\mathcal{A}, x(\mathcal{A}); V)}_{\text{total shape derivative}} = \underbrace{\partial_{\mathcal{B}} f(\mathcal{A}, x(\mathcal{A}); V)}_{\text{partial shape derivative}} + \partial_x f(\mathcal{A}, x(\mathcal{A})) \circ \underbrace{x'(\mathcal{A})[V]}_{\text{local shape derivative}}.$$

[15] Confer the first part of the proof of Theorem 5. It is possible to analogously define a Lagrangian for original set optimal control problem, too. However, this approach yields the sightly different necessary conditions which are obtained in Appendix B.

[16] It should be noted that applicability of the chain rule requires suitable notions of derivatives; Hadamard differentiability is required in particular. This topic is discussed in [44, p. 170 and Chp. 9 Sec. 2.3] in the context of shape calculus.

In particular, the derivative of an integral domain shape functional $f(\mathcal{B}) := \int_{\mathcal{B}} G(\mathcal{B})$ decomposes (see [151, Eq. (2.168)])

$$df(\mathcal{B};V) = \int_{\partial\mathcal{B}} G(\mathcal{B})|_{\partial\mathcal{A}}\, V\cdot\boldsymbol{n} + \int_{\mathring{\mathcal{B}}} G'(\mathcal{B})[V],$$

whereas the derivative of an integral boundary shape functional $f(\mathcal{B}) := \int_{\partial\mathcal{B}} g(\mathcal{B})$ (whose integrand is the trace of a distributed function $g(\mathcal{B}) = G(\mathcal{B})|_{\partial\mathcal{B}}$) yields (see [151, Eq. (2.174)])

$$df(\mathcal{A};V) = \int_{\partial\mathcal{A}} \big(\partial_{\boldsymbol{n}} G(\mathcal{A}) + \kappa\, g(\mathcal{A})\big) V\cdot\boldsymbol{n} + \int_{\partial\mathcal{A}} \big(G'(\mathcal{A})[V]\big)\,|_{\partial\mathcal{A}}.$$

In addition, one is often confronted with the situation, that the integrand of an integral boundary shape functional is the product of the trace of a distributed function and another function which cannot be seen as the trace of a distributed function, this is

$$f(\mathcal{B}) := \int_{\partial\mathcal{B}} G(\mathcal{B})|_{\partial\mathcal{B}}\, h(\mathcal{B}).$$

At this, the decomposition into partial and local shape derivative is given by[17]

$$\begin{aligned} df(\mathcal{A};V) &= \int_{\partial\mathcal{A}} \kappa\, g(\mathcal{A})\, h(\mathcal{A})\, V\cdot\boldsymbol{n} + g'(\mathcal{A})[V]\, h(\mathcal{A}) + g(\mathcal{A})\, h'(\mathcal{A})[V] \\ &= \int_{\partial\mathcal{A}} \kappa\, g(\mathcal{A})\, h(\mathcal{A})\, V\cdot\boldsymbol{n} + \big(\big(G'(\mathcal{A})[V]\big)|_{\partial\mathcal{A}} + \partial_{\boldsymbol{n}} G(\mathcal{A})\, V\cdot\boldsymbol{n}\big)\; h(\mathcal{A}) + g(\mathcal{A})\, h'(\mathcal{A})[V] \\ &= \int_{\partial\mathcal{A}} \Big(\partial_{\boldsymbol{n}} G(\mathcal{A})\, h(\mathcal{A}) + \kappa\, g(\mathcal{A})\, h(\mathcal{A})\Big) V\cdot\boldsymbol{n} + \int_{\partial\mathcal{A}} \big(G'(\mathcal{A})[V]\big)\,|_{\partial\mathcal{A}}\, h(\mathcal{A}) + g(\mathcal{A})\, h'(\mathcal{A})[V]. \end{aligned}$$

The situation gets even more involved if a normal derivative is part of the integrand

$$f(\mathcal{B}) := \int_{\partial\mathcal{B}} \partial_{\boldsymbol{n}} G(\mathcal{B}).$$

In order to derive its (partial) shape derivative one can make use of the oriented distance function to be introduced next. As a start define the *distance function* from a point $x \in \mathbb{R}^2$ to a set $M \subset \mathbb{R}^2$.

$$d_M(x) := \begin{cases} \inf_{y\in M} |x-y|, & M \neq \emptyset, \\ +\infty, & M = \emptyset. \end{cases} \tag{2.65}$$

The the *oriented distance function* from a point $x \in \mathbb{R}^2$ to a set $M \subset \mathbb{R}^2$ is defined as

$$b_M(x) := d_M(x) - d_{M^c}(x). \tag{2.66}$$

It is known that, when M is of class $C^{1,1}$, there is a radius $\rho > 0$ such that $b_M \in C^{1,1}(\overline{B_\rho(x)})$ for each $x \in \partial M$, cf. [44, Chp. 7 Thm. 8.5]. Furthermore, the gradient ∇b_M of the oriented distance function is an extension of the outer unit normal vector field locally in $\overline{B_\rho(x)}$, (ibidem). Consequently, the oriented distance function $b_{\mathcal{B}}$ exists for any $\mathcal{B} \in \mathcal{O}$, is $C^{1,1}$-regular in a tubular neighborhood of $\beta = \partial\mathcal{B}$ and its gradient is an extension of the unit normal vector field $\boldsymbol{n}_{\mathcal{B}}$. Due to Rademacher's theorem the second derivative $\mathrm{D}^2 b_{\mathcal{B}}$ exists almost everywhere in this neighborhood. The set index of $b_{\mathcal{B}}$ will be omitted in the following, since its connection with the set will be obvious.

With this notion at hand one can resume the derivation of the shape derivative of f

$$\begin{aligned} df(\mathcal{A};V) &= \int_{\partial\mathcal{A}} \kappa\, \partial_{\boldsymbol{n}} G(\mathcal{A})\, V\cdot\boldsymbol{n} + \big(\partial_{\boldsymbol{n}} G(\mathcal{A})\big)'[V] \\ &= \int_{\partial\mathcal{A}} \kappa\, \partial_{\boldsymbol{n}} G(\mathcal{A})\, V\cdot\boldsymbol{n} + \big(\big(\nabla G(\mathcal{A})\cdot\nabla b\big)|_{\partial\mathcal{A}}\big)'[V]. \end{aligned}$$

Thus, one has to analyze the second summand in more detail

$$\begin{aligned} &\big(\big(\nabla G(\mathcal{A})\cdot\nabla b\big)|_{\partial\mathcal{A}}\big)'[V] \\ &\quad \overset{(2.64)}{=} \big(\nabla G(\mathcal{A})\cdot\nabla b\big)'[V]|_{\partial\mathcal{A}} + \partial_{\boldsymbol{n}}\big(\nabla G(\mathcal{A})\cdot\nabla b\big)\, V\cdot\boldsymbol{n} \end{aligned}$$

[17] Due to [151, Eqs. (2.173), (2.163) and (2.169)] there holds

$$g'(\mathcal{A})[V] = G'(\mathcal{A})[V]|_{\partial\mathcal{A}} + \partial_{\boldsymbol{n}} G(\mathcal{A})\, V\cdot\boldsymbol{n}. \tag{2.64}$$

$$
\begin{aligned}
&= \Big(\big(\nabla G(\mathcal{A})\big)'[V] \cdot \nabla b + \nabla G(\mathcal{A}) \cdot (\nabla b)'[V] \Big) \Big|_{\partial\mathcal{A}} + \big(\nabla\big(\nabla G(\mathcal{A}) \cdot \nabla b\big) \cdot \nabla b\big) \,|_{\partial\mathcal{A}}\, V \cdot \boldsymbol{n} \\
&= \Big(\nabla\big(G'(\mathcal{A})[V]\big) \cdot \nabla b\Big)\Big|_{\partial\mathcal{A}} + \big(\nabla G(\mathcal{A})\big)\big|_{\partial\mathcal{A}} \cdot (\nabla b)'[V]\big|_{\partial\mathcal{A}} \\
&\quad + \Big(\mathrm{D}^2 G(\mathcal{A})\, \nabla b \cdot \nabla b + \underbrace{\mathrm{D}^2 b\, \nabla G(\mathcal{A}) \cdot \nabla b}_{\nabla G(\mathcal{A}) \cdot \mathrm{D}^2 b \nabla b = 0, \text{ since } D^2 b\, \nabla b = 0, \text{ cf. [44, p. 372]}} \Big)\Big|_{\partial\mathcal{A}} V \cdot \boldsymbol{n} \\
&= \partial_{\boldsymbol{n}} G'(\mathcal{A})[V] + \big(\nabla G(\mathcal{A})\big)\big|_{\partial\mathcal{A}} \cdot (\nabla b)'[V]\big|_{\partial\mathcal{A}} + \partial_{\boldsymbol{nn}} G(\mathcal{A}).
\end{aligned}
$$

Altogether this results in

$$
df(\mathcal{A}; V) = \int_{\partial\mathcal{A}} \big(\partial_{\boldsymbol{nn}} G(\mathcal{A}) + \kappa\, \partial_{\boldsymbol{n}} G(\mathcal{A})\big)\, V \cdot \boldsymbol{n} + \int_{\partial\mathcal{A}} \partial_{\boldsymbol{n}} G'(\mathcal{A})[V] + \big(\nabla G(\mathcal{A})\big)\big|_{\partial\mathcal{A}} \cdot (\nabla b)'[V]\big|_{\partial\mathcal{A}}.^{18}
$$

Another frequent case is that test functions are contained in the integrand. Test functions do not depend on the shape explicitly and henceforth their local shape derivative vanishes. With respect to the partial shape derivative they behave like shape dependent functions (since the explicit shape dependency is neglected then), and consequently the most common situations are covered within the above considered cases.

With these deliberations at hand it is possible to compute the derivative of the Lagrangian. Since all variables of the Lagrangian are independent – in particular, all functions space variables are independent of the choice of the set $\mathcal{B} \in \mathcal{O}$[19] – the (total) shape derivative of $\mathcal{L}$ coincides with its partial shape derivative

$$
d\mathcal{L}(\mathcal{B}; \ldots; V) = \partial_{\mathcal{B}} \mathcal{L}(\mathcal{B}; \ldots; V).
$$

2.4.3 New necessary conditions

In order to derive first order necessary conditions for the set optimal control problem (2.61), one needs to compute all partial derivatives of the Lagrangian $\mathcal{L}$ at the optimum (cf. [52, Prop. 1.6 on p. 170])

$$
\bar{o} := (\mathcal{A}; \bar{u}_{\mathcal{I}}, \bar{u}_{\mathcal{A}}, \bar{y}_{\mathcal{I}}, \bar{y}_{\mathcal{A}}; \bar{p}_{\mathcal{I}}, \bar{p}_{\mathcal{A}}, \bar{q}_{\mathcal{A}}, \bar{\sigma}_{\mathcal{I}}, \bar{\sigma}_{\mathcal{A}}).
$$

The experiences of Section 2.3 say, that the strict inequality constraint (2.61g) has no influence on local optimality and consequently the cone of admissible directions of variation is not restricted. The Lagrangian is a convex-concave functional with respect to the function space variables. To the best of the author's knowledge, the qualitative dependency on the set variable cannot be classified but nonlinear, which is due to the underlying manifold structure of the shape space, see Section 2.6. Nevertheless, $\mathcal{L}$ has a critical point a the optimum $\bar{o}$. Hence, each partial derivative of $\mathcal{L}$ evaluated at $\bar{o}$ has a null. These necessary conditions for optimality will be derived in the following.

As a start, regard the derivatives with respect to the control variables.

$$
\begin{aligned}
0 = \big(\partial_{u_{\mathcal{I}}} \mathcal{L}(\bar{o})\big)(h) &= \int_{\mathcal{I}} \lambda(\bar{u}_{\mathcal{I}} - u_d)\, h + h\, \bar{p}_{\mathcal{I}}, \quad \forall h \in L^2(\mathcal{I}), \\
0 = \big(\partial_{u_{\mathcal{B}}} \mathcal{L}(\bar{o})\big)(h) &= \int_{\mathring{\mathcal{A}}} \lambda(\bar{u}_{\mathcal{A}} - u_d)\, h + h\, \bar{p}_{\mathcal{A}} + h\, \bar{q}_{\mathcal{A}}, \quad \forall h \in L^2(\mathring{\mathcal{B}}).
\end{aligned}
$$

Consequently, the fundamental lemma of the calculus of variations yields

$$
0 = \lambda(\bar{u}_{\mathcal{I}} - u_d) + \bar{p}_{\mathcal{I}} \qquad \text{a. e. in } \mathcal{I}, \tag{2.68a}
$$

$$
0 = \lambda(\bar{u}_{\mathcal{A}} - u_d) + \bar{p}_{\mathcal{A}} + \bar{q}_{\mathcal{A}} \qquad \text{a. e. in } \mathring{\mathcal{A}}. \tag{2.68b}
$$

[18] The local shape derivative of the gradient of the oriented distance function is given by [44, Chp. 9 Eq. (4.38)]

$$
(\nabla b)'[V] = \big((\mathrm{D}V)\, \nabla b \cdot \nabla b\big)\, \nabla b - (\mathrm{D}V)^{\top} \nabla b - \mathrm{D}^2 b\, V. \tag{2.67}
$$

[19] However, Definition 7 reveals on closer examination, that the function spaces of the variables do dependent on the set $\mathcal{B}$. This implicit dependency can be overcome by regarding the variables as the restrictions of some else, which are defined on the holdall Ω. This method of *function space embedding* is used in [44, p. 565ff.]. Alternatively one can regard the Lagrangian as a functional on a vector bundle on a shape related manifold, see Paragraph 2.6.3. It is important to notice however, that the space dependency does not imply a predetermination of function space variables if the set variable is fixed.

The partial derivatives with respect to the state variables yield the adjoint equations with the help of Green's formula (cf. Remark to Lemma 3)

$$
\begin{aligned}
0 = \big(\partial_{y_{\mathcal{J}}} \mathcal{L}(\bar{o})\big)(h) &= \int_{\mathcal{I}} (\bar{y}_{\mathcal{I}} - y_d)\, h + (\Delta h - h)\, \bar{p}_{\mathcal{I}} - \langle \bar{p}_{\mathcal{I}} \,, \partial_{\boldsymbol{n}} h \rangle_{\Gamma} - \langle \bar{p}_{\mathcal{I}} \,, \partial_{\boldsymbol{n}}^{\mathcal{I}} h \rangle_{\gamma} + \langle\!\langle \bar{\sigma}_{\mathcal{I}} \,, h \rangle\!\rangle_{\gamma} \\
&= \int_{\mathcal{I}} (\Delta \bar{p}_{\mathcal{I}} - \bar{p}_{\mathcal{I}} + \bar{y}_{\mathcal{I}} - y_d)\, h - \langle\!\langle \partial_{\boldsymbol{n}} \bar{p}_{\mathcal{I}} \,, h \rangle\!\rangle_{\Gamma} + \langle\!\langle \bar{\sigma}_{\mathcal{I}} - \partial_{\boldsymbol{n}}^{\mathcal{I}} \bar{p}_{\mathcal{I}} \,, h \rangle\!\rangle_{\gamma}, \quad \forall h \in H^2(\mathcal{J}), \\
0 = \big(\partial_{y_{\mathcal{B}}} \mathcal{L}(\bar{o})\big)(h) &= \int_{\mathring{\mathcal{A}}} (\bar{y}_{\mathcal{A}} - y_d)\, h + (\Delta h - h)\, \bar{p}_{\mathcal{A}} + \langle\!\langle \bar{\sigma}_{\mathcal{A}} \,, h \rangle\!\rangle_{\gamma} \\
&= \int_{\mathring{\mathcal{A}}} (\Delta \bar{p}_{\mathcal{A}} - \bar{p}_{\mathcal{A}} + \bar{y}_{\mathcal{A}} - y_d)\, h + \langle \bar{p}_{\mathcal{A}} \,, \partial_{\boldsymbol{n}}^{\mathcal{A}} h \rangle_{\gamma} + \langle\!\langle \bar{\sigma}_{\mathcal{A}} - \partial_{\boldsymbol{n}}^{\mathcal{A}} \bar{p}_{\mathcal{A}} \,, h \rangle\!\rangle_{\gamma}, \quad \forall h \in H^2(\mathring{\mathcal{B}}).
\end{aligned}
$$

That is so say

$$
\begin{aligned}
-\Delta \bar{p}_{\mathcal{I}} + \bar{p}_{\mathcal{I}} &= \bar{y}_{\mathcal{I}} - y_d && \text{a. e. in } \mathcal{I}, && \text{(2.69a)} \qquad & -\Delta \bar{p}_{\mathcal{A}} + \bar{p}_{\mathcal{A}} &= \bar{y}_{\mathcal{A}} - y_d && \text{a. e. in } \mathring{\mathcal{A}}, && \text{(2.69d)} \\
\partial_{\boldsymbol{n}} \bar{p}_{\mathcal{I}} &= 0 && \text{a. e. on } \Gamma, && \text{(2.69b)} & \bar{p}_{\mathcal{A}}|_{\gamma} &= 0 && \text{a. e. on } \gamma, && \text{(2.69e)} \\
\partial_{\boldsymbol{n}}^{\mathcal{I}} \bar{p}_{\mathcal{I}} &= \bar{\sigma}_{\mathcal{I}} && \text{a. e. on } \gamma, && \text{(2.69c)} & \partial_{\boldsymbol{n}}^{\mathcal{A}} \bar{p}_{\mathcal{A}} &= \bar{\sigma}_{\mathcal{A}} && \text{a. e. on } \gamma. && \text{(2.69f)}
\end{aligned}
$$

Whereas the derivatives with respect to the multipliers yield the original constraints, as usual, the partial shape derivative of the Lagrangian can be simplified, with the help of different equations as indicated

$$
\begin{aligned}
0 &= \partial_{\mathcal{B}}\big(\mathcal{L}(\bar{o}); V\big) \\
&= \int_{\gamma} \Big(\frac{1}{2}(\bar{y}_{\mathcal{I}} - y_d)^2 + \frac{\lambda}{2}(\bar{u}_{\mathcal{I}} - u_d)^2\Big) V \cdot \boldsymbol{n}_{\mathcal{I}} + \int_{\gamma} \Big(\frac{1}{2}(\bar{y}_{\mathcal{A}} - y_d)^2 + \frac{\lambda}{2}(\bar{u}_{\mathcal{A}} - u_d)^2\Big) V \cdot \boldsymbol{n}_{\mathcal{A}} \\
&\quad - \int_{\gamma} (-\Delta \bar{y}_{\mathcal{I}} + \bar{y}_{\mathcal{I}} - \bar{u}_{\mathcal{I}})\, \bar{p}_{\mathcal{I}} V \cdot \boldsymbol{n}_{\mathcal{I}} - \int_{\gamma} \bar{p}_{\mathcal{I}}\, \big(\partial_{\boldsymbol{nn}}(\bar{y}_{\mathcal{I}} - y_{\min}^{\max}) + \kappa_{\mathcal{I}} \partial_{\boldsymbol{n}}^{\mathcal{I}}(\bar{y}_{\mathcal{I}} - y_{\min}^{\max})\big)\, V \cdot \boldsymbol{n}_{\mathcal{I}} \\
&\quad - \int_{\gamma} \underbrace{\partial_{\boldsymbol{n}}^{\mathcal{I}}(\bar{y}_{\mathcal{I}} - y_{\min}^{\max})}_{=0,\text{ use (2.61l)}}\, \partial_{\boldsymbol{n}}^{\mathcal{I}} \bar{p}_{\mathcal{I}}\, V \cdot \boldsymbol{n}_{\mathcal{I}} \\
&\quad - \int_{\gamma} (-\Delta \bar{y}_{\mathcal{A}} + \bar{y}_{\mathcal{A}} - \bar{u}_{\mathcal{A}})\, \underbrace{\bar{p}_{\mathcal{A}}}_{=0,\text{ use (2.69e)}}\, V \cdot \boldsymbol{n}_{\mathcal{A}} \\
&\quad + \int_{\gamma} (\underbrace{\Delta y_{\min}^{\max} - y_{\min}^{\max} + \bar{u}_{\mathcal{A}}}_{=0,\text{ use (2.61e)}})\, \bar{q}_{\mathcal{A}} V \cdot \boldsymbol{n}_{\mathcal{A}} + \int_{\gamma} \big(\underbrace{\partial_{\boldsymbol{n}}^{\mathcal{A}}(\bar{y}_{\mathcal{A}} - y_{\min}^{\max})}_{=0,\text{ use } \bar{y}_{\mathcal{A}} \equiv y_{\min}^{\max}} + \kappa_{\mathcal{A}}\,(\underbrace{\bar{y}_{\mathcal{A}} - y_{\min}^{\max}}_{=0,\text{ use (2.61f)}})\big)\, \bar{\sigma}_{\mathcal{A}}\, V \cdot \boldsymbol{n}_{\mathcal{A}} \\
&\quad + \int_{\gamma} \big(\underbrace{\partial_{\boldsymbol{n}}^{\mathcal{I}}(\bar{y}_{\mathcal{I}} - y_{\min}^{\max})}_{=0,\text{ use (2.61l)}} + \kappa_{\mathcal{I}}\,(\underbrace{\bar{y}_{\mathcal{I}} - y_{\min}^{\max}}_{=0,\text{ use (2.61k)}})\big)\, \bar{\sigma}_{\mathcal{I}}\, V \cdot \boldsymbol{n}_{\mathcal{I}} \\
&= \int_{\gamma} \frac{1}{2}\Big(\underbrace{(\bar{y}_{\mathcal{I}} - y_d)^2 - (\bar{y}_{\mathcal{A}} - y_d)^2}_{=0,\text{ use (2.61k), (2.61f) and } y_d \in H^1(\Omega)}\Big) V \cdot \boldsymbol{n}_{\mathcal{I}} + \int_{\gamma} \Big(\frac{\lambda}{2}(\bar{u}_{\mathcal{I}} - u_d)^2 - \frac{\lambda}{2}(\bar{u}_{\mathcal{A}} - u_d)^2\Big) V \cdot \boldsymbol{n}_{\mathcal{I}} \\
&\quad + \int_{\gamma} \big(\underbrace{\Delta \bar{y}_{\mathcal{I}} - \partial_{\boldsymbol{nn}} \bar{y}_{\mathcal{I}} - \kappa_{\mathcal{I}} \partial_{\boldsymbol{n}}^{\mathcal{I}} \bar{y}_{\mathcal{I}}}_{=\Delta\gamma \bar{y}_{\mathcal{I}},\text{ use (2.51)}} - \bar{y}_{\mathcal{I}} + \bar{u}_{\mathcal{I}}\big)\, \bar{p}_{\mathcal{I}} V \cdot \boldsymbol{n}_{\mathcal{I}} + \int_{\gamma} \big(\underbrace{\partial_{\boldsymbol{nn}} y_{\min}^{\max} + \kappa_{\mathcal{I}} \partial_{\boldsymbol{n}}^{\mathcal{I}} y_{\min}^{\max}}_{=\Delta y_{\min}^{\max} - \Delta_{\gamma} y_{\min}^{\max},\text{ use (2.51)}}\big)\, \bar{p}_{\mathcal{I}} V \cdot \boldsymbol{n}_{\mathcal{I}} \\
&= \int_{\gamma} \Big(\frac{\lambda}{2}(\bar{u}_{\mathcal{I}} - u_d)^2 - \frac{\lambda}{2}(\bar{u}_{\mathcal{A}} - u_d)^2 + (\underbrace{\Delta_{\gamma} \bar{y}_{\mathcal{I}} - \Delta_{\gamma} y_{\min}^{\max}}_{=0\ [20]} + \underbrace{\Delta y_{\min}^{\max} - \overbrace{\bar{y}_{\mathcal{I}}}^{=y_{\min}^{\max},\text{ use (2.61k)}}}_{=-\bar{u}_{\mathcal{A}},\text{ use (2.61e)}} + \bar{u}_{\mathcal{I}})\ \underbrace{\bar{p}_{\mathcal{I}}}_{=-\lambda(\bar{u}_{\mathcal{I}} - u_d),\text{ use (2.68a)}}\Big) V \cdot \boldsymbol{n}_{\mathcal{I}} \\
&= \int_{\gamma} -\frac{\lambda}{2}\big(-\bar{u}_{\mathcal{I}}^2 + \cancel{2\,\bar{u}_{\mathcal{I}}\, u_d} - \cancel{y_d^2} + \bar{u}_{\mathcal{A}}^2 - \cancel{2\,\bar{u}_{\mathcal{A}}\, u_d} + \cancel{y_d^2} - \cancel{2\,\bar{u}_{\mathcal{A}}\, \bar{u}_{\mathcal{I}}} + 2\,\bar{u}_{\mathcal{A}}\, u_d + 2\,\bar{u}_{\mathcal{I}}^2 - \cancel{2\,\bar{u}_{\mathcal{I}}\, u_d}\big) V \cdot \boldsymbol{n}_{\mathcal{I}} \\
&= \int_{\gamma} -\frac{\lambda}{2}(\bar{u}_{\mathcal{I}} - \bar{u}_{\mathcal{A}})^2 V \cdot \boldsymbol{n}_{\mathcal{I}}, \quad \forall V \in \mathcal{V}.
\end{aligned}
$$

Hence, one obtains weak continuity of the optimal control across the interface γ

$$
\bar{u}_{\mathcal{I}}|_{\gamma} = \bar{u}_{\mathcal{A}}|_{\gamma} \quad \text{a. e. on } \gamma. \tag{2.70}
$$

[20] Use the same reasoning as in the first part of the proof of Lemma 8 which yielded (2.49).

In summary, the vanishing partial shape derivative of the Lagrangian $\mathcal{L}$ is compatible with the inequality constraint (2.61f), since weak continuity of the optimal control actually is a necessary condition for the considered optimal control problem, cf. (2.5).

Remark:
Weak continuity of the optimal control across the interface between active and inactive set is the analog to the continuity of the Hamiltonian across junction points of boundary arcs for autonomous problems in OC-ODE, see [122, p. 22 (iii)], [75, Eq. (5.15)].

This Section ends with a corollary, in which some principle results of sections 2.3 and 2.4 are collected.

Corollary 4:
Let $\mathcal{A} \in \mathcal{O}$ and $(\bar{u}_{\mathcal{I}}, \bar{u}_{\mathcal{A}}, \bar{y}_{\mathcal{I}}, \bar{y}_{\mathcal{A}}) \in H^2(\mathcal{I}) \times H^2(\mathring{\mathcal{A}}) \times H^2(\mathcal{I}) \times H^2(\mathring{\mathcal{A}})$ be the unique solution of the set optimal control problems (2.30), and respectively (2.61).

Then there holds:

1. The necessary conditions of Corollary 3, which were obtained via the reduction technique of Section 2.3 coincide with the saddle point characterizing equations of the Lagrangian, i. e. the equality constraints of (2.61) and (2.68)–(2.70).
2. The necessary conditions are compatible with the strict inequality constraint $y_{\min} < \bar{y}_{\mathcal{I}} < y_{\max}$ in $\mathcal{I}$, i. e. (2.30g).

Proof. **1)** The first assertion is obvious, except that equation (2.69f) has no analog within the conditions of Corollary 3. However, this equation is unnecessary, since it only determines the additional multiplier $\bar{\sigma}_{\mathcal{B}}$ to be the Neumann trace of $\bar{p}_{\mathcal{A}}$.
2) The optimal state obviously respects the strict inequality constraint, since the inactive set $\mathcal{I}$ is defined such that this condition is fulfilled, cf. Definition 3. □

2.5 Second order analysis

After the analysis of first order necessary conditions for the set optimal control problem (2.30) has been presented in detail in the last three sections, this section is devoted to a brief analysis of the second order derivative of the reduced objective $\mathcal{F}$, which is defined in (2.38a) and (2.45a), respectively.
However, there are no second order sufficient conditions derived – or to be more precise, it is shown, that the second order shape semiderivative of $\mathcal{F}$ is not definite at the optimum. Nonetheless, the knowledge of the second order shape semiderivative is used for the design of efficient algorithms, cf. Chapter 3.

2.5.1 Second order shape semiderivative and lack of second order sufficiency

In the course of the lengthy derivation of Paragraph 2.3.6 it turned out (cf. Theorem 7), that the shape semiderivative of the reduced objective $\mathcal{F}$ (see (2.45a)) is given by

$$d\mathcal{F}(\mathcal{B}; V) = -\frac{1}{2\lambda} \int_{\beta} (p_{\min}^{\max} - \bar{p}_{\mathcal{J}})^2 V \cdot \boldsymbol{n}_{\mathcal{J}}, \quad \forall V \in \mathcal{V}.$$

The second order shape semiderivative is as follows.

Lemma 11 (Second order shape semiderivative of $\mathcal{F}$)**:**
Let the family of admissible sets $\mathcal{O}$ be given by Definition 4, let $\mathcal{B} \in \mathcal{O}$ and let $V, W \in \mathcal{V}$ – see (2.47) – be arbitrarily chosen. Furthermore, let $p'_{\mathcal{J}}[W] \in L^2(\mathcal{J}, \Delta)$ be the local shape derivative with respect to the velocity field W according to Lemma 8.

Then the second order shape semiderivative of the reduced objective $\mathcal{F}$ – see (2.45a) – with respect to V and W is given by

$$d^2\mathcal{F}(\mathcal{B}; V, W) = -\frac{1}{2\lambda} \int_{\beta} (\bar{p}_{\mathcal{J}} - p_{\min}^{\max}) \Big(2p'_{\mathcal{J}}[W] V \cdot \boldsymbol{n}_{\mathcal{J}} + \Big(2\nabla(\bar{p}_{\mathcal{J}} - p_{\min}^{\max}) \cdot V + (\bar{p}_{\mathcal{J}} - p_{\min}^{\max}) \operatorname{div} V \Big) W \cdot \boldsymbol{n}_{\mathcal{J}} \Big).$$

Proof. The second order shape semiderivative is obtained via repeated differentiation, cf., for instance, [146, 156] and the extensive presentation in [44, Chp. 9 Sec. 6], in particular page 508ff. and 516ff. Note, that V and W are autonomous vector fields, and consequently there hold some simplifications, e. g. $V' = 0$, which are used below.

Due to Gauß's divergence theorem and since one has $V \cdot \boldsymbol{n} = 0$ on Γ, the shape semiderivative can be expressed as a volume integral

$$d\mathcal{F}(\mathcal{B}; V) = -\frac{1}{2\lambda} \int_{\mathcal{J}} \operatorname{div}\left((\bar{p}_{\mathcal{J}} - p_{\min}^{\max})^2 \, V \right).$$

Note, that the coefficient function $p_{\min}^{\max}$, which was defined in (2.46) by means of an analog construction as in the proof of Lemma 4, depends on the set $\mathcal{B}$, but can be chosen unchanged, if the set is only slightly deformed; see the Remark to Lemma 4. Hence, the local shape derivative of $p_{\min}^{\max}$ is zero.

The rules of computation for shape semiderivatives of domain integrals (see [151, Eq. (2.167)], [44, Chp. 9 Eq. (4.6)] and Paragraph 2.4.2) and the Gauß divergence theorem yield

$$\begin{aligned}
&d^2\mathcal{F}(\mathcal{B}; V, W) \\
&\quad = -\frac{1}{2\lambda} \int_{\mathcal{J}} \left(\operatorname{div}\left((\bar{p}_{\mathcal{J}} - p_{\min}^{\max})^2 \, V \right) \right)' [W] + \operatorname{div}\left(\operatorname{div}\left((\bar{p}_{\mathcal{J}} - p_{\min}^{\max})^2 \, V \right) W \right) \\
&\quad = -\frac{1}{2\lambda} \int_{\mathcal{J}} \operatorname{div}\left(\left((\bar{p}_{\mathcal{J}} - p_{\min}^{\max})^2 \, V \right)' [W] + \operatorname{div}\left((\bar{p}_{\mathcal{J}} - p_{\min}^{\max})^2 \, V \right) W \right) \\
&\quad = -\frac{1}{2\lambda} \int_{\mathcal{J}} \operatorname{div}\left(2(\bar{p}_{\mathcal{J}} - p_{\min}^{\max})(p'_{\mathcal{J}}[W] - \underbrace{p_{\min}^{\max\prime}[W]}_{=0}) V + (\bar{p}_{\mathcal{J}} - p_{\min}^{\max})^2 \underbrace{V'}_{=0} \right) \\
&\qquad -\frac{1}{2\lambda} \int_{\mathcal{J}} \operatorname{div}\left(\left(2(\bar{p}_{\mathcal{J}} - p_{\min}^{\max}) \nabla(\bar{p}_{\mathcal{J}} - p_{\min}^{\max}) \cdot V + (\bar{p}_{\mathcal{J}} - p_{\min}^{\max})^2 \operatorname{div} V \right) W \right) \\
&\quad = -\frac{1}{2\lambda} \int_{\beta} (\bar{p}_{\mathcal{J}} - p_{\min}^{\max}) \left(2p'_{\mathcal{J}}[W] V \cdot \boldsymbol{n}_{\mathcal{J}} + \left(2\nabla(\bar{p}_{\mathcal{J}} - p_{\min}^{\max}) \cdot V + (\bar{p}_{\mathcal{J}} - p_{\min}^{\max}) \operatorname{div} V \right) W \cdot \boldsymbol{n}_{\mathcal{J}} \right).
\end{aligned}$$

□

As an easy consequence, every second order shape semiderivative has a null at the optimal configuration.

Corollary 5 ($\mathcal{A}$ is a null of the second order shape semiderivative)**:**
Let $\mathcal{A} \in \mathcal{O}$ be the (optimal) active set.

Then for each $V, W \in \mathcal{V}$ the second order shape semiderivative of the reduced objective vanishes

$$d^2\mathcal{F}(\mathcal{A}; V, W) = 0.$$

Proof. The assertion of Theorem 8 yields that the trace of $\bar{p}_{\mathcal{I}} - p_{\min}^{\max}$ on the boundary γ of the active set $\mathcal{A}$ is zero

$$(\bar{p}_{\mathcal{I}} - p_{\min}^{\max})|_{\gamma} \equiv 0.$$

Consequently, the second order shape semiderivative of the shape functional $\mathcal{F}$ given by Lemma 11 vanishes for all $V, W \in \mathcal{V}$, too. □

Remark:
The assertion of Corollary 5 still holds true, when the velocity fields V and W are chosen to be nonautonomous, since the additional term induced thereby contains a $\bar{p}_{\mathcal{I}} - p_{\min}^{\max}$-factor; cf. the proof of Lemma 11.

Hence, the unique minimum $\mathcal{A}$ of the shape optimization problem (2.45) is a critical point of the first *and* the second order shape semiderivative of $\mathcal{F}$. This result has some important consequences:

- In order to prove that the optimal configuration $\mathcal{A}$ is an isolated critical point of $\mathcal{F}$ in $\mathcal{O}$, with respect to an appropriate topology, one would usually apply positive definiteness – or more precisely uniform ellipticity – of the Hessian, cf. [159, Thm. 4.23]. Ellipticity is obviously not given in the present context. Approaches which are used by Dambrine et al. in [38, 37, 39] seem to be not applicable here, and consequently one requires some suitable substitute.

- There is not only a lack of ellipticity, but there are even descent directions in $\mathcal{A}$, cf. Paragraph 3.1.1. It is argued there, that descent directions can only be avoided by constraining the set of feasible directions of variation by respecting the neglected inequality constraint. In other words, the inequality constraint does not has to be respected as long as first order necessary conditions are considered, but the situation changes completely with respect to second order analysis.
- The presence of descent direction influences the choice of algorithms since descent algorithms cannot be applied. Thus, one has to look for approaches related to Newton's method. This topic is discussed in-depth in Chapter 3.

2.5.2 Remarks on isolated critical points

The goal of this paragraph is to investigate, whether the optimal active set $\mathcal{A}$ is an isolated critical point of the reduced objective $\mathcal{F}$ defined in (2.45a). Since the second order shape derivative has a null at $\mathcal{A}$, cf. Corollary 5, the typical reasoning, i. e. using positive definiteness in order to get a quadratic growth condition, is not applicable. One might argue that perhaps a fourth order shape derivative might provide positive definiteness. However, higher order shape derivatives are not easily achievable. Furthermore, the results of the Paragraph 3.1.1 show, that there are even descent directions in $\mathcal{A}$ and thus it is not possible to prove positive definiteness of any higher order shape derivative. These descent directions have to be infeasible with respect to the strict inequality constraint (2.45c); otherwise $\mathcal{A}$ would not be an optimum. Consequently, positive definiteness of any higher order shape derivative can only be achieved – if at all – in a suitable restricted set of admissible directions, which is prescribed by the strict inequality. However, it is by far not an easy task to characterize this set of directions. All in all, the approach via higher order derivatives seems not very promising.

In particular, it is sufficient but not necessary to prove any growth condition in order to ensure, that $\mathcal{A}$ is an isolated critical point. The idea of the following approach is to directly prove that $\mathcal{A}$ is an isolated null of the shape gradient $\nabla\mathcal{F}$. However, all efforts made in order to get a rigorous proof suffer from a lack of uniformity with respect to the direction of variation in the end. Henceforth, it can only be conjectured, that the optimal configuration is an isolated critical point of the reduced shape functional $\mathcal{F}$.

Conjecture 1 ($\mathcal{A}$ is an isolated critical point of $\mathcal{F}$)**:**
The optimal active set $\mathcal{A}$ is an isolated critical point of the reduced shape functional $\mathcal{F}$ with respect to the Courant metric (cf. Lemma 12) in the set $\mathcal{X}(\mathcal{A}) \subset \mathcal{O}$ (see Lemma 13).

Remark:
Note, that the set $\mathcal{X}(\mathcal{A})$ is used here, in order to restrict the assertion to a subset of $\mathcal{O}$ where the shape sensitive Courant metric is defined. This topic is discussed in-depth in Paragraph 2.6.1.

Moreover, the assertion does not contain any claim with respect to changes of the topology of $\mathcal{A}$.

The main reason for the null of the second order shape semiderivative of $\mathcal{F}$ is due to the fact, that the term $\bar{p}_{\mathcal{J}} - p_{\min}^{\max}$ appears in quadratic form in the gradient, see (2.55). As already argued above, such problems can typically be overcome by means of higher order derivatives; but this approach is very difficult in the context of shape calculus. The proposed remedy is to analyze a single $\bar{p}_{\mathcal{J}} - p_{\min}^{\max}$-factor. Obviously, the null of $\nabla\mathcal{F}$ is isolated, if it is possible to prove, that the Dirichlet trace of $\bar{p}_{\mathcal{J}} - p_{\min}^{\max}$ is nonzero in a small neighborhood of $\mathcal{A}$ (except at $\mathcal{A}$ itself of course).

Hence, let $V \in \mathcal{V}$ be arbitrary and define the transformed inactive set $T_t(V)(\mathcal{I})$ by means of the velocity method, which is presented in the 2nd item of the discussion on page 72. Assume that for all $V \in \mathcal{V}$ the mapping $t \mapsto \bar{p}_{T_t(V)(\mathcal{I})}$ is continuously differentiable for $t \in [0;\delta]$ and for a sufficiently small $\delta > 0$. Then a first order Taylor expansion yields

$$\bar{p}_{T_t(V)(\mathcal{I})} - p_{\min}^{\max}|_{T_t(V)(\mathcal{I})} = \bar{p}_{\mathcal{I}} - p_{\min}^{\max}|_{\mathcal{I}} + d\big(\bar{p}_{\mathcal{I}} - p_{\min}^{\max}|_{\mathcal{I}}; V\big)t + o(t^2).$$

This equation is still valid if one applies the Dirichlet trace operator $(.)|_t := (.)|_{\partial T_t(V)(\mathcal{I})\setminus\Gamma}$

$$\begin{aligned}(\bar{p}_{T_t(V)(\mathcal{I})} - p_{\min}^{\max})|_t &= \underbrace{\bar{p}_{\mathcal{I}}|_\gamma - p_{\min}^{\max}|_\gamma}_{=0} + \Big(\underbrace{p'_{\mathcal{I}}[V]}_{=0} - \underbrace{p_{\min}^{\max\prime}[V]}_{=0}\Big)\Big|_\gamma t + \big(\nabla(\bar{p}_{\mathcal{I}} - p_{\min}^{\max})\cdot V\big)|_\gamma\, t + o(t^2)\\ &= \Big(\underbrace{\nabla_\gamma(\bar{p}_{\mathcal{I}} - p_{\min}^{\max})}_{=0}\cdot V\Big)t + \underbrace{\partial_{\boldsymbol{n}}^{\mathcal{I}}(\bar{p}_{\mathcal{I}} - p_{\min}^{\max})}_{\mu_\gamma}\, V\cdot \boldsymbol{n}_{\mathcal{I}}\, t + o(t^2)\\ &= \mu_\gamma V\cdot \boldsymbol{n}_{\mathcal{I}}\, t + o(t^2).\end{aligned}$$

Now assume in addition that $\mu_\gamma \neq 0$ almost everywhere on γ. Then for each $V \in \mathcal{V}\setminus\{0\}$ there exists a $\delta = \delta(V) > 0$ such that

$$(\bar{p}_{T_t(V)(\mathcal{I})} - p_{\min}^{\max})|_t \not\equiv 0, \quad \forall t \in [0;\delta].$$

Consequently, $\nabla\mathcal{F}(T_t(V)(\mathcal{A})) \neq 0$ for all those t. However, there is no guarantee, that $\inf\{\delta(V)\,|\,V \in \mathcal{V}\} > 0$. This can only be assured, if the above Taylor expansion features a qualitative property similar to a uniform estimate of the quadratic remainder term.

2.5.3 Total linearization

Besides second order analysis of the reduced shape functional $\mathcal{F}$, it is useful to consider second order derivatives of the Lagrangian (2.63) in order to construct second order algorithms based upon the formal Lagrange technique of Section 2.4. This perspective is pursued in Paragraph 3.3.3 and Appendix E.

2.6 Shape calculus and calculus on manifolds

This section is devoted to give a rather abstract point of view on shape optimization and shape calculus. Hereby, it is possible to gain some insight in the general structure, which constitutes the basis of shape optimization in general and of the set optimal control problem (2.30) in particular. The following considerations shall be seen as a starting point for a deeper analysis of this topic, since many details are rather suggested and conjectured than rigorously proven. It is very likely that the abstract point of view is well-known within the community of people who work in the field of shape optimization. Nonetheless, a comprehensive comparison of shape calculus and calculus on (infinite dimensional) manifolds is not available to the best of the author's knowledge. Moreover, it seems, that specific details that have to be considered when constructing optimization algorithms on manifolds have not found entrance into shape optimization yet. Indeed, the content of this section was motivated originally to extract the second covariant derivative of the reduced shape functional, since this object is required for Newton methods; cf. Section 3.2.

2.6.1 Decomposition of $\mathcal{O}$ into manageable subsets $\mathcal{X}(.)$

This Paragraph 2.6.1 is concerned with the analysis of the family of feasible sets $\mathcal{O}$ given by Definition 4. It will turn out, that $\mathcal{O}$ can be seen as a manifoldlike object. Manifolds possess three hierarchical layers of inherent structure:

- the global layer describing the object as a whole,
- the local layer described by charts,
- the infinitesimal layer described by differential calculus.

These three layers of $\mathcal{O}$ will be considered in-depth in the following. In order to do so, there are some natural conditions to be respected when analyzing $\mathcal{O}$ with respect to shape optimization.

1. Shape optimization does not deal with changes of the topology of the set that should be optimized.
2. The set $\mathcal{O}$ must not be quit.

The first condition is related to the global description layer of $\mathcal{O}$. The family $\mathcal{O}$ decomposes into subsets, each collecting all elements of the same topology. In other words, for every $\mathcal{B} \in \mathcal{O}$ define the subset

$$\mathcal{O}(\mathcal{B}) := \{M \in \mathcal{O}\,|\,M \text{ is homeomorphically homotope to } \mathcal{B} \text{ in } \Omega\}. \tag{2.71}$$

Two sets $M, N \subset \Omega$ are called *homeomorphically homotope (isotopic) in* Ω here, if there exists a homotopy

$$\Phi : [0;1] \times \Omega \to \Omega$$

that fulfills

$$\begin{aligned}
&\Phi(0, M) = M, \\
&\Phi(1, M) = N \text{ and} \\
&\forall t \in [0;1] : \Phi(t, .) : \Omega \to \Omega \text{ is a homeomorphism.}
\end{aligned}$$

The decomposition (2.71) induces an equivalence relation in $\mathcal{O}$

$$\mathcal{A} \sim \mathcal{B} \quad :\Leftrightarrow \quad \mathcal{A} \in \mathcal{O}(\mathcal{B}).$$

Let $\{\mathcal{B}_i\}_{i \in I}$ be a *collection of representatives* of each *equivalence class* (i. e. a *partition* of $\mathcal{O}$). Then $\mathcal{O}$ decomposes disjointly

$$\mathcal{O} = \bigcup_{i \in I} \mathcal{O}(\mathcal{B}_i)$$

and each set $\mathcal{O}(\mathcal{B}_i)$ is a maximal set on which shape optimization can act reasonably. In this context it is important to note the semantic in the definition of $\mathcal{O}(\mathcal{B})$: it is not enough to demand that M is homeomorphic to $\mathcal{B}$ (in $\mathbb{R}^2$), since there are obvious cases where the topology of M and $\mathcal{B}$ are the same (in $\mathbb{R}^2$), but the two sets cannot be mapped onto each other without changing their topology in Ω; cf. Figure 2.8. In addition, the property of continuous deformation of the original set to the other (homotopy) is fundamental from the point of view of shape optimization, which only deals with such kind of set transformations.

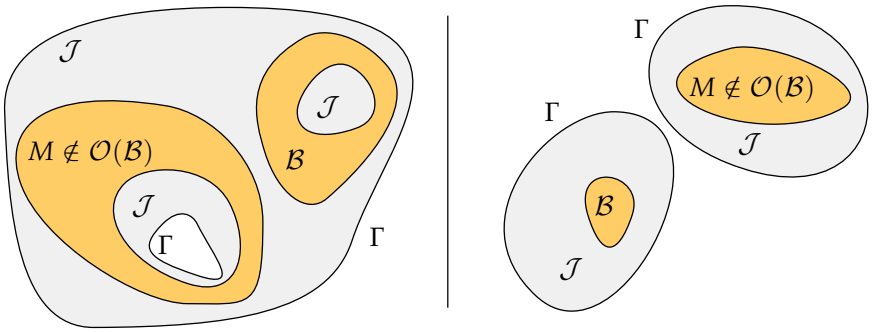


Figure 2.8: Illustration of sets $M \in \mathcal{O} \setminus \mathcal{O}(\mathcal{B})$: M is not the image of $\mathcal{B}$ with respect to any homeomorphism of Ω (left), whereas $\mathcal{B}$ can be mapped to M by means of a homeomorphism of Ω (right) but it cannot be continuously deformed into it while staying in Ω.

The second condition is related to the local description layer of $\mathcal{O}$. Shape optimization may only perform in such a way that it ensures that the image set $\mathcal{B}$ of a "deformation" of a set $\mathcal{A} \in \mathcal{O}$ is still contained in $\mathcal{O}$. Two different properties are important here. For one thing $C^{1,1}$-boundary regularity has to be preserved and for another thing the image set has to be located in Ω. "Deformations" can be modeled by transformations of $\mathbb{R}^2$, i. e. mappings $\mathbb{R}^2 \to \mathbb{R}^2$. Consequently, one has to choose a suitable class Θ_0 of such mappings.

Based upon Θ_0, which indeed is a Banach space of functions, it is possible to construct a set of charts, each of which is defined on a subset of the above constructed equivalence classes $\mathcal{O}(\mathcal{B}_i)$. Since the approach requires a bundle of notations and some notions of group theory, its various steps are illustrated in Figure 2.9 on page 62 for convenience.

The characterization is constructed in a series of three technical lemmas, which extensively use results from [44, Chp. 3 and Chp. 4] and which are followed by an in-depth discussion. The reader is also referred to the detailed presentation of Younes [161], in particular, chapters 8 and 12.

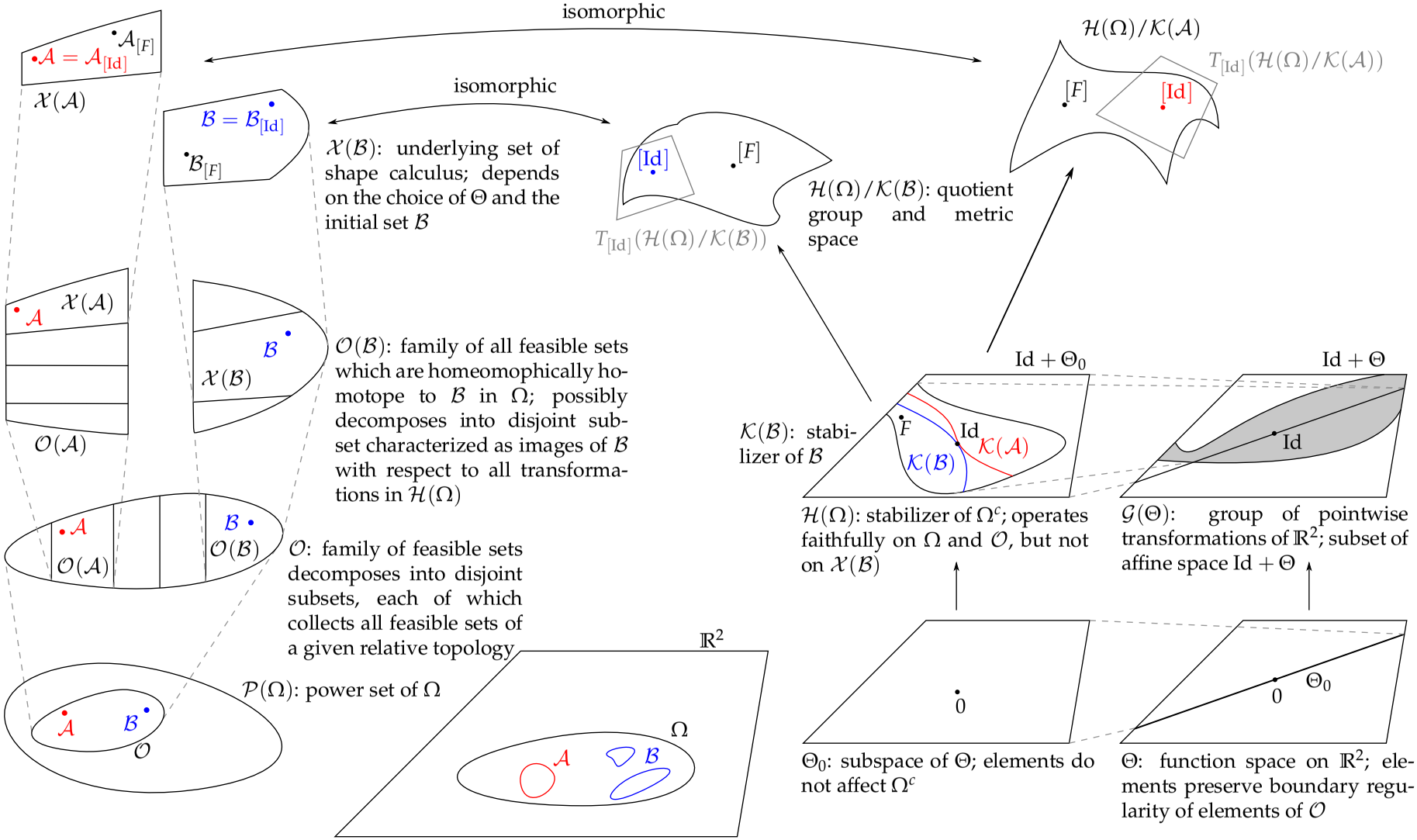


Figure 2.9: Illustration of the construction of $\mathcal{X}(\mathcal{A})$.

As a start let[21] $\Theta := C^{1,1}(\overline{\mathbb{R}^2}, \mathbb{R}^2)$ and consider the subspace

$$\mathcal{G}(\Theta) := \{\mathrm{Id} + f \,|\, f \in \Theta,\ \mathrm{Id} + f \text{ bijective and } (\mathrm{Id} + f)^{-1} - \mathrm{Id} \in \Theta\} \subset \mathrm{Id} + \Theta,^{22} \tag{2.72}$$

which has a group structure for the composition $(F \circ G)(x) := F(G(x))$. The elements of $\mathcal{G}(\Theta)$ are called *perturbations of identity*. Furthermore, $\mathcal{G}(\Theta)$ equipped with the right-invariant metric[23]

$$\begin{aligned} &d : \mathcal{G}(\Theta) \times \mathcal{G}(\Theta) \to \mathbb{R}, \quad (F, G) \mapsto d(F, G) := d(\mathrm{Id}, G \circ F^{-1}), \text{ where} \\ &d(\mathrm{Id}, F) := \inf_{\substack{n \in \mathbb{N} \\ F = F_1 \circ \cdots \circ F_n \\ F_i \in \mathcal{G}(\Theta)}} \sum_{i=1\ldots n} \|F_i - \mathrm{Id}\|_\Theta + \|F_i^{-1} - \mathrm{Id}\|_\Theta \end{aligned} \tag{2.73}$$

is a *complete right-invariant metric space*; cf. [44, Chp. 3 Thm. 2.9(i)]. Note, that the basic space Θ is chosen with respect to the abovementioned, local layer related condition. In particular, the elements of the group $\mathcal{G}(\Theta)$ preserve $C^{1,1}$-boundary regularity; cf. the proof of the 2nd item of Lemma 13.

Up to this point the considered diffeomorphisms do not ensure, that the images of a set $\mathcal{B} \in \mathcal{O}$ are located in Ω. Thus, the interest is focused on the set of all transformations in $\mathcal{G}(\Theta)$, which do not affect the complement Ω^c of the bounded domain Ω and which map some specific subset onto themselves:

Lemma 12:
Let the group $\mathcal{G}(\Theta)$ defined in (2.72) and let $\mathcal{B} \in \mathcal{O}$.
Then the sets of all transformation, which do not affect Ω^c and do not change $\mathcal{B}$ respectively

$$\mathcal{H}(\Omega) := \{F \in \mathcal{G}(\Theta) \,|\, F|_{\Omega^c} = \mathrm{Id}_{\Omega^c}\}, \tag{2.74}$$

$$\mathcal{K}(\mathcal{B}) := \{F \in \mathcal{H}(\Omega) \,|\, F(\mathcal{B}) = \mathcal{B}\} \tag{2.75}$$

are closed subgroups of $\mathcal{G}(\Theta)$ and $\mathcal{H}(\Omega)$ respectively.
Moreover, the function

$$d_\mathcal{K} : \mathcal{H}(\Omega) \times \mathcal{H}(\Omega) \to \mathbb{R}, \quad d_\mathcal{K}(F \circ \mathcal{K}(\mathcal{B}),\, H \circ \mathcal{K}(\mathcal{B})) := \inf_{K \in \mathcal{K}(\mathcal{B})} d(F, H \circ K),$$

induced by the metric d in $\mathcal{H}(\Omega)$, is a right-invariant metric on $\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{B})$, called *Courant metric*. The space $(\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{B}), d_\mathcal{K})$ is complete and the topology induced by $d_\mathcal{K}$ is equivalent to the quotient topology of $\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{B})$.

Remark:
The closed subgroup $\mathcal{K}(\mathcal{B})$ induces an equivalence relation in $\mathcal{H}(\Omega)$. Let $F, G \in \mathcal{H}(\Omega)$, then

$$F \sim_\mathcal{B} G \quad :\Leftrightarrow \quad F(\mathcal{B}) = G(\mathcal{B}).$$

The equivalence classes are denoted by $[.]_\mathcal{B}$ in the following and in most cases by $[.]$ if the choice of the defining set $\mathcal{B}$ is clear. Note, that $[\mathrm{Id}]_\mathcal{B} = \mathcal{K}(\mathcal{B})$.

Proof. **1)** This part is devoted to prove that $\mathcal{H}(\Omega) \subset \mathcal{G}(\Theta)$ is a closed subgroup.
Recall that $(\mathcal{G}(\Theta), d)$ is complete. As preliminary step, introduce an auxiliary metric. According to [44, Chp. 3 Thm. 2.3] the topologies induced by the metric d and the semimetric[24] $d_0 : \mathcal{G}(\Theta) \times \mathcal{G}(\Theta) \to \mathbb{R}$

$$(F, G) \mapsto d_0(F, G) := d_0(\mathrm{Id}, G \circ F^{-1}) := \|G \circ F^{-1} - \mathrm{Id}\|_\Theta + \|F \circ G^{-1} - \mathrm{Id}\|_\Theta,$$

are equivalent. Thus, they are equivalent metrics on $\mathcal{H}(\Omega)$ and one can start to prove the assertion.

[21] $C^1(\mathbb{R}^2, \mathbb{R}^2) := \{f : \mathbb{R}^2 \to \mathbb{R}^2 \,|\, \forall |\alpha| \le 1 : \partial^\alpha f \text{ is continuous in } \mathbb{R}^2\}$
$C^1(\overline{\mathbb{R}^2}, \mathbb{R}^2) := \{f \in C^1(\mathbb{R}^2, \mathbb{R}^2) \,|\, \forall |\alpha| \le 1 : \partial^\alpha f \text{ is bounded and uniformly continuous on } \mathbb{R}^2\}$
$C^{1,1}(\overline{\mathbb{R}^2}, \mathbb{R}^2) := \{f \in C^1(\overline{\mathbb{R}^2}, \mathbb{R}^2) \,|\, \forall |\alpha| \le 1\ \exists c > 0\ \forall x, y \in \mathbb{R}^2 : |\partial^\alpha f(x) - \partial^\alpha f(x)| \le c|x - y|\}$
[22] The somehow complicated notation is due to the fact, that $\mathrm{Id} \notin \Theta$, since Id is not bounded on $\mathbb{R}^2$.
[23] A metric d on a group $(\mathcal{G}, \circ)$ is said to be *right-invariant*, if for all $F, G, H \in \mathcal{G}$: $d(F \circ H, G \circ H) = d(F, G)$.
[24] A *semimetric* d_0 is a function with the following properties

(i) $d_0(F, G) \ge 0$
(ii) $d_0(F, G) = 0 \quad \Leftrightarrow \quad F = G$
(iii) $d_0(F, G) = d_0(G, F)$.

In other words, it lacks the triangle inequality in order to be a proper metric.

Obviously, $\mathcal{H}(\Omega)$ is a subgroup of $\mathcal{G}(\Theta)$ and it is sufficient to show, that any given sequence $(F_n)_{n\in\mathbb{N}} \subset \mathcal{H}(\Omega)$, which converges against an $F \in \mathcal{G}(\Theta)$, already converges in $\mathcal{H}(\Omega)$. Let $F \in \mathcal{G}(\Theta)$ be the limit of F_n, i. e. $d(F, F_n) \to 0$. Let $x \in \Omega^c$ be arbitrarily chosen. Since $F_n \in \mathcal{H}(\Omega)$ and by means of the equivalent metric, there holds

$$|F(x) - x| = |F \circ F_n^{-1}(x) - x| \leq \|F \circ F_n^{-1} - \mathrm{Id}\|_\Theta \leq d_0(F, F_n) \to 0\,,$$

thus $F|_{\Omega^c} = \mathrm{Id}_{\Omega^c}$. That is to say, $F \in \mathcal{H}(\Omega)$.

2) In this part it is proven, that $\mathcal{K}(\mathcal{B}) \subset \mathcal{H}(\Omega)$ is a closed subgroup and uses arguments similar to the proof of [44, Chp. 3 Lem. 2.3].
Obviously $\mathcal{K}(\mathcal{B})$ is a subgroup and it remains to show, that it is closed. Let $(F_n)_{n\in\mathbb{N}} \subset \mathcal{K}(\mathcal{B})$ be a sequence which converges against an $F \in \mathcal{H}(\Omega)$. By means of the equivalent metric d_0 there holds

$$\|F \circ F_n^{-1} - \mathrm{Id}\|_\Theta + \|F_n \circ F^{-1} - \mathrm{Id}\|_\Theta = d_0(F_n, F) \to 0.$$

Let $y \in \mathcal{B}$; then for each $n \in \mathbb{N}$ there is an $x_n \in \mathcal{B}$ with $F_n^{-1}(x_n) = y$, since $F_n \in \mathcal{K}(\mathcal{B})$. Moreover,

$$|F(y) - x_n| = |F \circ F_n^{-1}(x_n) - x_n| \leq \|F \circ F_n^{-1} - \mathrm{Id}\|_\Theta \to 0,$$

this is $F(y) \in \overline{\{x_n\}} \subset \bar{\mathcal{B}} = \mathcal{B}$. Consequently, $F(\mathcal{B}) \subset \mathcal{B}$ and there even holds $\overline{F(\mathcal{B})} \subset \bar{\mathcal{B}} = \mathcal{B}$, since $\mathcal{B}$ is closed.

On the other hand $F \circ F_n^{-1}(y) \in F(\mathcal{B})$ for each $n \in \mathbb{N}$ and $y \in \mathcal{B}$, since due to $F_n \in \mathcal{K}(\mathcal{B})$ there holds $x_n := F_n^{-1}(y) \in \mathcal{B}$. In addition, $F(x_n) \to y$ since

$$|F(x_n) - y| = |F \circ F_n^{-1}(y) - y| \leq \|F \circ F_n^{-1} - \mathrm{Id}\|_\Theta \to 0.$$

In other words, each $y \in \mathcal{B}$ is the limit of a sequence $F(x_n) \subset F(\mathcal{B})$. Consequently, $\mathcal{B} \subset \overline{F(\mathcal{B})}$.

All in all, one obtains $\mathcal{B} \subset \overline{F(\mathcal{B})} \subset \bar{\mathcal{B}} = B$. Continuity of F and closedness of $\mathcal{B}$ ensure that $F(\mathcal{B})$ is closed such that one finally gets $F(\mathcal{B}) = \mathcal{B}$. That is to say, $F \in \mathcal{K}(\mathcal{B})$.

3) The metric d is right-invariant on $\mathcal{G}(\Theta)$ and consequently on $\mathcal{H}(\Omega)$. Furthermore, $\mathcal{H}(\Omega)$ is complete as a closed subgroup of the complete metric space $\mathcal{G}(\Theta)$. The assertion now follows directly from [44, Chp. 3 Thm. 2.8]. □

By means of the metric group $\mathcal{G}(\Theta)$ and its subgroups defined in Lemma 12 some important subsets of the family $\mathcal{O}$ can be characterized:

Lemma 13:
Let $\mathcal{B} \in \mathcal{O}$ and let $\mathcal{H}(\Omega)$ and $\mathcal{K}(\mathcal{B})$ be given by Lemma 12. Consider the family of all images of $\mathcal{B}$ which can be obtained via transformations in $\mathcal{H}(\Omega)$

$$\mathcal{X}(\mathcal{B}) := \{F(\mathcal{B}) \subset \mathbb{R}^2 \,|\, F \in \mathcal{H}(\Omega)\}.$$

Then there holds:

1. $F \in \mathcal{H}(\Omega)$ is bijective on $\overline{\Omega}$ and $F(\Gamma) = \Gamma$,
2. $\mathcal{X}(\mathcal{B}) \subset \mathcal{O}$; in particular, there is $F(\mathcal{B}) \subset \Omega$ for $F(\mathcal{B}) \in \mathcal{X}(\mathcal{B})$,
3. there is a bijective map
$$\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{B}) \to \mathcal{X}(\mathcal{B}), \quad [F]_\mathcal{B} \mapsto \mathcal{B}_{[F]} := F(\mathcal{B}),$$
4. the Courant metric $d_\mathcal{K}$ in $\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{B})$ induces a metric in $\mathcal{X}(\mathcal{B})$
$$d(\mathcal{B}_{[F]}, \mathcal{B}_{[G]}) := d_\mathcal{K}([F]_\mathcal{B}, [G]_\mathcal{B}), \quad \forall\, \mathcal{B}_{[F]}, \mathcal{B}_{[G]} \in \mathcal{X}(\mathcal{B}).$$

In particular, the bijection given by the third assertion enables to identify the topological structure of $\mathcal{X}(\mathcal{B})$ with the one of the quotient group $\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{B})$.

Proof. **1)** Each $F \in \mathcal{H}(\Omega)$ fulfills $F|_{\Omega^c} = \mathrm{Id}_{\Omega^c}$ by definition. Hence, there holds $F(\Gamma) = \Gamma$ and moreover $F(\Omega) = \Omega$, since F is bijective as an element of $\mathcal{G}(\Theta)$.

2) Each $F(\mathcal{B}) \in \mathcal{X}(\mathcal{B})$ is an element of the family $\mathcal{O}$, i. e. it fulfills the Assumption 1. This can be seen as follows. Due to $F(\Omega) = \Omega$, there holds $F(\mathcal{B}) \subset \Omega$ for all $F \in \mathcal{H}(\Omega)$. Moreover, since $F \in \mathcal{H}(\Omega)$ is a

homeomorphism, the topologies of $\mathcal{B}$ and $F(\mathcal{B})$ (regarded as subsets of $\mathbb{R}^2$) are the same. Let $(x_n)_{n\in\mathbb{N}} \subset \mathcal{B}$ and $(y_n)_{n\in\mathbb{N}} \subset \mathcal{B}^c$ converge to an arbitrary but fixed $x \in \beta := \partial\mathcal{B}$, then

$$F(\mathcal{B}) \ni F(x_n) \to F(x) \quad \text{and} \quad F(\mathcal{B}^c) \ni F(y_n) \to F(x).$$

That is, $F(x) \in \overline{F(\mathcal{B})} \cap \overline{F(\mathcal{B}^c)} = \partial F(\mathcal{B})$ and thus $F(\beta) \subset \partial F(\mathcal{B})$. The inverse inclusion is proven with the same reasoning and the fact that $F^{-1} - \mathrm{Id} \in \Theta$, which yields F^{-1} is continuous on $\mathbb{R}^2$. In particular, the property $\overline{\mathring{\mathcal{B}}} = \mathcal{B}$ transfers to the image $F(\mathcal{B})$ and the boundaries of the connected components of $F(\mathcal{B})$ and Γ are pairwise disjoint. In addition, the boundary of $F(\mathcal{B})$ is of class $C^{1,1}$, since $\mathcal{B} \in \mathcal{O}$ is of class $C^{1,1}$ and since $F \in \mathcal{G}(\Theta)$, where $\Theta = C^{1,1}(\overline{\mathbb{R}^2}, \mathbb{R}^2)$. In other words, the boundary is locally the image of the $C^{1,1}$-regular map $F \circ \psi^q$, whereas ψ^q was used in the proof of Lemma 2; cf. Figure 2.1.

3) The map is onto: let $X \in \mathcal{X}(\mathcal{B})$, i. e. there is an $F \in \mathcal{H}(\Omega)$ with $F(\mathcal{B}) = X$. Consequently, $F \circ K(\mathcal{B}) = F(\mathcal{B}) = X$, for all $K \in \mathcal{K}(\mathcal{B})$. That is $X = G(\mathcal{B})$, for all $G = F \circ K \in [F]$.
Moreover, the map is injective: let $[F], [G] \in \mathcal{H}(\Omega)/\mathcal{K}(\mathcal{B})$ with $[F] \neq [G]$. In other words $F \neq G \circ K$ in $\mathcal{H}(\Omega)$, for all $K \in \mathcal{K}(\mathcal{B})$. Since $\mathcal{H}(\Omega)$ is a group, this is equivalent to $G^{-1} \circ F \neq K$, for all $K \in \mathcal{K}(\mathcal{B})$. Briefly worded $G^{-1} \circ F \notin \mathcal{K}(\mathcal{B})$. In particular, $G^{-1} \circ F(\mathcal{B}) \neq \mathcal{B}$ and consequently – since G is bijective on Ω and $G^{-1} \circ F(\mathcal{B}) \subset \Omega$ – there holds $F(\mathcal{B}) \neq G(\mathcal{B})$.

4) According to Lemma 12 $d_{\mathcal{K}}$ is a metric on the quotient group $\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{B})$. This metric is induced to $\mathcal{X}(\mathcal{B})$ via the constructed bijection. □

The subgroup $\mathcal{H}(\Omega)$ has the advantageous property that "small" transformation in $\mathcal{H}(\Omega)$ can be associated with a path, which lies in $\mathcal{H}(\Omega)$.

Lemma 14:
For $\varepsilon > 0$ define the ball around the identity

$$B_\varepsilon := \{F \in \mathcal{H}(\Omega) \,|\, d(\mathrm{Id}, F) < \varepsilon\} \tag{2.76}$$

and define the family of associated paths

$$\mathcal{T}(B_\varepsilon) := \{T_t(F) := \mathrm{Id} + t(F - \mathrm{Id}) \,|\, F \in B_\varepsilon,\, t \in [0,1]\}\,. \tag{2.77}$$

Then there exists an $\varepsilon > 0$ such that $\mathcal{T}(B_\varepsilon) \subset \mathcal{H}(\Omega)$. In other words, each $F \in B_\varepsilon$ is isotopic to Id within $\mathcal{H}(\Omega)$.

Proof. **1)** The first step is to prove that there is an $\varepsilon > 0$ such that $\mathcal{T}(B_\varepsilon) \subset \mathcal{G}(\Theta)$.
Thereto, use the equivalence between the topologies induced by the (semi-)metrics d and d_0, as it has been done in the proof of Lemma 12:

$$\exists\, \varepsilon > 0 \quad \forall F \in B_\varepsilon \,:\, d_0(\mathrm{Id}, F) < 1.$$

Consequently, there holds for $F \in B_\varepsilon$ and $t \in [0,1]$

$$\|t(F - \mathrm{Id})\|_\Theta \leq \|F - \mathrm{Id}\|_\Theta \leq \|F - \mathrm{Id}\|_\Theta + \|F^{-1} - \mathrm{Id}\|_\Theta = d_0(\mathrm{Id}, F) < 1. \tag{2.78}$$

That is to say, $t(F - \mathrm{Id}) \in B_1(0) \subset \Theta$.[25] Using [44, Chp. 3 Thm. 2.17], one obtains $\mathcal{T}(B_\varepsilon) \subset \mathcal{G}(\Theta)$.
2) It remains to show that that $\mathcal{T}(B_\varepsilon) \subset \mathcal{H}(\Omega)$.
Let $F \in B_\varepsilon$ (in particular $F|_{\Omega^c} = \mathrm{Id}_{\Omega^c}$) and let $t \in [0,1]$, then there holds

$$\forall x \in \Omega^c \,:\quad T_t(F)(x) = x + t\,(F - \mathrm{Id})(x) = x + t\,(\mathrm{Id}_{\Omega^c} - \mathrm{Id})(x) = x.$$

Thus, each $T_t(F) \in \mathcal{T}(B_\varepsilon)$ is in $\mathcal{H}(\Omega)$, too. □

It may be instructive to consider the assertions of lemmas 12 to 14 from the perspective of group theory. For convenience, the required and rather elementary notions from group theory are collected in the Appendix D in Definition 21.

1. The group $\mathcal{G}(\Theta)$ operates faithfully on the points $X \in \mathbb{R}^2$.
2. The subgroup $\mathcal{H}(\Omega)$ ensures, that the domain Ω is invariant with respect to the action of its elements. This is important, since one is interested in transformations of the (candidate) active set $\mathcal{B}$ while the holdall Ω (as a set) should remain unchanged. In other words, the subgroup $\mathcal{H}(\Omega) \subset \mathcal{G}(\Theta)$ is the isotropy group of Ω^c with respect to the group operation.

[25] $B_1(0)$ is the ball with radius 1 around the null function in Θ with respect to $\|.\|_\Theta$

3. Another reasonable choice for a subgroup $\mathcal{H}(\Omega)$ would be $\{F \in \mathcal{G}(\Theta) \mid F(\overline{\Omega}) = \overline{\Omega}\}$. This subgroup is larger than $\mathcal{H}(\Omega)$ and thus may yield richer quotient groups. However, it has the drawback that there may be $f \in B_1(0)$ such that $\mathrm{Id} + f$ is in this subgroup, whereas $\mathrm{Id} + tf$ is not. Consequently, the assertion of Lemma 14 cannot be verified in this situation. A prototypic counterexample is illustrated in Figure 2.10. In this respect, confer also the discussion of transformation- and flow approach of path following in Section 3.1 on page 92ff.

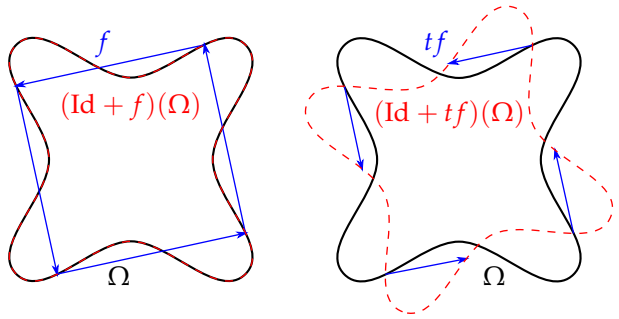


Figure 2.10: Illustration of a transformation $\mathrm{Id} + f$, which fixes Ω, whereas $\mathrm{Id} + tf$ does not.

4. For the purpose of optimization one is essentially interested in how $F \in \mathcal{H}(\Omega)$ transforms the elements of $\mathcal{O}$ interpreted as subsets of Ω (this is how they are deformed). Hence, one focuses on a group operation on the set of all feasible active sets $\mathcal{O}$, and not on a group operation on the points of $\mathbb{R}^2$. Fortunately, $\mathcal{H}(\Omega)$ operates faithfully on $\mathcal{O}$: Indeed, it operates on $\mathcal{O}$ as it is shown in the second part of the proof of Lemma 13. In addition, let $F \in \mathcal{H}(\Omega) \setminus \{\mathrm{Id}\}$, let $X_0 \in \Omega$ with $|(F - \mathrm{Id})(X_0)| = \max_{X \in \Omega} |(F - \mathrm{Id})(X)|$ and let $0 < \varepsilon < \min\{|(F - \mathrm{Id})(X_0)|/2, \mathrm{dist}(X_0, \Gamma)\}$. Then the ε-ball $B_\varepsilon(X_0)$ and its image $F(B_\varepsilon(X_0))$ are contained in $\mathcal{O}$, but $F(B_\varepsilon(X_0)) \neq B_\varepsilon(X_0)$, since $F(X_0) \notin B_\varepsilon(X_0)$. Consequently, only $\mathrm{Id} \in \mathcal{H}(\Omega)$ is the identity on $\mathcal{O}$.
5. $\mathcal{H}(\Omega)$ operates faithfully, but obviously not transitive, since $\mathcal{O}$ contains sets of different topology, which cannot be mapped onto each other by the elements of the group $\mathcal{H}(\Omega)$. Consequently, there are several orbits; actually there are at least countably many, since for each number of connected components of $\mathcal{B} \in \mathcal{O}$ there is at least one orbit of its own. Thus, one question occurs immediately:

   Is there more than one orbit for a fixed topological configuration? More precisely, is the equivalence class $\mathcal{O}(\mathcal{B})$, which is defined in (2.71), identically equal to the orbit $\mathcal{X}(\mathcal{B})$, defined in Lemma 13, or is it a proper superset?

   If $\mathcal{O}(\mathcal{B})$ is a proper superset of $\mathcal{X}(\mathcal{B})$, this fact has fundamental impact on algorithms that shall solve the set optimal control problem (2.30) and which are supposed to be constructed on the basis of shape calculus: Suppose the right topology of the active set $\mathcal{A}$ is known apriori, but its exact shape is not. Then it is (at least in principle) an easy task to start the algorithm with some initial guess in $\mathcal{O}(\mathcal{A})$. This is a mandatory choice since a shape calculus based algorithm cannot be expected to change the topology of the set which has to be optimized. However, actually one has to start the algorithm with some initial guess in the *apriori unknown* set $\mathcal{X}(\mathcal{A})$, in order to ensure that the algorithm, which uses transformations in $\mathcal{H}(\Omega)$ as iterative steps, at least has a chance to converge to $\mathcal{A}$. Otherwise there were no hope for success, since all the iterates that the algorithm can produce were located in the orbit of the initial guess, which is disjoint to the orbit $\mathcal{X}(\mathcal{A})$.

   Unfortunately, this question remains open, to the best of the author's knowledge. Apparently, this question has not gained much attention in the community of shape optimization. Perhaps this is due to the somehow unfavorable position between questions about solvability of shape optimization problems on the one hand and questions about how to design efficient numerics, which are located in the regime of finite dimension after discretization, on the other hand.

   Nonetheless, this question is related to the infinite dimensional framework and should not cause problems in the finite dimensional world after discretization, as long as questions concerning convergence of the discretized problem towards the continuous one are not addressed.
6. It is possible to repeat the derivation of the characterization of the subsets $\mathcal{X}(\mathcal{A})$ of $\mathcal{O}$ when starting with other function spaces Θ, as long as the requirement that the corresponding transformations yield images in $\mathcal{O}$ is fulfilled (cf. page 61). One can choose, for instance, $\Theta := \mathcal{D}(\mathbb{R}^2)$, the space of

all infinitely often differentiable (smooth) functions with compact support. This choice may lead to a smaller set $\mathcal{X}(\mathcal{A})$, since the set of transformations is smaller then. However, a higher order of differentiability of the functions in Θ may yield a higher order of differentiability of the objects $\mathcal{X}(.)$ regarded as manifolds (see below). This observation is relevant with respect to vector bundles which are constructed on $\mathcal{X}(.)$; cf. the 16th item on page 78.

7. The subgroup $\mathcal{K}(\mathcal{B}) \subset \mathcal{H}(\Omega)$ is the stabilizer of the set $\mathcal{B}$ with respect to the operation on the family $\mathcal{O}$, introduced in 4th item. Note, that the subgroups $\mathcal{H}(\Omega) \subset \mathcal{G}(\Theta)$ and $\mathcal{K}(\mathcal{B}) \subset \mathcal{H}(\Omega)$ are the stabilizers of the sets Ω^c and $\mathcal{B}$ with respect to different operations: a pointwise operation of $\mathbb{R}^2$ and a "setwise" operation on $\mathcal{O}$.
8. There are many different transformations in $\mathcal{H}(\Omega)$ which map $\mathcal{B}$ to a fixed image $F(\mathcal{B})$. However, the stabilizer $\mathcal{K}(\mathcal{B})$ induces an equivalence relation on the metric space $\mathcal{H}(\Omega)$

   $$\forall F, G \in \mathcal{H}(\Omega): \quad F \sim G :\Leftrightarrow F(\mathcal{B}) = G(\mathcal{B}) \Leftrightarrow F \circ \mathcal{K}(\mathcal{B}) = G \circ \mathcal{K}(\mathcal{B}) \Leftrightarrow [F] = [G] \text{ in } \mathcal{H}(\Omega)/\mathcal{K}(\mathcal{B}).$$

   Consequently, the transition from $\mathcal{H}(\Omega)$ to the quotient group ensures the unique identification of the image set $\mathcal{B}_{[F]} = F(\mathcal{B})$ with respect to a given equivalence class of transformations $[F]$. The equivalence class is constructed in such a way that all $F \in \mathcal{H}(\Omega)$, which transform $\mathcal{B}$ and $\mathcal{J}$ without moving the interface β are disregarded.

   In other words, the whole approach from $\mathcal{G}(\Theta)$ via $\mathcal{H}(\Omega)$ through to $\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{A})$ yields a characterization of the set $\mathcal{X}(\mathcal{A})$ by means of the bijection stated in Lemma 13. And the set $\mathcal{X}(\mathcal{A})$ is the relevant object, wherein shape optimization takes place. To be more precise, it would be favorable to have a detailed characterization of the set $\mathcal{O}(\mathcal{A})$, since it is exactly the subset of the admissible set $\mathcal{O}$, which is related to the shape optimization part of the set optimal control problem (2.30), but with regard to the 5th remark there is no deeper analysis available yet but for $\mathcal{X}(\mathcal{A})$.

Apart from the group theoretic point of view, it is paying to analyze $\mathcal{H}(\Omega)$ from the perspective of metric spaces.

9. Define the closed subspace

   $$\Theta_0 := \{f \in \Theta \,|\, f|_{\Omega^c} \equiv 0\} \tag{2.79}$$

   and consider the associated metric group $\mathcal{G}(\Theta_0)$. Obviously, there holds $\mathcal{G}(\Theta_0) = \mathcal{H}(\Omega)$. Although one is interested in $\mathcal{G}(\Theta_0)$ actually, since it is a rather natural choice for a space of transformations, that ought to map the family $\mathcal{O}$ into itself, the approach via $\mathcal{H}(\Omega)$ was chosen here, since it is straight forward to use the results of [44, Chp. 3].
10. Furthermore, it is known, that the tangent space $T_F\mathcal{G}(\Theta)$ to $\mathcal{G}(\Theta)$ is Θ for all $F \in \mathcal{G}(\Theta)$; cf. [44, Chp. 3 Thm. 2.17]. The proof can be carried over to $\mathcal{H}(\Omega)$ easily and one obtains that the tangent space $T_F\mathcal{H}(\Omega)$ to $\mathcal{H}(\Omega)$ is Θ_0 for each $F \in \mathcal{H}(\Omega)$. An illustration of the relation of the tangent space Θ_0, the affine space $\mathrm{Id} + \Theta_0$ and the subset $\mathcal{H}(\Omega)$ is given in Figure 3.1.
11. According to [44, Chp. 3 Rem. 2.6], the framework presented here, is similar to an infinite dimensional *Riemannian manifold* and hence there are similarities to the theory of *Lie groups*[26]. Note, that the notions of manifold as well as Lie group are typically defined in the finite dimensional case, and thus may only help to give a conception of the considered metric space with group structure. Nonetheless, there is literature available concerning the infinite dimensional case; see, for instance, [113, 106].
12. The quotient group of a Lie group and a closed subgroup carries a manifold structure; cf., for example [83, Thm. 10.1.10]. Moreover, the tangent space of the quotient manifold can be characterized by means of the *Lie algebras* of the Lie group and its subgroup; cf. [83, Cor. 10.2.13].

    In view of these results, it should be worthwhile having a closer look at the analogous, infinite dimensional situation of the tangent space of $\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{B})$. This goal is all the more interesting because shape calculus and algorithms of shape optimization essentially rely on the tangent spaces of the set of admissible shapes. This set is $\mathcal{X}(\mathcal{A}) \cong \mathcal{H}(\Omega)/\mathcal{K}(\mathcal{A})$ here. Although this topic goes beyond the scope of this thesis, and the analysis is confined to the tangent space of $\mathcal{H}(\Omega)$, a qualitative consideration shall give an expectation of the tangent space of $\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{B})$ here: Imagine a smooth deformation of the set $\mathcal{B}$ driven by $t \mapsto \mathrm{Id} + tf \in \mathcal{H}(\Omega)$ and regard its influence on the set $\mathcal{B}$ as t converges to zero. Due to the factorization with $\mathcal{K}(\mathcal{B})$ both the nature of f in the bulks

[26] For a precise definition of the notions of manifold and Riemannian manifold, cf., for instance, Definition 14 or [1, Sec. 3.1.1 and 3.6] and for those of Lie group, cf., for instance, [83, Def. 9.1.1].

of $\mathcal{B}$ and $\mathcal{J} = \Omega \setminus \mathcal{B}$, and any part of the transformation which acts in tangential direction on the boundary β are irrelevant. Thus, it can be expected that the tangent space of the quotient manifold $\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{B})$ is given by $\{(f \cdot \boldsymbol{n}_{\mathcal{B}})|_{\beta} \,|\, f \in \mathcal{H}(\Omega)\}$.

This observation is in accordance with the *Hadamard Structure Theorem* [44, Chp. 9 Thm.3.6], which says that shape gradients are concentrated on the boundary of the deformed set and are dependent on the normal component of the deformation vector field only. Confer also the 13th item on page 77.

13. In view of the manifoldlike character of $\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{B}) \cong \mathcal{X}(\mathcal{B})$, the orbits $\mathcal{X}(M)$, $M \in \mathcal{O}(\mathcal{B})$ might be regarded as domains of definition of charts of the (possibly larger) manifoldlike object $\mathcal{O}(\mathcal{B})$. In particular, the family of orbits $\{\mathcal{X}(M) \,|\, M \in \mathcal{O}(\mathcal{B})\}$ defines an *atlas* of $\mathcal{O}(\mathcal{B})$. However, it is not obvious how these charts should look like. A definition in the style of
$$\varphi_M : \mathcal{X}(M) \cong \mathcal{H}(\Omega)/\mathcal{K}(M) \to \Theta_0, \quad M_{[F]} \cong [F] \mapsto F - \mathrm{Id}$$
is too simple, since the representative F has to be specified in order to get a well-defined map; perhaps some kind of determination via a minimal norm property might help.

If $\mathcal{O}(\mathcal{B})$ is a proper superset of $\mathcal{X}(\mathcal{B})$ in the sense of the 5th item, it seems that $\mathcal{O}$ is not connected, since there is no overlap area of the different charts. It should be emphasized however, that the definition of a chart usually is based on the fact, that its image is an open set. This property may be fulfilled within the above considerations, but the question is left open here. It is possible to define charts with open image in the tangential space indeed by means of Lemma 14, but it is not decided definitely here, whether $\mathcal{O}(\mathcal{B})$ is connected or not.

14. According to [44, Chp. 3 Thm. 2.17(i)], the mapping $\Theta \supset B_1(0) \to \mathcal{G}(\Theta)$, $f \mapsto \mathrm{Id} + f$ is well-defined and continuous. This yields that the tangential space Θ can be mapped to the manifoldlike space $\mathcal{G}(\Theta)$, locally around $f = 0$. The mapping is surjective onto an ε-ball around $\mathrm{Id} \in \mathcal{G}(\Theta)$, which can be proven with the idea of the proof of Lemma 14. This assertion carries over to $B_1(0) \subset \Theta_0$ and $\mathcal{H}(\Omega)$. It would be desirable at this point, if this consideration could be advanced to the quotient space $\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{B})$. For one thing it would be possible to analyze shape functionals (at least locally) be means of the tangent space instead of regarding them on the set $\mathcal{X}(\mathcal{B})$, which is much easier, since the tangent space is expected to be a Banach space. For another thing it would no longer be necessary to analyze the local properties of shape functionals in the tangent space Θ_0 of $\mathcal{H}(\Omega)$, which is not perfectly adapted to the situation, since the corresponding group operation acts pointwisely and not setwisely; cf. the 4th remark.
15. The whole approach yields a metric on the feasible set $\mathcal{O}$. It is possible to introduce other metrics, cf. [44, Chp. 5, 6, 7] or [161], however the Courant metric is distinguished: If a shape functional is Hadamard semidifferentiable, then it is continuous with respect to the Courant metric; see [44, Chp. 9, Thm. 3.3]. Moreover, the method of perturbation of identity is directly connected with the approach.

Up to this point the global and the local description layer of the family of feasible sets $\mathcal{O}$ have been investigated. It remains to analyze the infinitesimal layer, which means putting shape calculus into the framework of differential calculus on manifolds.

2.6.2 Abstract view on shape calculus

Therefore, up next a very brief crash course in first and second order calculus on smooth (i. e. C^∞-) manifolds, which is based on [1, Chp. 3 and 5]. It is far beyond the scope of this thesis either to give a sufficient review of this topic or to prove that the results thereof can be carried over to shape calculus directly. In particular, the manifold $\mathcal{O}$ is not C^∞ if it is equipped with the structure, which was constructed in Paragraph 2.6.1, since the elements of Θ are not C^∞ (cf. the 6th item of the discussion above). The aim is to give some insight into the structure of calculus on manifolds and thereby to establish some understanding of shape calculus, without having to cope with technical details.

The first intermediate goal is the notion of a derivative of a smooth function on a manifold. The main difficulty is the fact that a direct generalization of directional derivatives in Banach spaces (i. e. *Gateaux semiderivative*, [44, Chp. 9 Def. 2.1(i)]) as
$$df(x, \xi) := \lim_{t \to 0} \frac{f(x + t\xi) - f(x)}{t}$$

is not possible, since the argument $x + t\xi$ has no meaning due to the lack of a linear vector space structure on a manifold. As a consequence, directional derivatives are constructed by means of curves on the manifold in manner of *Hadamard semiderivatives*; cf. [44, Chp. 9 Def. 2.1(ii)].

Definition 8 (directional derivative, tangent vector, tangent space)**:**
Let $\mathcal{M}$ be a (real) manifold, let $\mathfrak{F}_x(\mathcal{M})$ be the set of real-valued, smooth functions defined in a neighborhood of $x \in \mathcal{M}$. Furthermore, let $\alpha : \mathbb{R} \to M$ be a smooth mapping with $\alpha(0) = x$, called *smooth curve in* $\mathcal{M}$ *trough* x. Then

1. $\frac{\mathrm{d}}{\mathrm{d}t} f(\alpha(t))|_{t=0}$ is called the *directional derivative of* $f \in \mathfrak{F}_x(\mathcal{M})$ *in* x
2. the map $\dot{\alpha}(0) : \mathfrak{F}_x(\mathcal{M}) \to \mathbb{R}$, $f \mapsto \dot{\alpha}(0)\, f := \frac{\mathrm{d}}{\mathrm{d}t} f(\alpha(t))|_{t=0}$ is called *tangent vector* or *canonical lifting of the curve* α *at* $t = 0$
3. a *tangent vector* ξ_x *to the manifold* $\mathcal{M}$ *at point* x is defined as mapping $\mathfrak{F}_x(\mathcal{M}) \to \mathbb{R}$ such that there exists a curve α trough x ($\alpha(0) = x$) with
$$\xi_x f = \dot{\alpha}(0)\, f$$
4. the set $T_x\mathcal{M}$ of all tangent vectors ξ_x at point x is called *tangent space to* $\mathcal{M}$ *at* x and admits a linear vector space structure.

Definition 9 (tangent bundle, vector field, derivation)**:**
Let $\mathcal{M}$ be a (real) manifold and let $\mathfrak{F}(\mathcal{M})$ be the set of real-valued, smooth functions defined on $\mathcal{M}$. Then

1. the *tangent bundle* $T\mathcal{M}$ is the collection of all tangent vectors to $\mathcal{M}$[27]
$$T\mathcal{M} := \bigcup_{x \in \mathcal{M}} T_x\mathcal{M}$$
2. a smooth mapping $\xi : \mathcal{M} \to T\mathcal{M}$, $x \mapsto \xi_x \in T_x\mathcal{M}$ is called *vector field on* $\mathcal{M}$
3. the set of all vector fields on $\mathcal{M}$ is denoted by $\mathfrak{V}(\mathcal{M})$
4. a *derivation at* $x \in \mathcal{M}$ is defined as a mapping $D_x : \mathfrak{F}(\mathcal{M}) \to \mathbb{R}$, which fulfills
   a) $\mathbb{R}$-linearity: $\forall a, b \in \mathbb{R}, f, g \in \mathfrak{F}(\mathcal{M}) : \quad D_x(af + bg) = aD_x(f) + bD_x(g)$ and
   b) the product rule: $\forall f, g \in \mathfrak{F}(\mathcal{M}) : \quad D_x(fg) = D_x(f)g + fD_x(g)$
5. a *derivation on* $\mathfrak{F}(\mathcal{M})$ is a mapping $D : \mathfrak{F}(\mathcal{M}) \to \mathfrak{F}(\mathcal{M})$, which fulfills
   a) $\mathbb{R}$-linearity: $\forall a, b \in \mathbb{R}, f, g \in \mathfrak{F}(\mathcal{M}) : \quad D(af + bg) = aD(f) + bD(g)$ and
   b) the product rule: $\forall f, g \in \mathfrak{F}(\mathcal{M}) : \quad D(fg) = D(f)g + fD(g)$.

The notion of a derivation axiomizes the notion of vector fields and as a result of this it axiomizes the notion of the covariant derivative (cf. Definition 11). Each vector field $\xi \in \mathfrak{V}(\mathcal{M})$ defines a derivation
$$D : \mathfrak{F}(\mathcal{M}) \to \mathfrak{F}(\mathcal{M}), \quad f \mapsto D(f) := \xi f,$$
where (by means of Definition 8 for a suitable curve α)
$$\forall x \in \mathcal{M} : \quad (\xi f)(x) := \xi_x f = \dot{\alpha}(0) f = \frac{\mathrm{d}}{\mathrm{d}t} f(\alpha(t))|_{t=0} \in \mathbb{R}.$$
Vice versa, each derivation on $\mathfrak{F}(\mathcal{M})$ can be realized by a vector field. Consequently, it is sufficient to maintain the notion of a vector field and the notion of a derivation can be abandoned.

Definition 10 (covector, cotangent space, cotangent bundle)**:**
Let $\mathcal{M}$ be a (real) manifold. Then

1. a *covector at* $x \in \mathcal{M}$ is a linear functional $\mu_x : T_x\mathcal{M} \to \mathbb{R}$
2. the set of all covectors at $x \in \mathcal{M}$ form the *cotangent space* $T_x^*\mathcal{M}$ *to* $\mathcal{M}$ *at* x, which is the dual space of the tangent space $T_x\mathcal{M}$

[27] Even if two tangent spaces $T_x\mathcal{M}$ and $T_y\mathcal{M}$ ($x, y \in \mathcal{M}$) are isomorphic, their elements ξ_x and ξ_y are not identified with each other in $T\mathcal{M}$. Consequently, each element $\xi_x \in T\mathcal{M}$ is characterized by the tangent vector $\xi_x \in T_x\mathcal{M}$ itself and its *foot* $x \in \mathcal{M}$.

3. the *cotangent bundle* $T^*\mathcal{M}$ is the collection of all covectors to $\mathcal{M}$
$$T^*\mathcal{M} := \bigcup_{x\in\mathcal{M}} T_x^*\mathcal{M}.$$

Definition 11 (covector field, covariant derivative of functions)**:**
Let $\mathcal{M}$ be a (real) manifold. Then

1. a *covector field* is a smooth map $\mu : \mathcal{M} \to T^*\mathcal{M}$, $x \mapsto \mu_x$
2. the set of all covector fields is referred to as $\mathfrak{V}(\mathcal{M})^*$
3. a covector field μ acts on a vector field $\xi \in \mathfrak{V}(\mathcal{M})$ as follows
$$\forall x \in \mathcal{M}: \quad (\mu[\xi])(x) := \mu_x[\xi_x] \in \mathbb{R}$$
and consequently $\mu[\xi] \in \mathfrak{F}(\mathcal{M})$
4. for each $f \in \mathfrak{F}(\mathcal{M})$ there exists one distinct covector field, the *covariant derivative of* f, defined by
$$\mathrm{D}f : \mathcal{M} \to T^*\mathcal{M},\ x \mapsto (\mathrm{D}f[.])(x), \text{ where } \forall \xi \in \mathfrak{V}(\mathcal{M}) : (\mathrm{D}f[\xi])(x) := \xi_x f.$$

The covariant derivative is the generalization of the common concept of the first order derivative of a function $f : \mathbb{R}^N \to \mathbb{R}$, and thus the first intermediate goal is reached. The next step is to provide tools needed for a generalization of a second order derivative.

Second order derivatives are based on the notion of the first order derivative of vector fields. However, it is not possible to use the (manifold intrinsic) concept of curves to define directional derivatives of vector fields, as it was done in Definition 8 in order to introduce directional derivatives of real-valued functions:
$$\lim_{t\to 0} \frac{\xi_{\alpha(t)} - \xi_{\alpha(0)}}{t}$$
is not well-defined since the tangent vectors $\xi_{\alpha(t)}$ and $\xi_{\alpha(0)}$ live in different tangent spaces. Consequently, one has to choose an axiomatic approach to introduce the notion of a covariant derivative of a vector field, which is similar to the axiomatic definition of derivations in Definition 9.

Definition 12 (covariant derivative of (co-)vector fields)**:**
Let $\mathcal{M}$ be a (real) manifold. Then

1. a *covariant derivative of a vector field* $\xi \in \mathfrak{V}(\mathcal{M})$ *with respect to the direction* $\eta \in \mathfrak{V}(\mathcal{M})$ is a mapping $\nabla_\eta\xi : \mathcal{M} \to T\mathcal{M}$, $x \mapsto (\nabla_\eta\xi)(x)$ which fulfills
   a) $\mathfrak{F}(\mathcal{M})$-linearity in η: $\forall f, g \in \mathfrak{F}(\mathcal{M})$, $\eta, \chi, \xi \in \mathfrak{V}(\mathcal{M})$: $\nabla_{f\eta+g\chi}\xi = f\nabla_\eta\xi + g\nabla_\chi\xi$
   b) $\mathbb{R}$-linearity in ξ: $\forall a, b \in \mathbb{R}$, $\eta, \xi, \zeta \in \mathfrak{V}(\mathcal{M})$: $\nabla_\eta(a\xi + b\zeta) = a\nabla_\eta\xi + b\nabla_\eta\zeta$
   c) the product rule: $\forall f \in \mathfrak{F}(\mathcal{M})$, $\eta, \xi \in \mathfrak{V}(\mathcal{M})$: $\nabla_\eta(f\xi) = (\eta f)\xi + f\nabla_\eta\xi$,[28]
2. a *covariant derivative of a covector field* $\mu \in \mathfrak{V}(\mathcal{M})^*$ can be defined by
$$\forall \eta, \xi \in \mathfrak{V}(\mathcal{M}) : (\nabla_\eta\mu)[\xi] := \eta(\mu[\xi]) - \mu[\nabla_\eta\xi] = (\mathrm{D}(\mu[\xi]))[\eta] - \mu[\nabla_\eta\xi].$$

Remark:
The notion of a covariant derivative of a vector field is closely related to the notion of an *affine connection* on the manifold; cf., for instance, [1] or [113, p. 101ff.]. In fact each affine connection defines a covariant derivative and vice versa. Furthermore, it is known, that there are infinitely many affine connections on a manifold [1, p. 94]. One or another of them may distinguish itself with respect to computational accessibility or some other properties, for instance, the *Riemannian/Levi-Civita connection*. All in all, there is some freedom for the choice of a covariant derivative ∇ on a manifold. In particular, one has to choose, since there is none of them given apriori.

[28] Note, that there is an essential difference between $f\xi$ and ξf for $f \in \mathfrak{F}(\mathcal{M})$ and $\xi \in \mathfrak{V}(\mathcal{M})$: $(f\xi)(x) = f(x)\xi_x \in T_x\mathcal{M}$ is a simple multiplication, whereas $(\xi f)(x) = \xi_x f = (\mathrm{D}f[\xi])(x) \in \mathbb{R}$ is the application of a directional derivative.

Let $\mathcal{M}$ be a (real) manifold. Furthermore, let $f \in \mathfrak{F}(\mathcal{M})$ and let $\eta \in \mathfrak{V}(\mathcal{M})$. Hence, the covariant derivative of f applied to the vector field η is

$$F := \mathrm{D}f[\eta] \in \mathfrak{F}(\mathcal{M})$$

and it is possible to derive the covariant derivative of F itself by repeated differentiation. By doing so, it becomes apparent, that the definition of the covariant derivative of a covector field is chosen in such a way that the expected rules for differentiation are fulfilled. Thus, let $\xi \in \mathfrak{V}(\mathcal{M})$ be another vector and compute

$$\mathrm{D}F[\xi] = \mathrm{D}(\mathrm{D}f[\eta])[\xi] = (\nabla_\xi(\mathrm{D}f))[\eta] + \mathrm{D}f[\nabla_\xi \eta]. \tag{2.80}$$

This equation can be understood as product rule for the differentiation of the "product" $\mathrm{D}f[\eta]$. It accounts for some inherent, general properties of derivatives; cf. [1, p. 95f.]: A directional derivative depends locally on the object to be differentiated and pointwisely on the direction towards which is differentiated. Locally means, that the object has to be known in a local neighborhood around the point of evaluation. In particular, it is not enough if one has information about the vector field η at some point $x \in \mathcal{M}$ only in order to compute $\mathrm{D}(\mathrm{D}f[\eta])[\xi](x)$, since it is not possible to derive $\mathrm{D}f[\nabla_\xi \eta](x)$ then.

Definition 13 (second covariant derivative)**:**
Let $\mathcal{M}$ be a (real) manifold. Then the *second covariant derivative* ∇^2 of a function $f \in \mathfrak{F}(\mathcal{M})$ is defined as

$$\forall \eta, \xi \in \mathfrak{V}(\mathcal{M}) : \ \nabla^2 f[\xi, \eta] := (\nabla_\xi(\mathrm{D}f))[\eta].$$

By means of the notion of the second covariant derivative one recognizes that the second directional derivative $\mathrm{D}(\mathrm{D}f[\eta])[\xi]$ of a function $f \in \mathfrak{F}(\mathcal{M})$ decomposes into two parts: the second covariant derivative and one term which contains the (first order) covariant derivative.

Herewith, the second intermediate goal of the notion of a second order derivative is reached. However, the terms *shape gradient* and *shape Hessian* are frequently used in the context of shape calculus. Thus, it is worthwhile to introduce the notions of *(Riemannian) gradient* and *(Riemannian) Hessian* in the calculus on (Riemannian) manifolds in a third intermediate step.

Definition 14 (Riemannian metric, Riemannian manifold)**:**
Let $\mathcal{M}$ be a (real) manifold and let all tangent spaces $T_x\mathcal{M}$ be Hilbert spaces with an inner product (symmetric positive definite)

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

Then

1. the mapping $g : x \mapsto g_x$ is called a *Riemannian metric on* $\mathcal{M}$, if for all vector fields $\xi, \zeta \in \mathfrak{V}(\mathcal{M})$ the map
$$\mathcal{M} \to \mathbb{R}, \quad x \mapsto \langle \xi_x \ , \ \zeta_x \rangle_x$$
is smooth,
2. a pair $(\mathcal{M}, g)$ of a manifold and a Riemannian metric is called a *Riemannian manifold*.

Definition 15 (gradient, Hessian)**:**
Let $(\mathcal{M}, g)$ be a Riemannian manifold and let $f \in \mathfrak{F}(\mathcal{M})$ be a smooth function.

1. The *gradient* $\operatorname{grad} f$ *of* f *(with respect to the metric* g*)* is the Riesz representative of the covariant derivative $\mathrm{D}f$. In other words, it is defined as the unique vector field that fulfills
$$\forall x \in \mathcal{M}, \quad \forall \xi \in \mathfrak{V}(\mathcal{M}) : \quad \big\langle (\operatorname{grad} f)(x) \ , \ \xi_x \big\rangle_x = (\mathrm{D}f[\xi])(x)$$
2. Let ∇ be an affine connection on $\mathcal{M}$. Then the *Hessian* $\operatorname{Hess} f$ *of* f *(with respect to the metric* g *and the affine connection* ∇*)* is the mapping
$$\operatorname{Hess} f : \mathfrak{V}(\mathcal{M}) \to \mathfrak{V}(\mathcal{M}), \ \xi \mapsto \nabla_\xi \operatorname{grad} f.$$
This is the linear operator induced by the second covariant derivative
$$\forall \xi, \eta \in \mathfrak{V}(\mathcal{M}) : \quad \big\langle (\operatorname{Hess} f)[\xi] \ , \ \eta \big\rangle = \big\langle \nabla_\xi \operatorname{grad} f \ , \ \eta \big\rangle = \nabla^2 f[\xi, \eta].$$

By means of gradient and Hessian the second directional derivative (2.80) can be expressed as follows

$$\mathrm{D}(\mathrm{D}f[\eta])[\xi] = \nabla^2 f[\xi,\eta] + \mathrm{D}f[\nabla_\xi \eta] = \big\langle (\mathrm{Hess}\, f)[\xi]\ ,\ \eta \big\rangle + \big\langle \mathrm{grad}\, f\ ,\ \nabla_\xi \eta \big\rangle.$$

The last intermediate step is concerned with *vector bundles*, see for example [83, Def. 10.2.2] or [113, Chp. III §1]. Vector bundles are a generalization of tangent bundles and help to understand the idea of function space parametrization in shape calculus and of shape dependent functions in general. The definition is given for the finite dimensional case in order to hide some technical overhead, although it is used in the infinite dimensional setting later on.

Definition 16 (vector bundle)**:**
Let $\mathcal{M}$ and E be (real) manifolds and let B be a Banach space.

Then

1. E together with a smooth map $\pi : E \to \mathcal{M}$ is called a *smooth vector bundle on* $\mathcal{M}$ if the following conditions are fulfilled
   a) For each $x \in \mathcal{M}$, there is an open neighborhood U of x in $\mathcal{M}$ and a diffeomorphism
   $$\varphi_U : \pi^{-1}(U) \to U \times B$$
   commuting with the projection on U, $\mathrm{pr}_U : U \times B \to U$, $(x,v) \mapsto x$. That is, the following diagram is commutative
   $$\begin{array}{ccc} \pi^{-1}(U) & \xrightarrow{\varphi_U} & U \times B \\ & {\scriptstyle \pi}\searrow \quad \swarrow{\scriptstyle \mathrm{pr}_U} & \\ & U & \end{array}$$
   and in particular, by means of the projection $\mathrm{pr}_B : U \times B \to B$, $(x,v) \mapsto v$, one obtains an isomorphism for each $x \in U$
   $$\varphi_U^x := \mathrm{pr}_B \circ \varphi_U|_{\pi^{-1}(x)} : \pi^{-1}(x) \to B.$$
   b) For each $x \in \mathcal{M}$ the set $\pi^{-1}(x)$ carries the structure of a *Banachable space* (i. e. a complete topological space whose topology can be defined by a norm) and the maps
   $$\varphi_U^x : \pi^{-1}(x) \to B$$
   are linear and continuous.
2. The spaces E and B are called *total space* and *base space* of the vector bundle, respectively.
3. The sets $B_x := \pi^{-1}(x)$ are called *fibers* of the bundle.
4. The prototype Banach space B is often called *standard fiber* of the bundle.
5. The maps φ_U are called *trivializing maps* of the vector bundle.
6. A vector bundle E is called *trivial*, if it is isomorphic to $\mathcal{M} \times B$. (Note, that a vector bundle is always trivial on U, i. e. a vector bundle is always locally trivial.)

Remark:
Let $\mathcal{M}$ be a manifold and $T\mathcal{M}$ its tangent bundle. Then $T\mathcal{M}$ is a vector bundle on $\mathcal{M}$ together with the natural projection $\pi : T\mathcal{M} \to \mathcal{M}$, $T_x\mathcal{M} \ni \xi_x \mapsto x$.

The tight review of various notions known from the theory of manifolds is finished here. Following the approach of Delfour and Zolésio in [44, Chp. 9] one finds the following similarities among shape calculus and the differential calculus on manifolds, which is already slightly indicated in [42]:

1. The metric space $\mathcal{H}(\Omega)$ plays the role of an infinite dimensional manifold; cf. the 11th and the 13th item of the discussion on page 67.
2. Transformations $f := F - \mathrm{Id} \in \Theta_0$ and velocity fields $V \in \mathcal{V}$ define paths in $\mathcal{H}(\Omega)$ by means of transformation- and flow approach, respectively (see page 92). Let $F \in \mathcal{H}(\Omega)$, let $f \in \Theta_0$ and let

$V \in \mathcal{V}$ then there exists an interval $I \subset \mathbb{R}$ containing 0 such that one gets paths (straight lines and integral curves, cf. the 5th item) through F

$$I \to \mathcal{H}(\Omega), t \mapsto T_t(f) := F + tf$$

$$I \to \mathcal{H}(\Omega), t \mapsto T_t(V) := x(t,.), \text{ the solution of } \frac{\mathrm{d}}{\mathrm{d}t}x(t,X) = V(x(t,X)),\ x(0,X) := F(X),\ X \in \Omega.$$

Consequently, f and V define tangent vectors $f = \partial_t(T_t(f))|_{t=0}$ and $V = \partial_t(T_t(V))|_{t=0}$.

However, the spaces Θ_0 (see (2.79)) and $\mathcal{V}$ (see (2.47)) are not the same. In particular, $\Theta_0 \subset \mathcal{V}$ since each $f \in \Theta_0$ fulfills $f \cdot \boldsymbol{n} \equiv 0$ on Γ but not every $V \in \mathcal{V}$ is such that $V|_\Gamma = 0$. Hence, following Definition 8, the two approaches seem to induce different tangent spaces to $\mathcal{H}(\Omega)$. Several facts should be noted in this respect:

- The difference between Θ_0 and $\mathcal{V}$ is concentrated in a neighborhood of the boundary Γ, i. e.
$$\forall K \subset\subset \Omega, V \in \mathcal{V} \quad \exists f \in \Theta_0 : \quad V|_K \equiv f|_K.$$
- As already discussed in the 8th and the 12th item on page 67f. one actually is interested in optimization within the set $\mathcal{O}(\mathcal{A})$ or at least $\mathcal{X}(\mathcal{A}) \cong \mathcal{H}(\Omega)/\mathcal{K}(\mathcal{A})$. By means of the consequential fact that shape calculus is "set sensitive" only but not "pointwise sensitive"[29] (cf. the 4th item of the above mentioned discussion) and by means of the requirement that the distance between the boundary of the active set $\mathcal{B} \in \mathcal{O}$ and the boundary of the holdall Ω is positive $\operatorname{dist}(\gamma, \Gamma) \geq \delta > 0$ (cf. Assumption 1, and the proof of Lemma 4), one can deduce that the constitution of the transformation near Γ does not influence any result. Consequently, one should expect no disagreement when applying transformations $f \in \Theta_0$ instead of velocity fields $V \in \mathcal{V}$.
- These considerations find expression in two important results of shape calculus. For one thing the *method of perturbation of identity* (which corresponds to the usage of $f \in \Theta_0$) and the *velocity method* yield the same shape derivatives [44, Chp. 9 Thm. 3.1] and for another thing the relevant part of the shape derivative is concentrated on the perturbed interface [44, Chp. 9 Thm. 3.6, Cor. 3.1].
- Nonetheless, it should be emphasized that the two approaches are not equivalent in general and that only the velocity method is straight forward applicable in presence of a constraining holdall (see [44, Chp. 9 Sec. 3.3]). A detailed analysis of the structure of shape derivatives in the presence of a holdall can be found in [43].

3. The tangent bundle $T\mathcal{H}(\Omega)$ is trivial since $\mathcal{H}(\Omega)$ is an open submanifold of the affine space $\mathrm{Id} + \Theta_0$; cf. the reasoning on page 95. Moreover, it is naturally isomorphic to $\mathcal{H}(\Omega) \times \Theta_0$, in the sense, that one can use the identity as isomorphism.

   Thus, many notions concerning manifolds, which have to be distinguished in general denote the same object in the context of $\mathcal{H}(\Omega)$, since one rather is in framework of vector spaces than in those of manifolds. Especially the handling of vector fields and many therefrom deduced notions is considerably simplified. This finding is amplified by the fact that the manifold $\mathcal{H}(\Omega)$ can be covered by a single chart. Unfortunately, it is not possible to give a complete overview of simplifications here, and many interesting consequences are left to the reader. He is referred to the extensive textbook of Lang [113] once again.

4. One outcome of the (natural) triviality of the tangent bundle $T\mathcal{H}(\Omega) = \mathcal{H}(\Omega) \times \Theta_0$ is, that an $f \in \Theta_0$ (as well as an $V \in \mathcal{V}$) is a tangent vector to all $F \in \mathcal{H}(\Omega)$ simultaneously. In particular, there is a natural notion of *parallel transport* (see [108, Chp. II Sec. 3]) of tangent vectors. Two elements $f \in T_F\mathcal{H}(\Omega) = \Theta_0$ and $g \in T_G\mathcal{H}(\Omega) = \Theta_0$ are parallel if and only if $f = g \in \Theta_0$.

   This canonical parallel transport induces a canonical choice for a covariant derivative of vector fields on $\mathcal{H}(\Omega)$, cf. the 7th item.

5. Vector fields on $\mathcal{H}(\Omega)$ in the sense of Definition 8 are smooth maps
$$\mathcal{H}(\Omega) \ni F \mapsto \zeta_F \in T_F\mathcal{H}(\Omega) = \Theta_0 \quad \text{or respectively} \quad \mathcal{H}(\Omega) \ni F \mapsto \xi_F \in \mathcal{V}.$$

[29] All functions considered in this thesis are dependent on specific sets only ($\mathcal{F} = \mathcal{F}(\mathcal{B})$, $u_{\mathcal{J}} = u_{\mathcal{J}}(\mathcal{J})$, $\sigma = \sigma(\beta)$, ...) and are invariant with respect to pointwise reparametrization of the sets. This means the functions fulfill a so called *compatibility condition*; cf. [44, p. 202]

They are not common in shape calculus, but it is usual to work with nonautonomous velocity fields (these are *time-dependent vector fields*, cf. [113, Chp. IV §1 and §2]) and their corresponding paths (these are *integral curves*, ibidem). A nonautonomous velocity field $V \in C^1([0;\tau],\mathcal{V})$ induces an integral curve in the manifold $\mathcal{H}(\Omega)$ through $\mathrm{Id} \in \mathcal{H}(\Omega)$ by means of

$$T_\cdot(V) : [0;\tau] \to \mathcal{H}(\Omega), \quad t \mapsto T_t(V) := x(t,.), \text{ where}$$
$$\frac{\mathrm{d}}{\mathrm{d}t}x(t,X) = V(t,x(t,X)), \quad x(0,X) := X \in \Omega.$$

This finding can be generalized to an integral curve in $\mathcal{H}(\Omega)$ through arbitrary elements $F \in \mathcal{H}(\Omega)$ by imposing the initial condition $x(0,X) := F(X)$ for $X \in \Omega$, as it was done in the 2nd item.

In contrast, the paths defined via the transformation approach $t \mapsto F + tf$ define straight lines in $\mathcal{H}(\Omega)$ regarded as subset of the affine space $\mathrm{Id} + \Theta_0$.

6. There are two different canonical identifications of the tangent bundle of the manifold $\mathcal{H}(\Omega)$. One of them is induced by the underlying Banach space Θ_0, since $\mathcal{H}(\Omega)$ is an open subset of the affine space $\mathrm{Id} + \Theta_0$. The other one is due to the group structure of $\mathcal{H}(\Omega)$ with respect to composition. It will be come apparent in the following (in particular, cf. the 18th item), that the first one gives the notion of a shape derivative (cf. [151, Sec. 2.30], whereas the second one yields the *Eulerian* or *material derivative* (cf. [151, Sec. 2.11 and 2.25]).

   Let $\xi \in \mathfrak{V}(\mathcal{H}(\Omega))$ be a vector field. It is a smooth map

   $$\xi : \mathcal{H}(\Omega) \to T\mathcal{H}(\Omega), \quad F \mapsto \xi_F \in T_F\mathcal{H}(\Omega) = \Theta_0.$$

   The underlying Banach space structure permits the canonical identification of the tangent bundle. That is, for arbitrary $F \in \mathcal{H}(\Omega)$ identify $T_F\mathcal{H}(\Omega)$ with the standard fiber $T_{\mathrm{Id}}\mathcal{H}(\Omega) = \Theta_0$ by means of the trivializing map[30]

   $$\varphi_{\mathrm{Id}} : T\mathcal{H}(\Omega) \to \mathcal{H}(\Omega) \times \Theta_0,$$

   which is defined fiber-wise by means of the identity

   $$\varphi^F_{\mathrm{Id}} : T_F\mathcal{H}(\Omega) \to \Theta_0, \quad \xi_F \mapsto \xi_F.$$

   This results in the identification of different tangent spaces by means of the identity

   $$(\varphi^G_{\mathrm{Id}})^{-1} \circ \varphi^F_{\mathrm{Id}} : T_F\mathcal{H}(\Omega) \to T_G\mathcal{H}(\Omega), \quad \xi_F \mapsto (\varphi^G_{\mathrm{Id}})^{-1} \circ \varphi^F_{\mathrm{Id}}(\xi_F) = \mathrm{Id}(\xi_F) = \xi_F.$$

   The group structure permits another canonical identification of tangent bundle. That is, for arbitrary $F \in \mathcal{H}(\Omega)$ identify $T_F\mathcal{H}(\Omega)$ with the standard fiber $T_{\mathrm{Id}}\mathcal{H}(\Omega) = \Theta_0$ by means of the trivializing map

   $$\varphi_\circ : T\mathcal{H}(\Omega) \to \mathcal{H}(\Omega) \times \Theta_0,$$

   which is defined fiber-wise by means of the pull-back to Id

   $$\varphi^F_\circ : T_F\mathcal{H}(\Omega) \to \Theta_0, \quad \xi_F \mapsto \xi_F \circ F.$$

   This results in the identification of different tangent spaces by means of the composition of transformations

   $$(\varphi^G_\circ)^{-1} \circ \varphi^F_\circ : T_F\mathcal{H}(\Omega) \to T_G\mathcal{H}(\Omega), \quad \xi_F \mapsto (\varphi^G_\circ)^{-1} \circ \varphi^F_\circ(\xi_F) = \xi_F \circ F \circ G^{-1}.$$

7. There is a canonical covariant derivative of vector fields on $\mathcal{H}(\Omega)$ which is induced by the underlying Banach space Θ_0, since $\mathcal{H}(\Omega)$ is an open subset of the affine space $\mathrm{Id} + \Theta_0$. Let $\xi, \eta \in \mathfrak{V}(\mathcal{H}(\Omega)$ be two vector fields. They are smooth maps $\xi, \eta : \mathcal{H}(\Omega) \to T\mathcal{H}(\Omega)$, $F \mapsto \xi_F, \eta_F \in T_F\mathcal{H}(\Omega) = \Theta_0$. By means of the (global) chart $\phi : \mathcal{H}(\Omega) \to \Theta_0$, $F \mapsto \phi(F) := F - \mathrm{Id}$, the vector fields can be uniquely identified with smooth vector fields on (a subset of) the Banach space Θ_0:

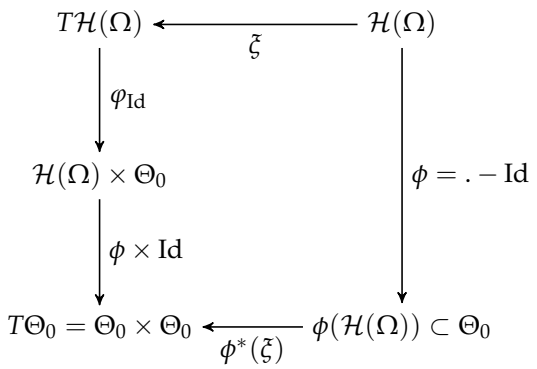
   $$\phi^* : \mathfrak{V}(\mathcal{H}(\Omega)) \to \mathfrak{V}(\Theta_0), \quad \xi \mapsto \phi^*(\xi),$$

   where for $f \in \mathcal{H}(\Omega) - \mathrm{Id} = \phi(\mathcal{H}(\Omega)) \subset \Theta_0$

   $$(\phi^*(\xi))_f := \varphi^{\phi^{-1}(f)}_{\mathrm{Id}}(\xi_{\phi^{-1}(f)}) = \xi_{\mathrm{Id}+f}.$$

   The following commutative diagram illustrates the situation.

[30] Note, that in defiance of the notation of Definition 16 the subscript of the trivializing map does not indicate the local chart here, but the connection to the identity in order to distinguish it from the trivializing map to be introduced below.



A canonical choice for a covariant derivative of ξ with respect to direction η at F is then given by the directional derivative of the projection $\phi^*(\xi)$ with respect to η_F at the projected point $\phi(F) = F - \mathrm{Id}$. In order to mark that this choice it related to the trivializing map φ_{Id} of 6th item, the covariant derivatives gets the superscript Id

$$\nabla^{\mathrm{Id}}_\eta \xi : \mathcal{H}(\Omega) \to T\mathcal{H}(\Omega), \quad F \mapsto (\nabla^{\mathrm{Id}}_\eta \xi)_F \in T_F\mathcal{H}(\Omega),$$

where for $F = \mathrm{Id} + f \in \mathcal{H}(\Omega)$

$$(\nabla^{\mathrm{Id}}_\eta \xi)_F := \lim_{t\to 0} \frac{(\phi^*(\xi))_{f+t(\phi^*(\eta))_f} - (\phi^*(\xi))_f}{t} = \lim_{t\to 0} \frac{\xi_{F+t\eta_F} - \xi_F}{t}.$$

Consequently, the definition of the covariant derivative corresponds to the usual directional derivative in the Banach space Θ_0. It is important here, that $F + t\eta_F = (\mathrm{Id} + t\eta_F \circ F^{-1}) \circ F$ is an element of $\mathcal{H}(\Omega)$ for sufficiently small $t \geq 0$. Fortunately this is true, cf. the 3rd item on page 66.

8. In order to see the relation of the covariant derivative from the 7th item to the derivative of nonautonomous velocity fields (cf. [44, Chp. 9 Sec. 6.3]), let $\alpha : I \subset \mathbb{R} \to \mathcal{H}(\Omega)$ be an *integral curve* to η through F. That is to say,

$$\forall t \in I : \quad \dot\alpha(t) = \eta_{\alpha(t)} \text{ and } \alpha(0) = F \quad (\text{in particular } 0 \in I).$$

($\dot\alpha : I \to T\mathcal{H}(\Omega)$ is called *canonical lifting* of α (see [114, Chp. IV §3]).) This defines a nonautonomous velocity field (cf. [113, Chp. IV §2])

$$W : I \times \Omega \to \mathbb{R}^2, \quad (t,x) \mapsto W(t,x) := \eta_{\alpha(t)}(x).$$

In particular, there holds

$$\forall t \in I : \quad W(t,.) \in T_{\alpha(t)}\mathcal{H}(\Omega)$$
$$\forall X \in \Omega, \forall t \in I : \quad \dot\alpha(t)(X) = W(t, \alpha(t)(X)), \quad \alpha(0)(X) = F(X).$$

The second vector field ξ defines a nonautonomous velocity field V (*lifting*) along α

$$\forall x \in \Omega, \forall t \in I : \quad V(t,x) := \xi_{\alpha(t)}(x).$$

Now one recognizes, that there holds[31]

$$\begin{aligned}
(\nabla^{\mathrm{Id}}_\eta \xi)_F(x) &= \lim_{t\to 0} \frac{\xi_{F+t\eta_F}(x) - \xi_F(x)}{t} \\
&= \lim_{t\to 0} \frac{\xi_{\alpha(t)}(x) - \xi_{\alpha(0)}(x)}{t} \\
&= \lim_{t\to 0} \frac{V(t,x) - V(0,x)}{t} \\
&= \frac{\partial}{\partial t} V(t,x)\Big|_{t=0} \\
&= V'(0,x).
\end{aligned}$$

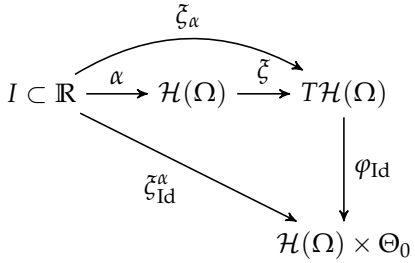
[31] A detailed analysis for the justification for the transition from Gateaux to Hadamard semiderivatives is left open here and the reader is referred to [44, Chp. 9 Sec. 3.1 and 3.3].

In other words, the derivative is given by

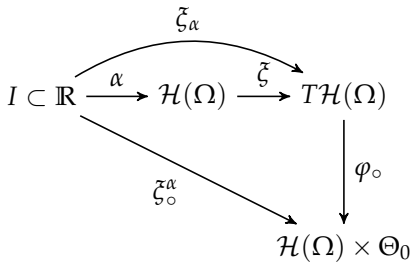
$$(\nabla^{\mathrm{Id}}_{\eta}\xi)_F = \frac{\mathrm{d}}{\mathrm{d}t}\xi^{\alpha}_{\mathrm{Id}}(t)\Big|_{t=0},$$

where $\xi^{\alpha}_{\mathrm{Id}}$ is the by means of φ_{Id} to the standard fiber Θ_0 transported lifting ξ_α

$$\xi^{\alpha}_{\mathrm{Id}} : I \to \Theta_0, \quad t \mapsto \varphi^{\alpha(t)}_{\mathrm{Id}}(\xi_{\alpha(t)}) = \xi_{\alpha(t)}.$$



9. The construction of the 8th item can be repeated with use of the trivializing map $\varphi_\circ$ from the 6th item instead of φ_{Id}. The transport of the lifting ξ_α to the standard fiber Θ_0 now reads



$$\xi^{\alpha}_{\circ} : I \to \Theta_0, \quad t \mapsto \varphi^{\alpha(t)}_{\circ}(\xi_{\alpha(t)}) = \xi_{\alpha(t)} \circ \alpha(t).$$

This gives rise to a second canonical covariant derivative

$$\begin{aligned}
(\nabla^{\circ}_{\eta}\xi)_F(x) \;&:= \frac{\mathrm{d}}{\mathrm{d}t}\xi^{\alpha}_{\circ}(t)\Big|_{t=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\xi_{\alpha(t)}\big(\alpha(t)(X)\big)\Big|_{t=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}t}V\big(t,\alpha(t)(X)\big)\Big|_{t=0} \\
&= \frac{\partial}{\partial t}V\big(t,\alpha(t)(X)\big)\Big|_{t=0} + \mathrm{D}_x V\big(t,\alpha(t)(X)\big)\,\dot{\alpha}(t)(X)\Big|_{t=0} \\
&= V'\big(0,\alpha(0)(X)\big) + \mathrm{D}_x V\big(0,\alpha(0)(X)\big)\,W\big(0,\alpha(0)(X)\big) \\
&= V'(0,x) + \mathrm{D}_x V(0,x)\,W(0,x) \\
&= (\nabla^{\mathrm{Id}}_{\eta}\xi)_F(x) + D_x\xi_F(x)\,\eta_F(x).
\end{aligned}$$

As it becomes apparent during the course of the 18th item, the approach via $\varphi_\circ$ is more general, since covariant derivatives in more general vector bundles on $\mathcal{H}(\Omega)$ can be constructed that way. Nonetheless, the covariant derivative ∇^{Id} is useful in order to understand the differences between material and shape derivatives, from the perspective of calculus on the manifold $\mathcal{H}(\Omega)$.

10. To the best of the author's knowledge, the usual shape calculus avoids the introduction of a covariant derivative of vector fields and consequently of covector fields, too. One confines oneself with the computation of first and second order directional derivatives of shape functionals. This approach is general enough to be able to extract the essence of shape calculus in terms of differential calculus on manifolds at least up to second order derivatives. In other words, one recognizes that these objects possess some intrinsic structure afterwards.

11. The notion of the *Hadamard semiderivative* [44, Chp. 9 Rem. 2.1] corresponds to the notion of the directional derivative from Definition 8.

12. In contrast, the directional derivative which corresponds to the method of perturbation of identity, is a *Gateaux derivative* [44, Chp. 9 Def. 2.1(i)]. Moreover, the transformation approach is essentially based on the fact, that the manifold $\mathcal{H}(\Omega)$ is embedded to the affine space $\mathrm{Id} + \Theta_0$.[32] Hence, it is possible (by means of the notation of the Remark on page 63 and of Lemma 13) to define

$$\mathcal{F}_{\mathcal{B}} : B_1(0) \subset \Theta_0 \to \mathbb{R}, f \mapsto \mathcal{F}_{\mathcal{B}}(f) := \mathcal{F}([\mathrm{Id} + f]_{\mathcal{B}}(\mathcal{B})) = \mathcal{F}(\mathcal{B}_{[\mathrm{Id}+f]})$$

for each $\mathcal{B} \in \mathcal{O}$. In other words, it is possible to locally transform the shape functional $\mathcal{F}$ defined on the manifold $\mathcal{X}(\mathcal{B}) \subset \mathcal{O}$ into a functional $\mathcal{F}_{\mathcal{B}}$ defined on the Banach space Θ_0. This notation is also used by Delfour and Zolésio; cf. [44, Chp. 9 Sec. 3.3].

However, it is important to notice that this kind of notation is not perfectly adapted to the situation of shape calculus, where the considered objects usually are set dependent only. The shape functional $\mathcal{F}$ is defined on $\mathcal{O}$ or at least on a the subset $\mathcal{X}(\mathcal{B})$ (for an arbitrarily chosen $\mathcal{B} \in \mathcal{O}$), but the local definition of $\mathcal{F}_{\mathcal{B}}$ is based on a subset of Θ_0 which corresponds to $\mathcal{H}(\Omega)$. In particular, all $f \in [F]_{\mathcal{B}} - \mathrm{Id} := \{G - \mathrm{Id} \in \Theta \,|\, G \in [F]_{\mathcal{B}}\}$ yield the same set $\mathcal{B}_{[F]} \in \mathcal{X}(\mathcal{B}) \subset \mathcal{O}$ and hence they evaluate $\mathcal{F}$ at the same point (i. e. set $\mathcal{B}_{[F]}$).

A remedy would be to transport the equivalence classes $[F]_{\mathcal{B}} \subset \mathcal{H}(\Omega)$ down to the Banach space Θ_0 – that is to define the quotient of Θ_0 and the equivalence relation induced by the classes $[F]_{\mathcal{B}} - \mathrm{Id}$. But the equivalence classes are no (affine) linear subspaces and consequently the quotient is no linear space. Hence, this reasoning contradicts the original goal of transporting shape calculus into a Banach space.

This aim can only be achieved by the introduction of a retraction (cf. Definition 20) or charts on the manifold $\mathcal{X}(\mathcal{B})$. However, this topic goes beyond the scope of this thesis.

13. The notion of *shape derivative* of a real-valued shape functional corresponds to the notion of a derivation on the manifold $\mathcal{X}(\mathcal{B})$ and the mapping "set $\mapsto$ shape derivative" corresponds to a covariant derivative. Hence, in the particular context of the shape functional $\mathcal{F}$, the mapping

$$\mathcal{X}(\mathcal{A}) \to T^*\mathcal{X}(\mathcal{A}), \quad \mathcal{B} \mapsto (\mathrm{D}\mathcal{F})(\mathcal{B})[.] \in T^*_{\mathcal{B}}\mathcal{X}(\mathcal{A})$$

is the covariant derivative (see (2.38a)), where (cf. Lemma 10 and Theorem 7)

$$(\mathrm{D}\mathcal{F})(\mathcal{B})[V \cdot \boldsymbol{n}_{\mathcal{J}}] = -\frac{1}{2\lambda}\int_{\beta}(\bar{p}_{\mathcal{J}} - p_{\min}^{\max})^2 V \cdot \boldsymbol{n}_{\mathcal{J}}, \quad V \cdot \boldsymbol{n}_{\mathcal{J}} \in T_{\mathcal{B}}\mathcal{X}(\mathcal{A}). \tag{2.81}$$

This point of view clarifies the notion of the shape semiderivative used in the sections above

$$d\mathcal{F}(\mathcal{B}; V) = (\mathrm{D}\mathcal{F})(\mathcal{B})[V \cdot \boldsymbol{n}_{\mathcal{J}}].$$

It is a bit of inconvenient that the notation $d\mathcal{F}(\mathcal{B}; V)$ mixes elements of different manifolds (also cf. the 12th item): on the one hand $d\mathcal{F}(\mathcal{B}; .)$ says, that shape derivatives are studied and therefore suggests that only set sensitive (and not pointwisely sensitive) operations are considered (which is the regime of $\mathcal{X}(\mathcal{B})$ and $\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{B})$), and on the other hand the usage of $V \in \mathcal{V}$ corresponds to pointwise vector fields and transformations (which is the regime of $\mathcal{H}(\Omega)$).

It should be noted, that working with $\mathcal{H}(\Omega)$ seems to be inevitable, when the feasible shapes do not have as pretty properties as those contained in $\mathcal{O}$. The subgroup $\mathcal{K}(\mathcal{A})$ may be singleton $\{\mathrm{Id}\}$ in this case.

14. An obvious drawback when working with the manifold $\mathcal{X}(\mathcal{A})$ is the fact that the (expected; see the 12th item on page 67) tangent bundle

$$T\mathcal{X}(\mathcal{A}) = \{V \cdot \boldsymbol{n}_{\mathcal{B}} \,|\, V \in \mathcal{V},\ \boldsymbol{n}_{\mathcal{B}} \text{ is the outer unit normal vector field of } \mathcal{B} \in \mathcal{X}(\mathcal{A})\}$$

is not trivial. Thus, it may be more convenient to stay in the context of $\mathcal{H}(\Omega)$ and mind that all considered functions are constant with respect to the submanifold $\mathcal{K}(\mathcal{A})$. This seems to be the common approach in shape calculus. It entails the disadvantage that extra structure which is induced by the constancy with respect to $\mathcal{K}(\mathcal{A})$ has to be recovered in each assertion. The structure theorems for first and second order shape derivatives [44, Chp. 9 Thm. 3.6, Cor. 3.1 and Thm. 6.3] represent this finding.

15. The framework of the manifold $\mathcal{X}(\mathcal{A})$ and of shape calculus is based on the Banach space Θ_0. The choice $\Theta_0 = \{f \in C^{1,1}(\overline{\mathbb{R}}^2, \mathbb{R}^2) \,|\, f|_{\Omega^c} \equiv 0\}$ is no Hilbert space and hence $\mathcal{X}(\mathcal{A})$ cannot be a

[32] According to the result of Henderson [79] every (separable) Banach manifold can be embedded into a lager Banach space and thus the considered imbedding is not surprising.

Riemannian manifold in sense of Definition 14. Consequently, the notion of a gradient (cf. Definition 15) should actually not be used in shape optimization. Nonetheless, there is some justification for the usage of this term in the sense of Lemma 10 and the associated Remark. As already addressed in the 6th item on page 66 it is possible to repeat the construction of the manifold based on the space $\mathcal{D}(\mathbb{R}^2)$. By doing so, one obtains the same term (2.81) for the covariant derivative of the shape functional $\mathcal{F}$. The covariant derivative $(D\mathcal{F}[.])(\mathcal{B})$ evaluated at set $\mathcal{B}$ is an element of $T^*_{\mathcal{B}}\mathcal{H}(\Omega) = \Theta_0^* = \mathcal{D}'(\Omega, \mathbb{R}^2)$ then. This is a distribution and its support is concentrated on the interface β. Moreover, there is a regular representative $-\frac{1}{2\lambda}(\bar{p}_{\mathcal{J}} - p_{\min}^{\max})^2 \in L^1(\beta)$ of the distribution, which is called the (shape) gradient here. All in all, the so-defined shape gradient fulfills

$$\int_\beta \nabla\mathcal{F}(\mathcal{B})V \cdot \boldsymbol{n}_{\mathcal{J}} = (\mathrm{D}\mathcal{F})(\mathcal{B})[V \cdot \boldsymbol{n}_{\mathcal{J}}].$$

16. The idea of *function space parametrization*, see [44, p. 565], which enables the comparison of two function spaces $B(\mathcal{A})$ and $B(\mathcal{B})$ defined on different sets $\mathcal{A}, \mathcal{B} \in \mathcal{O}$ can be formalized as *parallel transport* (cf. [109, Chp. III Thm. 9.8]) in a vector bundle (see Definition 16) on $\mathcal{H}(\Omega)$ in the following way:

    Let $\mathcal{A} \in \mathcal{O}$ and let $F \in \mathcal{H}(\Omega)$, then there is $F(\mathcal{A}) \in \mathcal{X}(\mathcal{A}) \subset \mathcal{O}$ ($\mathcal{A}$ does not have to be the optimal active set here, but the letter is used instead of $\mathcal{B}$, since $\mathcal{B}$ and B look quite similar). Moreover, let $\mathcal{S} \subset \Omega$ and let $B(\mathcal{S})$ be a Banach space on the set $\mathcal{S}$. One may think of $B(\mathcal{S}) \in \{L^2(\mathcal{S}), H^1(\mathcal{S})\}$, where $\mathcal{S} \in \{\Omega, \mathring{\mathcal{A}}, \mathcal{I}, \gamma\}$ for instance. Then the set

    $$E(B, \mathcal{S}) := \{B(F) := B(F(\mathcal{S})) \,|\, F \in \mathcal{H}(\Omega)\}$$

    can be given the structure of a vector bundle to the base space $\mathcal{H}(\Omega)$ (cf. [113, Chp. III §1]): The manifold $\mathcal{H}(\Omega)$ can be covered by a single chart, since it is an open subset of $\mathrm{Id} + \Theta_0$, which makes it all the way easier. Let

    $$\pi : E(B, \mathcal{S}) \to \mathcal{H}(\Omega), \quad B(F) \mapsto F$$

    and define the trivializing map (cf. the trivializing map $\varphi_\circ$ from the 6th item)

    $$\varphi_{\mathcal{H}(\Omega)} : \pi^{-1}(\mathcal{H}(\Omega)) \to \mathcal{H}(\Omega) \times B(\mathcal{S}), \quad B(F) \ni v_F \mapsto (F, v_F \circ F),$$

    Here $B(\mathcal{S})$ is the standard fiber. By this means all requirements for Definition 16 can be fulfilled, if the regularity of the transformations $F \in \mathcal{H}(\Omega)$ is suitable. Indeed, the regularity of F is a limiting factor, since the composed function $v_F \circ F$ used in the trivializing map can only be as regular as F is. When F is $C^{1,1}$, one cannot expect for instance that it is possible to identify $H^3(\mathcal{B} = F(\mathcal{A}))$ with $H^3(\mathcal{A})$ by means of the proposed idea.

    Consequently, if the Banach spaces $B(\mathcal{S})$ and Θ_0 fit together, the space $B(G)$ is linearly isomorphic to the space $B(F)$ (where $F, G \in \mathcal{H}(\Omega)$) by means of

    $$\varphi_G^{-1} \circ \varphi_F : B(F) \to B(G), \quad v_F \mapsto v_F \circ F \circ G^{-1}. \tag{2.82}$$

17. Here one recognizes once again, that Θ_0 has a fundamental impact on structure of $\mathcal{X}(\mathcal{A})$ and that it has to be chosen carefully. The more regularity is induced via the choice of a small function space Θ, the more structures (as e. g. vector bundles) can be build upon the manifold $\mathcal{X}(\mathcal{A})$. However, such a choice is accompanied with a possible shrink of $\mathcal{X}(\mathcal{A})$, which are the orbits of the group operation on the family of feasible sets $\mathcal{O}$ and hence $\mathcal{O}$ might decompose into even more separate components, cf. the 5th item on page 66. All in all, one has to balance different requirements:
    - Requirements on the choice of a class of feasible sets $\mathcal{O}$.
    - Requirements on the regularity of functions/solutions related to boundary value problems posed on sets which are elements of $\mathcal{O}$.
    - Requirements from shape calculus in order to be able to construct vector bundles of function spaces.

18. Now let $\alpha : [0; \tau] \to \mathcal{H}(\Omega)$ be a differentiable curve on $\mathcal{H}(\Omega)$. Then the above construction induces a *parallel translation along* α (cf. [1, p. 104])

    $$P_\alpha^{s \leftarrow t} = \varphi_{\alpha(s)}^{-1} \circ \varphi_{\alpha(t)} : B(\alpha(t)) \to B(\alpha(s)), \quad v_{\alpha(t)} \mapsto v_{\alpha(t)} \circ \alpha(t) \circ \alpha(s)^{-1}, \text{ for all } s, t \in [0; \tau].$$

    This parallel translation induces a covariant derivative (an affine connection, respectively) in the vector bundle $E(B, \mathcal{S})$; cf. [1, Sec. 8.1.1]. Figure 2.11 illustrates the whole setting for convenience.

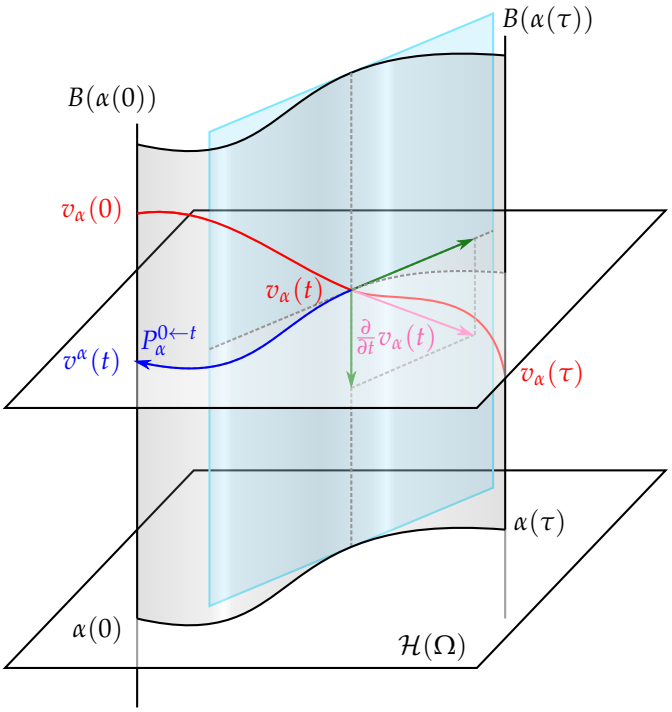


Figure 2.11: Illustration of a vector bundle $E(B,\mathcal{S})$, the parallel translation $P_\alpha^{0\leftarrow t}$ and the derivative of the section v.

Let v be a *section* of $E(B,\mathcal{S})$, this is a differentiable assignment $v : \mathcal{H}(\Omega) \to E(B,\mathcal{S})$, $F \mapsto v(F) \in B(F)$; cf. [109, Chp. I Def. 3.1]. Then

$$v_\alpha : [0;\tau] \to E(B,\mathcal{S}), \quad t \mapsto v_\alpha(t) := v_{\alpha(t)} \in B(\alpha(t))$$

is a section of $E(B,\mathcal{S})$ over α. By means of the parallel translation along α this section can be transported to the fiber above the starting point $\alpha(0)$

$$v^\alpha : [0;\tau] \to B(\alpha(0)), \quad t \mapsto v^\alpha(t) := P_\alpha^{0\leftarrow t}(v_\alpha(t)) = v_\alpha(t) \circ \alpha(t) = v_{\alpha(t)} \circ \alpha(t).^{33}$$

It is possible to derive the total time derivative of v^α then

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}v^\alpha(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\left(v_\alpha(t)\circ\alpha(t)\right)\\ &= \left(\frac{\partial}{\partial t}v_\alpha(t)\right)\circ\alpha(t) + \left(\mathrm{D}_x(v_\alpha(t))\circ\alpha(t)\right)\dot\alpha(t),\end{aligned}$$

where $\mathrm{D}_x v_\alpha(t)$ is the derivative of $v_\alpha(t) \in B(\alpha(t))$ with respect to the spacial variable x in $\alpha(t)(\mathcal{S}) \subset \Omega \subset \mathbb{R}^2$. This reasoning can be regarded as a definition for a covariant derivative of the section v with respect to the vector field ξ (cf. item 9)

$$(\nabla_\xi v)_F := \frac{\partial}{\partial t}v_\alpha(t)\Big|_{t=0} + \left(\mathrm{D}_x(v_\alpha(t))\circ\alpha(t)\right)\dot\alpha(t)\Big|_{t=0}, \text{ where } F = \alpha(0) \text{ and } \xi_F = \dot\alpha(0) \in T_F\mathcal{H}(\Omega).$$

In order to understand this object in terms of shape calculus, let α be the integral curve of a nonautonomous velocity field $V \in C^1([0;\tau],\mathcal{V})$. Then the following relations between the different perspectives hold true

$$\begin{aligned}\forall X \in \Omega,\ \forall t \in [0;\tau]: \quad &\alpha(t)(X) = T_t(V)(X), \text{ cf. item 5},\\ &\alpha(0)(X) = T_0(V)(X) = F(X) = x,\\ &\mathcal{S}_t := T_t(V)(\mathcal{S}),\\ &v_t := v_\alpha(t) \in B(\mathcal{S}_t),\\ &v^t := v_t \circ T_t(V) = v^\alpha(t) \in B(\mathcal{S}).\end{aligned}$$

[33] Note that $v^\alpha(t) = v_{\alpha(t)} \circ \alpha(t)$ must not be confused with $v \circ \alpha(t) \circ \alpha(t)$. In order to see this, it is necessary to understand, that the section v depends on both the footpoint $F \in \mathcal{H}(\Omega)$ and the variable $x \in F(\mathcal{S})$. In other words, $v = v(F,x)$ and in particular $v_\alpha(t)(x) = v(\alpha(t),x)$. Consequently, there is $v^\alpha(t)(X) = v(\alpha(t),\alpha(t)(X))$, where $x = \alpha(t)(X) \in \alpha(t)(\mathcal{S})$ and $X \in \mathcal{S}$.

Moreover, one has

$$\begin{aligned}(\nabla_V v)_F(x) &= \frac{\mathrm{d}}{\mathrm{d}t} v^t(x)\Big|_{t=0} \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\big(v_t \circ T_t(V)(X)\big)\Big|_{t=0} \\ &= \left(\frac{\partial}{\partial t} v_t\right) \circ T_t(V)(X)\Big|_{t=0} + \big((\mathrm{D}_x v_t) \circ T_t(V)(X)\big)\, \frac{\mathrm{d}}{\mathrm{d}t} T_t(V)(X)\Big|_{t=0} \\ &= \frac{\partial}{\partial t} v_t(x)\Big|_{t=0} + \big(\mathrm{D}_x v_0(x)\big) V(0,x)\end{aligned}$$

In particular, with the common notation (cf. for instance [151, Eq. (2.106)])

$$(\nabla_V v)_F(x) = v_0'[V(0)] + \mathrm{D}v_0 V(0).$$

All in all, the covariant derivative in direction $V \cong \xi$ corresponds to the material derivative of shape calculus and the partial time derivative (which can be seen as a second type of covariant derivative, cf. item 8th) is the shape derivative.

19. It is beyond the scope of this thesis to carry the considerations made in the 16th and 18th item over to the quotient manifold $\mathcal{X}(\mathcal{A})$, though this is an interesting and important task in order to get a complete overview of shape calculus from the perspective of calculus on manifolds.

    Some explaining remarks seem to be indicated. Specialize to the choice $\mathcal{S} \in \{\Omega, \mathcal{A}, \mathcal{I}, \gamma\}$ for the following considerations. The approach via the vector bundle $E(B, \mathcal{S})$ has the drawback that two fibers $B(F)$ and $B(G)$ can denote the "same" Banach space, although $F \neq G$ in the following sense: Let $[F]_{\mathcal{A}} = [G]_{\mathcal{A}}$, then $F(\mathcal{S}) = G(\mathcal{S})$ and consequently both $B(F)$ and $B(G)$ represent the space $B(F(\mathcal{S}))$. The situation is similar to (nonlinear) change of coordinates. This drawback can be overcome when regarding the corresponding vector bundle on the quotient manifold $\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{A})$

    $$\tilde{E}(B, \mathcal{S}) = \{B([F]_{\mathcal{A}}) := B([F]_{\mathcal{A}}(\mathcal{S})) \,|\, [F]_{\mathcal{A}} \in \mathcal{H}(\Omega)/\mathcal{K}(\mathcal{A})\}.$$

    However it not directly possible to carry over the idea of the linear isomorphisms (2.82), since the mapping

    $$v \mapsto v \circ F$$

    requires a pointwise interpretation of F, which is not given when trying to work with

    $$v \mapsto v \circ [F]_{\mathcal{A}}.$$

    Hence, basically the necessary construction needed here should be similar to one of the horizontal subbundle on page 93f.

20. Delfour and Zolésio show in a detailed analysis of the second order shape semiderivative of shape functionals that this object has a rich inherent structure, [44, Chp. 9 Sec. 6], [41, 42]. It is valuable to regard some of their results from the perspective of calculus on manifolds.

    To this, let $\mathcal{M}$ be a set of feasible shapes, let $J : \mathcal{M} \to \mathbb{R}$ be a differentiable shape functional and let V, W be two suitable nonautonomous velocity fields. They prove (cf. [44, Chp. 9 Thm. 6.5(iii)]), that the second order shape semiderivative decomposes into two distinct parts:

    $$d\big(dJ(M; V); W\big) = \nabla^2 J(M)[V(0), W(0)] + dJ\big(M; V'(0) + \mathrm{D}V(0)\, W(0)\big). \tag{2.83}$$

    This result is nothing but the decomposition already obtained in (2.80): the first term corresponds to the second covariant derivative of J and the second one fits as well since $V'(0) + \mathrm{D}V(0)\, W(0)$ is the covariant derivative of the vector field V with respect to W evaluated at $t = 0$; cf. the 9th item.

    At this point it becomes apparent, why the second order shape semiderivative for autonomous and nonautonomous velocity fields differ. It has been explained in the paragraph below of (2.80), that the computation of the second directional derivative of a function on a manifold requires information about the vector field η in a neighborhood of the evaluation point, since one has to compute its covariant derivative. Hence, that is why the $V'(0)$ term appears in (2.83) and which requires a nonautonomous velocity field.

    The second shape semiderivative of the reduced shape functional $\mathcal{F}$ defined in (2.38a) can be expressed in a different fashion than those of Lemma 11. It is worthwhile to analyze this formula, since it is possible to extract the second covariant derivative from it. With use of the oriented dis-

tance function $b = b_{\mathcal{B}}$ (defined in (2.66) on page 54) and by means of [44, Chp. 9 Eq. (6.21)] there holds for two nonautonomous velocity fields V and W

$$\begin{aligned}
&d^2\mathcal{F}(\mathcal{B};V,W)\\
&= -\frac{1}{2\lambda}\int_\beta \left((\bar p_{\mathcal{J}} - p_{\min}^{\max})^2\right)' [W(0)]V(0)\cdot \boldsymbol{n}_{\mathcal{J}}\\
&\quad -\frac{1}{2\lambda}\int_\beta \left(\partial_{\boldsymbol{n}}^{\mathcal{J}}\left((\bar p_{\mathcal{J}} - p_{\min}^{\max})^2\right) + \kappa_{\mathcal{J}}(\bar p_{\mathcal{J}} - p_{\min}^{\max})^2\right) V(0)\cdot\boldsymbol{n}_{\mathcal{J}} W(0)\cdot\boldsymbol{n}_{\mathcal{J}}\\
&\qquad\qquad +(\bar p_{\mathcal{J}} - p_{\min}^{\max})^2\left((\mathrm{D}^2 b\, V(0))\cdot W(0) - V(0)\cdot\nabla_\beta V(0)\cdot\boldsymbol{n}_{\mathcal{J}} - W(0)\cdot\nabla_\beta W(0)\cdot\boldsymbol{n}_{\mathcal{J}}\right)\\
&\quad -\frac{1}{2\lambda}\int_\beta (\bar p_{\mathcal{J}} - p_{\min}^{\max})^2 (V'(0) + \mathrm{D}V(0)\, W(0))\cdot\boldsymbol{n}_{\mathcal{J}}\\
&= -\frac{1}{2\lambda}\int_\beta 2(\bar p_{\mathcal{J}} - p_{\min}^{\max}) p'_{\mathcal{J}}[W(0)]V(0)\cdot\boldsymbol{n}_{\mathcal{J}}\\
&\quad -\frac{1}{2\lambda}\int_\beta (\bar p_{\mathcal{J}} - p_{\min}^{\max})\left(2\partial_{\boldsymbol{n}}^{\mathcal{J}}(\bar p_{\mathcal{J}} - p_{\min}^{\max}) + \kappa_{\mathcal{J}}(\bar p_{\mathcal{J}} - p_{\min}^{\max})\right) V(0)\cdot\boldsymbol{n}_{\mathcal{J}} W(0)\cdot\boldsymbol{n}_{\mathcal{J}}\\
&\qquad\qquad +(\bar p_{\mathcal{J}} - p_{\min}^{\max})^2\left((\mathrm{D}^2 b\, V(0))\cdot W(0) - V(0)\cdot\nabla_\beta V(0)\cdot\boldsymbol{n}_{\mathcal{J}} - W(0)\cdot\nabla_\beta W(0)\cdot\boldsymbol{n}_{\mathcal{J}}\right)\\
&\quad -\frac{1}{2\lambda}\int_\beta (\bar p_{\mathcal{J}} - p_{\min}^{\max})^2 \left(V'(0) + \mathrm{D}V(0)\, W(0)\right)\cdot\boldsymbol{n}_{\mathcal{J}}
\end{aligned}$$

If $V(0)$ and $W(0)$ have no tangential component on β, that is to say $V(0)|_\beta = V(0)\cdot\boldsymbol{n}_{\mathcal{J}}\boldsymbol{n}_{\mathcal{J}}$ and $W(0)|_\beta = W(0)\cdot\boldsymbol{n}_{\mathcal{J}}\boldsymbol{n}_{\mathcal{J}}$ the formula simplifies, since $\mathrm{D}^2 b\,\boldsymbol{n}_{\mathcal{B}} = \mathrm{D}^2 b\,\nabla b = 0$ and since $\boldsymbol{n}_{\mathcal{J}}\perp\nabla_\beta(.)$

$$\begin{aligned}
&d^2\mathcal{F}(\mathcal{B};V,W)\\
&= -\frac{1}{2\lambda}\int_\beta 2(\bar p_{\mathcal{J}} - p_{\min}^{\max}) p'_{\mathcal{J}}[W(0)]V(0)\cdot\boldsymbol{n}_{\mathcal{J}}\\
&\quad -\frac{1}{2\lambda}\int_\beta (\bar p_{\mathcal{J}} - p_{\min}^{\max})\left(2\partial_{\boldsymbol{n}}^{\mathcal{J}}(\bar p_{\mathcal{J}} - p_{\min}^{\max}) + \kappa_{\mathcal{J}}(\bar p_{\mathcal{J}} - p_{\min}^{\max})\right) V(0)\cdot\boldsymbol{n}_{\mathcal{J}} W(0)\cdot\boldsymbol{n}_{\mathcal{J}}\\
&\quad -\frac{1}{2\lambda}\int_\beta (\bar p_{\mathcal{J}} - p_{\min}^{\max})^2 \left(V'(0) + \mathrm{D}V(0)\, W(0)\right)\cdot\boldsymbol{n}_{\mathcal{J}}
\end{aligned}\tag{2.84}$$

In view of the result of the 9th item and (2.83) the second covariant derivative of the reduced functional $\mathcal{F}$ reads

$$\begin{aligned}
&\nabla^2\mathcal{F}(\mathcal{B})[V,W]\\
&= -\frac{1}{2\lambda}\int_\beta 2(\bar p_{\mathcal{J}} - p_{\min}^{\max}) p'_{\mathcal{J}}[W(0)]V(0)\cdot\boldsymbol{n}_{\mathcal{J}}\\
&\quad -\frac{1}{2\lambda}\int_\beta (\bar p_{\mathcal{J}} - p_{\min}^{\max})\left(2\partial_{\boldsymbol{n}}^{\mathcal{J}}(\bar p_{\mathcal{J}} - p_{\min}^{\max}) + \kappa_{\mathcal{J}}(\bar p_{\mathcal{J}} - p_{\min}^{\max})\right) V(0)\cdot\boldsymbol{n}_{\mathcal{J}} W(0)\cdot\boldsymbol{n}_{\mathcal{J}}.
\end{aligned}\tag{2.85}$$

This second covariant derivative plays an essential role when solving the original model problem (2.1) numerically; cf. sections 3.2 and 3.4.

21. In practice one typically knows the velocity fields $V(0)$, $W(0)$ only on the boundary and they are parallel to the normal vector field $\boldsymbol{n}_{\mathcal{J}}$; cf. Chapter 3 and in particular the 9th item of the discussion of Algorithm 1 on page 111. Thus, one has to introduce some artificial extension into the bulk of the domain in order to derive $(\mathrm{D}V(0)\, W(0))\cdot\boldsymbol{n}_{\mathcal{J}}$. A nearby idea to obtain such an extension is based on the distance projection onto the boundary; cf. [44, Chp. 6 Def. 3.1] and Figure 2.12. Thereto, let d_β be the distance function of the boundary $\beta = \partial\mathcal{B}$, which was defined in (2.65) and let $x\in\mathbb{R}^2$ be arbitrary but fix. Then the *distance projection* p_β of x is

$$p_\beta(x) := \begin{cases} y, \text{ iff } \{y\in\beta \,|\, |y-x| = d_\beta(x)\} \text{ is a singleton}\\ \text{not defined otherwise.}\end{cases}$$

Due to [44, Chp. 7 Thm. 8.3] the distance projection p_β is well-defined in a sufficiently small tubular neighborhood of β and is of class $C^{0,1}$ there.

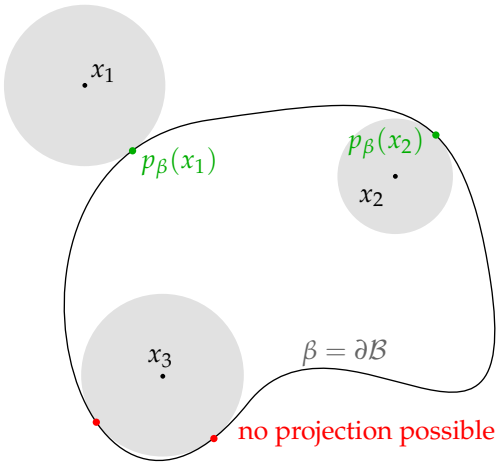


Figure 2.12: Illustration the projection p_β onto the boundary $\beta := \partial\mathcal{B}$.

Now let $V(0)|_\beta = v\,\boldsymbol{n}_\mathcal{J}$ and define its extension

$$V(0) = (v \circ p)\,\nabla b.$$

Then there holds

$$\begin{aligned}(\mathrm{D}V(0)W(0)) \cdot \boldsymbol{n}_\mathcal{J} &= \Big(\mathrm{D}\big((v \circ p)\,\nabla b\big)\,W(0)\Big) \cdot \nabla b\Big|_\beta \\ &= \Big(\big((\mathrm{D}v) \circ p\,\underbrace{\mathrm{D}p\,\nabla b}_{=0} + (v \circ p)\,\mathrm{D}^2 b\big)\,W(0)\Big) \cdot \nabla b\Big|_\beta \\ &= (v \circ p)\,W(0) \cdot \big(\underbrace{\mathrm{D}^2 b\,\nabla b}_{=0}\big)\big|_\beta \\ &= 0.\end{aligned}$$

Consequently, if the velocity field is autonomous, the last integral term of (2.84) vanishes and the second shape semiderivative coincides with the second covariant derivative (2.85). This reasoning excuses the use of second order shape semiderivatives in numerical calculations, although the second order covariant derivative should be used actually; cf. Section 3.2.

All in all the conclusion of the comparison of differential calculus on manifolds and shape calculus is threefold

1. Shape calculus embeds into the more general differential calculus on manifolds
2. the comparison enables valuable insight in the structure of shape calculus
3. the comparison presented here is far from complete. In particular, at least two urgent question remain unanswered
   a) How does a suitable atlas of $\mathcal{X}(\mathcal{A})$ and $\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{A})$ look like?
   b) How can shape derivatives and function space parametrization be expressed by means of intrinsic notions of $\mathcal{X}(\mathcal{A})$?

2.6.3 Abstract view on set optimal control problems

By means of the preceding discussion, it is possible to embed shape optimization and set optimal control problems into the general framework of *optimization on vector bundles*.

The most common situation in shape optimization is illustrated Figure 2.13 on the left hand side. One has to minimize a functional $J : E \to \mathbb{R}$, where E is a vector bundle on a manifold $\mathcal{M}$, and where $\mathcal{M}$ collects a suitable class of shapes. Moreover, there are explicit constraints on the admissibility of the shapes, which is reflected by a set $U \subset \mathcal{M}$. One may think of volume constraints for instance. Typically, one has to fulfill an additional constraint like a boundary value problem. This constraint is modeled by a function $f : \mathcal{M} \to E$, which associates an element $f(x) \in E_x$ with each set $x \in \mathcal{M}$. The particular situation enables the introduction of a reduced shape functional $\mathcal{F} : \mathcal{M} \to \mathbb{R}$, $x \mapsto J(x, f(x))$.

A more general situation is illustrated on the right hand side of Figure 2.13. It is not possible to introduce a reduced shape functional here, due to the lack of a function $f : \mathcal{M} \to E$. In this context, it is assumed that some fiber-wise constraints have to be fulfilled. That is to say, only those elements f_x of the fiber E_x for $x \in U$ are feasible, which fulfill $a_x \leq f_x \leq b_x$. In particular, this optimization problem does not fit into the abstract framework of optimal control, since there is no distinction between "free" controls and "dependent" states.

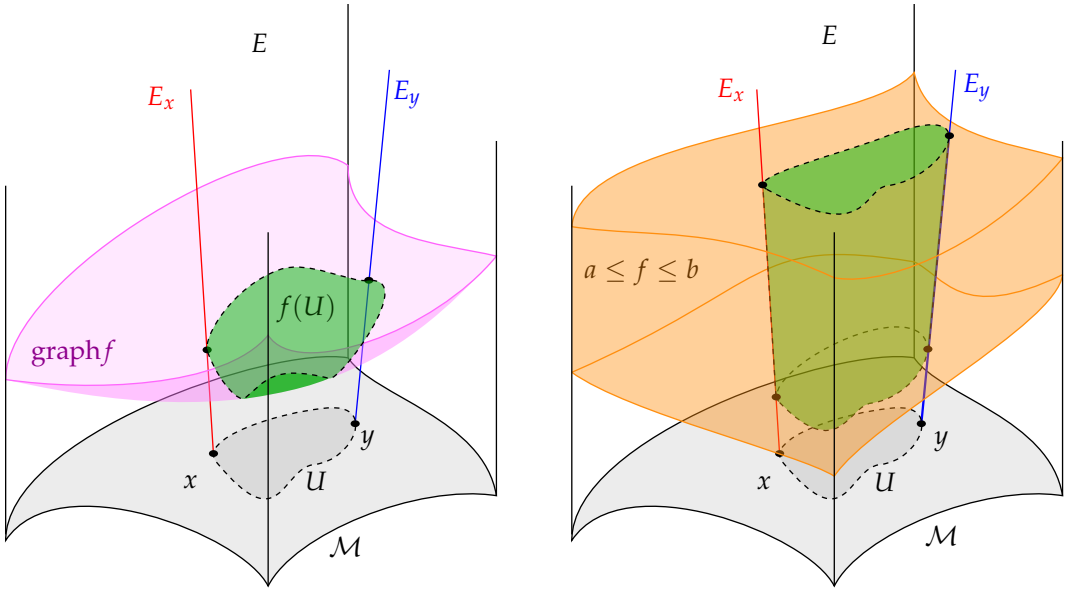


Figure 2.13: Illustration of two prototypic optimization problems on a vector bundle E: A typical shape optimization problem with state equation (left) and an optimization problem that is no optimal control problem (right).

With these deliberations at hand it is possible to class the set optimal control problem (2.30) within the abstract framework, see Figure 2.14. The objective $\mathfrak{J}$ (see (2.30a)) is defined on the vector bundle

$$E = E(L^2_{\mathcal{J}}, L^2_{\mathcal{B}}, H^1_{\mathcal{J}}\Delta, H^1_{\mathcal{B}}\Delta) \tag{2.86}$$

over the manifoldlike object $\mathcal{O}$. For $\mathcal{B} \in \mathcal{O}$ the fiber $E_{\mathcal{B}}$ of the vector bundle is given by $L^2(\mathcal{J}) \times L^2(\mathring{\mathcal{B}}) \times H^1(\mathcal{J}, \Delta) \times H^1(\mathring{\mathcal{B}}, \Delta)$. With respect to the shape optimization perspective it is allowed to confine oneself with the manifold $\mathcal{X}(\mathcal{A})$ instead of coping with the more complex set $\mathcal{O}$. The split state equation (2.30h)–(2.30l) and the BDD reformulation (2.30e)–(2.30f) of the state constraints define a subset in each fiber, namely the graph of the control-to-state operator $(u_{\mathcal{J}}, u_{\mathcal{B}}) \mapsto (y_{\mathcal{J}}, y^{\max}_{\min})$. In each fiber there is a unique point on the graph singled out by means of Theorem 3, which gives rise to the geometry-to-solution operator G (cf. Definition 6). This operator plays the role of a boundary value constraint in "common" shape optimization. Additionally, one has to fulfill the strict inequality constraint (2.30g), which defines a subset of the vector bundle in the fashion of the second prototypic situation presented above. The intersection of this subset and the graph of the geometry-to-solution operator implicitly defines the feasible set of shapes of the optimal control problem.

2.7 Remarks on optimal control and PDAE

This section is devoted to the analysis of the first order necessary conditions from the perspective of *partial differential-algebraic equations* (PDAE). This point of view reveals a fundamental idea behind the Bryson-Denham-Dreyfus approach: the reduction of the differentiation index.

Optimal control problems typically possess a relevant part of conditions which can be seen as PDAE-system. The dynamics of the process to be controlled are modeled by differential equations and they are typically accompanied by a set of algebraic constraints; state and control constraints, coupling equations, etc. Furthermore, first order necessary conditions are (P)DAE-systems or at least are composed of a

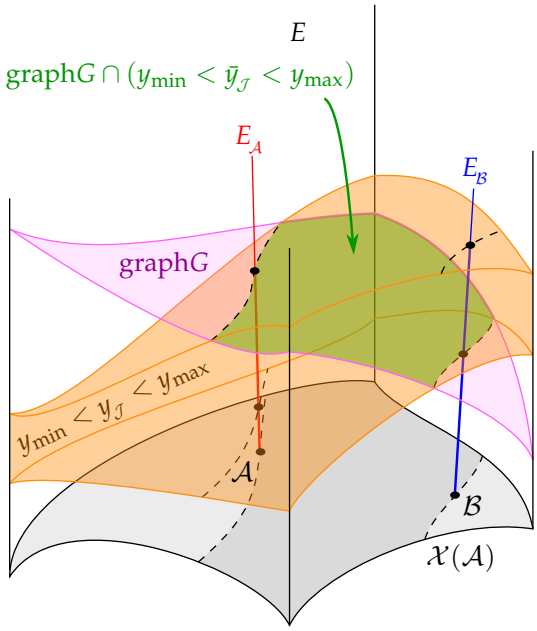


Figure 2.14: Illustration of the set optimal control problem in the vector bundle E and the implicitly defined subset of those elements of $\mathcal{X}(\mathcal{A})$ at which the strict inequality constraint $y_{\min} < \bar{y}_{\mathcal{J}} < y_{\max}$ is fulfilled.

(P)DAE-system and additional conditions. Hence, the theory of (P)DAE-systems has natural value when analyzing optimal control problems.

The essential benefit of the BDD approach in OC-ODE and the associated determination of the order (index) of the state constraint is twofold:[34]

- the approach turned out to be superior in view of its numerical realization, cf. [20, 21, 75]
- the index (i. e. the *order*) of the state constraint enables some prediction of possible active sets; see [73, 141, 75].[35]

Since the analysis of the BDD approach in the context of optimal control of PDEs within this thesis is very far from complete, such kind of assertions are not available there yet. Moreover, especially the question of apriori knowledge of certain properties of the active set is much more complex in OC-PDE, since a classification of the possible active sets (points, curves, ...) is considerably more sophisticated than in the regime of ODEs, since there is no canonical choice yet. It is even necessary to develop suitable notions of indices/orders of the state constraint.

This should be regarded with respect to the development in the field of (P)DAEs. The field of (ordinary) DAE has been investigated intensively during the last three decades and comes up with considerably insight, whereas the topic of PDAE has gained more and more attention in the last 15 years and its theory is far from complete.

2.7.1 Remarks on DAE

A short survey on DAEs can be found in [65] and in-depth introductions are due to Brenan et al. [16] and to Griepentrog and März [66]. A substantial overview on the numerical solution of DAEs can be found in the book of Hairer et al. [71, 72].

One central notion within the treatment of DAEs it the *index*, which comes in various variants: *differentiation index*, *perturbation index*, *strangeness index*, etc. The index plays a crucial role in both the theoretical and numerical treatment of DAEs. The concept of the differential index is presented briefly in the following and equipped with some remarks on its implications.

[34] A definition of the order of a state constraint can be found in [75, p. 183] or [122].

[35] In particular, if the order of a state constraint is odd, the active set consists of *contact points* only. In contrast, when it is even the active set can also be composed of intervals (*boundary arcs*).

Firstly define a DAE as an everywhere singular implicit ordinary differential equation; see [65]:

Definition 17 ((Ordinary) differential-algebraic equation (DAE))**:**
Let $I \subset \mathbb{R}$ be an interval, $m \in \mathbb{N}$, D, $D' \subset \mathbb{R}^m$ and $F : I \times D \times D' \to \mathbb{R}^m$. Then the equation

$$F(t, y(t), y'(t)) = 0, \quad t \in I$$

is called a *differential-algebraic equation*. Typically it is required that $\partial_{y'} F$ is singular in order to discriminate the notion from a "normal" implicitly defined ODE.

There are several prototypic subclasses of DAEs, for example the class of *semi-explicit* DAEs

$$\begin{aligned} y' &= f(t, y, z), \\ 0 &= g(t, y, z), \end{aligned} \tag{2.87}$$

and the class of *linear constant coefficient* DAEs

$$Ay' + By = f(t). \tag{2.88}$$

Both subclasses are instructive. For one thing one can easily recognize the part of differential equations $y' = f(t, y, z)$ and the part of algebraic equations $0 = g(t, y, z)$ of semi-explicit DAEs, which are of major interest, since various practical applications can be modeled by this class. For another thing linear constant coefficient DAEs are the best understood representatives and are closely related to the PDAE-system considered below.

The differentiation index is defined as follows (cf. for instance [72, Chp. VII Def. 2.1], [16, Def. 2.2.2])

Definition 18 (Differentiation index)**:**
Let $F(t, y(t), y'(t)) = 0$ be a DAE.

Then the minimal number s of differentiations

$$\frac{\mathrm{d}}{\mathrm{d}t} F(t, y(t), y'(t)) = 0, \quad \frac{\mathrm{d}^2}{\mathrm{d}t^2} F(t, y(t), y'(t)) = 0, \quad \ldots \frac{\mathrm{d}^s}{\mathrm{d}t^s} F(t, y(t), y'(t)) = 0 \tag{2.89}$$

required such that (2.89) allows for the extraction of an explicit ODE $y' = \Phi(t, y)$ be means of algebraic manipulations is called the *differentiation index* of the DAE.

The algebraic conditions of a DAE may be seen as description of a submanifold of the $\mathbb{R}^m$, in which the trajectories of the solutions have to be located; see [72, p. 454ff.]. For instance consider

$$\begin{aligned} y' &= f(t, y, z) \\ 0 &= g(y). \end{aligned}$$

Then $g(y) = 0$ implies that the trajectories $y(t)$ lie in the zero level set which is a submanifold indeed. Moreover, differentiation of $g(y) = 0$ with respect to time t yields

$$0 = \frac{\mathrm{d}}{\mathrm{d}t} g(y) = \mathrm{d}_y g(y)\, y' = \mathrm{d}_y g(y)\, f(t, y, z).$$

Consequently, the trajectories have to be located in an additional "hidden" submanifold defined by $0 = \mathrm{d}_y g(y)\, f(t, y, z)$ as well, if $\mathrm{d}_y g(y)$ is regular.

The mechanism of deriving the differentiation index reveals such conditions which are implicitly given by the DAE. For linear constant coefficient DAEs there is a constructive method available to determine the differentiation index (cf. [16, p. 20]), which will be used later on in Paragraph 2.7.3.

From the perspective of solving DAEs numerically the differentiation index plays a fundamental role. Whereas index-1 DAEs can be solved more or less as usual ODEs [72, Chp. VI], the situation chances essentially when solving higher index DAEs. There are several efficient methods, which cope with index-2 DAEs, [72, Chp. VII]. However, there seems to be no standard procedure available for higher index DAEs. Consequently, one has to use methods of index reduction which may suffer from drift-off effects and thus have to be stabilized or in some specific situations one can choose problem-adapted local coordinates [72, Sec. VII.2]. Drift-off means that numerical integration schemes tend to diverge from the manifold defined by the algebraic equations. Hence, there is need for mechanisms which either ensure that the algebraic

equations are satisfied or, if they are not, that the iterates are transported back onto the corresponding manifold. Minimal coordinates, known from multibody physics, are an example for problem-adapted local coordinates. They are characterized by the fact the each degree of freedom of motion is associated with exactly one coordinate (i. e. a variable) and consequently the model comes without additional algebraic constraints. However, such coordinates cannot be found in general and the resulting ODEs might be hard to solve.

2.7.2 Remarks on PDAE

As already mentioned, the topic of PDAE is relatively young and the understanding is far from being complete. The situation is much more complex than DAEs due to different variables and it may be conjectured that the transition from DAE to PDAE is as complicated as the passage from ordinary to partial differential equations. In particular, there are various notions of indices which try to capture the essence of its corresponding analogs in the world of DAEs; cf. for instance [23, 120, 140, 148] and the references therein.

Although it would be desirable to confront theory of OC-PDE with the results of PDAE in general, this goal is far beyond the scope of this thesis. The analysis is content with the specific analysis of the model problem and the corresponding results in view of PDAE. It should be taken to mean a first step towards a better understanding of state-constrained optimal control of partial differential equations.

2.7.3 First order necessary conditions as PDAE

The part of distributed (in contrast to boundary-) equations of the first order necessary conditions of Proposition 2 and Proposition 3 reads as follows, where $\mu_{\mathring{\mathcal{A}}}$ collects $\mu^{\max}$ and $-\mu^{\min}$ as in Proposition 5

$$-\Delta y_{\mathcal{J}} + y_{\mathcal{J}} - u_{\mathcal{J}} = 0 \quad \text{in } \mathcal{J}, \tag{2.90a}$$
$$-\Delta y_{\mathcal{B}} + y_{\mathcal{B}} - u_{\mathcal{B}} = 0 \quad \text{in } \mathring{\mathcal{B}}, \tag{2.90b}$$
$$-\Delta p^{trad}_{\mathcal{J}} + p^{trad}_{\mathcal{J}} - y_{\mathcal{J}} = -y_d \quad \text{in } \mathcal{J}, \tag{2.90c}$$
$$-\Delta p^{trad}_{\mathcal{B}} + p^{trad}_{\mathcal{B}} - y_{\mathcal{B}} - \mu_{\mathring{\mathcal{B}}} = -y_d \quad \text{in } \mathring{\mathcal{B}}, \tag{2.90d}$$
$$\lambda\, u_{\mathcal{J}} + p^{trad}_{\mathcal{J}} = \lambda\, u_d \quad \text{in } \mathcal{J}, \tag{2.90e}$$
$$\lambda\, u_{\mathcal{B}} + p^{trad}_{\mathcal{B}} = \lambda\, u_d \quad \text{in } \mathring{\mathcal{B}}, \tag{2.90f}$$
$$y_{\mathcal{B}} = y^{\max}_{\min} \quad \text{in } \mathring{\mathcal{B}}. \tag{2.90g}$$

By means of the classification (2.87) and (2.88) this is a linear constant coefficient PDAE in semi-explicit form. Moreover, there occurs only one single differential operator. Hence, though Δ involves partial derivatives with respect to different variable, this PDAE is pretty similar to the DAE situation.

In order to determine the differentiation index of the PDAE-system use the iterative two-step strategy presented in [16] on page 20:

- perform some algebraic equivalence transformation such that the PDAE becomes semi-explicit,
- differentiate the algebraic equations.

In terms of linear algebra applied on (2.88) this reads as follows. There is a nonsingular matrix P such that (2.88) premultiplied by P is

$$\begin{pmatrix} A_1 & A_2 \\ 0 & 0 \end{pmatrix} \Delta v + \begin{pmatrix} B_1 & B_2 \\ B_3 & B_4 \end{pmatrix} v = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}, \tag{2.91}$$

where $(A_1\ A_2)$ has full row rank and where $v := (u_{\mathcal{J}}, u_{\mathcal{B}}, y_{\mathcal{J}}, y_{\mathcal{B}}, p^{trad}_{\mathcal{J}}, p^{trad}_{\mathcal{B}}, \mu_{\mathring{\mathcal{B}}})^\top$. Differentiation of the second row (the algebraic equations) yields

$$\begin{pmatrix} A_1 & A_2 \\ B_3 & B_4 \end{pmatrix} \Delta v + \begin{pmatrix} B_1 & B_2 \\ 0 & 0 \end{pmatrix} v = \begin{pmatrix} f_1 \\ \Delta f_2 \end{pmatrix}.$$

If the matrix multiplied by Δv is still singular the process is repeated, otherwise the number of iterations equals the differentiation index.

In particular, the PDAE (2.90) is already in semi-explicit form (2.91). Differentiation of the algebraic equations yields

$$\lambda \Delta u_{\mathcal{J}} + \Delta p^{trad}_{\mathcal{J}} = \lambda \Delta u_d \quad \text{in } \mathcal{J}, \tag{2.92a}$$
$$\lambda \Delta u_{\mathcal{B}} + \Delta p^{trad}_{\mathcal{B}} = \lambda \Delta u_d \quad \text{in } \mathring{\mathcal{B}}, \tag{2.92b}$$
$$\Delta y_{\mathcal{B}} = \Delta y^{\max}_{\min} \quad \text{in } \mathring{\mathcal{B}}. \tag{2.92c}$$

Obviously one has reached differential equations for the control variables. However, there is no differential equation obtained for $\mu_{\mathring{\mathcal{A}}}$ and thus the matrix to be multiplied with Δv is still singular. It remains to manipulate (2.92c) again. By means of the state equation (2.90b) one obtains an algebraic equation

$$y_{\mathcal{B}} - u_{\mathcal{B}} = \Delta y_{\min}^{\max} \quad \text{in } \mathring{\mathcal{B}}. \tag{2.93}$$

Repeated differentiation and usage of (2.90b), (2.92b) and (2.90d) yields

$$\Delta y_{\mathcal{B}} - \Delta u_{\mathcal{B}} = \Delta^2 y_{\min}^{\max} \quad \text{in } \mathring{\mathcal{B}}, \tag{2.94a}$$

$$\Rightarrow \quad y_{\mathcal{B}} - u_{\mathcal{B}} + \frac{1}{\lambda}\Delta p_{\mathcal{B}}^{trad} = \Delta^2 y_{\min}^{\max} + \frac{1}{\lambda}\Delta u_d \quad \text{in } \mathring{\mathcal{B}}, \tag{2.94b}$$

$$\Rightarrow (\lambda-1)y_{\mathcal{B}} - \lambda u_{\mathcal{B}} + p_{\mathcal{B}}^{trad} - \mu_{\mathring{\mathcal{B}}} = \lambda\Delta^2 y_{\min}^{\max} + \Delta u_d - y_d \quad \text{in } \mathring{\mathcal{B}}. \tag{2.94c}$$

Still there is no differential equation obtained for $\mu_{\mathring{\mathcal{A}}}$ yet, but a third differentiation step is sufficient

$$(\lambda-1)\Delta y_{\mathcal{B}} - \lambda\Delta u_{\mathcal{B}} + \Delta p_{\mathcal{B}}^{trad} - \Delta\mu_{\mathring{\mathcal{B}}} = \lambda\Delta^3 y_{\min}^{\max} + \Delta^2 u_d - \Delta y_d \quad \text{in } \mathring{\mathcal{B}}. \tag{2.95}$$

Consequently, the PDAE-system that is the core of the common first order necessary conditions has differentiation index $s = 3$.

In marked contrast to this result the differentiation index of the PDAE-system related to the new first order necessary conditions of Corollary 3 equals $s = 1$: the PDAE-system reads

$$-\Delta y_{\mathcal{J}} + y_{\mathcal{J}} - u_{\mathcal{J}} = 0 \quad \text{in } \mathcal{J}, \tag{2.96a}$$

$$-\Delta y_{\mathcal{B}} + y_{\mathcal{B}} - u_{\mathcal{B}} = 0 \quad \text{in } \mathring{\mathcal{B}}, \tag{2.96b}$$

$$u_{\mathcal{B}} = -\Delta y_{\min}^{\max} + y_{\min}^{\max} \quad \text{in } \mathring{\mathcal{B}}, \tag{2.96c}$$

$$-\Delta p_{\mathcal{J}} + p_{\mathcal{J}} - y_{\mathcal{J}} = -y_d \quad \text{in } \mathcal{J}, \tag{2.96d}$$

$$-\Delta p_{\mathcal{B}} + p_{\mathcal{B}} - y_{\mathcal{B}} = -y_d \quad \text{in } \mathring{\mathcal{B}}, \tag{2.96e}$$

$$\lambda\, u_{\mathcal{J}} + p_{\mathcal{J}} = \lambda\, u_d \quad \text{in } \mathcal{J}, \tag{2.96f}$$

$$\lambda\, u_{\mathcal{B}} + p_{\mathcal{B}} + q_{\mathcal{B}} = \lambda\, u_d \quad \text{in } \mathring{\mathcal{B}}. \tag{2.96g}$$

Differentiation of the algebraic equations yields equations such that all variables occur in differentiated form

$$\Delta u_{\mathcal{B}} = -\Delta^2 y_{\min}^{\max} + \Delta y_{\min}^{\max} \quad \text{in } \mathring{\mathcal{B}}, \tag{2.97a}$$

$$\lambda\Delta u_{\mathcal{J}} + \Delta p_{\mathcal{J}} = \lambda\Delta u_d \quad \text{in } \mathcal{J}, \tag{2.97b}$$

$$\lambda\Delta u_{\mathcal{B}} + \Delta p_{\mathcal{B}} + \Delta q_{\mathcal{B}} = \lambda\Delta u_d \quad \text{in } \mathring{\mathcal{B}}. \tag{2.97c}$$

Hence, there holds

Proposition 6 (Index reduction by means of Bryson-Denham-Dreyfus approach)**:**
Let s be the differentiation index of DAEs given by Definition 18. This notion can by carried over to the specific PDAE (2.90) and (2.96), which yields

$s = 3$ for (2.90) and

$s = 1$ for (2.96).

One recognizes that the impact of the Bryson-Denham-Dreyfus approach of Paragraph 2.2.2, which basically consists in differentiating the original state constraint in order to obtain a control law, caused a double index reduction: for one thing the control law (2.96c) appears in the PDAE-system instead of the original state constraint and for another thing the whole approach yielded a multiplier $\bar{q}_{\mathcal{A}}$ which can be interpreted as an integrated version of the original multiplier $\mu_{\mathring{\mathcal{A}}}$; see Proposition 5. In other words, the double index reduction is due to an effect in the primal stage and due to an effect in the dual stage.

It has been explained in Paragraph 2.7.1, that the differentiation index has considerably impact on suitable numerical treatment. One may assume that this is valid for PDAE as well; see [51]. Consequently, the new necessary conditions of Corollary 3 might be more easily to solve than the original ones of Proposition 2 and Proposition 3.

However, due to the very simple structure of the considered model problem, it is nearby to reduce the PDAE to the inactive set; cf. the discussion in Paragraph 3.3.2 on page 104f.

The notion of the differentiation index is closely connected with the specification of *consistent initial conditions* [16, p. 19], [72, p. 456]. Such initial conditions respect all algebraic constraints – in particular the

hidden constraints which are revealed by determining the differentiation index – and hence yield unique solutions locally. In the field of PDAE, the determination of unique solution relies on specification of initial and/or boundary conditions, which respect all involved algebraic constraints [120]. Hence, one fundamental task in theory of (P)DAE is the specification of consistent initial/boundary data in order to be able to compute solutions.

The task in solving the state-constrained OCP is somehow the other way around: Consistent initial/boundary conditions are prescribed there. These are the boundary conditions on Γ and the interface conditions on γ of the first order necessary conditions of propositions 2 and 3 or of Corollary 3, respectively. But one has to determine the domain (this is the active set) such that this data is consistent.

This point of view it similar to *free boundary problems* where some boundary conditions are prescribed and one has to find the right domain such that these conditions permit a solution; cf. Paragraph 3.3.2.

Moreover, it would be worthwhile to analyze the considered PDAE with respect to the perturbation index, cf. [72, Chp. VII Def. 3.1], [22] and [118, 119, 140], since it gives valuable insight into the behavior of the discretized counterparts of the problem. However, this topic is beyond the scope of this thesis.

2.7.4 Order of a state constraint

It has been revealed in Paragraph 2.7.3 that the application of the BDD approach yields a double reduction of the differentiation index of the first order necessary conditions. This Paragraph is devoted to the closely related determination of the order of the state constraint in the style of the corresponding notion in OC-ODE; see, e. g., [122, 75]. The concept is basically the same as determining the differentiation index and means applying the first three steps of the recipe of determining the control law from page 23. This scheme yielded that one single differentiation step by means of applying the Laplacian to the state constraint in Paragraph 2.2.2 is sufficient.

In view of the usual reasoning in OC-ODE, where one differentiation step corresponds to a *first* order time derivative of the constraint, it seems reasonable to call the state constraint of the considered model problem (2.1) a constraint of second order.

Definition 19 (Order of a state constraint)**:**
The number of iterations needed to reach the stopping criterion in the 3rd step within the heuristic from page 23 multiplied by the order of the applied differential operator is called *the order of the state constraint*.

Obviously this definition has a limited range, since it is based upon a non-formalized heuristic, which deserves further investigation. Nonetheless, one obtains a first conjectured result in view of the topological possibilities of the active set associated with the state constraint of the considered model problem, which is based upon the knowledge from OC-ODE; cf. the citations on page 84.

Conjecture 2:
Since the order of the state constraint equals two, the topological possibilities of the active set are not prescribed. In particular, distributed components as well as isolated components (i. e. curves and single points) may occur.

Although this conjecture is barely an assertion, it implies the expectation, that constraints that possess an odd order behave analog to the OC-ODE case as well. In other words, such constraints can only produce active sets of lower dimension. A slight hint is due to [70], where the gradient-constrained numerical test example exhibits an active curve only.

2.8 Remarks on different necessary conditions

It has already be indicated at the beginning of Section 2.2 that the BDD approach and the treatment of the active set as a separate variable are independent ideas. In particular, the latter can also be found in the paper of Hintermüller and Ring [90], where first order necessary conditions are formulated upon the

basis of a direct application of the state constraints. Henceforth, their optimality system (see [90, Prop. 2]) is similar to the one of Corollary 3 but contains the multipliers $\mu_{\mathcal{A}}$ and μ_γ.

In view of the different optimality systems, which are obtained via different BDD approaches (see Appendix A), and in advance of Paragraph 3.3.2 the abstract view on set optimal control problems of Paragraph 2.6.3 provides the following insight. Each of the different approaches yields its own geometry-to-solution operator and thus induces its own graph in the vector bundle E over $\mathcal{X}(\mathcal{A})$. By means of Paragraph 3.3.1 one recognizes, that these different graphs can be used to construct different Newton algorithms. A comparison of the performance of those algorithms may rate the value of the underlying approaches. However, the idea of relaxation approaches, which is presented in Paragraph 3.3.2, reveals that there are even more algorithms to be taken into account, such that a concluding verdict is beyond the scope of this thesis.

CHAPTER 3

Algorithms

This Chapter is devoted to the derivation of algorithms to solve the model problem (2.1) based upon the approach of Chapter 2. As presented in Section 2.6 shape optimization cannot be regarded as optimization in a linear space, but on an infinite dimensional manifold. Hence, one has to be aware of some fundamental differences to optimization on Banach spaces from the algorithmic point of view. These are discussed in sections 3.1 and 3.2. Further details on algorithms designed for optimization on (finite dimensional) manifolds can be found in a comprehensible textbook due to Absil, Mahony and Sepulchre [1].

The optimality systems obtained in Chapter 2 are analyzed in more detail in Section 3.3 in order to explore different approaches how they can be solved. Based upon the experiences gained thereby some adapted algorithms are formulated and discussed in Section 3.4. Finally Section 3.5 is devoted to an analysis of well established primal-dual active set strategy in order to contrast some benefits and drawbacks of the new algorithms.

It is important to note, that the whole algorithmic analysis comes without profound results like proofs for convergence. Moreover, this Chapter and the numerical analysis of Chapter 4 are contented with the shape calculus based point of view. Questions concerning topology calculus/optimization are left unattended and may be a topic for future research. A first step towards topological analysis of state-constrained OCPs is due to Hintermüller and Laurain [89].

3.1 Descent algorithms in $\mathcal{H}(\Omega)$

Descent algorithms form probably the most elaborate and most important class of solution strategies in (unconstrained) nonlinear optimization, since they are typically the basis for more sophisticated approaches. Hence, it is worth addressing them in more detail.

In the context of optimization in Banach spaces descent algorithms perform with an iterative three-step strategy.

1. As a start, find a search direction originating at the current iterate,
2. afterwards determine a step size to go in this direction,
3. finally add the vector "step size times search direction" to the current iterate.

If this is done astutely enough one obtains, that the objective at the new iterate is smaller than at the starting point. Of course, one has to guarantee convergence as such and that the descent is large enough in order to prevent the algorithm from converging against an objective value that is not optimal; however this is not the focus of the present discussion. The keywords *gradient-related sequence* and *Armijo step length* should be sufficient here, cf. [46, 131]. Instead, the interpretation of the three steps is addressed for optimization problems that are not posed in Banach spaces but in the metric space $\mathcal{H}(\Omega)$. At this, $\mathcal{H}(\Omega)$ ought to be regarded as similar to a Riemannian manifold with tangent space $T_{\mathrm{Id}}\mathcal{H}(\Omega) = \Theta_0$; cf. the 10th item of the discussion on page 67.

The search directions f of step one are located in the tangent space Θ_0 to the manifold now and the tangent space does not coincide with the space $\mathcal{H}(\Omega)$ anymore, as one was used to in the framework of Banach spaces.

The metric space indeed is a subset of the affine space $\mathrm{Id} + \Theta_0$, however it is not convex and hence the line search of step two should not be interpreted as the minimization along the line $\mathrm{Id} + tf \subset \mathrm{Id} + \Theta_0$, $t \in \mathbb{R}$, since this straight line might abandon the manifold $\mathcal{H}(\Omega)$. This situation is illustrated in Figure 3.1. The line search should be performed on the geodesic through the current iterate in the search

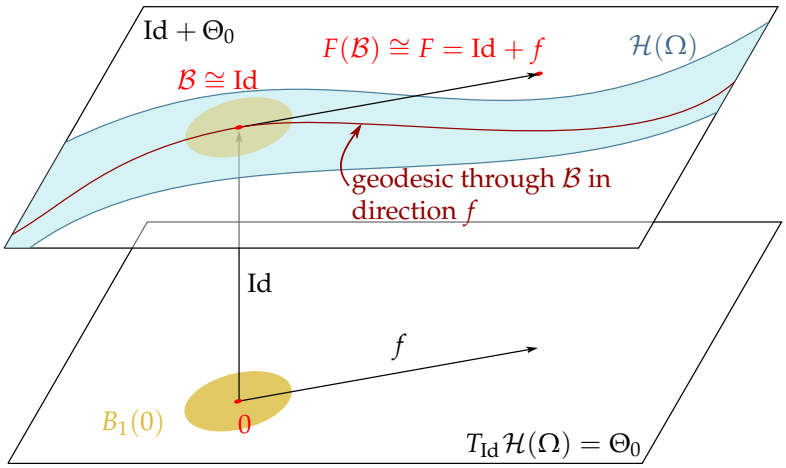


Figure 3.1: One step of a descent algorithm.

direction. Hence, the determination of a step size hosts two main difficulties. On the one hand it is a non-trivial task to establish a constructive characterization of geodesics in $\mathcal{H}(\Omega)$ and probably there is no efficient numerical scheme to evaluate them.[1] On the other hand, like in the Banach space situation, the (numerical) minimization along some path (for instance a geodesic or a straight line) is expensive and consequently one would use some well established line search algorithms like Armijo's step length rule or even more sophisticated ideas; cf. for instance [15, Sec. 4.2.1].

The third step of the general scheme is the update of the current iterate. This part is basically the same in the Banach space setting and in $\mathcal{H}(\Omega)$. Once, that the direction is identified, one follows the corresponding geodesic up to the determined step size.

The second step is scrutinized in the following. There are two nearby ideas in order to circumvent the difficulties arising from the determination of geodesics in $\mathcal{H}(\Omega)$. In other words, there are other retractions (cf. Definition 20) than the exponential map, which can be accessed computationally in a more efficient way. The first idea is about approximating geodesics via cheaply achievable paths inside $\mathcal{H}(\Omega)$, the second is to follow the descent direction in $\mathrm{Id} + \Theta_0$ without paying attention, whether $\mathcal{H}(\Omega)$ is left, and use a projection back into the manifold if required.

The first of them originates in the specific property of the metric space $\mathcal{H}(\Omega)$, that the unit ball $\mathcal{B}_1(0) \subset \Theta_0$ can be continuously embedded into the manifold, this is $\mathrm{Id} + B_1(0) \subset \mathcal{H}(\Omega)$; cf. the first step of the proof of Lemma 14 and Figure 3.1. Hence, it is possible to follow the vector field f in the affine space $\mathrm{Id} + \Theta_0$, without immediately leaving the manifold; henceforth denoted by *transformation approach (of path following)*. From the perspective of geometric deformations of the candidate active set $\mathcal{B}$ within the holdall Ω, that is nothing but choosing a suitable $\tau > 0$ and applying $X \mapsto (\mathrm{Id} + \tau f)(X)$ to all $X \in \Omega$. Actually it is sufficient to proceed likewise with all $X \in \beta$ and to determine the boundary of the image of $\mathcal{B}$ in this manner.

A more sophisticated approach uses the flow induced by the vector field f, henceforth denoted by *flow approach (of path following)*. Here $X \in \Omega$ is mapped to $T_\tau(f)(X) := x(\tau, X)$, where

$$\frac{\mathrm{d}}{\mathrm{d}t} x(t, X) = f(x(t, X)), \quad t \in [0, \tau], \quad x(0, X) := X. \tag{3.1}$$

[1] A well-known and often used connection between the tangent space and geodesics is the *(Riemannian) exponential map*; [1, Sec. 5.4]. However, it is mentioned there (page 103), that this map is not necessarily the best choice in view of computational efficiency.

Both approaches are illustrated in Figure 3.2, namely both from the perspective of the metric space $\mathcal{H}(\Omega)$ and from the perspective of geometric deformation of the boundary β of the active set $\mathcal{B}$.[2]

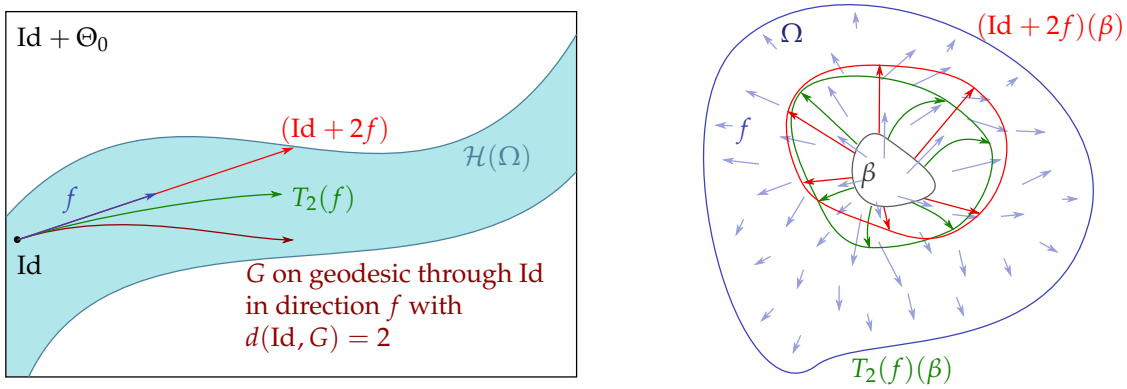


Figure 3.2: Transformation- and flow approach of approximate geodesic path following illustrated in the regime of the metric space $\mathcal{H}(\Omega)$ (left) and from the point of view of geometric deformation, respectively (right).

Obviously, the flow approach is more accurate but the price to pay is a more involved computation. In particular, it requires the knowledge of the vector field f in the whole domain Ω. It turns out (see the Remarks on the Newton update step on page 111), that f is known on the interface β only and consequently the flow approach demands an efficient scheme how to extend the vector field. Fortunately, there are several highly developed ideas and efficient numerics available, for instance *level set methods* and *fast marching methods*; cf. [149]. Additional ideas are discussed in Paragraph 4.1.3.

The question of extending the vector field from the current interface to the bulk of the domain is directly linked to the question of a *horizontal lift* of the tangent vector of the quotient manifold $\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{A})$ to a tangent vector of the manifold $\mathcal{H}(\Omega)$; cf. [1, Sec. 3.5.8]. The concept is the same as the construction of a *horizontal subbundle*, cf. [113, p. 101ff.], which is important in the context of *Ehresmann connections*.

From the perspective of $\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{A})$ the manifold $\mathcal{H}(\Omega)$ can be described as a *fiber bundle*, where the canonical projection

$$\pi_{\mathcal{A}} : \mathcal{H}(\Omega) \to \mathcal{H}(\Omega)/\mathcal{K}(\mathcal{A}), \quad F \mapsto [F]_{\mathcal{A}}$$

is the bundle projection. It induces a projection $T(\pi_{\mathcal{A}}) : T\mathcal{H}(\Omega) \to T(\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{A}))$ between the tangent bundles. This function can be understood as follows (cf. Figure 3.3). Let $\mathcal{B} \in \mathcal{X}(\mathcal{A})$. Then a normal vector field v defined on the interface $\beta = \partial\mathcal{B}$ can be interpreted as a tangent vector in $T_{\mathcal{B}}\mathcal{X}(\mathcal{A}) \cong T_{[F]_{\mathcal{A}}}\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{A})$, where $\mathcal{B} = [F]_{\mathcal{A}}(\mathcal{A})$ (see the Remark on page 63 for a definition of the notation). Choose an arbitrary representative G of the equivalence class $[F]_{\mathcal{A}}$. Hence, all tangent vectors $f \in T_G\mathcal{H}(\Omega)$ such that $T(\pi_{\mathcal{A}})(f) = v$ can be regarded as representatives of the tangent vector v. Each such f is an element of Θ_0 and fulfills $f|_\beta \cdot \boldsymbol{n}_{\mathcal{B}} = v$. In other words, the projection $T(\pi_{\mathcal{A}})$ is nothing but

$$T(\pi_{\mathcal{A}}) : T\mathcal{H}(\Omega) \to T(\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{A})), \quad (F, f) \mapsto f|_{\partial F(\mathcal{A})} \cdot \boldsymbol{n}_{F(\mathcal{A})} \in T_{[F]_{\mathcal{A}}}(\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{A})).$$

Obviously there are infinitely many f, which are elements of Θ_0, and which are projected on a single v, namely all extensions of v into the bulk of the domain Ω.

The equivalence class $[F]_{\mathcal{A}} \subset \mathcal{H}(\Omega)$ is a submanifold of $\mathcal{H}(\Omega)$ and consequently has its own tangent bundle $T[F]_{\mathcal{A}}$. This tangent bundle forms the kernel of projection $T(\pi_{\mathcal{A}})$ and is called the *vertical subbundle* $\boldsymbol{V}$ of $T\mathcal{H}(\Omega)$. Hence, there holds

$$\forall G \in [F]_{\mathcal{A}}, \forall f \in \boldsymbol{V}_G : \quad T(\pi_{\mathcal{A}})(f) = 0 \in T_{[F]_{\mathcal{A}}}(\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{A})).$$

[2] The two path following approaches are directly connected with the two different approaches of the *perturbation of identity* and the *velocity method* in shape calculus, cf. [44, Chp. 1 Sec. 10.6], and they define retractions

$$R_F^{trans} : T_F\mathcal{H}(\Omega) \to \mathcal{H}(\Omega), \quad f \mapsto \mathrm{Id} + F + f,$$
$$R_F^{flow} : T_F\mathcal{H}(\Omega) \to \mathcal{H}(\Omega), \quad f \mapsto T_1(f)(F),$$

as long as $\|f\|_{\Theta_0}$ is small enough.

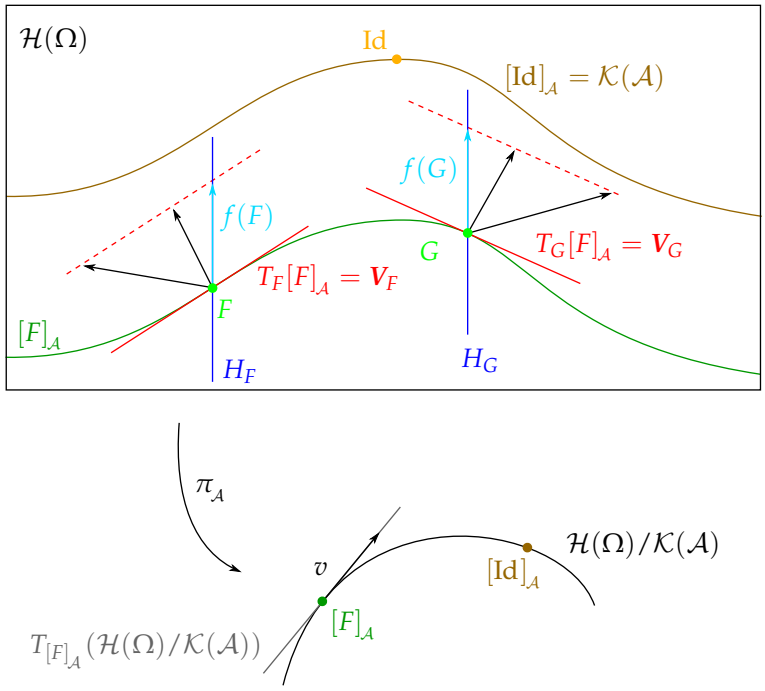


Figure 3.3: Illustration of the canonical projection $\pi_{\mathcal{A}}$ of $\mathcal{H}(\Omega)$ to $\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{A})$ and of vertical and horizontal subbundles $\boldsymbol{V}$ and $\boldsymbol{H}$.

Since each fiber $\boldsymbol{V}_G$ is a vector subspace of Θ_0, it is possible to define a corresponding direct complement H_G such that

$$\Theta_0 = \boldsymbol{V}_G \oplus \boldsymbol{H}_G.$$

If there are some regularity requirements fulfilled (concerning the choice of $\boldsymbol{H}_G$ in dependency of G, cf. [113, p. 101ff.]), the collection of all complements $\boldsymbol{H}_G$ form a subbundle of $T\mathcal{H}(\Omega)$, the *horizontal subbundle* $\boldsymbol{H}$. Note, that the vertical subbundle is uniquely determined, whereas there is some freedom to choose the horizontal subbundle.

The horizontal subbundle enables to choose a unique representative $f(G)$, called *horizontal lift*, of the tangent vector $v \in T_{[F]_{\mathcal{A}}}(\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{A}))$ for each representative $G \in [F]_{\mathcal{A}}$. It would be desirable now that all representatives $f(G)$ coincide, independent of G. Hence, all subspaces H_G must be equal. Such a choice is possible only, if all vertical spaces $\boldsymbol{V}_G$ are such that there exists a horizontal space $\boldsymbol{H}_F$ with

$$\Theta_0 = \boldsymbol{V}_G \oplus \boldsymbol{H}_F, \quad \text{for all } G \in [F]_{\mathcal{A}}.$$

It is beyond the scope of this thesis to prove that this condition can be satisfied. However, it is obvious, that there are various extensions of a normal vector field v and that these extension are only dependent on v and the boundary β. In particular, these extension are independent from the transformation "used" to map the set $\mathcal{A} \cong [\mathrm{Id}]_{\mathcal{A}}$ to $\mathcal{B} \cong [F]_{\mathcal{A}}$, which means they are independent of the choice $G \in [F]_{\mathcal{A}}$.

All in all, a specific algorithm which produces an extension of v to the holdall corresponds to a realization of a horizontal bundle whose fibers $\boldsymbol{H}_G$ are all equal.

As already mentioned, it is not possible to use the transformation approach with arbitrarily large step size $\tau > 0$, since one might leave the manifold $\mathcal{H}(\Omega)$. Regarded from the perspective of geometrical deformation of the interface β, this effect manifests itself as some pathological behavior. The image of interface might have corners and thus is not of class $C^{1,1}$ any more, or even worse, it might be self-intersecting.[3] Self-intersection implies, that there is no reasonable interpretation of the interior of the interface, and thus the responsible transformation of the active set is infeasible. These type of problems are illustrated in Figure 3.4 and cannot occur in the framework of the flow approach. For one thing the trajectories $x(., X)$ for $X \in \Omega$ do not intersect each other, which prevents self-intersection of the images of

[3] This topic is closely related to the introduction of *viscosity solutions* and *entropy solutions* and ideas of level set methods in the context of propagating interfaces; cf. [149].

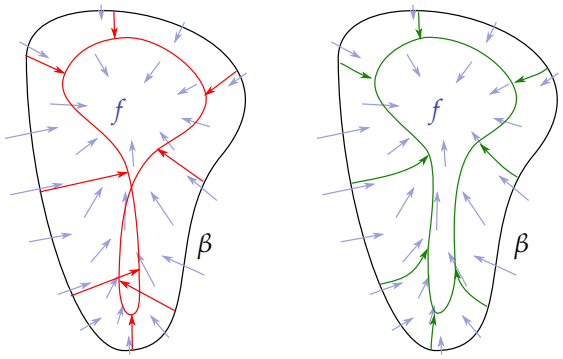


Figure 3.4: Self-intersection of the image of the interface produced by the transformation approach (left), whereas the flow approach does not produce problems (right).

the interface, and for another thing the trajectories cannot leave Ω, since $f|_{\Omega^c} \equiv 0$ and thus the maximal interval of existence is unbounded; see, for instance, [7, Satz 2.5.1] or [5, Thm. 7.6] and the manifold corresponding results in [113, Chp. 4 §1]. Finally, the flow $x(t, X)$ is $C^{1,1}$-regular with respect to $X \in \Omega$, since f is; cf. [5, Thm. 9.5, Rem. 9.6]. All in all, the images of the interface β transported by the flow are of class $C^{1,1}$ for all t.

However, the numerical practice of the flow approach is not immune from leaving the manifold $\mathcal{H}(\Omega)$. In fact, it is not possible to solve the flow equation (3.1) exactly and one must use approximate solutions produced by some ODE solver. There is no guarantee, that the approximations of the different analytical trajectories do not intersect each other, and hence self-intersection of the image of the interface β cannot be avoided in general. This problem is reinforced by the lack of knowledge of the perturbation vector field in the bulk of Ω and the resulting need for numerical extensions. Nonetheless, there is always a $\tau > 0$, such that the flow image $T_\tau(f)(\mathcal{B})$ of the active set remains in $\mathcal{H}(\Omega)$.

After having seen, that the numerical practice of both the transformation and the flow approach of path following can suffer from the fact that the next iterate may be outside the manifold, if one does not care about the maximal step size, it is nearby to analyze the second idea to circumvent geodesics, which uses a projection into the metric space $\mathcal{H}(\Omega)$. In the context of (finite dimensional) manifolds such kind of projections are called *retractions*, cf. [1, Def. 4.1.1].

Definition 20 (retraction)**:**
Let $\mathcal{M}$ be a manifold, let $T_x\mathcal{M}$ be the tangent space to $\mathcal{M}$ at point $x \in \mathcal{M}$ (cf. Definition 8) and let $T\mathcal{M}$ be the tangent bundle (cf. Definition 9).

Then a smooth mapping $R : T\mathcal{M} \to \mathcal{M}$ is called *retraction* on the manifold $\mathcal{M}$, iff it fulfills the following properties. Let R_x be the restriction of R to the tangent space $T_x\mathcal{M}$, then

1. $R_x(0_x) = x$, where 0_x is the zero element of $T_x\mathcal{M}$,
2. $\mathrm{D}R_x(0_x) = \mathrm{Id}_{T_x\mathcal{M}}$.

In the particular case of the metric space $\mathcal{H}(\Omega)$, it is known, that the tangent space $T_F\mathcal{H}(\Omega)$ is given by Θ_0 for each $F \in \mathcal{H}(\Omega)$. Consequently, due to the natural projection

$$T(\mathcal{H}(\Omega)) \to \mathcal{H}(\Omega), \quad T_F\mathcal{H}(\Omega) \ni f \mapsto F,$$

and since the manifold $\mathcal{H}(\Omega)$ can be covered by a single chart by construction, the tangent bundle becomes isomorphic to

$$T(\mathcal{H}(\Omega)) \cong \mathcal{H}(\Omega) \times \Theta_0.$$

This is to say, the tangent bundle $T\mathcal{H}(\Omega)$ is a trivial vector bundle on $\mathcal{H}(\Omega)$; cf. Definition 16.

By means of retractions it is possible to project paths located in the tangent bundle into the manifold. In other words, let $R : \mathcal{H}(\Omega) \times \Theta_0 \to \mathcal{H}(\Omega)$ be a retraction, let F be an element of $\mathcal{H}(\Omega)$ and let $f \in \Theta_0 = T_F\mathcal{H}(\Omega)$ be a tangent vector, then the mapping $t \mapsto R(t\,f)$ defines a path in the manifold $\mathcal{H}(\Omega)$, which can be interpreted as the projection of the path $t \mapsto t\,f$, which is located in $T\mathcal{H}(\Omega)$. This situation is illustrated in Figure 3.5. From the point of view of the geometric deformation of the interface β and the

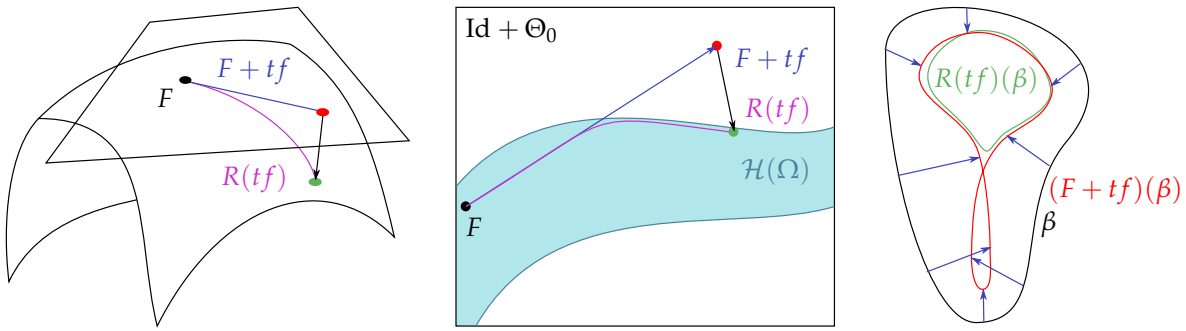


Figure 3.5: Different illustrations of a linear path $F + tf$ in the (affine) tangent space and its projection $R(tf)$ by means of a retraction: perspective from a typical manifold (left), from the metric space $\mathcal{H}(\Omega)$ and from the geometric deformation (right).

candidate active set $\mathcal{B}$, respectively, the retraction ensures, that the image keeps contained in the family $\mathcal{O}$ and therefore it avoids degeneracy phenomena.

Whereas Riemannian manifolds always admit a retraction given by the *Riemannian exponential map*, cf. [1, Sec. 5.4], it is not obvious whether the metric space $\mathcal{H}(\Omega)$ possesses a map that can be regarded as the generalization of a retraction to the infinite dimensional setting, and it goes beyond the scope of this thesis to answer this question. It should be emphasized, that mere existence of such a mapping is not enough – it is valuable only, if it can be computed efficiently. Nonetheless, there are ideas how to avoid degeneracy phenomena on the discretized level within the numerical realization of descent algorithms; cf. Paragraph 4.1.2. Indeed, these projections behave like a retraction. They do not influence the transformation of the active set, as long as the deformation is small enough and consequently the defining properties of Definition 20 are fulfilled.

As concluding remark it is worthwhile to comment on a similarity between steepest descent methods and the approximation of geodesics by means of the transformation approach of path following. The steepest descent algorithm tries to follow the path of steepest descent which originates at the initial guess and leads down to the minimum. However, this path is curved in general and hence it is inefficient/impossible to compute it exactly. Thus, one confines oneself with approximations by means of tangent direction on this path (similar to an Euler method for solving ODEs). This is obviously analog to the transformation approach.

This brief overview on descent algorithms must by sufficient here. The most prominent representative of this class is the method of steepest descent. Unfortunately, this algorithm is not applicable to the set optimal control problem (2.30) and its reduced shape optimization counterpart (2.45), since the critical points of the shape functional $\mathcal{F}$ are no (local) minima, which is shown next.

3.1.1 The optimal solution is no local minimum of $\mathcal{F}$

One possible idea for solving the bilevel optimization problem numerically is to apply a gradient-based algorithm on the shape functional $\mathcal{F}$. However, the bilevel optimization problem is constrained, since one has to respect the strict inequality constraint $y_{\min} < y_{\mathcal{J}} < y_{\max}$ in the candidate inactive set $\mathcal{J}$. Nonetheless, one of the most fundamental ideas behind the whole approach of this thesis is, that this constraint does not need to be respected rigorously. To be more precise, the strict inequality is not relevant if first order optimality conditions are investigated, but it has some impact on optimality from a global point of view; cf. Paragraph 2.2.4.

With regard to the construction of algorithms the question arises, whether the constraint has to be taken into account, or not. Sure enough, it has to be considered when a steepest descent algorithm is applied, since the active set $\mathcal{A}$ is no strict (local) minimum of the shape functional $\mathcal{F}$. This fact is illustrated from the algorithmic point of view now and it is neatly proven afterwards.

Due to Theorem 8 the shape gradient of $\mathcal{F}$ can be represented by the non-positive function

$$-\frac{1}{2\lambda}(\bar{p}_{\mathcal{J}} - p_{\min}^{\max})^2 \in L^1(\beta).$$

A steepest descent algorithm would choose the normal component of a perturbation vector field as a scalar multiple of the representative.[4] But the non-positivity of the gradient results in perturbation fields which can only shrink the current guess of the active set, since the step size is always positive. This behavior is due to the fact, that the unique global optimum of the reduced shape functional $\mathcal{F}$ is given by $\mathcal{B} = \emptyset$, which corresponds to the optimal solution of the state unconstrained version of the original optimal control problem (2.1). Therefore, such an algorithm can never reach the optimal active set $\mathcal{A}$, if the initial guess $\mathcal{B}$ is such that $\mathcal{A} \nsubseteq \mathcal{B}$ as illustrated in Figure 3.6. In the special case $\mathcal{B} \subset \mathcal{A}$ a steepest descent algorithm only has a chance to reach $\mathcal{A}$ if negative step sizes are allowed, too.

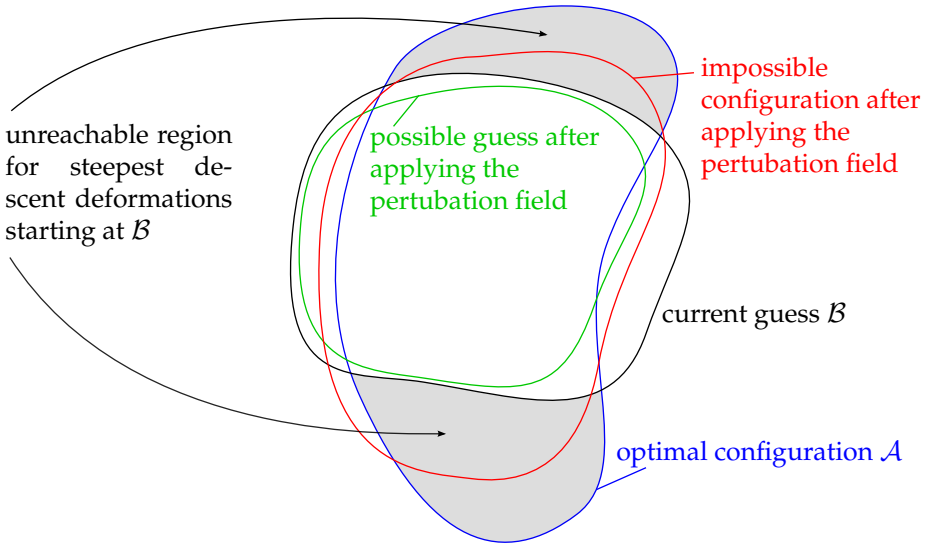


Figure 3.6: Illustration of an impossible step in a steepest descent algorithm.

Conversely this means that there are vector fields, which lead to perturbations of the optimal active set $\mathcal{A}$, such that the value of the shape functional decreases. Thus, $\mathcal{A}$ cannot be a strict local minimum of $\mathcal{F}$.

Proposition 7 ($\mathcal{A}$ is no strict local minimum of $\mathcal{F}$)**:**
Let $\mathcal{A} \in \mathcal{O}$ be the (optimal) active set for the original optimal control problem (2.1) and its equivalent reduced shape/topology optimization problem (2.45). Furthermore assume that $\mathcal{A} \neq \emptyset$, i. e. the state constraints are essential.

Then $\mathcal{A}$ is a critical point, but is no strict local minimum of the reduced objective $\mathcal{F}$.[5]

Proof. **1)** According to Theorem 8 the set $\mathcal{A}$ is a critical point of the reduced objective $\mathcal{F}$ with respect to shape calculus.

2) According to the 14th item of the discussion on page 68 the ball $B_1(0) \subset \Theta_0$ is continuously embedded in $\mathcal{H}(\Omega)$ by means of the mapping $f \mapsto \mathrm{Id} + f$. Moreover, the mapping is surjective onto a suitable ε-ball. Let $\varepsilon > 0$ be chosen adaptively. Furthermore, there exists an approximation $f \in \Theta_0$ of the outward unit normal vector field of the active set $\mathcal{A}$; see [69, Lem. 1.5.1.9]. That is, there exists a $\delta > 0$ and an $f \in \Theta_0$ such that

$$\forall x \in \gamma : \quad f(x) \cdot \boldsymbol{n}_{\mathcal{A}}(x) \geq \delta.$$

Choose $\eta > 0$ such that $F := \mathrm{Id} - \eta f \in B_\varepsilon \subset \mathcal{H}(\Omega)$ and define $\mathcal{B} := F(\mathcal{A}) \in \mathcal{X}(\mathcal{A}) \subset \mathcal{O}$. Then there holds $\mathcal{B} \subset \mathcal{A}$, since the transformation F maps the boundary γ towards the interior of $\mathcal{A}$, i. e. $F(\gamma) \subset \mathring{\mathcal{A}}$.

Let $(\bar{u}_{\mathcal{I}}, \bar{u}_{\mathcal{A}}, \bar{y}_{\mathcal{I}}, \bar{u}_{\mathcal{A}}) \in L^2(\mathcal{I}) \times L^2(\mathring{\mathcal{A}}) \times H^1(\mathcal{I}, \Delta) \times H^1(\mathring{\mathcal{A}}, \Delta)$ be the optimal solution of the set optimal control problem (2.30). This tuple is the optimal solution of the bilevel problem (2.36), (2.37), too. In

[4] A steepest descent algorithm based on a Sobolev gradient, which was introduced in the Remark to Theorem 7 would act similar, since the weak maximum principle still holds true for the corresponding surface PDE with the usual proof, and thus $\nabla_S \mathcal{F}$ has the same sign as $\nabla \mathcal{F}$.

[5] The assertion is related to shape calculus only, since the analysis of infinitesimal topology dependency goes beyond the scope of this thesis.

particular, it is the optimal solution of the inner optimization part (2.37) to the fixed set $\mathcal{A}$. Additionally, the tuple is feasible for the inner optimization problem to the fixed set $\mathcal{B}$: If one concatenates the states $\bar{y}_{\mathcal{A}}$ and $\bar{y}_{\mathcal{I}}$, one obtains a state on the whole domain Ω (cf. Proposition 4), which itself can be split again to $y_{\mathcal{J}} \in H^1(\mathcal{J}, \Delta)$ and $y_{\mathcal{B}} \in H^1(\mathring{\mathcal{B}}, \Delta)$. Assembling and anew dissection can be done with the optimal control as well. Furthermore, there holds $y_{\mathcal{B}} = \bar{y}_{\mathcal{A}}|_{\mathcal{B}} \equiv y_{\min}^{\max}$, since $\mathcal{B}$ is a subset of $\mathcal{A}$. Consequently, all constraints of the inner optimization problem (2.37) are fulfilled.[6]

Hence, there holds by definition of $\mathcal{F}$, cf. (2.38a)

$$\mathcal{F}(\mathcal{B}) = \min_{u_{\mathcal{J}}, u_{\mathcal{B}}, y_{\mathcal{J}}, y_{\mathcal{B}}} \mathfrak{J}(\mathcal{B}; u_{\mathcal{J}}, u_{\mathcal{B}}, y_{\mathcal{J}}, y_{\mathcal{B}}) \leq \mathfrak{J}(\mathcal{B}; \bar{u}_{\mathcal{I}}, \bar{u}_{\mathcal{A}}, \bar{y}_{\mathcal{I}}, \bar{y}_{\mathcal{A}}) = \mathfrak{J}(\mathcal{A}; \bar{u}_{\mathcal{I}}, \bar{u}_{\mathcal{A}}, \bar{y}_{\mathcal{I}}, \bar{y}_{\mathcal{A}}) = \mathcal{F}(\mathcal{A}).$$

In other words, for each (sufficiently small) $\varepsilon > 0$ there is a set $\mathcal{B} \in \mathcal{O}$ with $d(\mathcal{B}, \mathcal{A}) \leq \varepsilon$ (the metric induced in $\mathcal{X}(\mathcal{A})$, cf. Lemma 13) and $\mathcal{F}(\mathcal{B}) \leq \mathcal{F}(\mathcal{A})$. Hence, $\mathcal{A}$ is no strict local minimum of the reduced objective $\mathcal{F}$. □

Remark:
The proof shows that there even holds

$$\forall \mathcal{B}_1, \mathcal{B}_2 \in \mathcal{O} \text{ with } \mathcal{B}_1 \subset \mathcal{B}_2 : \quad \mathcal{F}(\mathcal{B}_1) \leq \mathcal{F}(\mathcal{B}_2).$$

Thus, it is necessary to use algorithms which either rely on the strict inequality constraint or which search for critical points of the shape functional $\mathcal{F}$.

It is not possible to satisfy the strict inequality constraint directly, as it is done, for instance, by projected gradient methods, since the constraint poses an implicit condition on the feasibility of a set $\mathcal{B}$. Here one recognizes the character of a state constraint. A common remedy is to fulfill such constraints iteratively by means of some penalization. However, this approach contradicts the original goal to get an algorithm which neither requires regularization nor penalization of the constraint.

At this point, it is worthwhile to comment on penalization of the strict inequality constraint $y_{\min} < \bar{y}_{\mathcal{J}} < y_{\max}$. One nearby idea is to augment $\mathcal{F}$ by a quadratic penalty term

$$A(\mathcal{B}) := \frac{c}{2} \int_{\mathcal{J}} \max\{0, \bar{y}_{\mathcal{J}} - y_{\max}\}^2 + \max\{0, y_{\min} - \bar{y}_{\mathcal{J}}\}^2 \tag{3.2}$$

where $c > 0$; cf. [90, Eq. (3.6)]. Afterwards one studies the behavior when c is send to infinity. However, this approach only ensures, that the inequality constraint $y_{\min} \leq \bar{y}_{\mathcal{J}} \leq y_{\max}$ is respected in the limit, whereas one requires the strict inequality counterpart. This approach is not recommended, since a loss of unique solvability and of the precise meaning of the (in-)active set is concomitant with it; cf. Paragraph 2.2.4 in particular page 29. A remedy would be to sharpen the penalty term to

$$A_\varepsilon(\mathcal{B}) := \frac{c}{2} \int_{\mathcal{J}} \max\{0, \bar{y}_{\mathcal{J}} - y_{\max} + \varepsilon\}^2 + \max\{0, y_{\min} - \bar{y}_{\mathcal{J}} + \varepsilon\}^2$$

and to drive ε to zero. However,

$$\bar{y}_{\mathcal{J}}(x) \to y_{\min}^{\max}(x), \quad x \to \bar{x} \in \beta,$$

since $\bar{y}_{\mathcal{J}}|_\beta = y_{\min}^{\max}|_\beta$ due to (2.45f). Consequently, this idea induces a conflict between the evaluation of $\mathcal{F}$ and its gradient. It manifests itself in the impossibility of convergence of the algorithm unless $\varepsilon = 0$, since otherwise the penalty term can never vanish and always gives a descent direction. Furthermore, the introduction of two penalty parameters always requires some smart coupling which is very likely to be problem dependent.

All in all, it is indicated to develop a method, which does without paying attention to the constraint $y_{\min} < y_{\mathcal{J}} < y_{\max}$ but searches for critical points of the reduced functional $\mathcal{F}$. An obvious choice is a Newton scheme, which is introduced in the next section 3.2.

However, despite the objections it seems to be reasonable to use the penalty approach as globalization strategy for a Newton algorithm. As long as the current guess of the active set is "far away" from the optimal one, an augmented functional may give the right idea how to deform the iterate. And when the guess is "near enough" one switches to a Newton scheme; cf. the remarks on the a posteriori step 3 on page 111.

[6] Note, that the state $y_{\mathcal{J}}$ obtained by this means is not expected to fulfill the strict inequality constraint (2.36c). This is just the substance of considerations: analyze the critical point $\mathcal{A}$ of the *unconstrained* functional $\mathcal{F}$.

3.2 Remarks on Newton techniques on manifolds

This paragraph is devoted to a short presentation of the concept of Newton's method on Riemannian manifolds. The main source is the recent textbook [1], in particular sections 5 and 6.

Given a sufficiently smooth objective f on a Euclidean vector space, the Newton's method is applied to find critical points of the gradient ∇f. It performs by iteratively solving the *Newton equation* and updating the old iterate (*Newton update*)

$$\begin{aligned}(\operatorname{Hess} f(x_k))[\delta_k] &= -\operatorname{grad} f(x_k),\\ x_{k+1} &= x_k + \delta_k.\end{aligned}$$

This rule requires the explanation of the different constituents in the context of a manifold $\mathcal{M}$:

- The Newton equation has to be handled with care, since one has to use the gradient and the Hessian operator (see Definition 15) and cannot simply use first and second order directional derivatives. In case of a non-Riemannian manifold there is neither a notion of a Hessian nor of a gradient available and a remedy is to apply first and second order covariant derivatives (cf. definitions 11 and 13) and to use the functional equation in $T^*_{x_k}\mathcal{M}$

  $$(\nabla^2 f)(x_k)[\delta_k, .] = -(\mathrm{D}f)(x_k)[.]$$

  or its variational form

  $$(\nabla^2 f)(x_k)[\delta_k, v] = -(\mathrm{D}f(x_k))[v], \quad \forall v \in T_{x_k}\mathcal{M}.$$

  In either case the increment δ_k is an element of the tangential space $T_{x_k}\mathcal{M}$.
- Moreover, the Newton update is not directly realizable on manifolds, since they lack a linear structure in general. However, the update can be understood by means of a retraction (see Definition 20)

  $$x_{k+1} := R_{x_k}(\delta_k).$$

Hence, it is an important step to identify the second covariant derivative of the reduced shape functional $\mathcal{F}$ in order to establish a Newton's method; cf. the 20th item of the discussion on page 80.

3.3 Different perspectives on first order optimality system

The basis of all algorithms analyzed within this thesis are the first order necessary conditions of the original model problem (2.1) related to the Bryson-Denham-Dreyfus approach (Paragraph 2.2.2) and which were stated in Corollary 3. However, the equivalent necessary conditions, which are derived in Appendix B, are used here for them being less tailored to the original model problem (cf. (2.40) for the specific reformulation which yields the necessary conditions of Corollary 3). They are put together for convenience:

$$-\Delta\bar{y}_{\mathcal{I}} + \bar{y}_{\mathcal{I}} = \bar{u}_{\mathcal{I}} \quad \text{in } \mathcal{I}, \tag{3.3a}$$
$$-\Delta\bar{y}_{\mathcal{A}} + \bar{y}_{\mathcal{A}} = \bar{u}_{\mathcal{A}} \quad \text{in } \mathring{\mathcal{A}}, \tag{3.3b}$$
$$\partial_{\boldsymbol{n}}\bar{y}_{\mathcal{I}} = 0 \quad \text{on } \Gamma, \tag{3.3c}$$
$$\bar{y}_{\mathcal{I}}|_\gamma - \bar{y}_{\mathcal{A}}|_\gamma = 0 \quad \text{on } \gamma, \tag{3.3d}$$
$$\partial^{\mathcal{I}}_{\boldsymbol{n}}\bar{y}_{\mathcal{I}} + \partial^{\mathcal{A}}_{\boldsymbol{n}}\bar{y}_{\mathcal{A}} = 0 \quad \text{on } \gamma, \tag{3.3e}$$
$$-\Delta y^{\max}_{\min} + y^{\max}_{\min} = \bar{u}_{\mathcal{A}} \quad \text{in } \mathring{\mathcal{A}}, \tag{3.3f}$$
$$y^{\max}_{\min}|_\gamma = \bar{y}_{\mathcal{A}}|_\gamma \quad \text{on } \gamma, \tag{3.3g}$$
$$y_{\min} < \bar{y}_{\mathcal{I}} < y_{\max} \quad \text{in } \mathcal{I}, \tag{3.3o}$$

$$-\Delta\hat{p}_{\mathcal{I}} + \hat{p}_{\mathcal{I}} = \bar{y}_{\mathcal{I}} - y_d \quad \text{in } \mathcal{I}, \tag{3.3h}$$
$$-\Delta\hat{p}_{\mathcal{A}} + \hat{p}_{\mathcal{A}} = \bar{y}_{\mathcal{A}} - y_d \quad \text{in } \mathring{\mathcal{A}}, \tag{3.3i}$$
$$\partial_{\boldsymbol{n}}\hat{p}_{\mathcal{I}} = 0 \quad \text{on } \Gamma, \tag{3.3j}$$
$$\hat{p}_{\mathcal{I}}|_\gamma - \hat{p}_{\mathcal{A}}|_\gamma = 0 \quad \text{on } \gamma, \tag{3.3k}$$
$$\partial^{\mathcal{I}}_{\boldsymbol{n}}\hat{p}_{\mathcal{I}} + \partial^{\mathcal{A}}_{\boldsymbol{n}}\hat{p}_{\mathcal{A}} = \hat{\sigma}_{\mathcal{I}} \quad \text{on } \gamma, \tag{3.3l}$$
$$\lambda\,(\bar{u}_{\mathcal{I}} - u_d) + \hat{p}_{\mathcal{I}} = 0 \quad \text{in } \mathcal{I}, \tag{3.3m}$$
$$\lambda\,(\bar{u}_{\mathcal{A}} - u_d) + \hat{p}_{\mathcal{A}} + \hat{q}_{\mathcal{A}} = 0 \quad \text{in } \mathring{\mathcal{A}}, \tag{3.3n}$$

$$\begin{cases} \hat{p}_{\mathcal{I}}|_\gamma - p^{\max}_{\min}|_\gamma = 0 & \text{on } \gamma, \quad (3.3p)\\ \bar{u}_{\mathcal{I}}|_\gamma - \bar{u}_{\mathcal{A}}|_\gamma = 0 & \text{on } \gamma, \quad (3.3q)\\ \hat{q}_{\mathcal{A}}|_\gamma = 0 & \text{on } \gamma, \quad (3.3r)\end{cases}$$

As in Corollary 3 the conditions (3.3p)–(3.3r) are different representatives of the necessary condition which is related to the variation of the active set.

The analysis of Paragraph 3.1.1 suggests, that it is unrewarding to use the first order optimality system in order to establish a steepest descent algorithm. Consequently, this thesis is focused on the presentation of Newton's method and some related approaches to solve (3.3). Hence, one is interested in derivatives of the system which are provided by the following paragraphs.

The strict inequality (3.3o) is an important part in this optimality system. It was argued in Paragraph 2.2.4, that it has no local impact on optimality. This point of view proved well-founded during the derivation of the optimality system in sections 2.3 and 2.4. This reasoning can also be used while deriving derivatives of the equation part of the optimality system. Nonetheless, the inequality must not be ignored, if an algorithm shall be designed for solving the original model problem (2.1) or the equivalent set optimal control problem (2.30), respectively.

3.3.1 Perspective from reduced/bilevel approach

The point of view of the reduced approach, which was pursued in Section 2.3 has already been used to illustrate descent and Newton algorithms in sections 3.1 and 3.2. It is based on the reformulation of the set optimal control problem as an equivalent bilevel optimization problem (2.36), (2.37), which introduces a hierarchical distinction between the set variable $\mathcal{B}$ and the function space variables $u_{\mathcal{J}}$, $u_{\mathcal{B}}$, $y_{\mathcal{J}}$ and $y_{\mathcal{B}}$.

In other words, this perspective focuses on the necessary condition of the outer optimization problem (2.36), which is given by a null of the covariant derivative of the reduced functional

$$\mathrm{D}\mathcal{F}(\mathcal{B})[V] = -\frac{1}{2\lambda}\int_{\beta}(\hat{p}_{\mathcal{J}} - p_{\min}^{\max})^2 V\cdot\boldsymbol{n}_{\mathcal{J}}.$$

In order to establish a Newton's method, these considerations and the ones of Section 3.2 give rise to derive the second covariant derivative of $\mathcal{F}$; cf. (2.85). This approach requires the evaluation of the different constituents of the first and second order covariant derivative. In particular, it requires $\hat{p}_{\mathcal{J}}|_{\beta}$, which can be accessed by solving the inner optimization problem. Due to the simple character of the original model problem, it could been shown (Theorem 6), that the inner optimization problem can equivalently be replaced by its first order necessary conditions (3.3a)–(3.3n).

This perspective emphasizes, that equation (3.3p) can be seen as the *whole* first order necessary condition of the bilevel optimization problem, and that (3.3a)–(3.3n) should rather be regarded as conditions to evaluate the reduced functional $\mathcal{F}$ than as necessary conditions. Consequently, this approach fits into the class of "first optimize, then discretize" black-box solvers, which was presented in the Introduction on page 1f. The inner optimization problem is handled as black box here and it is assumed that a corresponding solver is provided. Consequently, optimization takes place on the graph of the geometry-to-solution operator (cf. Figure 2.14) and approaching the optimum algorithmically means constructing/following some path on the graph.

It remains to analyze the facultative equations (3.3q) and (3.3r). As mentioned in the introducing text above Corollary 3, they are equivalent conditions for the vanishing of the first order derivative of the reduced functional $\mathcal{F}$. However, they are equivalent to (3.3p) at the optimum only. Hence, it is possible to construct different algorithms when trying to find nulls of[7]

$$\mathfrak{K}_u^u(\mathcal{B})[V] := \frac{1}{2}\int_{\beta}(\bar{u}_{\mathcal{J}} - \bar{u}_{\mathcal{B}})^2\, V\cdot\boldsymbol{n}_{\mathcal{J}}, \tag{3.4a}$$

$$\mathfrak{K}_q^q(\mathcal{B})[V] := \frac{1}{2}\int_{\beta}\hat{q}_{\mathcal{B}}^2\, V\cdot\boldsymbol{n}_{\mathcal{J}}. \tag{3.4b}$$

This idea basically means trying to find critical points of different functionals, which are all defined on the same graph of the geometry-to-solution operator, and which all have the same critical points. These ideas are special cases of the more general variational relaxation approach, which is introduced in the next Paragraph.

3.3.2 Perspective from free boundary problems: (variational) relaxation approaches

If one regards the optimality system (3.3) without keeping in mind how it has been derived – and therefore has no bias to treat the shape gradient equations (3.3p)–(3.3q) different than the others – one recog-

[7] The notation is explained in footnote 10 on page 102.

nizes typical properties of free boundary problems. The equation part of the system is overdetermined in general and solvable for very special sets $\mathcal{B} \in \mathcal{O}$ only. The strict inequality constraint then additional singles out the right active set $\mathcal{A}$. The equation part of (3.3) is denoted a *free boundary PDAE* here in order to distinguish it from typical (elliptic) free boundary problems, where one deals with (elliptic) boundary value problems equipped with an additional boundary condition for determination of the right domain.

An introduction to the theory of free boundary problems can be found in [107, 62]. A common strategy for solving free boundary problems – henceforth referred to as *relaxation approach* – consists in relaxing one boundary condition such that the remaining system is solvable for a given domain and then minimizing the residual in the relaxed equation; see, for instance, [53]. In other words, the free boundary problem is transformed into a shape/topology optimization problem, which is typically of least square type. A second idea of solving free boundary problems is based on total linearization of the system of equations and is discussed in Paragraph 3.3.3. Both approaches basically aim at solving a severely nonlinear equation[8].

The relaxation approach and total linearization are analog to the reduced approach and the Lagrange approach of Section 2.3 and Section 2.4, respectively. The reduced approach yields an algorithm which minimizes the residuum in one equation of the optimality system while (exactly) fulfilling the other equations (see Paragraph 3.3.1), whereas the Lagrange approach results in an algorithm that treats all necessary conditions as equal and simultaneously approximates the solution to all equations (see Paragraph 3.3.3).

When solving the free boundary PDAE by means of the relaxation approach the question arises which of the shape gradient equations (3.3p)–(3.3r) shall be used and which of the boundary conditions shall be relaxed. Moreover, it even seems to be reasonable to relax distributed algebraic equations, though this idea is not pursued within this thesis. Probably there is no universal recommendation, especially when problems and their necessary conditions become more complex than the simple model problem to be considered here.

The only interface condition, whose relaxation is unrewarding apriori, is the kink condition of the adjoint (3.3l), since its only value is the determination of $\hat{\sigma}_{\mathcal{J}}$. A second finding is, that after having decided which of the shape gradient equations (3.3p)–(3.3r) to use and which of the boundary conditions to relax, the remaining PDAE can be reduced.[9] Both in view of computational effort and in view of the experiences in solving DAEs (cf. Paragraph 2.7.1) it is recommended to reduce the PDAE as far as possible. By that means one can reduce the free boundary PDAE to a more common free boundary problem, due to the very simple structure of the model problem; in more complex situations (as for example problems with several states/controls, cf., e. g., [33, 145]) this would no longer be appropriate.

According to Eppler and Harbrecht [55, 54] it seems to be favorable to relax the Neumann boundary condition (3.3e) since relaxation of Dirichlet conditions (and tracking them in L^2) is ill-posed. Moreover, all other relaxation approaches suffer from difficulties with corresponding PDEs.

1. If the weak continuity condition (3.3d) is relaxed, the local shape derivatives of the reduced coupled PDE system contain the boundary condition
$$\partial_{\boldsymbol{n}}^{\mathcal{J}} y'[V] = -V \cdot \boldsymbol{n}_{\mathcal{J}}\, \partial_{\boldsymbol{nn}}(\bar{y}_{\mathcal{J}} - y_{\min}^{\max}) + \nabla_\beta(\bar{y}_{\mathcal{J}} - y_{\min}^{\max}) \cdot \nabla_\beta(V \cdot \boldsymbol{n}_{\mathcal{J}}), \quad \text{on } \beta$$
which is hard to get access to by means of standard finite element discretizations and which cannot be simplified as in (2.50), since the condition $(\bar{y}_{\mathcal{J}} - y_{\min}^{\max})|_\beta = 0$ is relaxed.
2. If the BDD condition (3.3g) is relaxed, it is not possible to reduce the remaining PDE system to the candidate inactive set $\mathcal{J}$, since $y_{\mathcal{B}}$ does not need to be equal to $y_{\min}^{\max}$ any longer. Consequently, the computational effort would increase significantly in comparison to other relaxation approaches.
3. All other possible choices of relaxation yield the boundary value problem (3.14), which is non-standard due to its asymmetric distribution of boundary conditions. Numerical practice shows, that the solution $\hat{p}_{\mathcal{J}}$ tends to have oscillations near the interface β if the finte-element mesh yields a zigzagging polygon representation of the interface; see Paragraph 4.1.2 on page 123. It would be interesting to analyze this behavior in order to understand whether these problems are due to theoretical reasons – the results of Conjecture 3 and Theorem 9 indicate that one has to expect

[8] Even if the boundary value problem is linear, the shape variable induces an intrinsic nonlinear behavior, since it is not located in a linear vector space; cf. Section 2.6.

[9] It is important to mind the ordering of the different steps. If one reduces the system after having chosen one of the shape gradient equations, but before having relaxed one condition, the first choice is irrelevant, since the shape gradient equations are equivalent then.

poor regularity – or due to the fact that an unsuitable finite element approximation was applied. However, a deeper investigation is beyond the scope of this thesis.

In order to illustrate the idea of the relaxation approach, the consequences of the relaxation of the Neumann boundary condition (3.3e) and of the shape gradient condition (3.3p) are analyzed in more detail. One ends up with the following reformulation no matter what representative of the shape gradient is chosen. One aims at the minimization of the L^2 cost functional (*merit functional*)[10]

$$K_p^{\partial_n}(\mathcal{B}) := \frac{1}{2}\|\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{J}}^* - y_{\min}^{\max})\|_{L^2(\beta)}^2 = \frac{1}{2}\int_\beta \left(\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{J}}^* - y_{\min}^{\max})\right)^2 \tag{3.5}$$

subject to the strict inequality constraint

$$y_{\min} < y_{\mathcal{J}}^* < y_{\max} \quad \text{in } \mathcal{J}$$

where $(y_{\mathcal{J}}^*, p_{\mathcal{J}}^*)$ fulfills

$$-\Delta y_{\mathcal{J}}^* + y_{\mathcal{J}}^* + \frac{1}{\lambda}p_{\mathcal{J}}^* = u_d \quad \text{in } \mathcal{J}, \tag{3.6a}$$
$$\partial_{\boldsymbol{n}} y_{\mathcal{J}}^* = 0 \quad \text{on } \Gamma, \tag{3.6b}$$
$$y_{\mathcal{J}}^*|_\beta = y_{\min}^{\max}|_\beta \quad \text{on } \beta, \tag{3.6c}$$
$$-\Delta p_{\mathcal{J}}^* + p_{\mathcal{J}}^* - y_{\mathcal{J}}^* = -y_d \quad \text{in } \mathcal{J}, \tag{3.6d}$$
$$\partial_{\boldsymbol{n}} p_{\mathcal{J}}^* = 0 \quad \text{on } \Gamma, \tag{3.6e}$$
$$p_{\mathcal{J}}^*|_\beta = p_{\min}^{\max}|_\beta \quad \text{on } \beta. \tag{3.6f}$$

Once having solved this shape optimization problem, the remaining variables of the original first order necessary conditions (3.3) can be obtained by assignments more or less

$$-\Delta y_{\min}^{\max} + y_{\min}^{\max} = \bar{u}_{\mathcal{B}} \quad \text{in } \mathring{\mathcal{B}}, \tag{3.7a}$$
$$-\Delta \bar{y}_{\mathcal{B}} + \bar{y}_{\mathcal{B}} = \bar{u}_{\mathcal{B}} \quad \text{in } \mathring{\mathcal{B}}, \tag{3.7b}$$
$$y_{\min}^{\max}|_\beta = \bar{y}_{\mathcal{B}}|_\beta \quad \text{on } \beta, \tag{3.7c}$$
$$-\Delta \hat{p}_{\mathcal{B}} + \hat{p}_{\mathcal{B}} - \bar{y}_{\mathcal{B}} = -y_d \quad \text{in } \mathring{\mathcal{B}}, \tag{3.7d}$$
$$\hat{p}_{\mathcal{B}}|_\beta = p_{\min}^{\max}|_\beta \quad \text{on } \beta, \tag{3.7e}$$
$$\bar{u}_{\mathcal{J}} = -\frac{1}{\lambda}p_{\mathcal{J}}^* + u_d \quad \text{in } \mathcal{J}, \tag{3.8a}$$
$$\hat{q}_{\mathcal{B}} = -\lambda(\bar{u}_{\mathcal{B}} - u_d) - \hat{p}_{\mathcal{B}} \quad \text{in } \mathring{\mathcal{B}}, \tag{3.8b}$$
$$\hat{\sigma}_{\mathcal{J}} = \partial_{\boldsymbol{n}}^{\mathcal{J}} p_{\mathcal{J}}^* + \partial_{\boldsymbol{n}}^{\mathcal{B}} \hat{p}_{\mathcal{B}} \quad \text{on } \beta. \tag{3.8c}$$

Hence, it is necessary to provide a constructive scheme to minimize the merit function only. Its critical points (i. e. sets) are characterized by nulls of its gradient. According to Paragraph 2.4.2 there holds

$$\begin{aligned} dK_p^{\partial_n}(\mathcal{B}; V) = {} & \frac{1}{2}\int_\beta \left(2\,\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{J}}^* - y_{\min}^{\max})\,\partial_{\boldsymbol{nn}}(y_{\mathcal{J}}^* - y_{\min}^{\max}) + \kappa_{\mathcal{J}}\left(\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{J}}^* - y_{\min}^{\max})\right)^2\right) V\cdot\boldsymbol{n}_{\mathcal{J}} \\ & + \frac{1}{2}\int_\beta 2\,\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{J}}^* - y_{\min}^{\max})\Big(\underbrace{\nabla(y_{\mathcal{J}}^* - y_{\min}^{\max})'[V]\cdot\boldsymbol{n}_{\mathcal{J}}}_{\partial_{\boldsymbol{n}}^{\mathcal{J}} y_{\mathcal{J}}^{*\prime}[V]} + \nabla(y_{\mathcal{J}}^* - y_{\min}^{\max})\cdot(\nabla b_{\mathcal{J}})'[V]|_\beta\Big) \end{aligned}$$

and with

$$\begin{aligned} \partial_{\boldsymbol{nn}}(y_{\mathcal{J}}^* - y_{\min}^{\max}) &= \Delta(y_{\mathcal{J}}^* - y_{\min}^{\max})|_\beta - \underbrace{\Delta_\beta(y_{\mathcal{J}}^* - y_{\min}^{\max})}_{=0} - \kappa_{\mathcal{J}}\,\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{J}}^* - y_{\min}^{\max}) \\ &= \left(y_{\mathcal{J}}^* + \frac{1}{\lambda}p_{\mathcal{J}}^* - u_d + \bar{u}_{\mathcal{A}} - y_{\min}^{\max}\right)|_\beta - \kappa_{\mathcal{J}}\,\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{J}}^* - y_{\min}^{\max}) \\ &= \left(y_{\mathcal{J}}^* + \frac{1}{\lambda}p_{\mathcal{J}}^* - u_d - \frac{1}{\lambda}p_{\min}^{\max} + u_d - y_{\min}^{\max}\right)|_\beta - \kappa_{\mathcal{J}}\,\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{J}}^* - y_{\min}^{\max}) \\ &= -\kappa_{\mathcal{J}}\,\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{J}}^* - y_{\min}^{\max}) \end{aligned}$$

and due to (2.67)

$$\begin{aligned} \nabla(y_{\mathcal{J}}^* - y_{\min}^{\max})\cdot(\nabla b_{\mathcal{J}})'[V]|_\beta &= \nabla(y_{\mathcal{J}}^* - y_{\min}^{\max})\cdot\Big(\left((\mathrm{D}V)\,\nabla b_{\mathcal{J}}\cdot\nabla b_{\mathcal{J}}\right)\nabla b_{\mathcal{J}} - (\mathrm{D}V)^\top\nabla b_{\mathcal{J}} - \mathrm{D}^2 b_{\mathcal{J}}\,V\Big)\Big|_\beta \\ &= \underbrace{\nabla(y_{\mathcal{J}}^* - y_{\min}^{\max})}_{\parallel \boldsymbol{n}_{\mathcal{J}},\ \text{cf. (3.6c)}}\cdot\Big(\underbrace{\left(\boldsymbol{n}_{\mathcal{J}}\cdot(\mathrm{D}V)^\top\boldsymbol{n}_{\mathcal{J}}\right)\boldsymbol{n}_{\mathcal{J}} - (\mathrm{D}V)^\top\boldsymbol{n}_{\mathcal{J}}}_{\perp \boldsymbol{n}_{\mathcal{J}}} - \mathrm{D}^2 b_{\mathcal{J}}|_\beta\,V\Big) \\ &= -\nabla(y_{\mathcal{J}}^* - y_{\min}^{\max})\,\mathrm{D}^2 b_{\mathcal{J}}|_\beta\,V = 0 \quad \text{since } \nabla b_{\mathcal{J}}\,\mathrm{D}^2 b = 0 \end{aligned}$$

[10] The notation of the functionals is according to the following scheme: the subscript (p, u or q) indicates which of the shape gradient equations (3.3p)–(3.3r) were used; the superscript tells which interface condition is relaxed.

one obtains

$$dK_p^{\partial_n}(\mathcal{B};V) = \int_\beta -\frac{1}{2}\kappa_{\mathcal{J}} \left(\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{J}}^* - y_{\min}^{\max})\right)^2 V\cdot\boldsymbol{n}_{\mathcal{J}} + \partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{J}}^* - y_{\min}^{\max})\,\partial_{\boldsymbol{n}}^{\mathcal{J}} y_{\mathcal{J}}^{*\prime}[V], \tag{3.9}$$

where the local shape derivatives fulfill

$$-\Delta y_{\mathcal{J}}^{*\prime}[V] + y_{\mathcal{J}}^{*\prime}[V] = -\frac{1}{\lambda}p_{\mathcal{J}}^{*\prime}[V] \text{ in } \mathcal{J}, \tag{3.10a}$$
$$\partial_{\boldsymbol{n}} y_{\mathcal{J}}^{*\prime}[V] = 0 \text{ on } \Gamma, \tag{3.10b}$$
$$y_{\mathcal{J}}^{*\prime}[V]|_\beta = -\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{J}}^* - y_{\min}^{\max})V\cdot\boldsymbol{n}_{\mathcal{J}} \text{ on } \beta, \tag{3.10c}$$
$$-\Delta p_{\mathcal{J}}^{*\prime}[V] + p_{\mathcal{J}}^{*\prime}[V] = y_{\mathcal{J}}^{*\prime}[V] \text{ in } \mathcal{J}, \tag{3.10d}$$
$$\partial_{\boldsymbol{n}} p_{\mathcal{J}}^{*\prime}[V] = 0 \text{ on } \Gamma, \tag{3.10e}$$
$$p_{\mathcal{J}}^{*\prime}[V]|_\beta = -\partial_{\boldsymbol{n}}^{\mathcal{J}}(p_{\mathcal{J}}^* - p_{\min}^{\max})V\cdot\boldsymbol{n}_{\mathcal{J}} \text{ on } \beta. \tag{3.10f}$$

In order to get rid of the local shape derivative in (3.9) one can introduce adjoint variables. By means of the heuristic of Remark on page 45 one derives an adjoint system

$$-\Delta Y_{\mathcal{J}}^* + Y_{\mathcal{J}}^* + \frac{1}{\lambda}P_{\mathcal{J}}^* = 0 \text{ in } \mathcal{J}, \tag{3.11a}$$
$$\partial_{\boldsymbol{n}} Y_{\mathcal{J}}^* = 0 \text{ on } \Gamma, \tag{3.11b}$$
$$Y_{\mathcal{J}}^*|_\beta = 0 \text{ on } \beta, \tag{3.11c}$$
$$-\Delta P_{\mathcal{J}}^* + P_{\mathcal{J}}^* - Y_{\mathcal{J}}^* = 0 \text{ in } \mathcal{J}, \tag{3.11d}$$
$$\partial_{\boldsymbol{n}} P_{\mathcal{J}}^* = 0 \text{ on } \Gamma, \tag{3.11e}$$
$$P_{\mathcal{J}}^*|_\beta = -\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{J}}^* - y_{\min}^{\max}) \text{ on } \beta. \tag{3.11f}$$

Using the adjoint variables one obtains the following representation of the shape semiderivative

$$dK_p^{\partial_n}(\mathcal{B};V) = \int_\beta \left(-\frac{1}{2}\kappa_{\mathcal{J}}\, P_{\mathcal{J}}^{*2} + P_{\mathcal{J}}^*\,\partial_{\boldsymbol{n}}^{\mathcal{J}} P_{\mathcal{J}}^* + \partial_{\boldsymbol{n}}^{\mathcal{J}}(p_{\mathcal{J}}^* - p_{\min}^{\max})\,\partial_{\boldsymbol{n}}^{\mathcal{J}} Y_{\mathcal{J}}^*\right) V\cdot\boldsymbol{n}_{\mathcal{J}}. \tag{3.12}$$

With this Hadamard form of the shape semiderivative of the merit functional at hand it is possible to construct a steepest descent algorithm for minimizing the cost functional (3.5). A Newton type method would require the second order semiderivative of the cost functional $K_p^{\partial_n}$ whose derivation goes beyond the scope of this thesis. In particular, there occur some difficulties due to differentiation of the curvature in (3.12). They are avoided in [54] by restricting the considerations to star shaped domains, which would be too strong an assumption in the present context.

As mentioned above there are good reasons to relax the Neumann boundary condition (3.3e). Nonetheless, it is worthwhile regarding the approach of relaxing the shape gradient condition (3.3p) and of minimizing

$$K_p^p(\mathcal{B}) := \frac{1}{2}\int_\beta (\hat{p}_{\mathcal{J}} - p_{\min}^{\max})^2, \tag{3.13}$$

subject to the usual inequality constraint, where $\hat{p}_{\mathcal{J}}$ is given by

$$-\Delta\bar{y}_{\mathcal{J}} + \bar{y}_{\mathcal{J}} + \frac{1}{\lambda}\hat{p}_{\mathcal{J}} = u_d \text{ in } \mathcal{J}, \tag{3.14a}$$
$$\partial_{\boldsymbol{n}}\bar{y}_{\mathcal{J}} = 0 \text{ on } \Gamma, \tag{3.14b}$$
$$\bar{y}_{\mathcal{J}}|_\beta = y_{\min}^{\max}|_\beta \text{ on } \beta, \tag{3.14c}$$
$$\partial_{\boldsymbol{n}}^{\mathcal{J}}\bar{y}_{\mathcal{J}} = \partial_{\boldsymbol{n}}^{\mathcal{J}} y_{\min}^{\max} \text{ on } \beta, \tag{3.14d}$$
$$-\Delta\hat{p}_{\mathcal{J}} + \hat{p}_{\mathcal{J}} - \bar{y}_{\mathcal{J}} = -y_d \text{ in } \mathcal{J}, \tag{3.14e}$$
$$\partial_{\boldsymbol{n}}\hat{p}_{\mathcal{J}} = 0 \text{ on } \Gamma, \tag{3.14f}$$

A comparison with the approach of Paragraph 3.3.1 reveals significant similarities, since (3.3a)–(3.3n) can be reduced to (3.14). In particular, one aims at minimizing the merit function K_p^p on the one hand, whereas one searches nulls of the very similar object $D\mathcal{F}(.)[.]$ on the other hand. Hence, a Newton scheme for minimizing the functional K_p^p is to be understood as a *third* order method with respect to minimization of $\mathcal{F}$. Applying this insight to the minimization of $K_p^{\partial_n}$ explains the difficulties when trying to derive its second order shape semiderivative: in fact, this is similar to derive a third order shape derivative of the functional $\mathcal{F}$.

All in all, many of the relaxation approaches for solving the free boundary PDAE (3.3) are very similar to the bilevel/reduced approach of Paragraph 3.3.1 from the analytical point of view. Both ideas are based on fulfilling all equations of the first order necessary conditions (3.3) but one, whose defect is to be brought to zero. From this perspective the relaxation approaches introduce other sections than the graph of the geometry-to-solution operator into the vector bundle E on the manifold $\mathcal{X}(\mathcal{A})$; cf. Paragraph 2.6.3. These sections may have beneficial properties as for instance guaranteeing higher regularity and therefore

may yield higher stability and efficiency of thereon based algorithms. However, the algorithmic access to relaxation approaches and to the bilevel approach are different. The classical idea of solving a free boundary problem by means of the relaxation approach, which is carried over to a free boundary PDAE here, uses minimization of quadratic merit functions, whereas the bilevel approach calls for solving a variational equation. It has been demonstrated in the case of $K_p^{\partial_n}$, that there are severe problems to establish a Newton's method for minimization of merit functions.

The proposed remedy is a new hybrid approach (henceforth denoted by *variational relaxation approach*). It benefits from the freedom of the relaxation approach to choose the equation which is relaxed and from the freedom to choose different norms for the merit function, while combining the efficiency of a Newton's method of the reduced/bilevel approach. The goal is to find solutions (these are sets $\mathcal{B} \in \mathcal{O}$) to variational equations of the following type (also cf. (3.4))

$$0 = \mathfrak{K}_p^p(\mathcal{B})[V] := \frac{1}{2}\int_\beta (\hat{p}_\mathcal{J} - p_{\min}^{\max})^2 V\cdot \boldsymbol{n}_\mathcal{J} = \mathrm{D}\mathcal{F}(\mathcal{B})[V], \quad \forall V \in \mathcal{V}, \tag{3.15a}$$

$$0 = \mathfrak{K}_u^u(\mathcal{B})[V] := \frac{1}{2}\int_\beta (\bar{u}_\mathcal{J} - \bar{u}_\mathcal{B})^2 V\cdot \boldsymbol{n}_\mathcal{J}, \qquad \forall V \in \mathcal{V}, \tag{3.15b}$$

$$0 = \mathfrak{K}_p^{\partial_n}(\mathcal{B})[V] := \frac{1}{2}\int_\beta \left(\partial_{\boldsymbol{n}}^\mathcal{J}(y^* - y_{\min}^{\max})\right)^2 V\cdot \boldsymbol{n}_\mathcal{J}, \qquad \forall V \in \mathcal{V}. \tag{3.15c}$$

Unfortunately, the indefiniteness of the second order shape semiderivative $d^2\mathcal{F}$ at the optimal configuration $\mathcal{A}$ (cf. Corollary 5 and (2.85)) reenters the considerations again, since any $d\mathfrak{K}_\cdot$ has a null at $\mathcal{A}$ as well. Hence, typical *quasi Newton schemes* based on positive definiteness preserving update rules like the *BFGS method* seem to be not suitable; however *symmetric rank-1 update methods (SR1)* should be applicable [131]. Another remedy is to solve the variational equations related to nonquadratic relaxation

$$0 = \tilde{\mathfrak{K}}_p^p(\mathcal{B})[V] := \int_\beta (\hat{p}_\mathcal{J} - p_{\min}^{\max}) V\cdot \boldsymbol{n}_\mathcal{J}, \qquad \forall V \in \mathcal{V}, \tag{3.16a}$$

$$0 = \tilde{\mathfrak{K}}_u^u(\mathcal{B})[V] := \int_\beta (\bar{u}_\mathcal{J} - \bar{u}_\mathcal{B}) V\cdot \boldsymbol{n}_\mathcal{J}, \qquad \forall V \in \mathcal{V}, \tag{3.16b}$$

$$0 = \tilde{\mathfrak{K}}_p^{\partial_n}(\mathcal{B})[V] := \int_\beta \partial_{\boldsymbol{n}}^\mathcal{J}(y^* - y_{\min}^{\max}) V\cdot \boldsymbol{n}_\mathcal{J}, \quad \forall V \in \mathcal{V}. \tag{3.16c}$$

This idea is pursued in the 8th item of the discussion on page 110.

After having commented on some selected ideas about how to solve the free boundary PDAE (3.3), the remaining part of this Paragraph is concerned with the perspective from the first order necessary conditions of Casas (Proposition 2) which are enhanced by Bergounioux and Kunisch (Proposition 3). Due to the simple structure of the model problem (2.1) it is possible to reduce that optimality system as well, such that the major effort has to be spend on solving (3.17):

$$-\Delta \bar{y}_\mathcal{I} + \bar{y}_\mathcal{I} + \frac{1}{\lambda} p_\mathcal{I}^{trad} = u_d \quad \text{in } \mathcal{I}, \tag{3.17a}$$

$$\partial_{\boldsymbol{n}} \bar{y}_\mathcal{I} = 0 \quad \text{on } \Gamma, \tag{3.17b}$$

$$\bar{y}_\mathcal{I}|_\gamma = y_{\min}^{\max}|_\gamma \quad \text{on } \gamma, \tag{3.17c}$$

$$\partial_{\boldsymbol{n}}^\mathcal{I} \bar{y}_\mathcal{I} = \partial_{\boldsymbol{n}}^\mathcal{I} y_{\min}^{\max} \quad \text{on } \gamma, \tag{3.17d}$$

$$-\Delta p_\mathcal{I}^{trad} + p_\mathcal{I}^{trad} - \bar{y}_\mathcal{I} = -y_d \quad \text{in } \mathcal{I}, \tag{3.17e}$$

$$\partial_{\boldsymbol{n}} p_\mathcal{I}^{trad} = 0 \quad \text{on } \Gamma, \tag{3.17f}$$

$$p_\mathcal{I}^{trad}|_\gamma = p_{\min}^{\max}|_\gamma \quad \text{on } \gamma, \tag{3.17g}$$

$$-\Delta y_{\min}^{\max} + y_{\min}^{\max} = \bar{u}_\mathcal{A} \quad \text{in } \mathring{\mathcal{A}}, \tag{3.18a}$$

$$y_{\min}^{\max} = \bar{y}_\mathcal{A} \quad \text{in } \mathring{\mathcal{A}}, \tag{3.18b}$$

$$p_\mathcal{A}^{trad} = p_{\min}^{\max} \quad \text{in } \mathring{\mathcal{A}}, \tag{3.18c}$$

$$\mu_{\mathring{\mathcal{A}}} = \mu_{\mathring{\mathcal{A}}}^{\max} - \mu_{\mathring{\mathcal{A}}}^{\min} \quad \text{in } \mathring{\mathcal{A}}, \tag{3.18d}$$

$$\bar{u}_\mathcal{I} = -\frac{1}{\lambda} p_\mathcal{I}^{trad} + u_d \quad \text{in } \mathcal{I}, \tag{3.19a}$$

$$\mu_\gamma = \partial_{\boldsymbol{n}}^\mathcal{I} p_\mathcal{I}^{trad} + \partial_{\boldsymbol{n}}^\mathcal{A} p_{\min}^{\max} \quad \text{on } \gamma, \tag{3.19b}$$

$$y_{\min} < \bar{y}_\mathcal{I} < y_{\max} \quad \text{in } \mathcal{I}. \tag{3.20a}$$

By means of Corollary 6 and Proposition 5 one obtains that $p_\mathcal{I}^{trad} = \hat{p}_\mathcal{I}$. Consequently the free boundary problem (3.17) and the one to be solved by (3.13)–(3.14) are the same. In addition, the relaxation approaches, which are presented above, applied to (3.17) are very similar to the idea of Hintermüller and Ring [90].

Hence, one has to pose the question, whether the perspectives presented in this thesis essentially differ from their approach. For one thing the Lagrange approach related ideas (see Paragraph 3.3.3), the reduced/bilevel approach of Paragraph 3.3.1 and the ideas of the variational relaxation approach are new and yield efficient numerics (see Chapter 4), and for another thing the Bryson-Denham-Dreyfus approach is expected to be beneficial. In order to see that, it is necessary to consider the context:

Up to now the considerations were based upon the following insight. By choosing one of the shape gradient conditions (3.3p)–(3.3q) and by relaxing one boundary condition, auxiliary shape optimization problems can be formulated. Evaluation of the corresponding merit functions requires to solve a remaining PDAE, which fortunately can be reduced to a more simple coupled boundary problem (cf. (3.6) and (3.14)). However, this reduction is due to the very simple structure of the model problem and cannot be expected to be possible in more complex situations. In the context of optimal control problems with several states and/or controls, one usually has to make the assumption, that within the active set only *one* state/control constraint is active. Hence, the active constraint does not provide sufficient information to determine *all* primal variables in the active set.[11] Consequently, it is neither possible to reduce the optimality system to a subproblem in the inactive set $\mathcal{J}$ nor to eliminate the algebraic conditions then.

These considerations indicate, that the significant reduction of the free boundary PDAE (3.3) is representative for a small class of optimal control problems only. Neglecting the strict inequality constraint one recognizes that the optimality system possesses properties both of free boundary problems and of PDAE (cf. Section 2.7) and thus one has to cope with the specific difficulties of both disciplines. In particular, it was shown in Proposition 6 that the BDD approach yields a double index reduction and therefore the corresponding optimality system may be solved more easily. Or, to put it in a nutshell, the real capability of the BDD approach cannot be illustrated by means of the very simple model problem, which was chosen here in order to keep theory as easy a possible.

Moreover, the ideas of the (variational) relaxation approach and the reduced/bilevel approach essentially benefit from the fact that the optimality system (3.3) cannot only be reduced at all, but that reduction yields *linear* systems like (3.6) and (3.14). If, for instance, the state equation of the original model problem were semilinear, the corresponding reduced free PDAE would contain this semilinear equation. Hence, assembling one of the cost functionals $K_\cdot$ or $\mathfrak{K}_\cdot$ and its derivative would require to solve a semilinear boundary value problem in each iteration of a solution algorithm. From the point of view of optimization in vector bundles (cf. Paragraph 2.6.3) this means that the computation of the next iterate, which has to lie in the graph the geometry-to-solution operator (or some similar object), is very expensive in this situation. Consequently, it might be more suitable to solve the equation inexactly or even to treat all variables as equal and use a total linearization method which does not bother whether the boundary value problem is linear or not; cf. Paragraph 3.3.3.

3.3.3 Perspective from Lagrange approach

As already indicated another possible approach to construct a Newton scheme which solves the equation part of optimality system (3.3) is the idea of *total linearization* (sometimes called *shape linearization*). This method was originally invented to solve free boundary problems; cf. [111, 105, 104]. In view of the analysis of Paragraph 3.3.2 there are two different starting points for a total linearization. For one thing it is possible to linearize the whole free boundary PDAE (this is the equation part of (3.3)) (*full total linearization approach*) and for another thing one can reduce the free PDAE as far as possible, hence obtains a free boundary problem and linearizes afterwards (*reduced total linearization approach*). Both ideas are addressed in the following. They can be classified as "first discretize, then optimize" all-at-once solvers, which were presented in the Introduction on page 1f.

The total linearization approach is based on the idea of treating function space variables and the shape variable as equal. This is similar to the Lagrange approach of Section 2.4, which does not induce a hierarchical distinction between the variables, as the reduced approach does. Hence, applying total linearization to the whole free boundary PDAE is the same as constructing a *Lagrange-Newton method* for the Lagrangian (2.63). In order to do so, one requires second order derivatives of the Lagrangian; cf. for instance [131]. As already discussed in Paragraph 2.4.2 the variables of the Lagrangian are independent

[11] Note that, in contrast, the state constraint of the simple model problem fixes the state $\bar{y}_{\mathcal{B}} = y_{\min}^{\max}$ and the corresponding BDD control law fixes the control $\bar{u}_{\mathcal{B}} = -\Delta y_{\min}^{\max} + y_{\min}^{\max}$ in the candidate active set $\mathcal{B}$ and consequently all primal variables are fully determined.

and hence one has to derive partial derivatives only. This fact considerably simplifies the computation, but nonetheless the produced expressions are longish. Since their exact wording is of minor interest here, they are derived in Appendix E.

The derivatives can be simplified by taking into account some properties of the optimal solution and substituting them into the formulae. In particular, one can get rid of terms which are very difficult if not impossible to access by means of finite element discretizations as for instance $\partial^{\mathcal{J}}_{\boldsymbol{nnn}}(y_{\mathcal{J}} - y^{\max}_{\min})$. A Newton scheme based upon such kind of simplified second order information proved to be efficient in the context of free boundary problems (cf. [104, ISL Algorithm, p. 61]) and is expected to perform comparably in the context of set optimal control problems. This approach is not followed up within this thesis though being worth it.

A major difference to the perspective of the reduced approach (Paragraph 3.3.1) and the relaxation approach (Paragraph 3.3.2) is the series of iterates produced by a Lagrange-Newton scheme. The iterates are not constrained to the graph of the geometry-to-solution operator and consequently the optimum can be approached from additional directions in the vector bundle E on the manifold $\mathcal{X}(\mathcal{A})$; cf. Paragraph 2.6.3.

The reduced total linearization approach, which starts from a fully reduced reformulation of the free boundary PDAE, is a hybrid of both perspectives. The reduced degrees of freedom, i. e. certain algebraic constraints, are always satisfied, whereas all remaining conditions are simultaneously relaxed. For convenience, this reduced total linearization method is illustrated in more detail here. The fully reduced reformulation of the free boundary PDAE is given by

$$
\begin{aligned}
-\Delta\bar{y}_{\mathcal{I}} + \bar{y}_{\mathcal{I}} + \frac{1}{\lambda}\hat{p}_{\mathcal{I}} &= u_d && \text{in } \mathcal{I}, && \text{(3.21a)}\\
\partial_{\boldsymbol{n}}\bar{y}_{\mathcal{I}} &= 0 && \text{on } \Gamma, && \text{(3.21b)}\\
\bar{y}_{\mathcal{I}}|_{\gamma} &= y^{\max}_{\min}|_{\gamma} && \text{on } \gamma, && \text{(3.21c)}\\
\partial^{\mathcal{I}}_{\boldsymbol{n}}\bar{y}_{\mathcal{I}} &= \partial^{\mathcal{I}}_{\boldsymbol{n}} y^{\max}_{\min} && \text{on } \gamma, && \text{(3.21d)}
\end{aligned}
\qquad
\begin{aligned}
-\Delta\hat{p}_{\mathcal{I}} + \hat{p}_{\mathcal{I}} - \bar{y}_{\mathcal{I}} &= -y_d && \text{in } \mathcal{I}, && \text{(3.21e)}\\
\partial_{\boldsymbol{n}}\hat{p}_{\mathcal{I}} &= 0 && \text{on } \Gamma, && \text{(3.21f)}\\
\hat{p}_{\mathcal{I}}|_{\gamma} &= p^{\max}_{\min}|_{\gamma} && \text{on } \gamma, && \text{(3.21g)}
\end{aligned}
$$

A variational formulation reads

$$
\begin{aligned}
\int_{\mathcal{I}} \nabla\bar{y}_{\mathcal{I}}\cdot\nabla\phi + (\bar{y}_{\mathcal{I}} + \frac{1}{\lambda}\hat{p}_{\mathcal{I}} - u_d)\,\phi - \int_{\gamma}\partial^{\mathcal{I}}_{\boldsymbol{n}} y^{\max}_{\min}\,\phi = 0, &\quad \forall\phi\in H^1(\mathcal{I}),\\
\int_{\mathcal{I}} \nabla\hat{p}_{\mathcal{I}}\cdot\nabla\varphi + (\hat{p}_{\mathcal{I}} - \bar{y}_{\mathcal{I}} + y_d)\,\varphi = 0, &\quad \forall\varphi\in H^1_{\gamma}(\mathcal{I}) := \{\varphi\in H^1(\mathcal{I}) \mid \varphi|_{\gamma} = 0\},\\
\int_{\gamma}(\bar{y}_{\mathcal{I}} - y^{\max}_{\min})\,\psi = 0, &\quad \forall\psi\in H^{1/2}(\gamma),\\
\int_{\gamma}(\hat{p}_{\mathcal{I}} - p^{\max}_{\min})\,\psi = 0, &\quad \forall\psi\in H^{1/2}(\gamma).
\end{aligned}
$$

Hence, one has to solve the following Newton equation in the variables V, δ_y and δ_p in each iteration of a total linearization method, where the current iterate is given by $(\mathcal{J}, y, p)$

$$
\begin{aligned}
&\int_{\mathcal{J}} \nabla\delta_y\cdot\nabla\phi + (\delta_y + \frac{1}{\lambda}\delta_p)\,\phi\\
&\qquad + \int_{\beta}\Big(\nabla y\cdot\nabla\phi + (y + \frac{1}{\lambda}p - u_d)\,\phi - \partial_{\boldsymbol{nn}} y^{\max}_{\min}\,\phi - \partial^{\mathcal{J}}_{\boldsymbol{n}} y^{\max}_{\min}\,\partial^{\mathcal{J}}_{\boldsymbol{n}}\phi - \kappa_{\mathcal{J}}\,\partial^{\mathcal{J}}_{\boldsymbol{n}} y^{\max}_{\min}\,\phi\Big)V\cdot\boldsymbol{n}_{\mathcal{J}}\\
&\qquad\qquad = -\Big(\int_{\mathcal{J}}\nabla y\cdot\nabla\phi + (y + p - u_d)\,\phi - \int_{\beta}\partial^{\mathcal{J}}_{\boldsymbol{n}} y^{\max}_{\min}\,\phi\Big),\\
&\int_{\mathcal{J}}\nabla\delta_p\cdot\nabla\varphi + (\delta_p - \delta_y)\,\varphi + \int_{\beta}\Big(\underbrace{\nabla p\cdot\nabla\varphi}_{\partial^{\mathcal{J}}_{\boldsymbol{n}} p\,\partial^{\mathcal{J}}_{\boldsymbol{n}}\varphi} + (p - y + y_d)\,\underbrace{\varphi}_{=0}\Big)V\cdot\boldsymbol{n}_{\mathcal{J}}\\
&\qquad\qquad = -\int_{\mathcal{J}}\nabla p\cdot\nabla\varphi + (p - y + y_d)\,\varphi, \qquad \text{(3.22)}\\
&\int_{\beta}\delta_y\,\psi + \Big(\partial^{\mathcal{J}}_{\boldsymbol{n}}(y - y^{\max}_{\min}) + \kappa_{\mathcal{J}}\,(y - y^{\max}_{\min})\Big)\,\psi\,V\cdot\boldsymbol{n}_{\mathcal{J}} = -\int_{\beta}(y - y^{\max}_{\min})\,\psi),\\
&\int_{\beta}\delta_p\,\psi + \Big(\partial^{\mathcal{J}}_{\boldsymbol{n}}(p - p^{\max}_{\min}) + \kappa_{\mathcal{J}}\,(p - p^{\max}_{\min})\Big)\,\psi\,V\cdot\boldsymbol{n}_{\mathcal{J}} = -\int_{\beta}(p - p^{\max}_{\min})\,\psi.
\end{aligned}
$$

The left hand side of this system can be simplified when using relations that are satisfied at the optimal configuration $(\mathcal{I}, \bar{y}_{\mathcal{I}}, \hat{p}_{\mathcal{I}})$. In particular, these are $y|_\beta = y_{\min}^{\max}|_\beta$, $\partial_{\boldsymbol{n}}^{\mathcal{J}} y = \partial_{\boldsymbol{n}}^{\mathcal{J}} y_{\min}^{\max}$ and $p|_\beta = p_{\min}^{\max}|_\beta$, and one obtains

$$\begin{aligned}
\int_{\mathcal{J}} \nabla\delta_y \cdot \nabla\phi + (\delta_y + \frac{1}{\lambda}\delta_p)\,\phi \\
+ \int_\beta \Big(\nabla_\beta y_{\min}^{\max} \cdot \nabla_\beta \phi + (y_{\min}^{\max} + \frac{1}{\lambda} p_{\min}^{\max} - u_d)\,\phi - \partial_{nn} y_{\min}^{\max}\,\phi - \kappa_{\mathcal{J}}\,\partial_{\boldsymbol{n}}^{\mathcal{J}} y_{\min}^{\max}\,\phi \Big) V\cdot\boldsymbol{n}_{\mathcal{J}} \\
= -\Big(\int_{\mathcal{J}} \nabla y \cdot \nabla\phi + (y + p - u_d)\,\phi - \int_\beta \partial_{\boldsymbol{n}}^{\mathcal{J}} y_{\min}^{\max}\,\phi \Big), \\
\int_{\mathcal{J}} \nabla\delta_p \cdot \nabla\varphi + (\delta_p - \delta_y)\,\varphi + \int_\beta \partial_{\boldsymbol{n}}^{\mathcal{J}} p\,\partial_{\boldsymbol{n}}^{\mathcal{J}} \varphi\, V\cdot\boldsymbol{n}_{\mathcal{J}} = -\int_{\mathcal{J}} \nabla p\cdot\nabla\varphi + (p - y + y_d)\,\varphi, \\
\int_\beta \delta_y\,\psi = -\int_\beta (y - y_{\min}^{\max})\,\psi, \\
\int_\beta \Big(\delta_p + \partial_{\boldsymbol{n}}^{\mathcal{J}} (p - p_{\min}^{\max})\, V\cdot\boldsymbol{n}_{\mathcal{J}} \Big)\,\psi = -\int_\beta (p - p_{\min}^{\max})\,\psi).
\end{aligned}$$

Furthermore, due to integration by parts on the interface β [151, Pro. 2.67 and 2.68], there holds

$$\int_\beta \nabla_\beta y_{\min}^{\max} \cdot \nabla_\beta \phi - \partial_{nn} y_{\min}^{\max}\,\phi - \kappa_{\mathcal{J}}\,\partial_{\boldsymbol{n}}^{\mathcal{J}} y_{\min}^{\max}\,\phi = \int_\beta -\Delta y_{\min}^{\max}\,\phi.$$

If this relation is used formally (this means ignoring the $V\cdot\boldsymbol{n}_{\mathcal{J}}$-factor above)[12] and if one applies

$$0 = -\Delta y_{\min}^{\max} + y_{\min}^{\max} - \bar{u}_{\mathcal{A}} = -\Delta y_{\min}^{\max} + y_{\min}^{\max} + \frac{1}{\lambda} p_{\min}^{\max} - u_d, \quad \text{in } \mathcal{A}$$

one ends up with

$$\begin{aligned}
\int_{\mathcal{J}} \nabla\delta_y \cdot \nabla\phi + (\delta_y + \frac{1}{\lambda}\delta_p)\,\phi = -\Big(\int_{\mathcal{J}} \nabla y \cdot \nabla\phi + (y + p - u_d)\,\phi - \int_\beta \partial_{\boldsymbol{n}}^{\mathcal{J}} y_{\min}^{\max}\,\phi \Big), \\
\int_{\mathcal{J}} \nabla\delta_p \cdot \nabla\varphi + (\delta_p - \delta_y)\,\varphi + \int_\beta \partial_{\boldsymbol{n}}^{\mathcal{J}} p\,\partial_{\boldsymbol{n}}^{\mathcal{J}} \varphi\, V\cdot\boldsymbol{n}_{\mathcal{J}} = -\int_{\mathcal{J}} \nabla p\cdot\nabla\varphi + (p - y + y_d)\,\varphi), \\
\int_\beta \delta_y\,\psi = -\int_\beta (y - y_{\min}^{\max})\,\psi, \\
\int_\beta \Big(\delta_p + \partial_{\boldsymbol{n}}^{\mathcal{J}} (p - p_{\min}^{\max})\, V\cdot\boldsymbol{n}_{\mathcal{J}} \Big)\,\psi = -\int_\beta (p - p_{\min}^{\max})\,\psi.
\end{aligned} \tag{3.23}$$

Hereby, a system is found, which seems to be accessible by means of standard finite elements.

3.4 Algorithms for set optimal control problems

The in-depth discussion of the first order optimality conditions of the set optimal control problem (2.30) in Section 3.3 contains different ideas of algorithms. Roughly speaking they can be divided in three groups

- reduced Newton methods,
- trial methods and
- total linearization methods.

Each of them is addressed in the subsequent paragraphs. Of course there are many more suitable algorithms and in particular more sophisticated ones, but the goal of the algorithmic and numerical analysis within this thesis is only to demonstrate, that it is possible to construct efficient numerics for solving state constrained optimal control problems based upon the ideas of BDD approach and shape/topology optimization. In particular, important questions like convergence analysis and error analysis remain unattained and globalization is touched on within the discussion of the a posteriori step on page 111f. only.

[12] The result can be achieved rigorously when homogenizing the free boundary problem (3.21) before shape linearizing it.

All in all, advantages and drawbacks of the algorithmic approaches are:

- The algorithms can be formulated without discretization in contrast to the primal-dual active set strategy, where the measure nature of the multiplier μ inhibits a formulation in function space.
- The algorithms do not contain regularization parameters and hence additional regularization loops are not necessary.
- Since the algorithms apply elements of shape calculus only, one cannot expect that they are capable to detect the right topology of the active set and therefore they cannot be globally convergent. Despite this, numerical tests show that the algorithms are able to handle certain changes of the topology of the active set during the iteration.
- The algorithms are essentially based on Assumption 1 of the active set. Thus, they cannot be applied to problems, e. g., where (parts of) the active set consists of sets of zero measure.

3.4.1 Reduced Newton methods

The group of reduced Newton methods shall collect schemes, which are based upon a hierarchy of variables and hence it is related to the reduced approach of Section 2.3, i. e. black-box solvers. In particular, these algorithms form the ideas from paragraphs 3.3.1 and 3.3.2. At this, the Newton's method, which is designed to solve the bilevel optimization problem, and the variational relaxation approaches emphasize different perspectives. The first approach emphasizes the bilevel optimization structure and thus does not bother how the inner optimization problem is solved and does not depend on an equivalent characterization of the set parametrized optima of the inner optimization problem by means of an equation system. In contrast, the latter approaches emphasize the free boundary PDAE character of the optimality system, which yields more flexibility to choose an equation to be relaxed.

Algorithm 1 (Newton scheme for bilevel optimization problem)**:**
Let $\mathcal{F}$ be the objective of the bilevel optimization problem (2.37), (2.36).

1. Set $i := 1$ and choose an initial guess $\mathcal{B}_i \in \mathcal{O}$.
2. Loop on i
   a) Solve the inner optimization problem (2.37) for the set B_i and extract $\hat{p}_{\mathcal{J}_i}|_{\beta_i} = -\lambda\,(\bar{u}_{\mathcal{J}_i} - u_d)|_{\beta_i}$ in order to be able to assemble $\nabla\mathcal{F}(\mathcal{B}_i)$ and $\nabla^2\mathcal{F}(\mathcal{B}_i)$; cf. (2.55) and (2.85).
   b) Solve the (variational) Newton equation
   $$\nabla^2\mathcal{F}(\mathcal{B}_i)[W_i, V] = -\langle \nabla\mathcal{F}(\mathcal{B}_i)\,,\, V\rangle \quad \forall V \in \mathcal{V} \tag{3.24}$$
   in the variable $W_i \in \mathcal{V}$.
   c) Perform the Newton update
   $$\mathcal{B}_{i+1} := R(W_i)$$
   where $R : T\mathcal{H}(\Omega) \to \mathcal{H}(\Omega)$ is a retraction.
   d) Stop the loop, if the update velocity field W_i is small enough; otherwise set $i := i + 1$.
   End of loop
3. Check if the strict inequality constraint $y_{\min} < \bar{y}_{\mathcal{J}_{i+1}} < y_{\max}$ and the sign conditions of the multipliers (cf. Corollary 6) are fulfilled. Stop if the check is passed; otherwise make another initial guess $\mathcal{B}_i \in \mathcal{O}$ and start all over again.

Algorithm 2 (Newton scheme for variational relaxation approaches)**:**
Choose one of the shape gradient conditions (3.3p)–(3.3r) to use and one of the boundary conditions of the corresponding free boundary PDAE (3.3) to be relaxed. In addition, choose a functional $\mathfrak{K}$ in the style of (3.15) or (3.16). Reduce the remaining part of the system and provide the shape semiderivative $D\mathfrak{K}(.)[W,.]$ (which either requires a shape adjoint system in case of a Hadamard derivative or a BVP for the local shape derivatives).

1. Set $i := 1$ and choose an initial guess $\mathcal{B}_i \in \mathcal{O}$.
2. Loop on i

a) Solve the remaining boundary value problem (and if available the corresponding adjoint system) for the set $\mathcal{B}_i$ in order to assemble $\mathfrak{K}(\mathcal{B}_i)[.]$ (and $\mathrm{D}\mathfrak{K}(\mathcal{B}_i)[.,.]$).

b) Solve the variational Newton equation

$$\mathrm{D}\mathfrak{K}(\mathcal{B}_i)[W_i, V] = -\mathfrak{K}(\mathcal{B}_i)[V], \quad \forall V \in \mathcal{V} \tag{3.25}$$

in the variable $W_i \in \mathcal{V}$.

c) Perform the Newton update

$$\mathcal{B}_{i+1} := R(W_i)$$

where $R : T\mathcal{H}(\Omega) \to \mathcal{H}(\Omega)$ is a retraction.

d) Stop the loop, if the update velocity field W_i is small enough; otherwise set $i := i + 1$.

End of loop

3. Solve the remaining part of the first order optimality system and check if the strict inequality constraint $y_{\min} < \bar{y}_{\mathcal{J}_{i+1}} < y_{\max}$ and the sign conditions of the multipliers (cf. Corollary 6) are fulfilled. Stop if the check is passed; otherwise make another initial guess $\mathcal{B}_i \in \mathcal{O}$ and start all over again.

Remarks on the initialization of the algorithms:

1. In order to obtain a reasonable initial guess $\mathcal{B}_1$ it may be useful to solve the state-unconstrained optimal control problem and to mark those subregions as active, where the state violates the constraints. In particular, one can stop the algorithm if the optimal solution of the unconstrained problem is already feasible for the constrained version.
2. Alternatively, one can start with a *candidate set* $\mathcal{C}$, which is obtained by means of the formula of $\mu_{\hat{\mathcal{A}}}$, cf. (2.4c) and (2.4d):

$$\mathcal{C}_{\max} := \left\{ x \in \Omega \,\middle|\, c_{\max} = \lambda(-\Delta^2 y_{\max} + 2\Delta y_{\max} - \Delta u_d + u_d) - y_{\max} + y_d \geq 0 \right\}, \tag{3.26a}$$

$$\mathcal{C}_{\min} := \left\{ x \in \Omega \,\middle|\, c_{\min} = \lambda(\quad \Delta^2 y_{\min} - 2\Delta y_{\min} + \Delta u_d - u_d) + y_{\min} - y_d \geq 0 \right\}. \tag{3.26b}$$

These sets contain the (optimal) active sets $\mathcal{A}_{\max}$ and $\mathcal{A}_{\min}$, respectively (at least for strictly complementary problems), since they are the maximal subsets of Ω where $\mu_{\hat{\mathcal{A}}}^{\max}$ and $\mu_{\hat{\mathcal{A}}}^{\min}$ can fulfill their sign conditions.

Remarks on the preparing step 2a:

3. Solving the inner optimization problem in step 2a is equivalent to solve its first order necessary conditions in the present context, since the optimization problem is strictly convex.
4. It is not necessary to solve the whole inner optimality system (3.3a)–(3.3n) in step 2a. It suffices to know $\hat{p}_{\mathcal{J}_i}$ ($= \bar{p}_{\mathcal{J}_i}$, cf. Corollary 6) in order to assemble the shape semiderivatives. This requires to solve the reduced inner optimality system (3.21) only (cf. Lemma 7 and Paragraph 3.3.2). Doing so, Algorithm 1 and Algorithm 2 are more or less equal.
5. If the model problem were more complicated – for instance equipped with a semilinear state equation – sufficiency and linearity of the first order conditions would be lost. Nonetheless, it is possible to use them, but it may be more efficient to solve them inexactly then. Moreover, it may be suitable to use the (possibly inexact) solution of the optimality system from the previous iteration as initial guess. However, it is necessary to transport it to the current geometry; also cf. step 2c of Algorithm 5 and the corresponding remarks.

Remarks on solving the Newton equation in step 2b:

6. The Newton equation (3.24) does not contain the second order shape semiderivative $d^2\mathcal{F}(\mathcal{B}_i; W_i, V)$ as one might expect at first glance; cf. Section 3.2 and the detailed analysis of Newton's method in [1, Chp. 6]. In particular, the Newton algorithm is proposed with the Hessian operator there instead of the second covariant derivative. A comparison with (2.85) reveals however, that there is no direct access to the shape Hessian nor to the second covariant derivative $\nabla^2\mathcal{F}(\mathcal{B}_i)$ due to the $p'_{\mathcal{J}}[.]$-term in it. Consequently, solving the Newton equation (3.24) actually means solving the equation simultaneously to the coupled system (2.48) which determines the local shape derivatives. The variational Newton equation (3.24) reads in detail

$$
\begin{aligned}
-\Delta y'_{\mathcal{J}_i}[W_i] + y'_{\mathcal{J}_i}[W_i] &= -\frac{1}{\lambda} p'_{\mathcal{J}_i}[W_i] && \text{in } \mathcal{J}_i, \qquad -\Delta p'_{\mathcal{J}_i}[W_i] + p'_{\mathcal{J}_i}[W_i] = y'_{\mathcal{J}_i}[W_i] \text{ in } \mathcal{J}_i,\\
\partial_{\boldsymbol{n}} y'_{\mathcal{J}_i}[W_i] &= 0 && \text{on } \Gamma, \qquad \partial_{\boldsymbol{n}} p'_{\mathcal{J}_i}[W_i] = 0 \quad \text{on } \Gamma,\\
y'_{\mathcal{J}_i}[W_i] &= 0 && \text{on } \beta_i,\\
\partial^{\mathcal{J}}_{\boldsymbol{n}} y'_{\mathcal{J}_i}[W_i] &= W_i \cdot \boldsymbol{n}_{\mathcal{J}_i} \frac{1}{\lambda} (p^{\max}_{\min}|_{\beta_i} - \hat{p}_{\mathcal{J}_i}|_{\beta_i}) && \text{on } \beta_i,
\end{aligned}
$$

$$
\begin{aligned}
\int_{\beta_i} \Big(2\, p'_{\mathcal{J}_i}[W_i] + \Big(2\, \partial^{\mathcal{J}}_{\boldsymbol{n}} (\hat{p}_{\mathcal{J}_i} - p^{\max}_{\min}) + \kappa_{\mathcal{J}_i} (\hat{p}_{\mathcal{J}_i} - p^{\max}_{\min}) \Big) W_i \cdot \boldsymbol{n}_{\mathcal{J}_i} \Big) (\hat{p}_{\mathcal{J}_i} - p^{\max}_{\min}) V \cdot \boldsymbol{n}_{\mathcal{J}_i} \\
= - \int_{\beta_i} (\hat{p}_{\mathcal{J}_i} - p^{\max}_{\min})^2 V \cdot \boldsymbol{n}_{\mathcal{J}_i}, \quad \forall V \in \mathcal{V}. \quad (3.27)
\end{aligned}
$$

7. This reasoning can be carried over to the Newton equation of Algorithm 2. The left hand side $D\mathfrak{K}$ is to be understood similar to a second covariant derivative. In particular, the parts of the shape semiderivative, which are due to differentiation of the variational vector field V are omitted, as it was done in order to obtain (2.85). For instance in the case of $\mathfrak{K}^p_p$ the Newton equation is the same as in Algorithm 1 and hence one obtains (3.27). Moreover, using $\mathfrak{K}^{\partial_n}_p$ yields (3.10) and

$$
\begin{aligned}
\int_{\beta_i} \Big(\partial^{\mathcal{J}}_{\boldsymbol{n}} y_i^{*\prime}[W_i] - \frac{1}{2} \kappa_{\mathcal{J}_i} \partial^{\mathcal{J}}_{\boldsymbol{n}} (y_i^* - y^{\max}_{\min})\, W_i \cdot \boldsymbol{n}_{\mathcal{J}_i} \Big) \partial^{\mathcal{J}}_{\boldsymbol{n}} (y_i^* - y^{\max}_{\min})\, V \cdot \boldsymbol{n}_{\mathcal{J}_i} \\
= -\frac{1}{2} \int_{\beta_i} \big(\partial^{\mathcal{J}}_{\boldsymbol{n}} (y_i^* - y^{\max}_{\min}) \big)^2 V \cdot \boldsymbol{n}_{\mathcal{J}_i}.
\end{aligned}
$$

   The situation is reversed in Algorithm 2, if a Hadamard form of the derivative has been computed by means of shape adjoint variables; cf. for instance (3.11). In that case, assembling the Newton equation (3.25) requires solving the shape adjoint system in advance and thus saves some computational effort in direct comparison to an approach, where the system of local shape derivatives has to be solved simultaneously to the Newton equation.

8. The proposed Newton Algorithm 1 is interfered with the not definite second covariant derivative of the reduced objective $\mathcal{F}$ (cf. Paragraph 2.5.1), since at the optimum both the left and the right hand side of the Newton equation are equal to zero. As long as the equation can be solved without discretization, there should arise no problems in determining the update $W_i \cdot \boldsymbol{n}_{\mathcal{J}_i}$, since the convergence rate of the gradient to zero is higher, than those of the second covariant derivative.[13] But this fact is not necessarily reflected on the finite dimensional level, if the (finite element) approximation is not chosen carefully enough. Moreover, one is confronted with a reduced convergence speed; see Paragraph 4.2.4. Nonetheless, there are two nearby workarounds, which are already effective on the continuous level.

   The first is to informally cancel out one $(\hat{p}_{\mathcal{J}_i} - p_{\max})$-factor in the Newton equation (3.24) and respectively in (3.27). This yields

$$
\begin{aligned}
\int_{\beta_i} \Big(2\, p'_{\mathcal{J}_i}[W_i] + \Big(2\, \partial^{\mathcal{J}}_{\boldsymbol{n}} (\hat{p}_{\mathcal{J}_i} - p^{\max}_{\min}) + \kappa_{\mathcal{J}_i} (\hat{p}_{\mathcal{J}_i} - p^{\max}_{\min}) \Big) W_i \cdot \boldsymbol{n}_{\mathcal{J}_i} \Big) V \cdot \boldsymbol{n}_{\mathcal{J}_i} \\
= - \int_{\beta_i} (\hat{p}_{\mathcal{J}_i} - p^{\max}_{\min}) V \cdot \boldsymbol{n}_{\mathcal{J}_i}, \quad \forall V \in \mathcal{V}. \quad (3.28)
\end{aligned}
$$

   A second idea is to compute the shape semiderivative of

$$
-\frac{1}{2\lambda} \int_{\beta} (\hat{p}_{\mathcal{J}} - p_{\max}) V \cdot \boldsymbol{n}_{\mathcal{J}}
$$

   instead of $d\mathcal{F}(\mathcal{B}; V)$ and afterwards extract the second covariant derivative. This reasoning results in

$$
\begin{aligned}
\int_{\beta_i} \Big(p'_{\mathcal{J}_i}[W_i] + \Big(\partial^{\mathcal{J}}_{\boldsymbol{n}} (\hat{p}_{\mathcal{J}_i} - p^{\max}_{\min}) + \kappa_{\mathcal{J}_i} (\hat{p}_{\mathcal{J}_i} - p^{\max}_{\min}) \Big) W_i \cdot \boldsymbol{n}_{\mathcal{J}_i} \Big) V \cdot \boldsymbol{n}_{\mathcal{J}_i} \\
= - \int_{\beta_i} (\hat{p}_{\mathcal{J}_i} - p^{\max}_{\min}) V \cdot \boldsymbol{n}_{\mathcal{J}_i}, \quad \forall V \in \mathcal{V}. \quad (3.29)
\end{aligned}
$$

   One immediately recognizes the variational relaxation approach due to (3.16a) here.

[13] This is similar to when applying a Newton's method to minimize the function $x \mapsto x^4$. The Newton equation then reads $12\, x_i^2\, \delta_i = -4 x_i^3$.

Both ideas aim at coping with the difficulties that arise out of the square in the first order shape semiderivative, namely that the second covariant derivative vanishes at the optimum. Since $p'_{\mathcal{J}}[W]$ and $\hat{p}_{\mathcal{J}} - p_{\min}^{\max}$ vanish there, the workarounds are expected to work unless the crucial Neumann derivative term in the left hand side of (3.28), and respectively of (3.29), is not zero there, too. This condition can be traced to a condition to μ_γ

$$\begin{aligned} & \partial_{\boldsymbol{n}}^{\mathcal{I}}(\hat{p}_{\mathcal{I}} - p_{\min}^{\max}) = 0 \\ \overset{\text{(B.9a)}}{\Longleftrightarrow} \quad & \partial_{\boldsymbol{n}}^{\mathcal{I}}(\bar{p}_{\mathcal{I}} - p_{\min}^{\max}) = 0 \\ \overset{\substack{\text{(2.60a), (2.46)} \\ \text{(2.4a), (2.4b)}}}{\Longleftrightarrow} \quad & \partial_{\boldsymbol{n}}^{\mathcal{I}}(p_{\mathcal{I}}^{trad} - p_{\mathcal{A}}^{trad}) = 0 \\ \overset{\substack{\text{(2.4j)} \\ \text{(2.4i)}}}{\Longleftrightarrow} \quad & \mu_\gamma = 0. \end{aligned}$$

Hence paradoxically, one has to expect convergence problems of Algorithm 1 and of the version of Algorithm 2 which is based on $\mathfrak{K}_p^p$, if the original model problem possesses a *regular* Lagrange multiplier μ. By means of a special trial algorithm this behavior can be confronted; cf. the 3rd item on page 113.

In view of the detailed analysis of different ideas of path following in Section 3.1, it is indicated to comment on the Newton update step 2c.

9. Solving the Newton equation in step 2b yields an increment of the set $\mathcal{B}_i$ in terms of a velocity field W_i. However, one has little information about this vector field. In particular, one gets the normal component on the current interface β_i only. This results fits in the abstract perspective on shape calculus of Section 2.6; cf. the 14th item of the discussion on page 77 in particular.
10. It is very common in the fields of shape optimization and free boundary problems to apply schemes for tracking the interface, which is driven by some velocity field W. The time dependent evolution of the interface is of particular interest in many practical applications like the simulation of free surfaces in fluid dynamics or of phase boundaries in the context of melting and solidification processes (*Stefan problem*) and phase separation (*Cahn-Hilliard equation*); see for instance [50]. However, tracking/evolution of the interface is not an issue in the present context, since the aim is to get the optimal active set no matter how the intermediate iterates look like. Consequently, there is a bias for a retraction based Newton update, for its lower computational cost in comparison with level set or fast marching methods. However, it is beyond the scope of this thesis to develop efficient retractions on the continuous level and the developed ideas remain restricted to the finite element discretization; cf. Paragraph 4.1.2.

It is a matter of fact, that Newton's method is only locally convergent and that one has to expect more than a unique critical shape of the reduce shape functional $\mathcal{F}$ and respectively of the functional $\mathfrak{K}$, though the original model problem (2.1) has a unique optimal solution. Consequently, the a posteriori step 3 is mandatory.

11. In view of the possibility to solve the reduced inner optimality system (3.21) instead of the whole inner optimality system (3.3) (cf. the 4th item above) the remaining parts (3.7)–(3.8) are to be evaluated only, if the stop criterion of step 2d is fulfilled and the a posteriori criteria are to be checked in step 3.
12. Numerical practice indicates that it is sufficient to solve the equation part of the first order necessary conditions (3.3) and to check afterwards if the inequality constraints are fulfilled (see Paragraph 4.2.3). This finding should be regarded in the context of the attempt to prove, that the critical points of the reduced objective are isolated (Paragraph 2.5.2) and that the neglected strict inequality constraint (3.3o) has a global impact on optimality (Paragraph 2.2.4) only.

    This global impact can be done justice to, when augmenting the corresponding functional as discussed in Paragraph 3.1.1. However, it seems to be inefficient to augment the reduced shape functional $\mathcal{F}$ with a penalty term of type (3.2), since this term induces additional terms of the sum within the first order shape derivative, which cause additional inhomogeneities in the shape adjoint system (2.53).[14] Hence, the first order shape semiderivative cannot be formulated without shape adjoints

[14] In particular, there occurs $c\,\big(\max\{0, \bar{y}_{\mathcal{J}} - y_{\max}\} - \max\{0, y_{\min} - \bar{y}_{\mathcal{J}}\}\big)$ on the left hand side of (2.53e).

any longer (cf. Theorem 7) and local shape derivatives of the shape adjoint variables are required in order to derive the second order shape semiderivative of the augmented shape functional. All in all, the Newton equation would call for solving the systems of local shape derivatives $(y'_{\mathcal{J}}, p'_{\mathcal{J}})$ and of local shape derivatives of the shape adjoint variables simultaneously, which is expected to be too expensive.

Hence, it is indicated to use augmentation in the context of a steepest descent method for globalization and to switch over to the Newton's method when the iterates have come close enough to the optimum.

13. If the algorithm restarts for an a posteriori criterion being not fulfilled, it is typically reasonable to choose the subset where the current state violates or meets the state constraint as guess for the active set. This reasoning is comparable to the primal part of the update of a primal-dual active set strategy (see Section 3.5) and enables topology changes of the active set. In particular, new connection components can be added to the current iterate.

    A shrink of the active set (for instance remove a connection component) can be achieved by means of two different ideas. For one thing intersection of the current active set with the candidate set $\mathcal{C}$ (cf. the 2nd item) eliminates all parts which cannot be optimal (at least for strictly complementary problems) and for another thing withdrawal of subregions of negative sign of the Lagrange multipliers, which corresponds to the dual part of a step of a primal-dual active set strategy.

3.4.2 Trial methods

Obviously it would be desirable to reduce the computational effort of the Newton step 2b in algorithms 1 and 2 to the computation of W_i without having to solve the system of local shape derivatives (see e. g. (3.27) and (3.10)). On closer examination of the different boundary value problems of local shape derivatives one recognizes that the variables typically are zero at the optimum, since the system is homogeneous then. Consequently – provided that local shape derivatives of the variables depend continuously on the shape – the corresponding terms on the left hand side of the Newton equations are close to zero, if $\mathcal{B}_i$ is close to the optimal active set.

These considerations give rise to so called *trial algorithms*; cf. [104, 156, 157]. They can be interpreted as fix point methods, which are characterized by neglecting the implicit shape dependency of the objects whose critical point ought to be found. Thus, they can also be interpreted as using the partial shape derivatives that were introduced in Paragraph 2.4.2. One obtains the following simplified algorithms

Algorithm 3 (Trial algorithm for bilevel optimization problem)**:**
Let $\mathcal{F}$ be the objective of the bilevel optimization problem (2.37), (2.36).

1. Set $i := 1$ and choose an initial guess $\mathcal{B}_i \in \mathcal{O}$.
2. Loop on i
   a) Solve the inner optimization problem (2.37) for the set $\mathcal{B}_i$ and extract $\hat{p}_{\mathcal{J}_i}|_{\beta_i} = -\lambda\,(\bar{u}_{\mathcal{J}_i} - u_d)|_{\beta_i}$ in order to be able to assemble the trial equation (3.30).
   b) Solve the trial equation in (3.30) in the variable $W_i \in \mathcal{V}$.
   c) Perform the trial update
   $$\mathcal{B}_{i+1} := R(W_i)$$
   where $R : T\mathcal{H}(\Omega) \to \mathcal{H}(\Omega)$ is a retraction.
   d) Stop the loop, if the update velocity field W_i is small enough; otherwise set $i := i + 1$.

   End of loop
3. Check if the strict inequality constraint $y_{\min} < \bar{y}_{\mathcal{J}_{i+1}} < y_{\max}$ and the sign conditions of the multipliers (cf. Corollary 6) are fulfilled. Stop if the check is passed; otherwise make another initial guess $\mathcal{B}_i \in \mathcal{O}$ and start all over again.

Algorithm 4 (Trial algorithm for variational relaxation approaches)**:**
Choose one of the shape gradient conditions (3.3p)–(3.3r) and relax one of the boundary conditions of the corresponding free boundary PDAE (3.3). In addition, choose a functional $\mathfrak{K}$ in the style of (3.15) or (3.16). Reduce the remaining part of the system and provide the shape semiderivative $\mathrm{D}\mathfrak{K}(.)[W,.]$.

1. Set $i := 1$ and choose an initial guess $\mathcal{B}_i \in \mathcal{O}$.
2. Loop on i
   a) Solve the remaining boundary value problem for the set $\mathcal{B}_i$ in order to assemble $\mathfrak{K}(\mathcal{B}_i)[.]$.
   b) Solve the trial equation (see e. g. (3.31)) in the variable $W_i \in \mathcal{V}$.
   c) Perform the trial update
   $$\mathcal{B}_{i+1} := R(W_i)$$
   where $R : T\mathcal{H}(\Omega) \to \mathcal{H}(\Omega)$ is a retraction.
   d) Stop the loop, if the update velocity field W_i is small enough; otherwise set $i := i + 1$.

   End of loop
3. Solve the remaining part of the first order optimality system and check if the strict inequality constraint $y_{\min} < \bar{y}_{\mathcal{J}_{i+1}} < y_{\max}$ and the sign conditions of the multipliers (cf. Corollary 6) are fulfilled. Stop if the check is passed; otherwise make another initial guess $\mathcal{B}_i \in \mathcal{O}$ and start all over again.

Likewise the Newton methods, the different steps of these two algorithms require some analysis.

1. The *trial equation* (i. e. the simplified Newton equation) of Algorithm 3, which corresponds to (3.27), then reads
$$\int_{\beta_i} \left(2\,\partial_{\boldsymbol{n}}^{\mathcal{J}}(\hat{p}_{\mathcal{J}_i} - p_{\min}^{\max}) + \kappa_{\mathcal{J}_i}(\hat{p}_{\mathcal{J}_i} - p_{\min}^{\max})\right)(\hat{p}_{\mathcal{J}_i} - p_{\min}^{\max}) W_i \cdot \boldsymbol{n}_{\mathcal{J}_i} V \cdot \boldsymbol{n}_{\mathcal{J}_i}$$
$$= -\int_{\beta_i} (\hat{p}_{\mathcal{J}_i} - p_{\min}^{\max})^2\, V \cdot \boldsymbol{n}_{\mathcal{J}_i}, \quad \forall V \in \mathcal{V}, \tag{3.30}$$
whereas the trial equations of the relaxed approaches due to $\mathfrak{K}_p^{\partial_n}$, $\mathfrak{K}_q^q$ and $\mathfrak{K}_p^p$ are given by
$$\int_{\beta_i} -\kappa_{\mathcal{J}_i} \left(\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_i^* - y_{\min}^{\max})\right)^2 W_i \cdot \boldsymbol{n}_{\mathcal{J}_i}\, V \cdot \boldsymbol{n}_{\mathcal{J}_i} = -\int_{\beta_i} \left(\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_i^* - y_{\min}^{\max})\right)^2 V \cdot \boldsymbol{n}_{\mathcal{J}_i}, \quad \forall V \in \mathcal{V} \tag{3.31a}$$
$$\int_{\beta_i} \left(2\,\partial_{\boldsymbol{n}}^{\mathcal{J}} \hat{q}_{\mathcal{B}_i} + \kappa_{\mathcal{J}_i} \hat{q}_{\mathcal{B}_i}\right) \hat{q}_{\mathcal{B}_i} W_i \cdot \boldsymbol{n}_{\mathcal{J}_i}\, V \cdot \boldsymbol{n}_{\mathcal{J}_i} = -\int_{\beta_i} \hat{q}_{\mathcal{B}_i}^2\, V \cdot \boldsymbol{n}_{\mathcal{J}_i}, \quad \forall V \in \mathcal{V} \tag{3.31b}$$
and (3.30) since $\mathfrak{K}_p^p = \mathrm{D}\mathcal{F}$. Obviously, the trial equation (3.31a) of the relaxation approach due to $\mathfrak{K}_p^{\partial_n}$ is not applicable, since the normal component of velocity field W_i is always reciprocal to the curvature $\kappa_{\mathcal{J}_i}$ and thus no reasonable update can be expected. Consequently, there is no guarantee for a given relaxation approach to yield a working trial algorithm and the corresponding trial equations have to be analyzed carefully.
2. Of course trial algorithms can be combined with the two ideas of avoiding the indefinite second order derivative at the optimum, which were developed in the 8th item on page 110.
3. The paradox situation, that the Newton algorithms 1 and 2 may get into trouble if the multiplier μ is regular (cf. the 8th item on page 110), can be encountered by means of another variational relaxation approach. Thereto, choose the shape gradient condition (3.3p) and use the weak continuity condition (3.3k) to deduce
$$\hat{p}_{\mathcal{A}}|_\gamma = p_{\min}^{\max}|_\gamma.$$
Now relaxing the weak continuity of the adjoint states, one has to solve
$$0 = \mathfrak{K}_p^\tau(\mathcal{B})[V] := \frac{1}{2}\int_\beta (\hat{p}_{\mathcal{J}} - \hat{p}_{\mathcal{B}})^2 V \cdot \boldsymbol{n}_{\mathcal{J}}, \quad \forall V \in \mathcal{V} \tag{3.32}$$
(subject to the usual strict inequality constraint) where $\hat{p}_{\mathcal{J}}$ and $\hat{p}_{\mathcal{B}}$ are the solutions of the boundary value problems (3.14) and
$$\begin{aligned} -\Delta\hat{p}_{\mathcal{B}} + \hat{p}_{\mathcal{B}} &= y_{\min}^{\max} - y_d && \text{in } \mathring{\mathcal{B}}, \\ \hat{p}_{\mathcal{B}}|_\beta &= p_{\min}^{\max}|_\beta && \text{on } \beta. \end{aligned}$$

The shape semiderivative of the functional is given by

$$\mathrm{D}\mathfrak{K}_p^\tau(\mathcal{B})[W,V] = \frac{1}{2}\int_\beta \Big(2\,(p'_\mathcal{J}[W] - p'_\mathcal{B}[W]) + \big(2\,\partial_{\boldsymbol{n}}^\mathcal{J}(\hat{p}_\mathcal{J} - \hat{p}_\mathcal{B}) + \kappa_\mathcal{J}(\hat{p}_\mathcal{J} - \hat{p}_\mathcal{B})\big) W\cdot\boldsymbol{n}_\mathcal{J}\Big)(\hat{p}_\mathcal{J} - \hat{p}_\mathcal{B})\, V\cdot\boldsymbol{n}_\mathcal{J}, \quad (3.33)$$

and the local shape derivatives $p'_\mathcal{J}[W]$ and $p'_\mathcal{B}[W]$ fulfill (2.48) and respectively

$$\begin{aligned} -\Delta p'_\mathcal{B}[W] + p'_\mathcal{B}[W] &= 0, && \text{in } \mathcal{B},\\ p'_\mathcal{B}[W]|_\beta &= -\partial_{\boldsymbol{n}}^\mathcal{B}(\hat{p}_\mathcal{B} - p_{\min}^{\max})W\cdot\boldsymbol{n}_\mathcal{B}, && \text{on } \beta. \end{aligned}$$

By means of (3.3f), (3.3n) and the definition of $p_{\min}^{\max}$ (see (2.46)) there holds

$$\hat{p}_\mathcal{B} - p_{\min}^{\max} = \hat{q}_\mathcal{B} \quad \text{in } \mathring{\mathcal{B}}.$$

Hence, one obtains

$$p'_\mathcal{B}[W]|_\beta = \partial_{\boldsymbol{n}}^\mathcal{J}\hat{q}_\mathcal{B} W\cdot\boldsymbol{n}_\mathcal{J}.$$

Consequently, the shape semiderivative can equivalently be written as

$$\begin{aligned} \mathrm{D}\mathfrak{K}_p^\tau(\mathcal{B})[W,V] &= \frac{1}{2}\int_\beta \Big(2\,p'_\mathcal{J}[W] \\ &\qquad + \big(2\,\partial_{\boldsymbol{n}}^\mathcal{J}(\hat{p}_\mathcal{J} - \underbrace{(\hat{p}_\mathcal{B} + \hat{q}_\mathcal{B})}_{=p_{\min}^{\max}}) + \kappa_\mathcal{J}(\hat{p}_\mathcal{J} - p_{\min}^{\max})\big) W\cdot\boldsymbol{n}_\mathcal{J}\Big)(\hat{p}_\mathcal{J} - p_{\min}^{\max})\, V\cdot\boldsymbol{n}_\mathcal{J} \\ &= \mathrm{D}\mathfrak{K}_p^p(\mathcal{B})[W,V]. \end{aligned}$$

Hence, the Newton Algorithm 2 applied to $\mathfrak{K}_p^p$ and to $\mathfrak{K}_p^\tau$, respectively, is the same. However, there is a difference when using the Algorithm 4, since the trial equations differ. Neglecting the local shape derivative terms in (3.33), one ends up with the trial equation

$$\begin{aligned} \int_{\beta_i} \big(2\,\partial_{\boldsymbol{n}}^\mathcal{J}(\hat{p}_{\mathcal{J}_i} - \hat{p}_{\mathcal{B}_i}) + \kappa_{\mathcal{J}_i}(\hat{p}_{\mathcal{J}_i} - p_{\min}^{\max})\big)(\hat{p}_{\mathcal{J}_i} - p_{\min}^{\max})\, W_i\cdot\boldsymbol{n}_{\mathcal{J}_i}\, V\cdot\boldsymbol{n}_{\mathcal{J}_i} \\ = -\int_{\beta_i}(\hat{p}_{\mathcal{J}_i} - p_{\min}^{\max})^2 V\cdot\boldsymbol{n}_{\mathcal{J}_i}, \quad \forall V\in\mathcal{V}. \end{aligned}$$

Now applying the two ideas of avoiding indefinite left hand sides (cf. the 8th item on page 110) yields

$$\begin{aligned} \int_{\beta_i} \big(\left\{\begin{matrix}2\\1\end{matrix}\right\}\partial_{\boldsymbol{n}}^\mathcal{J}(\hat{p}_{\mathcal{J}_i} - \hat{p}_{\mathcal{B}_i}) + \kappa_{\mathcal{J}_i}(\hat{p}_{\mathcal{J}_i} - p_{\min}^{\max})\big)\, W_i\cdot\boldsymbol{n}_{\mathcal{J}_i}\, V\cdot\boldsymbol{n}_{\mathcal{J}_i} \\ = -\int_{\beta_i}(\hat{p}_{\mathcal{J}_i} - p_{\min}^{\max})\, V\cdot\boldsymbol{n}_{\mathcal{J}_i}, \quad \forall V\in\mathcal{V}. \quad (3.34) \end{aligned}$$

At this point the original goal of creating an algorithm, which can cope with a regular Lagrange multiplier is reached. In order to see this, the simplified trial equation shall be analyzed at the optimal configuration $\mathcal{A}$. There holds

$$\begin{aligned} \partial_{\boldsymbol{n}}^\mathcal{I}(\hat{p}_\mathcal{I} - \hat{p}_\mathcal{A}) &= \partial_{\boldsymbol{n}}^\mathcal{I}(\hat{p}_\mathcal{I} - (p_{\min}^{\max} - \hat{q}_\mathcal{A})) \\ &= \mu_\gamma + \partial_{\boldsymbol{n}}^\mathcal{I}\hat{q}_\mathcal{A}. \end{aligned}$$

Hence, the Neumann term on the left hand side of (3.34) vanishes only if both μ_γ and $\partial_{\boldsymbol{n}}^\mathcal{I}\hat{q}_\mathcal{A}$ are zero (note, that both terms are nonnegative). Consequently, the corresponding simplified trial algorithm is expected to work even if the multiplier μ is regular (i. e. $\mu_\gamma = 0$).

Note, that it is very unlikely that $\partial_{\boldsymbol{n}}^\mathcal{I}\hat{q}_\mathcal{A}$ vanishes as well: according to Corollary 6 the multiplier $\hat{q}_\mathcal{A}$ is given as the solution of the boundary value problems

$$\begin{aligned} -\Delta\hat{q}_\mathcal{A} + \hat{q}_\mathcal{A} &= \mu_{\mathring{\mathcal{A}}}^{\max} && \text{in } \mathring{\mathcal{A}}_{\max},\\ \hat{q}_\mathcal{A}|_\gamma &= 0 && \text{on } \gamma_{\max}, \end{aligned} \qquad \begin{aligned} -\Delta\hat{q}_\mathcal{A} + \hat{q}_\mathcal{A} &= -\mu_{\mathring{\mathcal{A}}}^{\min} && \text{in } \mathring{\mathcal{A}}_{\min},\\ \hat{q}_\mathcal{A}|_\gamma &= 0 && \text{on } \gamma_{\min}. \end{aligned}$$

Hence, $\partial_n^{\mathcal{I}} \hat{q}_{\mathcal{A}}$ is zero only, if the optimal active set is the solution of a nontrivial free boundary problem or if $c_{\max}$ and $c_{\min}$ are zero, too. The first case is very unlikely and the second case means that both the regular and the singular part $\mu_{\mathcal{A}}$ and μ_γ vanish simultaneously. This can only happen, if the state constraint is not active or is completely non strictly complementary.

All in all, the trial method of the discussed relaxation approach is expected to work unless very specific situations which probably occur only in constructed test examples.

3.4.3 Total linearization methods

The idea of total linearization form Paragraph 3.3.3 suggest the following all-at-once algorithm.

Algorithm 5 (Total linearization method)**:**
Choose the full or the reduced total linearization approach, cf. Paragraph 3.3.3.

1. Set $i := 1$ and choose an initial guess for the active set $\mathcal{B}_i \in \mathcal{O}$ and for all function variables symbolized as $f_i = (y_{\mathcal{J}_i}, p_{\mathcal{J}_i}, \dots)$.
2. Loop on i
   a) Solve the Shape-Newton equation (this is (3.22) in case of the reduced total linearization approach) in the variables W_i, $\delta_{f,i}$.
   b) Perform the Newton update of the set
   $$\mathcal{B}_{i+1} := R(W_i)$$
   where $R : T\mathcal{H}(\Omega) \to \mathcal{H}(\Omega)$ is a retraction.
   c) Perform a Newton update of the function variables. That is to say, at first set
   $$\tilde{f}_i := f_i + \delta_{f,i}$$
   and then transport $\tilde{f}_i$ from $\mathcal{B}_i$ to $\mathcal{B}_{i+1}$, which yields f_{i+1}.
   d) Stop the loop, if the increment $(W_i, \delta_{f,i})$ is small enough; otherwise set $i := i + 1$.

   End of loop.
3. Solve the remaining part of the first order optimality system if necessary and check if the strict inequality constraint $y_{\min} < \bar{y}_{\mathcal{J}_{i+1}} < y_{\max}$ and the sign conditions of the multipliers (cf. Corollary 6) are fulfilled. Stop if the check is passed; otherwise make another initial guess $\mathcal{B}_i \in \mathcal{O}$ and start all over again.

In view of the analysis of Paragraph 3.3.3 it is possible to use a simplified Shape-Newton equation in step 2a in order to circumvent problems in view of finite element discretization. This is (3.23) in case of the reduced total linearization approach.

The Newton update step 2a has to be regarded in the context of function space parametrization and parallel transport, which were discussed in the 16th item on page 78. Let $F_i \in \mathcal{H}(\Omega)$ be a transformation induced by W_i such that $F_i(\mathcal{B}_i) = \mathcal{B}_{i+1}$. Then f_{i+1} is given by $f_{i+1} := \tilde{f}_i \circ F_i^{-1}$. However, F_i is not determined uniquely since only the normal component of the velocity field W_i at the interface β_i is known and hence an extension into the bulk of the domain is required; cf. the 19th item on page 80 and the discussion in Section 3.1 on page 93ff. Consequently, one has to provide a method how to construct the transformation; cf. the discussion in Paragraph 4.1.3, in particular, on page 125.

3.5 Analysis of the primal-dual active set strategy

The primal-dual active set strategy (PDAS) is a standard algorithm for solving inequality constrained quadratic optimization problems which typically occur as subproblems in sequential quadratic programming (SQP) [98]. In the context of optimal control of PDEs this algorithm was initially used to solve control-constrained problems [12]. Later on the method was applied to state-constrained OCPs [13]. However, due to lack of pointwise interpretation of the multiplier μ, this idea requires discretization beforehand and one is confronted with a mesh dependent behavior of the algorithm [11]. This disadvantage

can be handled on the continuous level by means of regularization of OCPs. In particular, there are two major ideas introduced, namely *Lavrentiev regularization*, see [158, 126, 142, 125, 143, 29, 92] and *Moreau-Yosida regularization*; cf. [94, 95, 96, 97, 87, 127, 88]. Equipped with these enhancements the ordering of the SQP- and the PDAS-loops can be reversed [97] and the PDAS turned out to be mesh independent [91, 85] for it being a *semi-smooth Newton method* and to be very efficient [11, 125, 10].

However, regularization requires to choose a suitable regularization parameter. On the one hand regularization shall be effective and thus the problem shall become easier to solve, but on the other hand the influence of the regularization on the OCP shall be as weak as possible. Between these two priorities one has to adjust the regularization parameter and there is no golden rule available. Fortunately, it is possible to use *path-following methods* with respect to Moreau-Yosida regularization [86, 87], this means starting with a strongly regularized problem and iteratively decrease regularization, while using the intermediate optimal solutions as initial guesses for the next less regular subproblem.

The goal of this Section is to provide some insight into how the PDAS works form a geometrical perspective and to give a hint, why the Moreau-Yosida regularization performs well. Both aspects allow for a deeper understanding of the shape optimization/calculus based algorithms from Section 3.4. The basis of the analysis is the following version of the PDAS.

Algorithm 6 (Unregularized PDAS (informally on continuous level))**:**

1. Set $i = 1$ and choose an initial guesses for the active set $\mathcal{B}_1^{\max}$ and $\mathcal{B}_1^{\min}$.
2. Loop on i
   a) Solve the coupled system

$$\begin{aligned}
-\Delta y_i + y_i &= u_i && \text{in } \Omega, & \qquad \mu_i^{\max} &= 0 && \text{in } \mathcal{J}_i^{\max} := \Omega \setminus \mathcal{B}_i^{\max},\\
-\Delta p_i + p_i &= y_i - y_d + \mu_i^{\max} - \mu_i^{\min} && \text{in } \Omega, & \qquad \mu_i^{\min} &= 0 && \text{in } \mathcal{J}_i^{\min} := \Omega \setminus \mathcal{B}_i^{\min},\\
\lambda\,(u_i - u_d) + p_i &= 0 && \text{in } \Omega, & \qquad y_i &= y_{\max} && \text{in } \mathcal{B}_i^{\max},\\
\partial_{\boldsymbol{n}} y_i &= 0 && \text{on } \Gamma, & \qquad y_i &= y_{\min} && \text{in } \mathcal{B}_i^{\min}.\\
\partial_{\boldsymbol{n}} p_i &= 0 && \text{on } \Gamma, &&&&
\end{aligned}$$

   b) Generate new active sets

$$\begin{aligned}
\mathcal{B}_{i+1}^{\max} &:= \left(\mathcal{B}_i^{\max} \cup \left\{x \in \Omega \,\middle|\, y_i(x) > y_{\max}(x)\right\}\right) \setminus \left\{x \in \Omega \,\middle|\, \mu_i^{\max}(x) < 0\right\}\\
\mathcal{B}_{i+1}^{\min} &:= \left(\mathcal{B}_i^{\min} \cup \left\{x \in \Omega \,\middle|\, y_i(x) < y_{\min}(x)\right\}\right) \setminus \left\{x \in \Omega \,\middle|\, \mu_i^{\min}(x) < 0\right\}.
\end{aligned}$$

   c) Stop if $\mathcal{B}_{i+1}^{\max} = \mathcal{B}_i^{\max}$ and $\mathcal{B}_{i+1}^{\min} = \mathcal{B}_i^{\min}$.
   Otherwise set $i := i + 1$ and restart the loop.

   End of loop.

This algorithm is to be understood informally since the generation of new active sets in step 2b cannot be carried out on the continuous level, because the multipliers $\mu_i^{\max}$ and $\mu_i^{\min}$ do not possess a pointwise interpretation for them being measures. Nonetheless, one can recognize the essential mechanisms.

3.5.1 Two drawbacks of the primal-dual active set strategy

In order to understand strengths and weaknesses of the PDAS it is valuable to have an insight in is qualitative behavior. The most relevant fact is to know how the update of the active sets works. Obviously, primal conditions (these are the state constraints) yield growth of the current guess of the active sets, whereas a shrink is due to the sign conditions of the multipliers.

Due to Proposition 3 one can expect both multipliers $\mu_i^{\max}$ and $\mu_i^{\min}$ to decompose into a regular part in the interior of the current active set and a singular part which is concentrated on the interface. Moreover, the regular parts are known to be equal to $c_{\max}$ and $c_{\min}$ (cf. (2.3a) and (2.3b)). Consequently, the regular parts of the multipliers can yield a shrink of the sets $\mathcal{B}_i^{\max}$ and $\mathcal{B}_i^{\min}$ only if they contain points which are outside the candidate sets $\mathcal{C}_{\max}$ and $\mathcal{C}_{\min}$; see (3.26a) and (3.26b). Vice versa, if the optimal active sets $\mathcal{A}_{\max}$ and $\mathcal{A}_{\min}$ are proper subsets of candidate sets the PDAS will work as follows. Within the first iteration

the sign conditions of the regular part of the multipliers restrict the initial guess to the candidate set. In all following iterations further diminishment of the active sets are only due to the singular part of the multipliers (unless some points outside the candidate sets were added by mistake). Hence, the informal Algorithm 6 gets into big trouble then. For one thing the singular part cannot be evaluated pointwisely and even if an evaluation were possible the algorithm would stagnate since the current guess of the active set can only be decremented by null sets (i. e. the interface).

Hence, it is worthwhile to have a closer look on a finite element discretized counterpart of Algorithm 6. Actually, it would be appropriate to discretize the singular and the regular component of the multipliers separately. However, this would yield too many degrees of freedom, since there is no additional condition to compensate for the doubled number of variables on the discretized interfaces. Thus, singular and regular part have to be discretized jointly. This approach results in a big deflection of the discretized multiplier at the interfaces; cf. [10, Kap. 12]. Interestingly enough this deflection does not only possess the qualitative interpretation to be the singular part of the multiplier. The integral over a cross section (this is a cut in normal direction on the interface) is roughly equal to the normal-kink of the adjoint state at the interface [10, Kap. 12.2]. Consequently, the character of the singular part μ_γ of being a measure, which is responsible for the kink of the adjoint state (see (2.4i), (2.4j)), is reflected on the discretized level. Hence, it is justified to check the sign of the deflection within update step 2b of the discretized counterpart of Algorithm 6. All in all, it is possible to shrink the size of the candidate active sets $\mathcal{B}_i^{\max}$ and $\mathcal{B}_i^{\min}$ on the discretized level even if they are proper subsets of the candidate sets, since the singular part is at least blurred to one mesh size.

This finding explains both why the discrete PDAS indeed has the ability to converge and why it is mesh dependent: the algorithms has to iterate mesh layer by mesh layer in order to find the right active set within the candidate set.

It is also possible to give plausible arguments why the Moreau-Yosida regularization is successful. Numerical practice confirms that it basically blurs the singular part of the multiplier, such that it is no longer concentrated on one or two mesh layers along the interface. The amount of this smoothing is independent of the chosen mesh size and consequently checking the sign condition of the (regularized) multiplier may yield a diminishment of the current guess for the active set which is not only concentrated near the interface. In other words, more progress – and, in particular, mesh size independent progress – can be achieved within each iteration.

Furthermore, path following ideas can be understood from this point of view as well. Starting with a strong regularization – that is starting with strongly blurred μ_γ – enables quick and large deformations of the candidate active sets. However, it inhibits an exact localization of the interface, which is hidden somewhere underneath the approximation of the singular part of the multipliers. Cutting back regularization sharpens the multiplier and yields a more precise localization of the interface. Thus, the path following methods represent highly sophisticated schemes of balancing fast progress and accurate identification of the active set.

However, the primal-dual active set strategy has difficulties with another type of phenomenon called *degeneracy*; cf. [11, 12]. This is the inability to determine if some grid points are part of the active set or not. This situation occurs if the distance between the optimal state and the state constraint is very small in some subregions of the optimal inactive set and typically yields chattering of the active sets $\mathcal{B}_i$. Typically, this behavior prevents the algorithm from converging fast. The autors of the cited papers present coping strategies which basically are enhanced stopping criteria.

3.5.2 Benefits of the new approach

The specific drawbacks of the PDAS, in particular the need for regularization and the treatment of the state constraint, which may lead to a slowdown of convergence (degeneracy), are not an issue for the shape optimization/calculus based algorithms from Section 3.4.

For one thing those algorithms can be formulated on the continuous level (that is to say in function and shape space) without regularization and thus may exhibit a mesh independent behavior; see Paragraph 4.2.5.

For another thing they rely on information which is not used in the context of PDAS, namely the differentiability with respect to deformation of the active set. This finding may be interpreted as follows. The pointwise state constraints are not completely independent. Once, that the right topology of the

active set is found (which of course is assumed throughout this thesis and should be a focus of further research), there is a bias which elements of the current inactive set $\mathcal{J}_i$ could be active and which elements of the current active set could be inactive. Namely those points neighboring the current interface. In contrast, those points which are in the bulk of $\mathcal{J}_i$ and $\mathcal{B}_i$ are more unlikely to be mistakenly assumed to be in-/active. This introduces a high amount of "sparsity" in comparison to the idea of a PDAS. The latter approach treats all elements of Ω as equal with respect to the state constraint. On the one hand this enables a global convergence, but on the other hand it induces the difficulties with respect to degeneracy.

All in all, it seems to be beneficial, to solve a highly to moderately regularized approximation of the original model problem by means of a PDAS for globalization and then switch over to one of the presented shape based algorithms.

CHAPTER 4

Numerics

This Chapter shall give some insight into the implementation of the algorithms of Chapter 3 (see Section 4.1) and is devoted to the presentation of numerical results; cf. Section 4.2. The numerical results are to be understood as a first small step in order to analyze the full capability of the presented ideas. For one thing the expected benefits of the Bryson-Denham-Dreyfus approach (i. e. index reduction of the optimality system) can only be validated by means of more complex OCPs than the simple model problem (2.1). For another thing the algorithms are expected to exhibit high performance on fine meshes and in a severely nonlinear context (which is the standard case in the optimal control of ordinary differential equations), since they come without regularization and are built up from the bottom in the nonlinear context of optimization on vector bundles (see Paragraph 2.6.3). Evaluation of the behavior of the algorithms in that regime is beyond the scope of this thesis as well.

Henceforth, the analysis is restricted to reduced Newton and trial methods (see paragraphs 3.4.1 and 3.4.2). Especially the promising total linearization method (i. e. the all-at-once solver), which supposedly is best fitted to a full nonlinear OCP, remains unattained. Nonetheless, it becomes apparent that shape calculus based algorithms perform well in comparison with a primal-dual active set strategy, which shall be a good starting point for future investigations.

4.1 Finite element discretization

Within this section only a brief overview on selected topics concerning *finite element* (FE) implementation of the shape calculus based algorithms from Chapter 3 is given. A comprehensive introduction to *finite elements methods* (FEM) in context of shape and topology optimization is due to Haslinger and Neittaanmäkki [78].

The algorithms of Chapter 3 are discretized by means of a standard FEM which uses continuous, piecewise linear elements on an unstructured triangular mesh. This choice and in particular the approximation of the $C^{1,1}$ boundaries Γ and γ by means of polygons may induce considerable problems. For one thing results of the theory of elliptic partial differential equations, which where applied in Chapter 2 may not be valid any more[1] and for another thing typical tools like integration by parts on boundaries, which are essential in shape calculus, have additional contributions when the boundaries have kinks.[2] Although these effects shall be considered when discretizing, they are disregarded in the implementation. This might cause reduced stability and convergence rates and ought to be investigated.

In order to distinguish between discretized and nondiscretized entities, the former are tagged with an underline $\underline{(.)}$.

[1] A comprehensive collection of results in polygonal domains can be found in [69], in particular paragraphs 1.4.5 and 1.5.2 and chapter 4.

[2] See [44, Chp. 10 Rem. 2.3] and the references therein; e. g. [40]. Moreover, confer [151, Sec. 3.8], which unfortunately refers to a nonexistent section.

4.1.1 Approximation of normal vector field and mean curvature

Shape calculus necessitates to discretize some geometric entities at the interface β, in particular the outer unit normal vector fields $\boldsymbol{n}_{\mathcal{J}}$ and $\boldsymbol{n}_{\mathcal{B}}$ and the curvatures $\kappa_{\mathcal{J}}$ and $\kappa_{\mathcal{B}}$. The curvature is required to assemble the corresponding terms, which occur in the Newton- and trial equations of the algorithms from Section 3.4. The unit normal vector field is important with respect to deformation of the active sets in the course of the algorithms. There is no "natural" approximation of these notions in the context of triangular finite elements, and hence one has to choose an approach which should preserve the order of convergence of the FEM. Since numerical error analysis is beyond the scope of this thesis, an approach of [104] is used and slightly adapted here.

The following considerations are related to one connection component of the boundary of a triangulated set (in $\mathbb{R}^2$). It is assumed that the boundary (i. e. a polygon) has an orientated parametrization and that the set lies locally on the left hand side of the edges. Thus, let B be an arbitrary interface node with previous orientated edge $\vec{u}$ and following orientated edge $\vec{v}$. Then, the curvature κ and the outer unit normal vector $\vec{n}$ of the interface at B can be approximated by means of the circumcircle of the triangle defined by B, $\vec{u}$ and $\vec{v}$; see Figure 4.1.

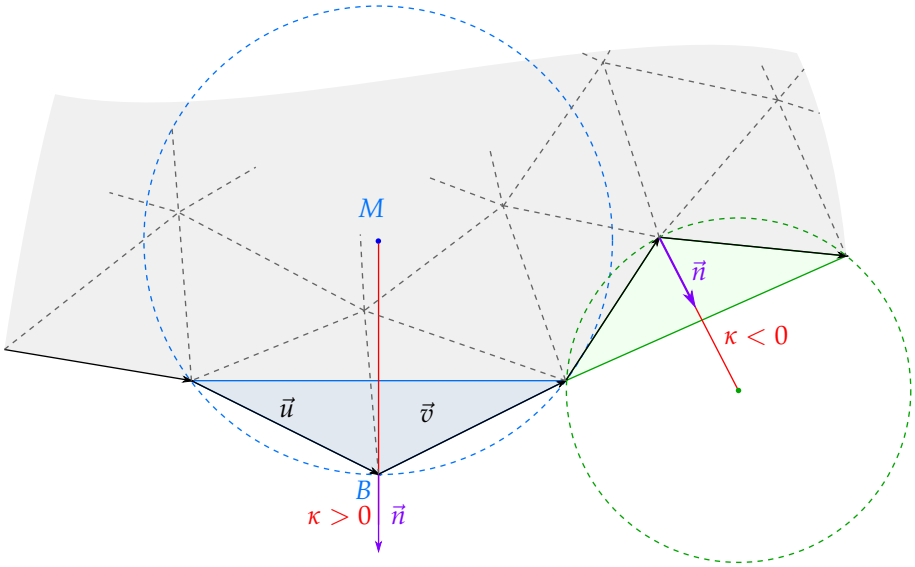


Figure 4.1: Discrete approximation of mean curvature and outer unit normal vector field.

Hence, one requires the distance between B and the center M of the circumcircle, which is uniquely determined as the intersection of the perpendicular bisectors. That is to say

$$M = B - \frac{1}{2}\vec{u} + \alpha\, Q\,\vec{u} \overset{!}{=} B + \frac{1}{2}\vec{v} + \delta\, Q\,\vec{v}, \quad \text{where } Q := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Thus, there holds

$$\frac{1}{2}\,(\vec{v}+\vec{u}) \overset{!}{=} \alpha\, Q\,\vec{u} - \delta\, Q\,\vec{v} = Q\,(\vec{u},\vec{v}) \begin{pmatrix} \alpha \\ -\delta \end{pmatrix}.$$

Cramer's rule yields

$$\alpha \overset{!}{=} \frac{\det(\vec{u}+\vec{v}, 2\,Q\,\vec{v})}{\det(2\,Q\,(\vec{u},\vec{v}))} = \frac{(\vec{u}+\vec{v})\cdot\vec{v}}{2\underbrace{\det Q}_{=1}\det(\vec{u},\vec{v})}$$

Finally, one obtains

$$|\kappa| = |M-B|^{-1} = \left|-\frac{1}{2}\vec{u} + \frac{(\vec{u}+\vec{v})\cdot\vec{v}}{2\det(\vec{u},\vec{v})}\,Q\,\vec{u}\right|^{-1} \underset{\vec{u}\perp Q\,\vec{u}}{=} 2\sqrt{|\vec{u}|^2 + \frac{\left((\vec{u}+\vec{v})\cdot\vec{v}\right)^2}{\det(\vec{u},\vec{v})^2}|\vec{u}|^2}^{-1}$$

$$= \frac{2|\det(\vec{u},\vec{v})|}{|\vec{u}|\sqrt{\det(\vec{u},\vec{v})^2 + \left((\vec{u}+\vec{v})\cdot\vec{v}\right)^2}}.$$

The sign of κ is defined as follows. The curvature is positive at B, if the triangulated set is locally convex at B and negative if it is locally concave. Moreover, if the set lies locally on the left of $\vec{u}$ there holds

$$\begin{aligned}\kappa > 0 \text{ at } B &\Leftrightarrow \text{ convexity at } B\\ &\Leftrightarrow \text{ interface is curved to the left at } B\\ &\Leftrightarrow \det(\vec{u},\vec{v}) > 0.\end{aligned}$$

Thus, one obtains

$$\kappa = \frac{2\,\det(\vec{u},\vec{v})}{|\vec{u}|\sqrt{\det(\vec{u},\vec{v})^2 + \big((\vec{u}+\vec{v})\cdot\vec{v}\big)^2}}.$$

Moreover, an outer unit normal vector $\vec{n}$ at node B can be defined as

$$\vec{n} := \operatorname{sgn}(\kappa)\,\frac{B-M}{|B-M|} = \kappa\,(B-M) = \kappa\left(\frac{1}{2}\,\vec{u} - \frac{(\vec{u}+\vec{v})\cdot\vec{v}}{2\,\det(\vec{u},\vec{v})}\,Q\,\vec{u}\right) = \frac{\det(\vec{u},\vec{v})\,\vec{u} - \big((\vec{u}+\vec{v})\cdot\vec{v}\big)\,Q\,\vec{u}}{|\vec{u}|\sqrt{\det(\vec{u},\vec{v})^2 + \big((\vec{u}+\vec{v})\cdot\vec{v}\big)^2}}.$$

This formula is even stable if $\vec{u}$ and $\vec{v}$ are approximately parallel and κ tends to zero. However, there is a situation where the whole reasoning yields somehow bad results. If the lengths of $\vec{u}$ and $\vec{v}$ are very different and if these vectors form an obtuse angle, one is confronted with the geometric setting of the left hand side of Figure 4.2. A reasonable remedy is to stretch the shorter one of both vectors to the length of the other; cf. the right hand side of Figure 4.2.

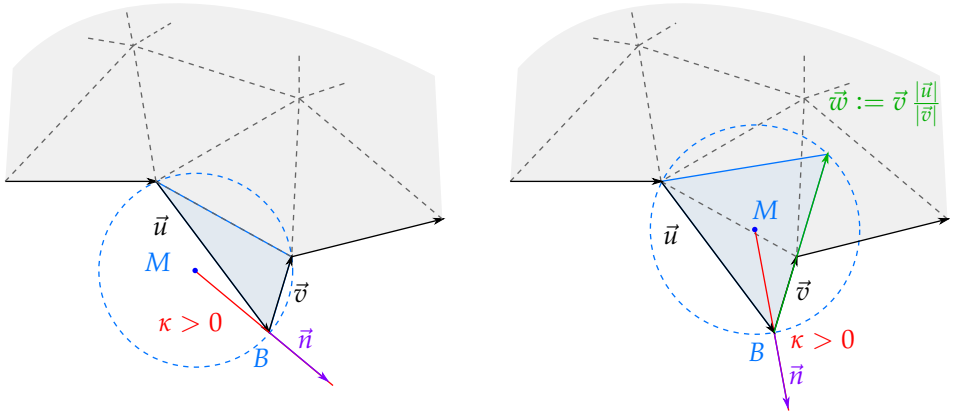


Figure 4.2: Remedy for a better approximation of curvature and outer unit normal vector field.

In view of the fact, that the interface β is an interior boundary of Ω it suffices to compute the discrete approximation of $\kappa_{\mathcal{B}}$ and $\boldsymbol{n}_{\mathcal{B}}$, since one obtains the approximation of $\kappa_{\mathcal{J}}$ and $\boldsymbol{n}_{\mathcal{J}}$ by reversing the signs node-by-node.

4.1.2 Splines and tracking the interface

Each of the algorithms presented in Section 3.4 relies on the iterative update of the active set $\mathcal{B}$ by means of a retraction; cf., for instance, step 2c of Algorithm 1. As already discussed in detail in Section 3.1 on page 93ff. the deformation of the active set is a topic of its own, whose comprehensive analysis is beyond the scope of this thesis. Hence, one confines oneself with some simple ideas, which perform adequately for numerical testing of the algorithms.

Due to the approximation of the mean curvature, one has access to reasonable unit normal vectors at each interface node. Hence, it is nearby to use the discrete normal component of the velocity field $\underline{W}$, which is obtained in each iteration of the different algorithms, and use it for the transformation approach of path following (see page 92). In other words, each interface node B is moved to $B + \underline{W}(B)\cdot\underline{\boldsymbol{n}}(B)$; cf. Figure 4.3.

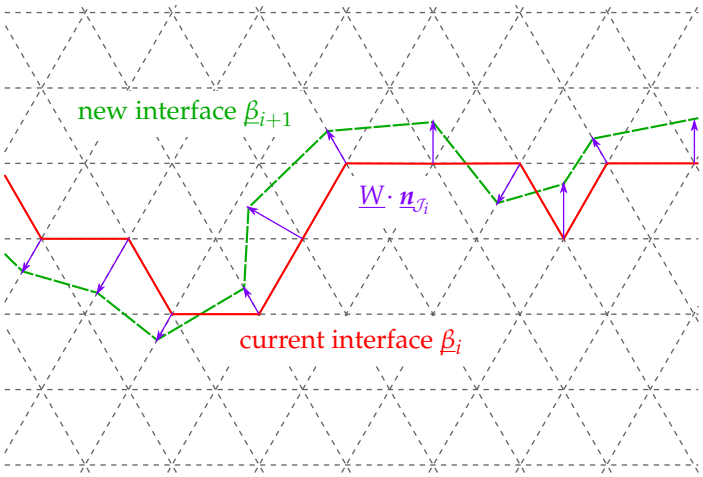


Figure 4.3: Discrete transformation approach of path following.

However, this reasoning has two drawbacks. For one thing the new discrete interface $\underline{\beta}_{i+1}$ may be degenerated (i. e. self-intersecting) and for another thing it does not lie on grid points.[3] Both problems can be circumvented by means of an idea, which is illustrated in Figure 4.4 and which works essentially like *Huygens' principle*. Each node is assigned with an integer $z = 1$, if and only if the node is in the discrete

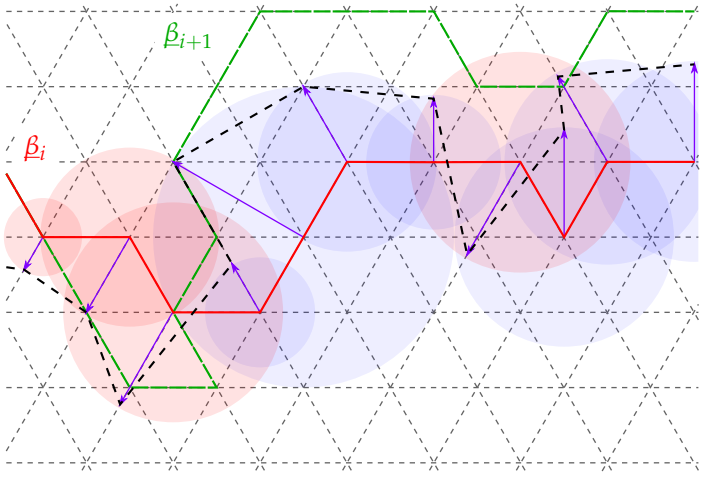


Figure 4.4: Comparison of nodal deformation of the interface corresponding to Huygens' principle (green) and pure discrete transformation approach (black).

active set $\underline{\mathcal{B}}_i$ and with $z = 0$, if and only if the node is in the discrete inactive set $\underline{\mathcal{J}}_i$ (by definition interface nodes are ascribed to the active set). Then, a circle is drawn around each interface node B, whose radius is equal to the absolute value of the velocity field $\underline{W}_i(B)$, and z is modified for each node of the grid in the following way: increment z by 1 for each covering circle, which is due to a deformation of the interface which increases the size of $\underline{\mathcal{J}}_i$, and decrement z if the circle stems from a deformation which increases the size of $\underline{\mathcal{B}}_i$. If the result for a given node is positive ($z > 0$), it remains/becomes an element of $\underline{\mathcal{B}}_{i+1}$ and otherwise an element of $\underline{\mathcal{J}}_{i+1}$. The new interface $\underline{\beta}_{i+1}$ is then given as the boundary polygon of $\underline{\mathcal{B}}_{i+1}$.[4]

Obviously, all vertices of $\underline{\mathcal{B}}_{i+1}$ are grid points. Moreover, this approach yields a well-defined polygon even if the transformation due to $\underline{W}_i$ would result in a self-intersecting curve as illustrated in Figure 4.5. The ability of coping with self-intersection can be used to induce changes in the topology of the active set. Suppose the active set has linkage which connects two larger components. If the transformation $\underline{W}_i$

[3] In particular, the first difficulty can be dealt be means of level set methods, which are not applied here, since a time dependent tracking of the evolution of the interface is not an issue in the present context; see the 10th item of the discussion on page 111.

[4] This idea can easily be generalized to situations, where different types of active set have to be distinguish, as for example when upper and lower state constraints are active simultaneously.

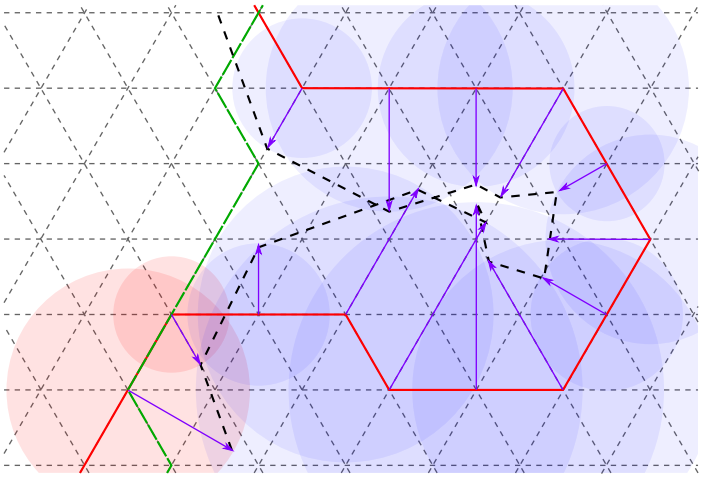


Figure 4.5: Huygens' principle prevents the interface from self-intersection.

moves the neighboring interface parts towards and even through each other the connection is cut and the original connection component of the active set is split into two separate parts; see Figure 4.6. Certainly, these considerations are also valid from the perspective of the inactive set, and hence union of connection components of the active set is possible as well.

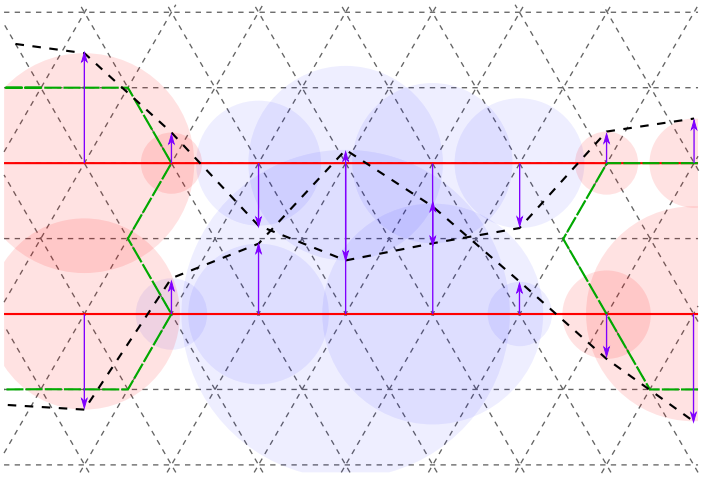


Figure 4.6: Huygens' principle allows changes of topology.

Those appealing benefits of Huygens' principle are accompanied by two drawbacks. For one thing it is expensive to compute the z-balance for each node, since the distance to every interface node has to be computed and compared with the value of $\underline{W}_i \cdot \underline{\boldsymbol{n}}_{\mathcal{J}_i}$. However, this task is perfectly parallelizable and hence the effort is expected to play a minor role within the whole algorithm. For another thing the new interface $\underline{\beta}_{i+1}$ typically has many sharp vertices even if the original $\underline{\beta}_i$ is smooth. This fact yields bad behavior of the FE approximation when solving the Newton-/trial equations of the algorithms from Section 3.4. In particular, some functions (e. g. the adjoint state) tend to have oscillations at the interface; see Figure 4.7. This in turn results in a very unregular velocity field $\underline{W}_{i+1}$ such that the thereon based deformation of the active set would have a much more zigzagging boundary. All in all, this effect is self-amplifying.

The proposed remedy is a twofold adaptive smoothing, which proved successful in practice. On the one hand a multilevel smoothing of the deformation function $\underline{W}_i \cdot \underline{\boldsymbol{n}}_{\mathcal{J}_i}$ is applied, and on the other hand the interface is smoothed down. Due to Assumption 1 each connection component of the interface is a closed curve, which results in closed polygons on the discretized level. Hence, the coordinates of the vertices of such an ordered polygon form a discrete periodic signal and it is suitable to use a *fast Fourier transformation* (FFT). If high frequency components of the transformed coordinate signal are damped, the

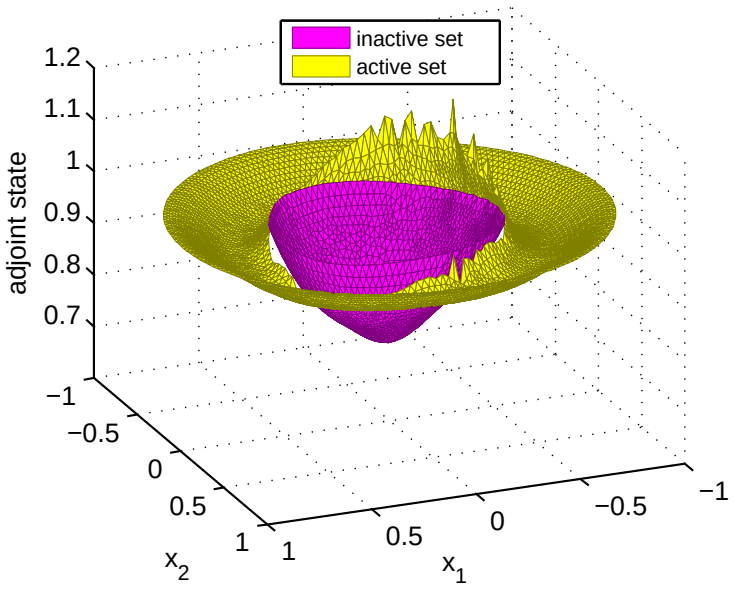


Figure 4.7: Oscillations of the adjoint state $\underline{\hat{p}}$ at the interface. See also the appertaining non-smooth interface of the fifth iteration in Figure 4.16.

retransformation is a smoothed version of the original polygon. The quality of smoothing can easily be adapted to different parameters. For instance the size of the deformation, which had yielded the current interface which has to be smoothed out, is a good index for the distance between the current iterate and the critical point (i. e. set) to be seeked. Thus, it indicates if only little smoothing is necessary. However, Assumption 1 cannot ensure that all iterates of the active set $\underline{\mathcal{B}_i}$ stay away form the outer boundary $\underline{\Gamma}$. Consequently there is no guarantee that the interfaces indeed are closed polygons and that FFT is a suitable method. Moreover, starting points and endpoints of such nonclosed components have to remain located at $\underline{\Gamma}$ after smoothing. In that situation it is possible to use spline related smoothing operations, which yield to least square problems. Note, that those optimization problems (as well as FFT) are of minor computational cost, since the degrees of freedom are related to the number of interface nodes only. The second starting point is the adaptive smoothing of the normal component of the deformation vector field. Firstly outliers are detected and replaced by means of interpolated data. This is an important step, since disproportional big values at isolated nodes may destroy the whole geometric setting; for instance, if the suggested, faulty deformation is bigger than Ω. Afterwards use FFT or some smoothing spline routine in order to adaptively reduce the variation of $\underline{W_i} \cdot \underline{\boldsymbol{n}_{\mathcal{J}_i}}$. One has to ensure at this, that the smoothing is ineffectual (i. e. it does not intervene) if the function is small enough, which indicates that the current iterate is near a critical point.

It should be noted, that the adaptive smoothing of the interface is successful, as long as Assumption 1 is fulfilled; in particular as long as the interface γ is of class $C^{1,1}$. In more general situations, for instance if the interface is allowed to be a curvilinear polygon in the sense of [69, Def. 1.4.5.1] (i. e. a curve, which is piecewise smooth and with finitely many kinks) the smoothing operation may inhibit convergence of the algorithms, since real kinks a smoothed out in each iteration. Hence, suitable workarounds have to be made in such situations.

4.1.3 Mesh deformation and mesh generation

Once, that these two smoothing approaches are used the resulting polygon is no longer fixed to the original mesh. Thus, three different approaches are suited.

- Remesh the domain Ω such that the new interface is contained in the new grid.
- Move the nodes of the current mesh such that the new interface is contained in the new grid.
- Use a finite element method that is capable to cope with that situation.

Although the third idea is probably the best fitted approach, it is beyond the scope of this thesis and the reader is referred to literature on *unfitted finite element methods* [8, 9, 74] and *extended finite element methods (XFEM)* [63, 30, 28]. Moreover, there is considerably progress in the *arbitrary Lagrangian-Eulerian (ALE)* methods; cf. the survey article [45] and the references therein. Those methods are widely used in the context of *computational fluid dynamics* and in the simulation of *structural mechanics*.

Movement of the mesh nodes is an efficient approach as long as deformation is of moderate size and this approach corresponds to the Lagrangian description used in ALE methods. Large deformations typically yield mesh entanglement and therefore require remeshing of the domain. However, as long as distortion is small enough, the movement of the mesh nodes is efficient and can be implemented by means of a three step strategy:

1. extend the velocity field $(\underline{W}_i \cdot \underline{\boldsymbol{n}}_{\mathcal{J}_i})\,\underline{\boldsymbol{n}}_{\mathcal{J}_i}$ to the bulk of the domain (or at least to a narrow band around the interface),
2. move the nodes and
3. regularize the mesh.

The first step is mandatory, if the displacement of the interface nodes is larger than the mesh size, since the nodes in the bulk of the domain have to be moved too in order to prevent entanglement of the mesh. Hence, one requires efficient schemes for extending the velocity field. As already mentioned in Section 3.1 on page 93 one can make use of ideas which are developed in the context of level set and fast marching methods. Furthermore, it is possible to apply methods of linear elasticity, see [93, 48], where the mesh is regarded as an elastic solid, whose outer boundary $\underline{\Gamma}$ is fixed while the interior is deformed in such a way that the interface $\underline{\beta}_i$ is mapped to $\underline{\beta}_{i+1}$. By that means one obtains a displacement field for all nodes of the mesh. A less sophisticated and less robust but easier approach was chosen for the computations in this thesis. The extension of the velocity field is obtained by means of interpolation. For this purpose, the spacial coordinates are treated separately. All boundary nodes are fixed, i. e. they have zero displacement and thus it is possible to compute the coordinates of the extended vector field at any node of the mesh by interpolation of the normal component of the velocity field at the interface nodes.

Applying the transformation approach of path following (see page 92), the movement of every node is nothing but adding the extension velocity field to its position. When the deformation of the mesh is completed successfully, a regularization of the mesh is typically indicated. The quality of the obtained mesh may be low due to sharp angles of some elements. Hence common strategies which jiggle the mesh, while interface and outer boundary nodes remain fixed, can be applied.

Another benefit of mesh deformation is a simple implementation of the transport of discretized function variables to the new mesh, which is required by total linearization methods (see step 2c of Algorithm 5). Since the whole mesh topology is preserved, nothing has to be done when using continuous and piecewise linear FE. The function values are attached to their corresponding nodes and are transported by means of the displacement of the nodes. However, the movement of nodes which is due to the mesh regularization step, which is not necessary from the perspective of the algorithms, but only for reasons of numerical stability, has to be applied independently. Thus there is need for interpolation actually (and, if required, extrapolation too). These effects are neglected in the implementation of the algorithms of this thesis, since mesh jiggling in order to increase the quality of the mesh has a minor impact on location of the grid points.

However, mesh deformation is not always possible. In particular, if topology changes of the above described type occur or when distortion is too large, a complete remesh of the domain is used. Since mesh generation is costly, this situation should be avoided as often as possible. However, it is typically necessary during the first iterations of the algorithms from Section 3.4, since the updates are large. Especially in that situation the current guess $\underline{\mathcal{B}}_i$ is far from optimal and high accuracy is of minor interest. Hence, it is reasonable to use coarse grids then and refine them during the course of iteration. However, the shape calculus based methods call for a sufficiently amount of nodes on the interface $\underline{\beta}_i$ such that update velocity fields are reliable. This fact constraints the mesh size from above and has to be taken into account when remeshing. In particular, if remeshing cannot be avoided, it should be carried out such that an anew need for mesh generation is unlikely. That is to say, use a smooth interface and ensure that the interface nodes are arranged regularly. Consequently, it is appropriate to use a smoothed spline interpolation of the current interface as input for the mesh generator.

Moreover, small connection components of the active set may occur, which consist of one single node in an extreme case. This is typically caused, when a protuberance is cut off but not completely eliminated

(like in Figure 4.8). For reasons of stability and efficiency such small artificial connection components are deleted: otherwise, unreliable FE approximations would be produced on the one hand if the mesh remained as coarse, and on the other hand a pointlessly fine mesh would have to be generated in order to get a suitable resolution of the very small connection component.[5] If the optimal in-/active set has such small connection components actually an appropriately fine mesh is needed anyway and thus the small components are not small in relation to mesh size any more.

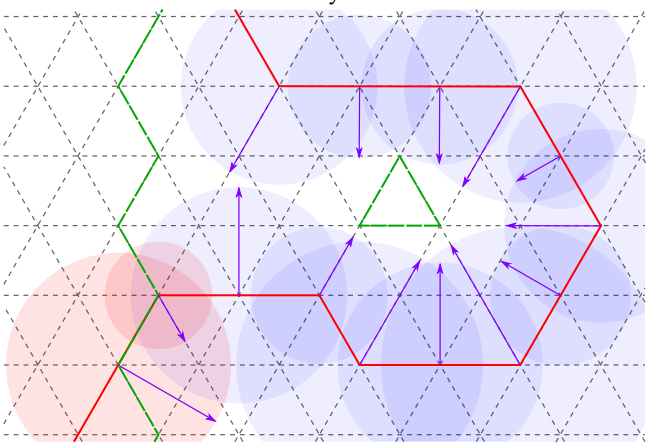


Figure 4.8: Incomplete cut off of a protuberance.

All in all, the implemented mesh update routine roughly works as follows:

- If the normal component $\underline{W}_i \cdot \underline{\boldsymbol{n}}_{\mathcal{J}_i}$ is large or if there is an interface node B whose distortion $(\underline{W}_i(B) \cdot \boldsymbol{n}_{\mathcal{J}_i}(B))\, \boldsymbol{n}_{\mathcal{J}_i}(B)$ is not significantly smaller than $\underline{\kappa}_{\mathcal{J}_i}(B)$ (this means that self-intersection of the interface may occur), Huygens' principle is used for the update. In addition, FFT or smoothing spline methods are applied to smooth the new interface nodes. Finally, a remesh is performed.
- If the normal component $\underline{W}_i \cdot \underline{\boldsymbol{n}}_{\mathcal{J}_i}$ is of moderate size and local self-intersection can be excluded the interface is updated by means of the transformation approach (see Figure 4.3). In addition, FFT or smoothing spline methods are applied to smooth the new interface nodes and a remesh is performed.
- Otherwise a mesh deformation strategy as described above is used. If mesh entanglement occurs one of the other two branches of the routine are on hand as fall-back option.

The second branch is used for several reasons. On the one hand it is more robust than the mesh deformation strategy and thus is a good alternative. On the other hand it is less robust than Huygens' principle, but cheaper (no need for nodal computation of z) and more accurate, since the update does not have to be bigger than the mesh size. Note, that Huygens' principle update can only be applied if the nodes (at least one, to be more precise) are shifted more than one mesh size. Moreover, it should be mentioned that it is a nontrivial task to detect self-intersection of the interface. It has been illustrated that Huygens' principle is capable to cope with such situations, but usage of smoothing methods induce additional difficulties illustrated in Figure 4.9. The smoothed version of a given connection component of the interface may intersect itself, or intersect with another connection component or with the outer boundary $\underline{\Gamma}$. Those different incidents have to be detected and handled adequately.

All those more or less sophisticated ideas are mainly devoted to one goal, namely to increase stability of the implementation of the algorithms. Actually, they help to cope with problems which are related to the lack of global convergence of shape calculus based algorithms. Questions like changing the topology of the active set during the iteration are actually not an issue of those methods. Moreover, the numerical schemes should be initialized with sets that are not only of the right topological type but with sets that actually are "near" the optimal set. The numerical practice shows that most of the discussed issues only occur during the "pseudo-global" phase of iteration and that once the current guess is near the optimal active set everything works fine. In particular, all smoothing and miscellaneous strategies intervene only if they are necessary and they typically do not influence if the current iterate is sufficiently near a critical shape. Nonetheless, the presented coping strategies enable (though not guaranteeing) convergence even if the initial guess is far away from any critical point; cf. paragraphs 4.2.3 and 4.2.6.

[5] Note, that the resolution has to be high enough such that the discrete polygon has no sharp vertices, since it has to mirror the $C^{1,1}$ regularity of the interface adequately.

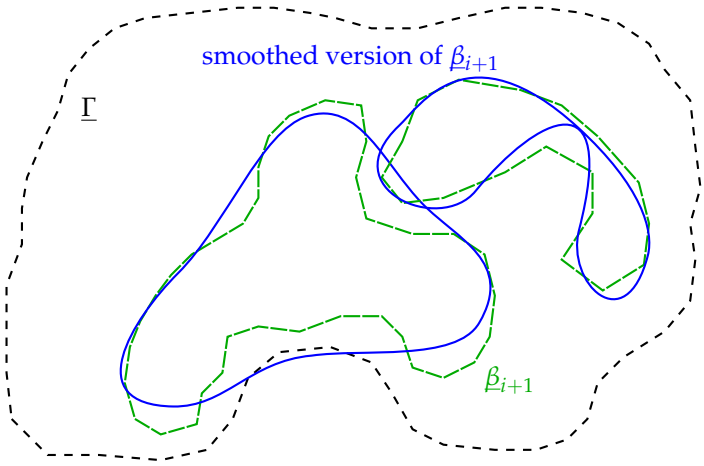


Figure 4.9: Prototypic intersections of the smoothed interface.

4.2 Numerical results

As already indicated in the introducing text of this chapter the numerical results presented in this section deal with the reduced Newton and trial algorithms of paragraphs 3.4.1 only – that is to say, the total linearization method from Paragraph 3.4.3 is not discussed. Moreover, only some selected topics are investigated. However, the claim to demonstrate that the ideas of the chapters 2 and 3 yield efficient solution strategies can be answered.

4.2.1 Test examples

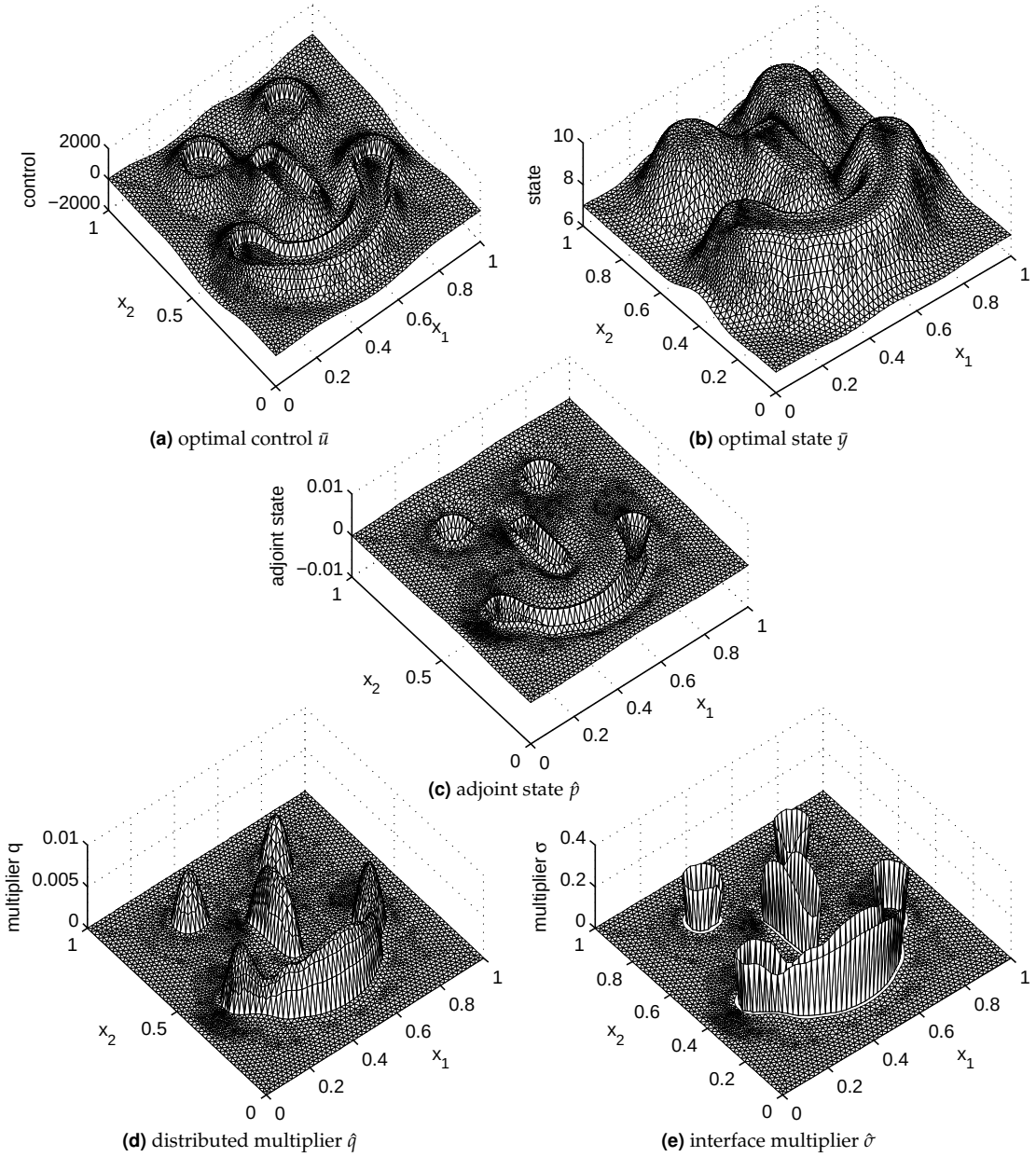
The algorithms are tested on various examples, which are developed by Simon Bechmann. They are constructed such that the regularity assumptions of coefficient functions of model problem (2.1) are fulfilled. Moreover, it is ensured that the optimal active set $\mathcal{A}$ is a strict subset of the candidate set $\mathcal{C}$ (see (3.26)). Especially this property induced considerably difficulties when constructing OCPs, whose optimal variables shall be known analytically. Hence, the corresponding Example 4 is radially symmetric.

The presented test cases are chosen such that each of them inherits a specific difficulty: the active set of Example 1 has four connection components, whereas those of Example 2 is not simply connected. Consequently, these two test cases are predestined to investigate the ability of changes of topology. Example 3 exhibits a very small area of convergence such that stability of the algorithms can be tested. Finally, the optimal variables of Example 4 are known exactly such that convergence can be checked.

Example 1 (Smiley): Despite the requirement of Ω to be a $C^{1,1}$ domain it is chosen to be a unit square. It turns out that this choice has no negative impact, since triangulation of Ω yields a polygonal domain anyway. Moreover, coefficients read as follows.

$$
\begin{aligned}
\Omega &=]0;1[\times]0;1[,\\
\lambda &= 10^{-6},\\
u_d &= 0,\\
y_d &= 7 + 10\Big(\sum_{s\in S_1} e^{-100\,(x-s)^2} + \sum_{s\in S_2} e^{-200\,(x-s)^2}\Big),\\
S_1 &= \left\{\begin{pmatrix}0.3\\0.8\end{pmatrix}, \begin{pmatrix}0.7\\0.8\end{pmatrix}, \begin{pmatrix}0.2\\0.4\end{pmatrix}, \begin{pmatrix}0.8\\0.4\end{pmatrix}\right\}\\
S_2 &= \left\{\begin{pmatrix}0.5\\0.45\end{pmatrix}, \begin{pmatrix}0.5\\0.55\end{pmatrix}, \begin{pmatrix}0.5\\0.65\end{pmatrix}, \begin{pmatrix}0.5\\0.1838\end{pmatrix}, \begin{pmatrix}0.3\\0.2551\end{pmatrix}, \begin{pmatrix}0.7\\0.2551\end{pmatrix}, \begin{pmatrix}0.6\\0.2\end{pmatrix}, \begin{pmatrix}0.4\\0.2\end{pmatrix}\right\},\\
y_{\max} &= 10.
\end{aligned}
$$

The active set $\mathcal{A}$ is smiley shaped and the optimal variables are given by Figure 4.10.



(a) optimal control $\bar{u}$ **(b)** optimal state $\bar{y}$

(c) adjoint state $\hat{p}$

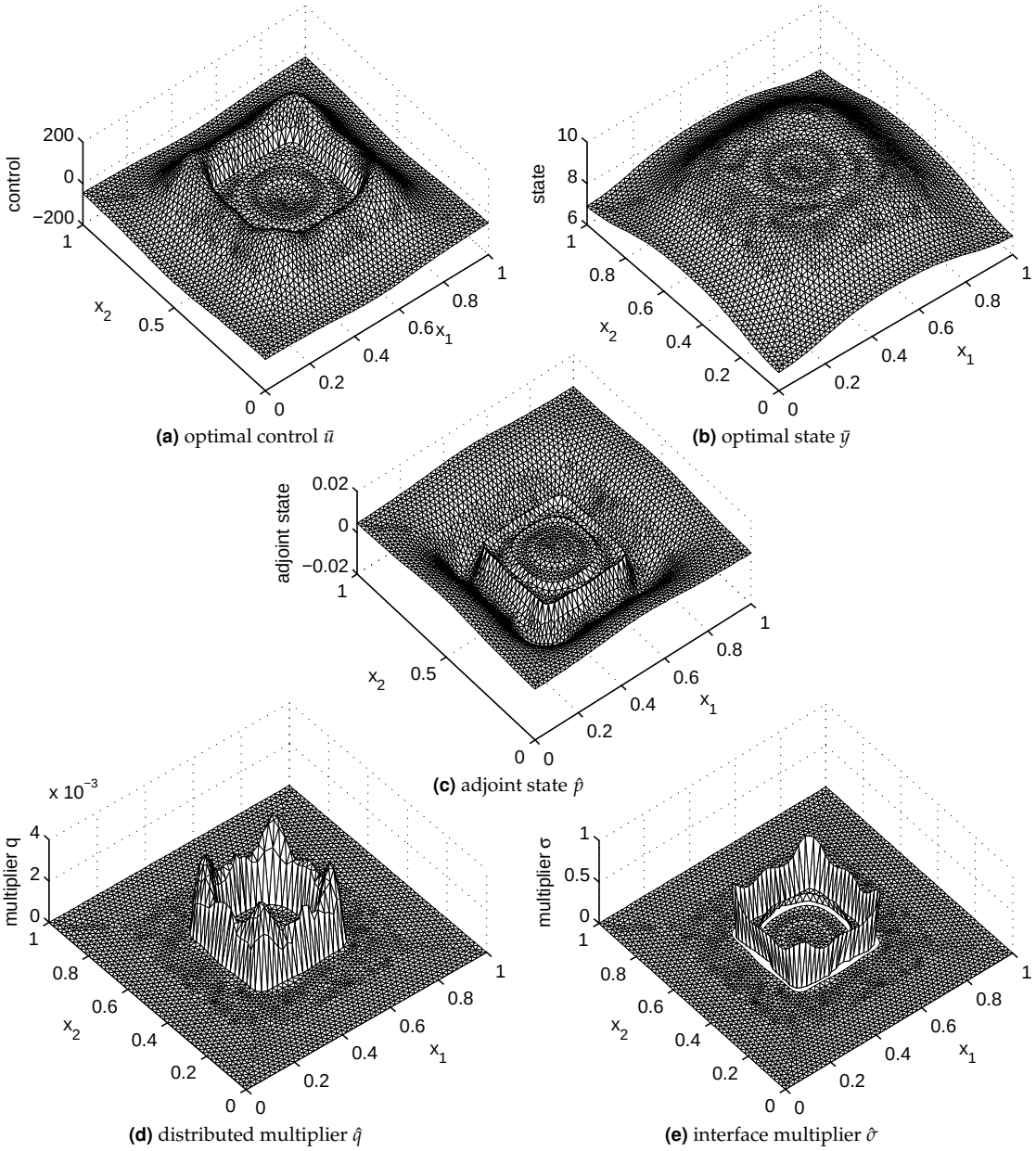
(d) distributed multiplier $\hat{q}$ **(e)** interface multiplier $\hat{\sigma}$

Figure 4.10: Optimal variables of Example 1.

Example 2 (Ring): This OCPs is constructed such that the candidate set $\mathcal{C}$ is simply connected, whereas the optimal active set $\mathcal{A}$ is not. Its coefficients are as follows

$$\begin{aligned}
\Omega &=]0;1[\times]0;1[\\
\lambda &= 10^{-4} \\
u_d &= 0 \\
y_d &= 7 + 10 \sum_{s\in S} e^{-100\,(x-s)^2} \\
S &= \left\{ \begin{pmatrix}0.3\\0.3\end{pmatrix}, \begin{pmatrix}0.3\\0.5\end{pmatrix}, \begin{pmatrix}0.3\\0.7\end{pmatrix}, \begin{pmatrix}0.5\\0.3\end{pmatrix}, \begin{pmatrix}0.5\\0.7\end{pmatrix}, \begin{pmatrix}0.7\\0.3\end{pmatrix}, \begin{pmatrix}0.7\\0.5\end{pmatrix}, \begin{pmatrix}0.7\\0.7\end{pmatrix} \right\} \\
y_{\max} &= 9.5\,.
\end{aligned}$$

The active set $\mathcal{A}$ is annular and the optimal variables are given by Figure 4.11.



(a) optimal control $\bar{u}$ **(b)** optimal state $\bar{y}$

(c) adjoint state $\hat{p}$

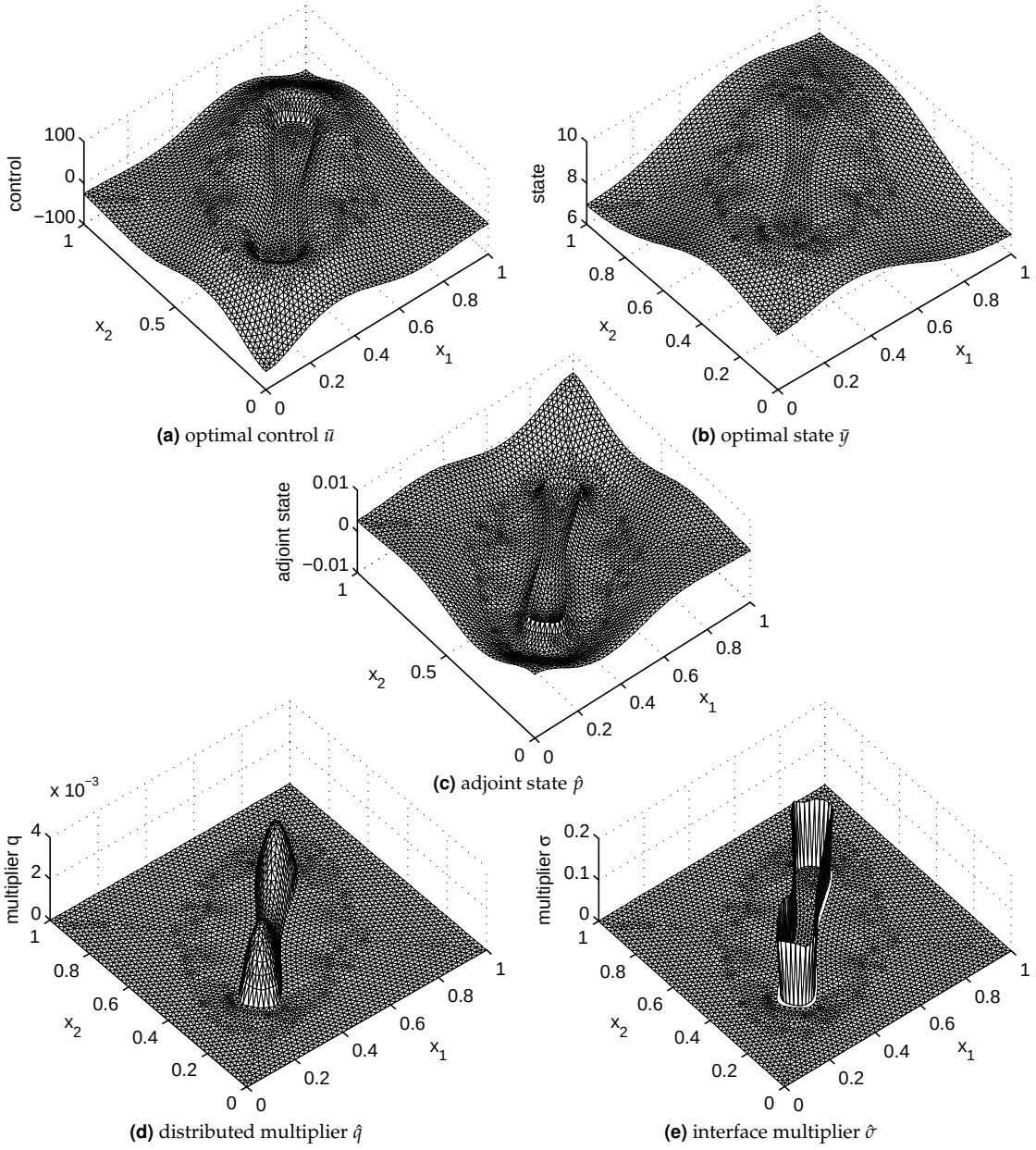
(d) distributed multiplier $\hat{q}$ **(e)** interface multiplier $\hat{\sigma}$

Figure 4.11: Optimal variables of Example 2.

Example 3 (Bone): This OCPs is situated in a unit square as well. Its coefficient are defined as follows.

$$\begin{aligned}
\Omega &=]0;1[\times]0;1[\\
\lambda &= 10^{-4} \\
u_d &= 0 \\
y_d &= 7 + 4 \sum_{s \in S} e^{-10\,(x-s)^2} \\
S &= \left\{ \begin{pmatrix} 0.3 \\ 0.3 \end{pmatrix}, \begin{pmatrix} 0.7 \\ 0.7 \end{pmatrix} \right\} \\
y_{\max} &= 9.5
\end{aligned}$$

The shape of the active set $\mathcal{A}$ is reminiscent of a bone and the optimal variables are illustrated in Figure 4.12.



(a) optimal control $\bar{u}$ **(b)** optimal state $\bar{y}$

(c) adjoint state $\hat{p}$

(d) distributed multiplier $\hat{q}$ **(e)** interface multiplier $\hat{\sigma}$

Figure 4.12: Optimal variables of Example 3.

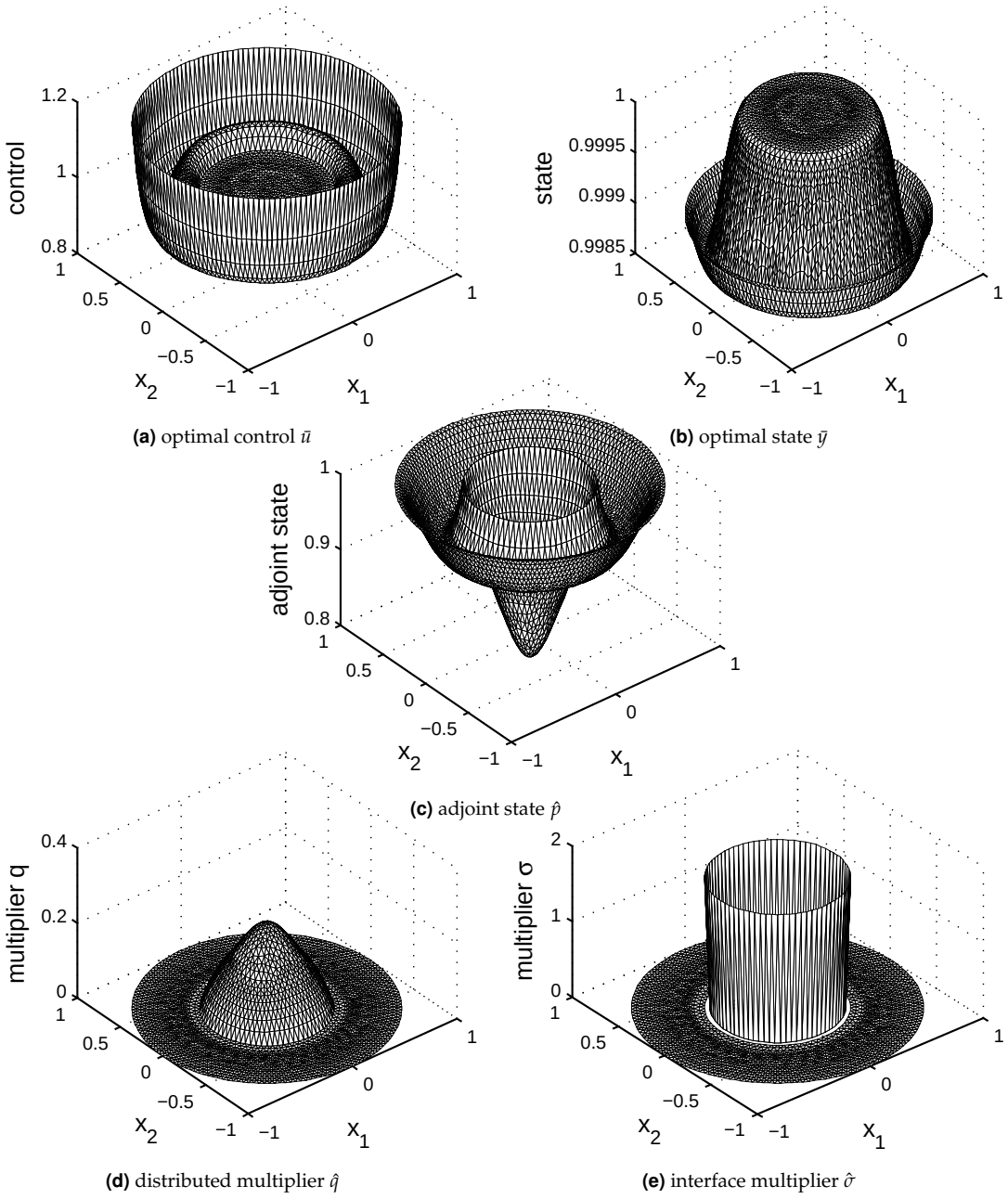
Example 4 (analytic): As already indicated it is beneficial to test the algorithms with examples whose optimal solutions are known exactly. Due to the specific requirements of the coefficients the example depicted here is rotationally symmetric and henceforth the domain Ω is a unit circle. The coefficients can be derived by means of sophisticated ideas and are polynomial.

$$\Omega = B_1(0)$$
$$\lambda = 1$$
$$u_d = \begin{cases} 2, & \text{if } |x| < 0.5 \\ -\frac{1}{6000\,|x|}\Big(4080\,|x|^7 - 13088\,|x|^6 - 131960\,|x|^5 + 344880\,|x|^4 \\ \qquad -299085\,|x|^3 + 105850\,|x|^2 - 22731\,|x| - 970\Big), & \text{if } |x| \geq 0.5 \end{cases}$$

$$y_d = \begin{cases} \frac{1}{3}\Big(64\,|x|^7 - 3232\,|x|^5 + 52\,|x|^4 + 2400\,|x|^3 - 832\,|x|^2 + \frac{93}{4}\Big), & \text{if } |x| < 0.5 \\ -\frac{1}{6000\,|x|}\Big(4080\,|x|^7 - 13088\,|x|^6 + 14920\,|x|^5 - 30320\,|x|^4 \\ \qquad +59635\,|x|^3 + 168970\,|x|^2 - 228191\,|x| + 48000\Big), & \text{if } |x| \geq 0.5 \end{cases}$$

$y_{\max} = 1.$

The optimal active set $\mathcal{A}$ is known to be a circle with radius 0.5: $\mathcal{A} = B_{0.5}(0)$. The optimal variables are illustrated in Figure 4.13.



(a) optimal control $\bar{u}$ **(b)** optimal state $\bar{y}$

(c) adjoint state $\hat{p}$

(d) distributed multiplier $\hat{q}$ **(e)** interface multiplier $\hat{\sigma}$

Figure 4.13: Optimal variables of Example 4.

4.2.2 Accuracy of detecting the active set

It turns out, that the shape calculus based algorithms are capable to detect the active set with an accuracy that is some factors higher than the mesh size. In order to get an impression of the possible accuracy the stopping criterion of Algorithm 1 is omitted while solving the analytic Example 4. The results for different mesh sizes h is presented in Figure 4.14. On the left hand side the median of size of the Newton

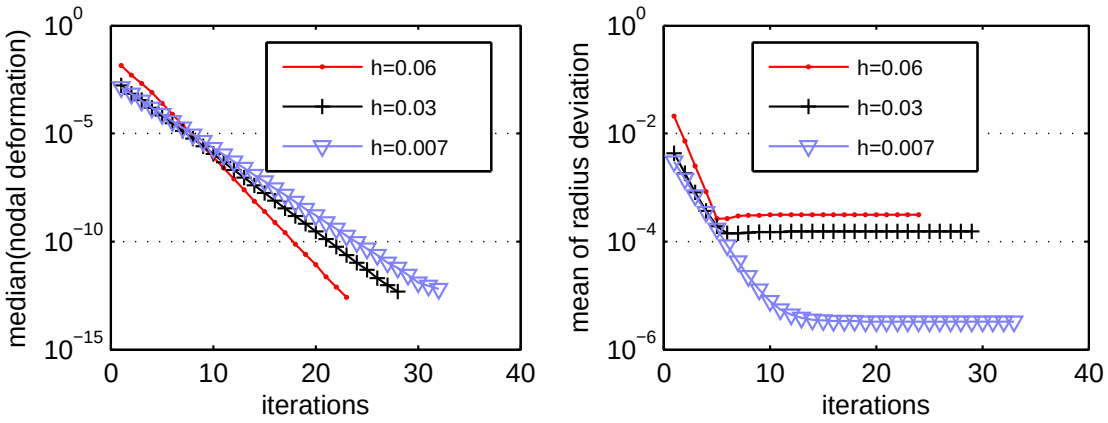


Figure 4.14: Accuracy of detecting the optimal active set.

update is plotted against the iterations; that is to say, how far the interface nodes are displaced from one step to the next. One recognizes that the deformation has a linear convergence rate and that it is not bounded from below by the mesh size. On the right hand side the mean of the deviation of interface nodes from the radius of the optimal set ($R = 0.5$)

$$\text{mean}_{i \text{ is interface node}} \big| |x_i| - 0.5 \big|$$

is plotted against the number of iterations. Obviously the accuracy of the approximation of the active set increases with smaller mesh sizes, but stagnates in each case – in contrast to the deformation. This is due to the fact that the null of the discretized shape gradient is not obtained when the interface nodes are all located on the exact interface, but somewhat outward the analytic active set. Otherwise the discretized active set would systematically underestimate the size of the continuous set; cf. Figure 4.15.

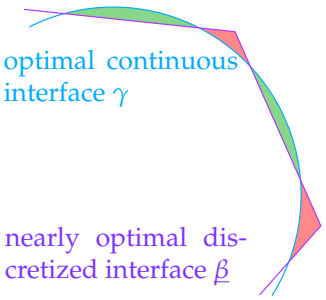


Figure 4.15: Vertices of the interface polygon lie outside of the active set $\mathcal{A}$.

4.2.3 Stability and area of convergence

It turns out that the trial Algorithm 3 is much more stable than the Newton Algorithm 1. Actually the Newton scheme converges only if the initial guess is very close to the optimum. In marked contrast the trial algorithm is stable and converges even if the initial guess for the active set is far away from the optimal shape or even has a different topology; cf. Figure 4.16 and Paragraph 4.2.6. Moreover, the figures illustrate that this method is able to make considerable progress in moving the interface in one iteration.

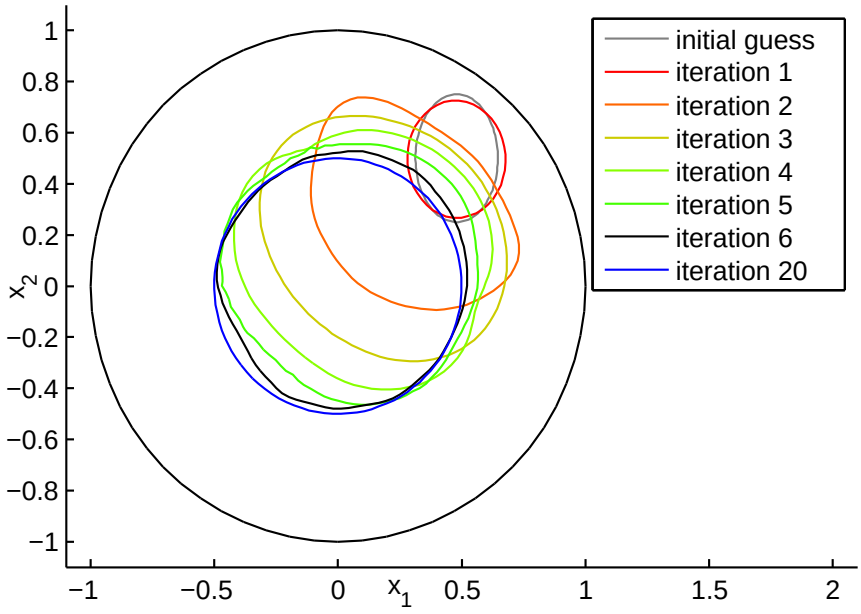


Figure 4.16: Progress of the interface at different iterates.

The best results with respect to stability and progress are obtained when the original trial equation (3.30) is substituted by the second version of (3.34). Likewise the Newton scheme profits form the ideas presented in the 8th item of the discussion on page 110. It turns out that using (3.28) as Newton equation is the best choice.

Though a profound analysis of stability and convergence of trial and Newton algorithms is beyond the scope of this thesis, their behavior shall be illustrated for a specific situation here. The analytical test Example 4 possesses (at least) two different critical shapes, namely the optimal active set $\mathcal{A} = B_{0.5}(0)$ and another set $\mathcal{B}^* = B_R(0)$ whose radius R is approximately 0.76 but not known exactly. To get an impression of the shape functional, its value and its first and second order covariant derivatives are plotted in Figure 4.17 for a one dimensional path through $\mathcal{O}$. At this, the three entities are computed for guesses

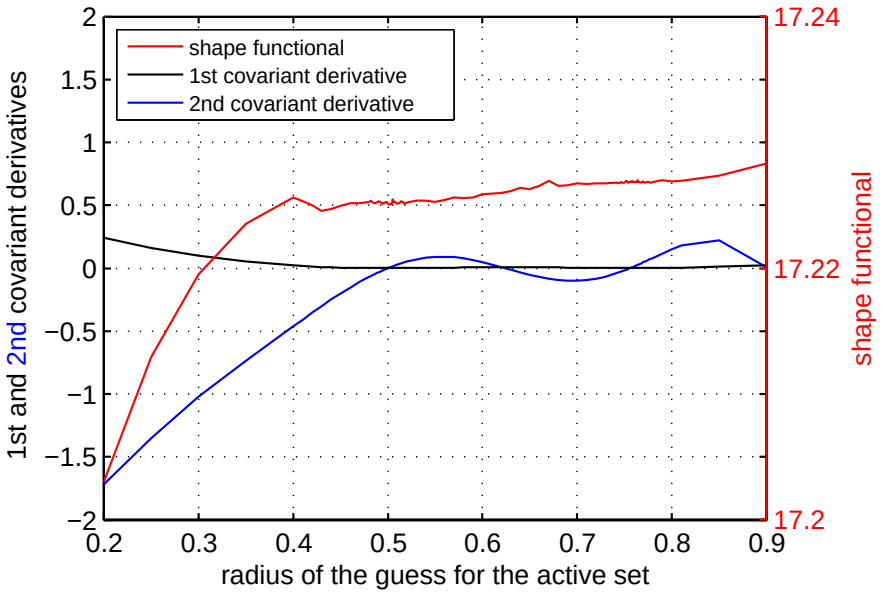


Figure 4.17: Behavior of the shape functional and its covariant derivatives along a one dimensional path through $\mathcal{O}$.

of the active set of type $\mathcal{B} = B_r(0)$, $r \in [0.2; 0.9]$ and the semiderivatives are given as $\mathrm{D}\mathcal{F}(\mathcal{B})[\boldsymbol{n}_\mathcal{B}]$ and $\nabla^2\mathcal{F}(\mathcal{B})[\boldsymbol{n}_\mathcal{B}, \boldsymbol{n}_\mathcal{B}]$. Hence, they can be seen as one dimensional and real-valued functions of the variable r and the latter two can indeed be regarded as (ordinary) derivatives of the first one. Obviously, the approximation of $\mathcal{F}$ is less accurate than those of the derivatives. This is due to the fact that evaluation of the shape functional requires integration on the whole triangulated domain $\underline{\Omega}$ which is a more mesh sensitive procedure than the evaluation of the derivatives, since the latter requires integration on the precisely determined interface only. One recognizes two double zeros of the first semiderivative at $r = 0.5$ and $r \approx 0.76$ which are accompanied by nulls of the second derivative. These are presented in Figure 4.18 once again. In order to understand the behavior of the trial and the Newton algorithms, their

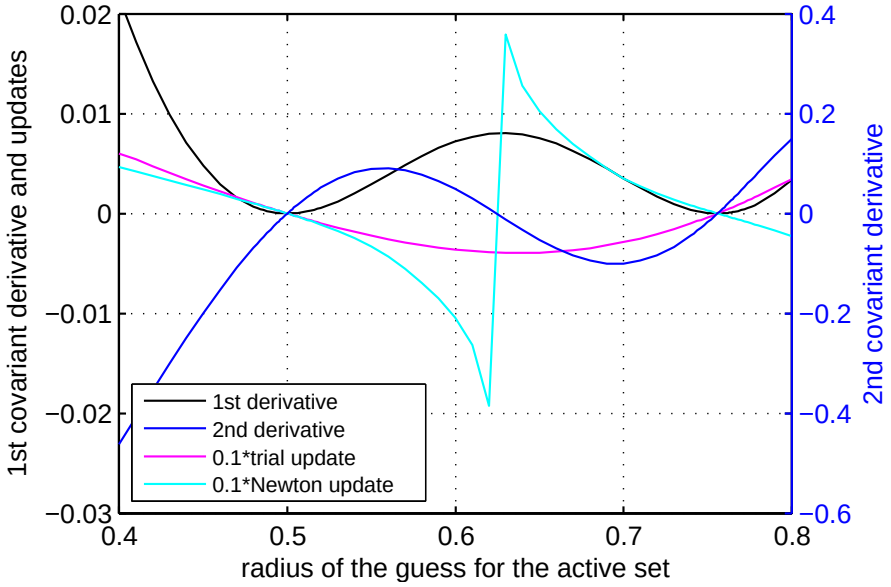


Figure 4.18: Behavior of the covariant derivatives of the shape functional and the deformation of the interface near the two critical points.

updates (scaled by the factor 0.1) are plotted as well. These graphs have to be read as follows: the value for a given radius tells how the radius of the guess for the active set would be modified in the current iteration. It is apparent that the trial algorithm reduces the radius for any guess whose radius lies between the two nulls. Hence, it converges to the optimal active set if its initial guess $B_r(0)$ has a radius smaller than 0.76. In other words, it succeeds even if it starts arbitrarily near the second critical radius. In contrast – as one would expect – the Newton scheme converges to the nonoptimal critical shape $\mathcal{B}^*$ if the initial guess has a radius bigger than approximately 0.63, where the second order shape derivative changes the sign. Moreover, the Newton update is very unreliable if the initial guess has a radius in the neighborhood of 0.63. For one thing the proposed update can be very large and for another thing it may have different sign for each interface node. Note in this respect, that the update of the algorithm is not one dimensional in practice, since the displacement of the interface nodes is not perfectly coupled. It may happen that some interface nodes are contained in $B_{0.63}(0)$ whereas some others are not, and consequently these two groups are moved in opposite direction such that the original shape of an approximate circle is completely destroyed.

As a result the trial algorithm is used as long as the current guess for the active set is far away from a critical shape and the Newton algorithm is applied only if the size of the update comes below a suitable threshold.

In addition, Figure 4.18 shows that the updates of the trial and the Newton algorithm converge to each other, when the active set approaches the optimal configuration with radius $r = 0.5$. Hence, the assertion that the local shape derivatives vanish at the optimum (see Corollary 5) is confirmed, since Newton and trial update just differ in this term; see Paragraph 3.4.2.

4.2.4 Convergence rate

A two-level strategy of using a trial algorithm first and then switching over to a Newton scheme has two advantages. For one thing one obtains higher stability (cf. Paragraph 4.2.3) and for another thing one profits from the higher speed of convergence of the Newton method. A prototypic result is given by Figure 4.19. The trial Algorithm 3 is started with the candidate set as initial guess for the test Example 1. Initially, the maximum of the update, i. e. $\max\{|\underline{W}_i(B) \cdot \underline{\boldsymbol{n}}_{\mathcal{J}_i}(B)| \,|\, B \text{ is interface node}\}$, is reduced during

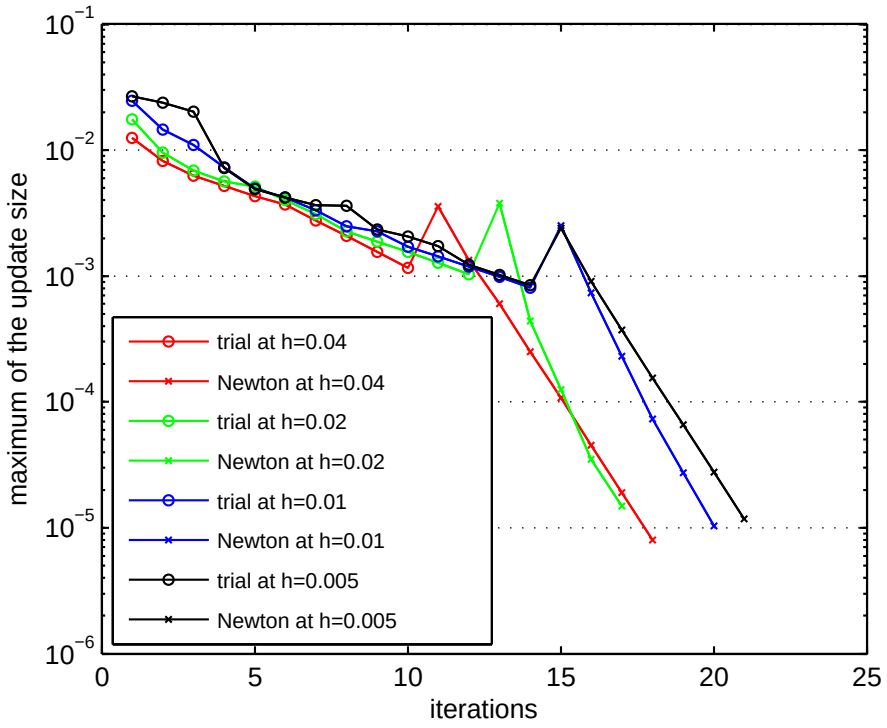


Figure 4.19: Convergence of trial and Newton scheme for Example 1 at different mesh sizes.

the trial iteration. If this update size comes below the value $2 \cdot 10^{-3}$, the Newton scheme is used. The decay of the step size is considerably faster then, though an superlinear convergence rate cannot be verified. The latter finding is due to several reasons.

- The second order shape derivative is singular at the optimal configuration; see Paragraph 2.5.1. Suitable workarounds to deal with that difficulty have been investigated by Griewank and Osbourne [68] and Griewank [67], but they are not applied here.
- Deforming the mesh (or even worse construction of a new mesh) from one iterate to the next implies changing the discrete optimization problem, and respectively the discrete optimality system. Consequently, the Newton algorithm gets a (slightly) modified objective in each iteration.
- The implementation does not care about specific terms which come into play when using polygons instead of $C^{1,1}$ curves; see the introducing text of Section 4.1.
- The discretization error may be dominant, since the movement of the interface is a fraction of the mesh size only.

Each of those reasons deserves further investigation, which is not contained in this thesis.

4.2.5 Mesh (in-)dependency

It is a nontrivial task to analyze whether the shape calculus based algorithms from Section 3.4 behave mesh independent. Since those algorithms are not globally convergent, it is necessary to start them with an initial guess which is inside the area of convergence which itself is not known (even not in retrospect when an algorithm has terminated successfully). Moreover, implementation has a major impact. In particular, the treatment of deforming the interface plays an important role; cf. paragraphs 4.1.2 and 4.1.3.

It may happen that the algorithm uses the mesh deformation strategy, but cannot reach the stopping criterion since mesh quality is not sufficient. Hence, a remesh is indicated but induces some distortion such that the current guess of the active set is deteriorated. As a result, the algorithm needs some extra iterations, which are due to implementational aspects only. It is expected that only finite element methods which move the interface through a fixed mesh (as for instance unfitted FEM) can exhibit the full capability of the presented approaches.

Nonetheless, the considered implementation exhibits a mesh independent behavior when the initial guess is good enough. For instance the variational relaxation approach due to the function $\mathfrak{K}_p^\tau$ (see (3.32)) enhanced with the informal method of canceling one $p_{\mathcal{J}} - p_{\min}^{\max}$ factor (cf. the 8th item of the discussion on page 110, in particular the Newton equation (3.28)) yields the following number of iterations.

mesh size	0.06	0.05	0.04	0.03	0.02	0.01	0.007
iterations starting at radius 0.4	27	5	6	5	6	7	6
iterations starting at radius 0.66	5	7	5	5	5	7	5

Table 4.1: Numbers of iterations needed to reach the stopping criterion for Example 4: displacement of the interface nodes is less than $1.5 \cdot 10^{-2}$.

In particular, the 27 iterations needed to converge for mesh size 0.06 when starting from a circle with radius 0.4 are due to the abovementioned implementation aspects. For coarse meshes even remeshing is not very successful to obtain higher accuracy and hence several remeshing/converging cycles have to be performed up to a point where the mesh generator accidentally produces an appropriate mesh. Fortunately this behavior can only be observed for large mesh size where the resolution of the interface is insufficient.

Another confirmation for a mesh independent behavior can be found in Figure 4.19. Although the overall algorithm tends to require more iterations when the mesh size gets smaller, the decay of the Newton algorithm is very similar for the different mesh sizes. Actually, the algorithm switches to the Newton scheme at approximately the same guess for the active set for the different mesh sizes. Consequently, the different numbers of iterations are mainly due to the bad initial guess, which is not located inside the area of convergence, what calls for stabilization.

4.2.6 Changes of topology

As already indicated several times, shape calculus based algorithms cannot be expected to be able to change the topology of the current guess of the active set in the course of iteration. Nevertheless, there are four criteria which help to change topology

- intersection with candidate set $\Rightarrow$ remove connection component of $\mathcal{B}$
- checking the state constraint $\Rightarrow$ add connection component of $\mathcal{B}$
- checking the sign conditions of multipliers $\Rightarrow$ remove connection component of $\mathcal{B}$
- self-intersection of the interface $\Rightarrow$ both remove and add connection component of $\mathcal{B}$.

These different types of topology changes are illustrated in more detail within this paragraph. An intersection with the candidate set is responsible for the change from the initial guess to the first iteration in Figure 4.20. The connection component of the left eye is lost after the second iteration, since it gets to small and hence is deleted mistakenly. The interfaces is evolved up to iteration 47 then, where the stopping criterion is met the first time. However the current state is not below the upper state constraint, which is why additional components of the active set are appended (iteration 48). Afterwards the Newton loop restarts and within the next six iterations the optimal configuration is found.

Figure 4.21 shows another change of the topology of the active set. After the first four iterations, a candidate optimal set is found, such that the a posteriori criteria are checked. In particular, the sign conditions of the multipliers are not fulfilled, such that a hole is inserted to the active set where the sign condition is not valid (iteration five). Finally, the algorithm converges within additional five iterations.

Topology changes due to self-intersection of the interface can be seen in Figure 4.22. The initial guess consists of two separate connection components which are moved towards each other. After iteration nine the interfaces intersect each other such that the two connection components are unified in iteration

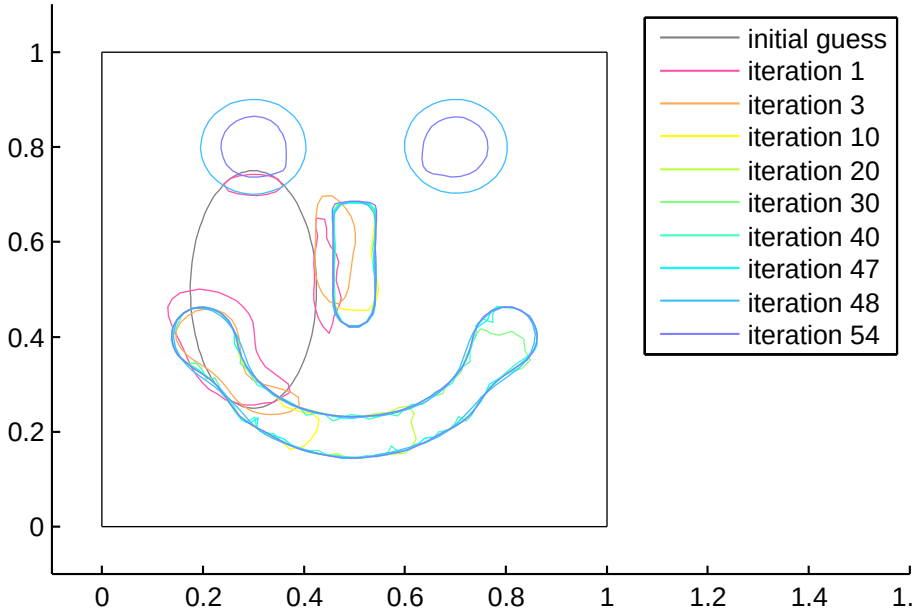


Figure 4.20: Changes of the topology due to the candidate set and the state constraint.

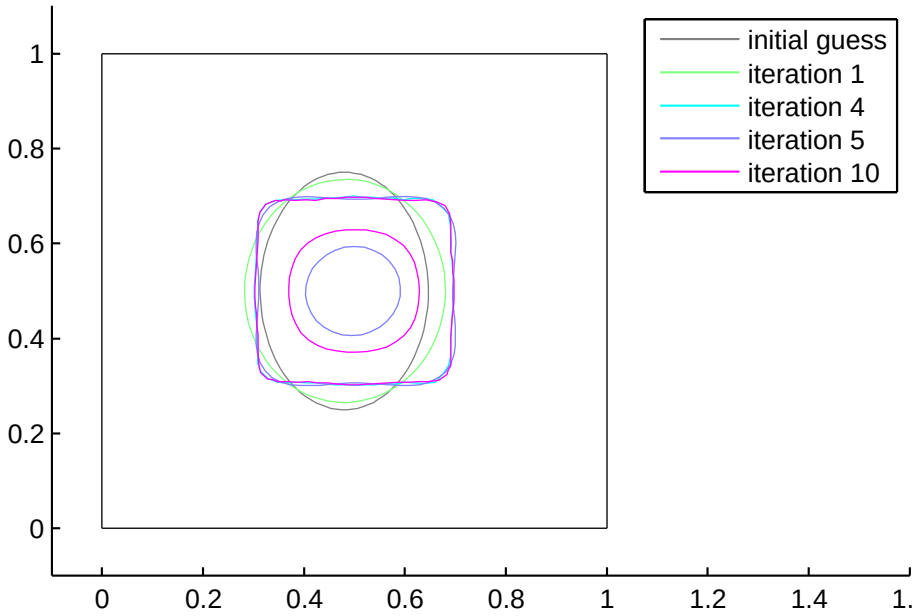


Figure 4.21: Changes of the topology due to the sign conditions of the multipliers.

ten and the current guess is now "u-shaped" (simply connected). After some additional iterations the 'u'-tips of the set intersect such that the active set consists of one not simply connected component at iteration 15. The inactive set inclusion is diminished and is finally eliminated at iteration 20.

All in all, the algorithms are capable to cope with changes of topology during iteration, but there are configurations where their iterations stagnate and an indispensable change of the topology is not performed. This happens typically when the interface of two connection components approach each other such that they are separated by very view mesh layers only. The finite element approximation may get poor then since on the one hand there are only view degrees of freedom concentrated in such gaps and on the other hand the functions may tend to large curvature there. Otherwise, if the functions behave well, the residuum in the defect equation (this is shape gradient or merit functional equals zero) is typically very small in the approach area and hence the update (i. e. movement of the interface) converges to zero there.

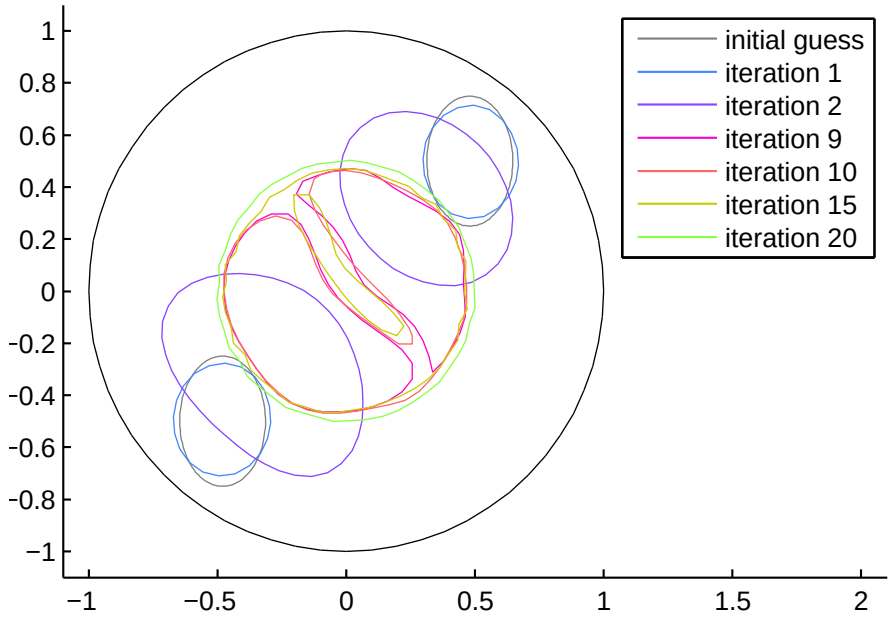


Figure 4.22: Changes of the topology due to self-intersection and small components.

4.2.7 Comparison with primal-dual active set methods

In order to assess the performance of the shape calculus based algorithms, different comparing tests with a Moreau-Yosida regularized primal-dual active set strategy, which is equipped with exact path-following (PDAS-EPF) (see [86, 87]), are run. The results are composed in Table 4.2. Here the two-level

	mesh size	0.04	0.02	0.01	0.005
Example 1	Trial/Newton	14	15	18	22
	(PDAS-EPF)	16	20	25	34
Example 2	Trial/Newton	9	15	16	26
	(PDAS-EPF)	24	29	37	39
Example 3	Trial/Newton	32	10	67	21
	(PDAS-EPF)	22	26	32	39

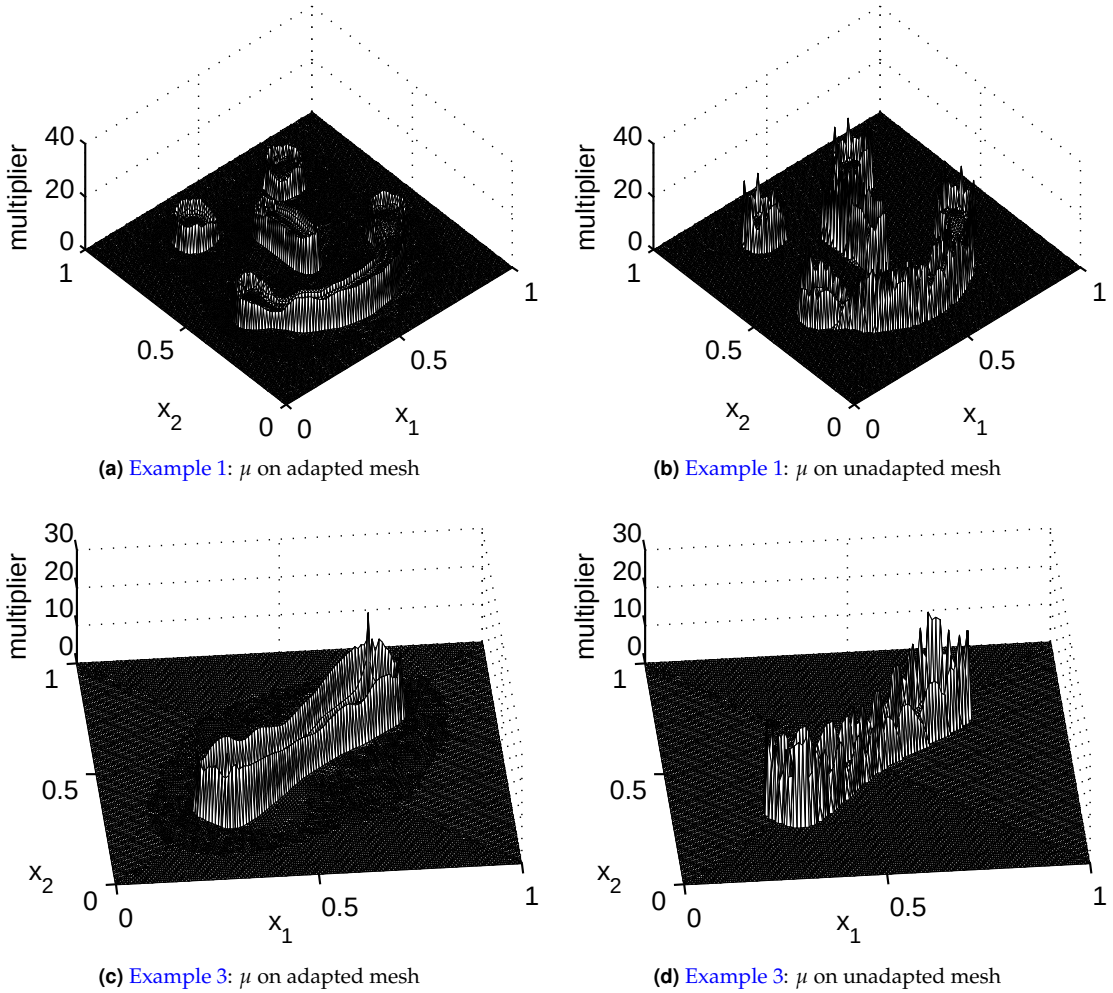
Table 4.2: Number of iterations needed to converge for different mesh sizes.

strategy of using Algorithm 3 (with second version of the simplified trial equation (3.34)) as pseudo-globalizer for the Newton scheme from Paragraph 4.2.4 is applied with a mesh dependent stopping criterion. In particular, the iteration stops if the Newton update (this is the maximum of the nodal displacement of the interface) is smaller than 10^{-2} times the mesh size. The (PDAS-EPF) is stopped if two subsequent iterations yield the same active nodes or if some more sophisticated criteria hold in order to prevent additional iterations caused by degeneracy. Basically the same implementation as in [10] was used. To guarantee comparability of the results, the (PDAS-EPF) uses the final mesh produced by the shape calculus based algorithm. It turns out that both algorithms always end up with the same set of active nodes, which is a reliable hint that shape calculus based algorithms indeed do converge to the right active set in more complex situations than that of analytical test examples; see Paragraph 4.2.2.

Both algorithms exhibit a moderate mesh dependent behavior. This is due to the mesh size dependent stopping criterion with respect to the trial/Newton scheme. The surprisingly high numbers for Example 3 at mesh size 0.04 and 0.01 are consequences of prototypic problems: when using the bigger mesh size the curvature of the interface is too high at some interface nodes in order to get a proper mesh, and when using the smaller mesh size convergence is slowed down since the criterion for using the mesh

deformation instead of remeshing is not sharp enough. Moreover, the (PDAS-EPF) needs the more iterations the smaller the mesh size gets, since finer meshes allow for advanced path-following.[6]

Besides comparability of the algorithmic results, using the final mesh produced by means of the trial enhanced Newton algorithm has another interesting consequence. The essence of those meshes is a very good approximation of the interface by means of a polygon. This quality cannot be expected from an apriori generated mesh that is not adapted to the optimal active set. Due to that property the singular part μ_γ of the Lagrange multiplier has a much more regular appearance than on typical meshes; cf. Figure 4.23.



(a) Example 1: μ on adapted mesh

(b) Example 1: μ on unadapted mesh

(c) Example 3: μ on adapted mesh

(d) Example 3: μ on unadapted mesh

Figure 4.23: Comparison of the multiplier μ on different meshes.

The analysis of Paragraph 3.5.1 revealed, that the singular part μ_γ of the multiplier is blurred by means of Moreau-Yosida regularization. This behavior can be exploited in order to use (PDAS-EPF) as globalizing algorithm for shape calculus based methods. Namely solving a highly or moderately regularized approximation of the original problem by applying the PDAS yields a rough but reliable determination of the active set. In particular, the right topology is typically obtained within very view iterations. The guess of the active set obtained this way is an excellent initial guess to start a shape calculus based algorithm. The combination of the algorithms yield the results which are composed in Table 4.3. For this purpose the path-following strategy was terminated after two iterations which yielded regularization parameters less than 100 (the parameter started with value 10), which has to be interpreted as strongly regularizing. A comparison with the pure trial/Newton algorithm (see Table 4.2) reveals a similar number of iterations.

[6] Note, that it is not reasonable to apply additional path-following when the regularization error gets smaller than the discretization error, see [84]. Hence the applied path-following strategy is mesh size dependent.

	mesh size	0.04	0.02	0.01	0.005
Example 1	trial/Newton	$7+16$	$8+13$	$9+13$	$9+15$
Example 2	trial/Newton	$8+14$	$8+13$	$8+19$	$8+20$
Example 3	trial/Newton	$8+6$	$9+42$	$8+9$	$9+10$

Table 4.3: Number of iterations needed by (PDAS-EPF) and trial/Newton together for different mesh sizes.

Hence, the performance of the combined algorithms is roughly the same, but stability is considerably higher.

CHAPTER 5

Conclusions and Outlook

Bryson-Denham-Dreyfus approach

This work basically presents how to obtain a new kind of first order necessary conditions for the state-constrained elliptic model problem (2.1). It is motivated by the so-called Bryson-Denham-Dreyfus approach, which aims at revealing a control law, that is induced by the state constraint. The corresponding general recipe known in theory of OC-ODE is transferred to the elliptic problem under consideration. At this, different reasonable, equivalent choices are suggested, among which one specific is pursued in detail. However, the presented concept rather remains on the level of a heuristic, as long as more comprehensive understanding of the underlying connection to the theory of partial differential-algebraic equations is not available. In particular, frequently encountered pointwise constraints on the (Euclidean) norm of the gradient

$$|\nabla y|^2 = (\partial_{x_1} y)^2 + (\partial_{x_2} y)^2 \leq y_{\max}$$

pose an (unsolved and unattained) difficulty. It seems to be adequate to rewrite the second order elliptic state equation (2.1b) as a system of first order

$$\begin{aligned} -\operatorname{div} z + y &= u \quad \text{a. e. in } \Omega, \\ \nabla y &= z \quad \text{a. e. in } \Omega. \end{aligned}$$

For one thing, this reasoning is quite standard in the field of OC-ODE, where the idea of the BDD approach was invented, and for another thing one gets a more direct access to the state constraint then. However, the non-linearity of the constraint prohibits a direct computation of a control law.

Moreover, the control u acts on the boundary Γ of the domain Ω in probably most of the real-world applications of OC-PDE. In this situation differentiation of the state constraint equation on the active set (which is still expected to lie in the interior of Ω) cannot yield a boundary expression; henceforth a control law is out of reach. It might be helpful to apply the concept of virtual distributed control then, which was introduced by Krumbiegel and Rösch [110].

Another crucial point is concerned with the restrictive Assumption 1. It is assumed throughout this work that the boundaries of the active set are required to be smooth enough in order to apply shape calculus. In particular, the treatment of corners is excluded here, although associated difficulties may reenter the considerations by means of discretization. Moreover, the active set may not contain any lower dimensional component, although it is well-known, that isolated active curves or points do occur indeed. From the theoretical point of view, isolated active point are not an issue, since there is no chance to get information by means of the BDD approach, since differentiation within such components is not possible. However, the situation changes when considering active curves (as long as this curve is a piecewise differentiable submanifold of Ω). It is expected, that one has to distinguish between normal and tangential directions then. Differentiation in tangential direction is possible by means of tangential calculus, and one might obtain hidden algebraic conditions this way. With respect to normal directions on the curve one probably reobtains the results of Bergounioux and Kunisch [14, Thm. 5].[1] Any starting- and endpoints of the active curve are expected to be the most challenging detail, since tangential calculus

[1] Note, that shape calculus typically deals with shapes, whose boundaries are (sufficiently smooth) submanifolds with codimension one. In particular, any vector field, which is defined on these boundaries can be decomposed into a normal and the tangential

fails there. Intuition tells, that the Lagrange multipliers possess an additional Dirac measure there. This view is encouraged by a result of Rund [144, Satz 3.4.13].

From this perspective, the analysis of the BDD approach in the context of optimal control of parabolic PDEs is appealing and interesting. On the one hand, one can restrict the analysis to simple OCPs, where spacial symmetry ensures a spacial predetermination of the active set. Consequently, one can focus initially on the behavior of the different variables at starting and endpoint of the active set, which is very close to the theory of state constrained OC-ODE, when the parabolic PDE is regarded as a formal system of ODEs. In particular, the well-established knowledge of the topological possibilities of the active set (i. e. contact point or boundary arc in the way of speaking in OC-ODE), which is determined by means of the order of the state constraint, could be used then. This might be a first step towards classifying the active set by means of the order of the state constraint in the context of OC-PDE. It seems natural to introduce different notions of order of the state constraint with respect to time and space. In a next step, the more complex situation could be investigated, where the spacial spread of the active set may vary in time. However, this is expected to be very challenging, since on the one hand the efficient numerical treatment of state-constrained parabolic OCP still faces fundamental problems, which a due to limitation of memory, and on the other hand one has to recourse to time dependent shape calculus then, since one deals with transient problems.

Time optimal control of PDEs

The presented approach of treating the active set, which is associated with the state constraint, as an equal variable strongly parallels the treatment of time optimal control of PDEs; see [138, 80, 116, 99, 112]. There one tries to control a time dependent process, which is modeled by parabolic or hyperbolic PDEs, to a prescribed final state in minimal time. Henceforth, the exact shape of the space-time cylinder is to be found. Obviously, the set of admissible shapes is a one parameter set here. Consequently, from the perspective of shape calculus this type of problem is much easier than finding the optimal active set of an elliptic OCP. Optimization with respect to the topology in particular is not an issue there. Nonetheless, time optimal control can be regarded as another representative of set optimal control and optimization on a vector bundle. It should be noted, that the frequently applied (nonlinear) transformation of the optimal control problem with free end-time to a fixed time interval is essentially the same as the idea of function space parametrization, which was considered in Paragraph 2.6.2. In particular, the additional condition, which is required to compensate the additional variable of free end-time, is the analog of the interface BDD condition, which is necessary to fix the shape of the active set.[2] Moreover, the derivative of the Lagrangian/Hamiltonian (or the time-parametrized minimal value functional) with respect to the parameter associated with the free end-time yields a necessary condition (often called *transversality condition*), which is the perfect analog to the weak continuity condition across the optimal interface for the control in the presented approach.

Set optimal control and optimization on vector bundles

The considerations of this work lead to a new type of optimization problems, which was called set optimal control. It is a class of hybrid problems, which contain elements of shape/topology optimization, since a set variable occurs, and which are optimal control problems, since a function space control is involved, which determines a state. Later on, this class is strongly generalized to optimization on vector bundles.

Although this is a very general framework, there is hope that this perspective is valuable for optimal control, since many different applications incorporate intrinsic nonlinear behavior, which is due to variables that are not elements of a linear space. Besides introducing the active set as an equal variable or time optimal control, another intrinsic nonlinear behavior is obtained, when angles are used as variables. They are treated typically as elements of a linear space, which yields problems like 2π periodicity and severe non-linearities by means of trigonometric functions. The perspective of optimization on vector bundles suggests to treat angle related rotations in their natural, nonlinear environment, e. g. the sphere S^{N-1} or the special orthogonal group $\mathrm{SO}(3)$. Moreover, there a countless applications, where shape/topology optimization and function space optimal control meet: technical constraints (which can be interpreted as

component. However, this is no possible any more if the active set is a submanifold of Ω with codimension greater than zero, as for instance a curve in $\mathbb{R}^3$. Henceforth, a more comprehensive treatment of lower dimensional active sets at least requires additional work at the fundament of shape calculus – not to speak about a "new shape calculus".

[2] This topic is discussed in more detail in Appendix C.

control and/or state constraints) inhibit a sufficient result, such that a (shape/topological) redesign of the considered structural component is applied. For instance, the design part may be the number, placement and exact shape of a conductor coil of a furnace, whereas the optimal control part may be control of the electric current in order to achieve a certain distribution of temperature. Simultaneous optimization with respect to shape and control might be the all-in-all approach in such applications.

Numerical approach and inherent structure of the Lagrange multiplier

The treatment of the OCP within this thesis enables a new numerical approach. The reformulation of the state constraint by means of the BDD approach yields a natural splitting into a distributed control law and an interface condition. It turns out, that this splitting leads to two Lagrange multipliers which can be associated with the regular and the singular part of the well-known multiplier. For one thing this reasoning emphasizes the PDAE character of the necessary conditions, and for another thing it suggests a numerical treatment, which uses this intrinsic structure of the dual variables algorithmically. Thus, there is no need for regularization, in order to be able to formulate algorithms on the infinite dimensional level. In particular, a shape optimization based reduced approach/algorithm, which can be classified into the middle branch of the illustration within the Introduction 1 (*"black-box approach"*), is presented. Moreover, an *"all-at-one approach"* approach, which fits into the right branch of the illustration, and which is based upon total linearization, is discussed, though not numerically tested.

The algorithms lack a profound convergence analysis yet, but exhibit an encouraging performance in direct comparison with Moreau-Yosida regularized PDAS, which is equipped with an exact path-following scheme. Nonetheless, a more sophisticated handling of finite element discretization (ALE methods, unfitted/extended FEM, etc.) is indicated. Moreover, it is shown, that the algorithms can cope with certain changes of the topology of the active set on the run. However, this is no satisfying substitute for a fully developed theoretical and algorithmic handling of the topology optimization component of the set optimal control problem, which is not attended within this work.

All in all, this thesis is only a small step and the research is open now for further investigations in very different directions.

Appendix

A Results of different Bryson-Denham-Dreyfus approaches

In order to give some insight to the usage of different BDD approaches, the optimality systems of the two additional approaches of Paragraph 2.2.2 are summarized here.[1] At the optimum there holds in either case

$$\begin{aligned}
-\Delta \bar{y}_{\mathcal{I}} + \bar{y}_{\mathcal{I}} &= \bar{u}_{\mathcal{I}} && \text{in } \mathcal{I},\\
-\Delta \bar{y}_{\mathcal{A}} + \bar{y}_{\mathcal{A}} &= \bar{u}_{\mathcal{A}} && \text{in } \mathring{\mathcal{A}},\\
\partial_{\boldsymbol{n}} \bar{y}_{\mathcal{I}} &= 0 && \text{on } \Gamma,\\
\bar{y}_{\mathcal{I}}|_{\gamma} - \bar{y}_{\mathcal{A}}|_{\gamma} &= 0 && \text{on } \gamma,\\
\partial_{\boldsymbol{n}}^{\mathcal{I}} \bar{y}_{\mathcal{I}} + \partial_{\boldsymbol{n}}^{\mathcal{A}} \bar{y}_{\mathcal{A}} &= 0 && \text{on } \gamma,
\end{aligned}
\qquad
\begin{aligned}
-\Delta p_{\mathcal{I}} + p_{\mathcal{I}} &= \bar{y}_{\mathcal{I}} - y_d && \text{in } \mathcal{I},\\
-\Delta p_{\mathcal{A}} + p_{\mathcal{A}} &= \bar{y}_{\mathcal{A}} - y_d && \text{in } \mathring{\mathcal{A}},\\
\partial_{\boldsymbol{n}} p_{\mathcal{I}} &= 0 && \text{on } \Gamma,\\
\lambda\,(\bar{u}_{\mathcal{I}} - u_d) + p_{\mathcal{I}} &= 0 && \text{in } \mathcal{I},\\
\lambda\,(\bar{u}_{\mathcal{A}} - u_d) + p_{\mathcal{A}} + q_{\mathcal{A}} &= 0 && \text{in } \mathring{\mathcal{A}},
\end{aligned}$$

$$\begin{aligned}
-\Delta y_{\min}^{\max} + y_{\min}^{\max} &= \bar{u}_{\mathcal{A}} && \text{in } \mathring{\mathcal{A}},\\
y_{\min} < \bar{y}_{\mathcal{I}} &< y_{\max} && \text{in } \mathcal{I},
\end{aligned}$$

and furthermore

BDD interface condition	adjoint interface condition	facultative shape gradient equations: either $\bar{u}_{\mathcal{I}}\|_{\gamma} - \bar{u}_{\mathcal{A}}\|_{\gamma} = 0$ or
$y_{\min}^{\max}\|_{\gamma} = \bar{y}_{\mathcal{A}}\|_{\gamma}$	$p_{\mathcal{I}} - p_{\mathcal{A}} = 0$ $\partial_{\boldsymbol{n}}^{\mathcal{I}} p_{\mathcal{I}} + \partial_{\boldsymbol{n}}^{\mathcal{A}} p_{\mathcal{A}} = \sigma_{\mathcal{I}}$	$p_{\mathcal{I}}\|_{\gamma} - p_{\mathcal{A}}\|_{\gamma} = 0$ or $q_{\mathcal{A}}\|_{\gamma} = 0$
$\partial_{\boldsymbol{n}}^{\mathcal{A}} y_{\min}^{\max} = \partial_{\boldsymbol{n}}^{\mathcal{A}} \bar{y}_{\mathcal{A}}$	$p_{\mathcal{I}}\|_{\gamma} - p_{\mathcal{A}}\|_{\gamma} = \sigma_{\mathcal{I}}$ $\partial_{\boldsymbol{n}}^{\mathcal{I}} p_{\mathcal{I}} + \partial_{\boldsymbol{n}}^{\mathcal{A}} p_{\mathcal{A}} = 0$	$p_{\mathcal{I}}\|_{\gamma} - (p_{\mathcal{A}}\|_{\gamma} + q_{\mathcal{A}}\|_{\gamma}) = 0$ or $q_{\mathcal{A}}\|_{\gamma} - \sigma_{\mathcal{I}} = 0$
$\partial_{\boldsymbol{n}}^{\mathcal{A}} y_{\min}^{\max} + y_{\min}^{\max} = \partial_{\boldsymbol{n}}^{\mathcal{A}} \bar{y}_{\mathcal{A}} + \bar{y}_{\mathcal{A}}$	$p_{\mathcal{I}}\|_{\gamma} - p_{\mathcal{A}}\|_{\gamma} = \sigma_{\mathcal{I}}$ $\partial_{\boldsymbol{n}}^{\mathcal{I}} p_{\mathcal{I}} + \partial_{\boldsymbol{n}}^{\mathcal{A}} p_{\mathcal{A}} = \sigma_{\mathcal{I}}$	$p_{\mathcal{I}}\|_{\gamma} - (p_{\mathcal{A}}\|_{\gamma} + q_{\mathcal{A}}\|_{\gamma}) = 0$ or $q_{\mathcal{A}}\|_{\gamma} - \sigma_{\mathcal{I}} = 0$

The approaches contain different interface conditions within the reformulation of the state constraint. This yields different interface conditions of the adjoint state. Consequently, the adjoint state $p_{\mathcal{A}}$ and the multipliers $\sigma_{\mathcal{I}}$ and $q_{\mathcal{A}}$ are different for each approach, though this is not marked by the notation.

These different interface conditions yield different additive decompositions of the original adjoint state $p_{\mathcal{A}}^{trad}$ into a new adjoint state $p_{\mathcal{A}}$ and a Lagrange multiplier $q_{\mathcal{A}}$. At this, $p_{\mathcal{A}}$ solves the same PDE in each case and this adjoint equation is only dependent on $\bar{y}_{\mathcal{A}}$. In contrast, $p_{\mathcal{A}}^{trad}$ solves an equation which is dependent on $\bar{y}_{\mathcal{A}}$ *and* $\mu_{\mathring{\mathcal{A}}}$, such that it mixes influences of the state equation and the state constraint. Thus, the BDD approach helps to distinguish between the impacts of the state equation and the state constraint.

Moreover, in particular, the BDD ansatz via the Neumann boundary condition yields a multiplier $\sigma_{\mathcal{I}}$, which is determined as the Dirichlet jump of the adjoint state across the interface. In a similar way as in

[1] In the case of the BDD approach that is based upon the Dirichlet boundary condition, the optimality system from Appendix B is used here.

the proof of Corollary 6, it should be possible to prove H^2-regularity of $p_{\mathcal{I}}$ and $p_{\mathcal{A}}$ such that the Lagrange multiplier $\sigma_{\mathcal{I}}$ is in $H^{3/2}(\gamma)$. In view of the 3rd item of the Remarks on page 51, one recognizes again, that differentiation of the primal condition in the BDD ansatz yields higher regularity of the corresponding multiplier.

B Existence of Lagrange multipliers

This section is devoted to prove existence of Lagrange multipliers for the inner optimization problem (2.37) in analogy to Theorem 5 on page 36. However, this goal could not be reached rigorously, and hence some conjectures are necessary.

The proof of the mentioned theorem relies on an equivalent reformulation of the constraints, i. e. (2.40), such that they can be decomposed in two separate/independent parts on $\mathcal{J}$ and $\mathcal{B}$. In consequence of this reformulation, the adjoint states $\bar{p}_{\mathcal{J}}$ and $\bar{p}_{\mathcal{B}}$ are not connected via interface conditions. This is a big advantage, since an assertion in the style of Proposition 4 is not required to claim existence of the adjoint states. Admittedly, it is possible to generalize this result to the situation, where the solutions of a geometrically split BVP has a kink (i. e. a jump in the normal derivative) across the interface between the domains $\mathcal{J}$ and $\mathring{\mathcal{B}}$, which is induced by a $H^{-1/2}$-function; see Proposition 8. But actually an analog result for kinks which are induced by $H^{-3/2}$-functions is required. Since the regularity of BVP solutions are expected to be elements of $L^2(.,\Delta)$, it is not possible to work with variational formulations then. Consequently, in order to prove the corresponding result, one requires other ideas than those which are applied in the proofs of propositions 4 and 8.

Proposition 8 (Unique solvability of an elliptic BVP with a kink in $H^{-1/2}$)**:**
Let $\mathcal{B} \in \mathcal{O}$, where $\mathcal{O}$ is given by Definition 4 and use the notations from Definition 5. Moreover, let $\sigma \in H^{-1/2}(\beta)$, let $f_{\mathcal{J}} \in L^2(\mathcal{J})$ and let $f_{\mathcal{B}} \in L^2(\mathring{\mathcal{B}})$ be arbitrary.
Then the boundary value problem

$$-\Delta v_{\mathcal{J}} + v_{\mathcal{J}} = f_{\mathcal{J}} \quad \text{a. e. in } \mathcal{J}, \tag{B.1a}$$
$$\partial_{\boldsymbol{n}} v_{\mathcal{J}} = 0 \quad \text{a. e. on } \Gamma, \tag{B.1b}$$
$$v_{\mathcal{J}}|_\beta - v_{\mathcal{B}}|_\beta = 0 \quad \text{a. e. on } \beta, \tag{B.1c}$$
$$v_{\mathcal{J}} \in H^1(\mathcal{J},\Delta), \tag{B.1d}$$
$$-\Delta v_{\mathcal{B}} + v_{\mathcal{B}} = f_{\mathcal{B}} \quad \text{a. e. in } \mathring{\mathcal{B}}, \tag{B.1e}$$
$$\partial_{\boldsymbol{n}}^{\mathcal{J}} v_{\mathcal{J}} + \partial_{\boldsymbol{n}}^{\mathcal{B}} v_{\mathcal{B}} = \sigma \quad \text{a. e. on } \beta, \tag{B.1f}$$
$$v_{\mathcal{B}} \in H^1(\mathring{\mathcal{B}},\Delta), \tag{B.1g}$$

is uniquely solvable and there exists a constant $c > 0$ independent of σ, $f_{\mathcal{J}}$ and $f_{\mathcal{B}}$ such that

$$\left(\|v_{\mathcal{J}}\|^2_{H^1(\mathcal{J})} + \|v_{\mathcal{B}}\|^2_{H^1(\mathring{\mathcal{B}})}\right)^{\frac{1}{2}} \leq c\left(\left(\|f_{\mathcal{J}}\|^2_{L^2(\mathcal{J})} + \|f_{\mathcal{B}}\|^2_{L^2(\mathring{\mathcal{B}})}\right)^{\frac{1}{2}} + \|\sigma\|_{H^{-1/2}(\beta)}\right). \tag{B.2}$$

Proof. The proof is basically along the lines of the proof of Proposition 4, but is given for convenience. It is based on the idea to show that (B.1) is equivalent to a variational formulation: Look for v satisfying

$$a_\Omega(v,\varphi) = F(\varphi), \quad \forall \varphi \in H^1(\Omega), \tag{B.3a}$$
$$v \in H^1(\Omega), \tag{B.3b}$$

where (with a piecewise defined function $f|_{\mathcal{J}} := f_{\mathcal{J}}$ and $f|_{\mathcal{B}} := f_{\mathcal{B}}$)

$$a_\Omega(v,\varphi) := \int_\Omega \nabla v \cdot \nabla \varphi + v\,\varphi,$$
$$F(\varphi) := (f\ ,\ \varphi|_\beta)_{L^2(\Omega)} + \left\langle \sigma\ ,\ \varphi|_\beta \right\rangle_{H^{-\frac{1}{2}}(\beta),H^{\frac{1}{2}}(\beta)} := \int_\Omega f\,\varphi + \left\langle \sigma\ ,\ \varphi \right\rangle_{H^{-\frac{1}{2}}(\beta),H^{\frac{1}{2}}(\beta)}.$$

The bilinear form $a(.,.)$ is known to be continuous and coercive on $H^1(\Omega) \times H^1(\Omega)$, and, moreover, there holds $F \in H^1(\Omega)^*$, since $\varphi \to \varphi|_\beta$ is continuous from $H^1(\Omega)$ to $H^{1/2}(\beta)$ (cf. Lemma 1). Consequently, the theorem of Lax and Milgram guarantees existence and uniqueness of a solution v of (B.3) and the existence of $c > 0$ such that (B.2) is fulfilled. To shorten the notation

$$\langle .\ ,\ . \rangle_M := \langle .\ ,\ . \rangle_{H^{-\frac{1}{2}}(M),H^{\frac{1}{2}}(M)}, \quad \text{for } M \in \{\beta,\Gamma\}$$

is used for the remainder of the proof.

1) (B.3) implies (B.1), which will be proven in this part. Due to Lemma 2 the space $H^1(\Omega)$ can be identified with $W := \{(v_{\mathcal{J}}, v_{\mathcal{B}}) \in V \,|\, v_{\mathcal{J}}|_\beta = v_{\mathcal{B}}|_\beta\}$ and thus (B.3) is equivalent to look for $(v_{\mathcal{J}}, v_{\mathcal{B}}) \in W$ satisfying

$$a_\Omega(v, \varphi) = F(\varphi), \quad \forall \varphi := (\varphi_{\mathcal{J}}, \varphi_{\mathcal{B}}) \in W. \tag{B.4}$$

In particular, there holds (B.1c), since $v \in H^1(\Omega) = W$. The next step is to apply the abstract Green's formula of Lemma 3. In order to check the assumptions, the following notations will be useful:

$$V := H^1(\mathcal{J}) \times H^1(\mathring{\mathcal{B}})$$
$$H := L^2(\mathcal{J}) \times L^2(\mathring{\mathcal{B}})$$
$$T := H^{\frac{1}{2}}(\partial\mathcal{J}) \times H^{\frac{1}{2}}(\partial\mathring{\mathcal{B}}) \cong H^{\frac{1}{2}}(\Gamma) \times H^{\frac{1}{2}}(\beta) \times H^{\frac{1}{2}}(\beta)$$
$$\tau : V \to T, \quad (v_{\mathcal{J}}, v_{\mathcal{B}}) \mapsto (\tau_{\mathcal{J}}(v_{\mathcal{J}}), \tau_{\mathcal{B}}(v_{\mathcal{B}})) \equiv (v_{\mathcal{J}}|_\Gamma, v_{\mathcal{J}}|_\beta, v_{\mathcal{B}}|_\beta)$$
$$a : V \times V \to \mathbb{R}, \quad (v, w) \mapsto a_{\mathcal{J}}(v_{\mathcal{J}}, w_{\mathcal{J}}) + a_{\mathcal{B}}(v_{\mathcal{B}}, w_{\mathcal{B}}) := \int_{\mathcal{J}} \nabla v_{\mathcal{J}} \cdot \nabla w_{\mathcal{J}} + v_{\mathcal{J}} w_{\mathcal{J}} + \int_{\mathring{\mathcal{B}}} \nabla v_{\mathcal{B}} \cdot \nabla w_{\mathcal{B}} + v_{\mathcal{B}} w_{\mathcal{B}}$$
$$V_0 := H_0^1(\mathcal{J}) \times H_0^1(\mathring{\mathcal{B}})$$
$$\Lambda = (-\Delta + \mathrm{Id}_{H^1(\mathcal{J})}, -\Delta + \mathrm{Id}_{H^1(\mathring{\mathcal{B}})}) : V \mapsto V_0^* = (H^{-1}(\mathcal{J}), H^{-1}(\mathring{\mathcal{B}})).$$

Then there holds

(i) τ is onto according to Lemma 1

(ii) $V \subset H$ according to the Sobolev embedding theorem and has a stronger topology

(iii) $C_0^\infty(\mathcal{J}) \times C_0^\infty(\mathring{\mathcal{B}})$ is dense in H and V_0; consequently $V_0 \subset H$ is dense, too.

Since Λ is the formal operator associated with the continuous bilinear form a, there holds

$$a(v, \varphi) = \langle \Lambda v \,,\, \varphi \rangle_{V_0^*, V_0}, \quad \forall \varphi \in V_0,$$
$$\xRightarrow[(B.4)]{V_0 \subset W} \langle \Lambda v \,,\, \varphi \rangle_{V_0^*, V_0} = F(\varphi) = (f \,,\, \varphi)_H, \quad \forall \varphi \in V_0,$$
$$\xRightarrow[\text{dense}]{V_0 \subset H} \Lambda v = f \text{ in } H, \text{ i. e. } \Lambda y \in H.$$

Consequently, $v \in V(\Lambda) := \{v \in V \,|\, \Lambda v \in H\} = H^1(\mathcal{J}, \Delta) \times H^1(\mathring{\mathcal{B}}, \Delta)$; in other words (B.1a), (B.1d), (B.1e) and (B.1g) are fulfilled as well as the assumptions of Lemma 3. That is to say, there exists a unique operator

$$\delta = (\delta_\Gamma, \delta_\beta^{\mathcal{J}}, \delta_\beta^{\mathcal{B}}) : V(\Lambda) \to T^* \cong H^{-\frac{1}{2}}(\Gamma) \times H^{-\frac{1}{2}}(\beta) \times H^{-\frac{1}{2}}(\beta),$$

such that there holds

$$a(v, \varphi) = (\Lambda v, \varphi)_H + \langle \delta v \,,\, \tau\varphi \rangle_{T^*, T}, \quad \forall \varphi \in V.$$

This equation is also fulfilled if φ only ranges in $W \subset V$ and a comparison with (B.4) yields

$$\langle \delta v \,,\, \tau\varphi \rangle_{T^*, T} = \langle \sigma \,,\, \varphi|_\beta \rangle_\beta, \quad \forall \varphi \in W,$$
$$\Leftrightarrow \quad \langle \delta_\Gamma v_{\mathcal{J}} \,,\, \varphi_{\mathcal{J}}|_\Gamma \rangle_\Gamma + \langle \delta_\beta^{\mathcal{J}} v_{\mathcal{J}} \,,\, \varphi_{\mathcal{J}}|_\beta \rangle_\beta + \langle \delta_\beta^{\mathcal{B}} v_{\mathcal{B}} \,,\, \varphi_{\mathcal{B}}|_\beta \rangle_\beta = \langle \sigma \,,\, \varphi_{\mathcal{B}}|_\beta \rangle_\beta, \quad \forall (\varphi_{\mathcal{J}}, \varphi_{\mathcal{B}}) \in W.$$

Since $(\varphi_{\mathcal{J}}, \varphi_{\mathcal{B}}) \in W$ one can make use of $\varphi_{\mathcal{J}}|_\beta = \varphi_{\mathcal{B}}|_\beta$ yielding

$$\langle \delta_\Gamma v_{\mathcal{J}} \,,\, \varphi_{\mathcal{J}}|_\Gamma \rangle_\Gamma + \left\langle \delta_\beta^{\mathcal{J}} v_{\mathcal{J}} + \delta_\beta^{\mathcal{B}} v_{\mathcal{B}} \,,\, \varphi_{\mathcal{B}}|_\beta \right\rangle_\beta = \langle \sigma \,,\, \varphi_{\mathcal{B}}|_\beta \rangle_\beta, \quad \forall (\varphi_{\mathcal{J}}, \varphi_{\mathcal{B}}) \in W.$$

Finally the stepwise variation $\varphi \in H_0^1(\Omega) \subset H^1(\Omega) \cong W$ and $\varphi \in W$ reveals

$$\left\langle \delta_\beta^{\mathcal{J}} v_{\mathcal{J}} + \delta_\beta^{\mathcal{B}} v_{\mathcal{B}} \,,\, \varphi|_\beta \right\rangle_\beta = \langle \sigma \,,\, \varphi|_\beta \rangle_\beta, \quad \forall \varphi \in H_0^1(\Omega)$$
$$\langle \delta_\Gamma v_{\mathcal{J}} \,,\, \varphi|_\Gamma \rangle_\Gamma = 0, \quad \forall \varphi \in W.$$

Since the trace operator $(.)|_\Gamma : W \to H^{1/2}(\Gamma)$ is onto (cf. Lemma 1) and referring to the Remark on page 16, one has

$$\partial_{\boldsymbol{n}} v_{\mathcal{J}} = \delta_\Gamma = 0 \text{ in } H^{-\frac{1}{2}}(\Gamma), \text{ i. e. (B.1b).}$$

The analog property of the trace operator $(.)|_\beta$ yields

$$\partial_{\boldsymbol{n}}^{\mathcal{J}} v_{\mathcal{J}} + \partial_{\boldsymbol{n}}^{\mathcal{B}} v_{\mathcal{B}} = \delta_\beta^{\mathcal{J}} v_{\mathcal{J}} + \delta_\beta^{\mathcal{B}} v_{\mathcal{B}} = \sigma \text{ in } H^{-\frac{1}{2}}(\beta), \text{ i. e. (B.1f).}$$

Altogether (B.3) implies (B.1).

2) This part is devoted to prove that (B.1) implies (B.3).
Let $\varphi \in H^1(\Omega)$ be arbitrary. Lemma 2 yields that $\varphi_{\mathcal{J}} := \varphi|_{\mathcal{J}}$ and $\varphi_{\mathcal{B}} := \varphi|_{\mathcal{B}}$ are H^1-functions with $\varphi_{\mathcal{J}}|_\beta = \varphi_{\mathcal{B}}|_\beta$. Multiplying the PDEs (B.1a) and (B.1e) with $\varphi_{\mathcal{J}}$ and $\varphi_{\mathcal{B}}$ respectively, integration, and integration by parts results in

$$\begin{aligned}\int_{\mathcal{J}} \nabla v_{\mathcal{J}} \cdot \nabla \varphi_{\mathcal{J}} + v_{\mathcal{J}}\, \varphi_{\mathcal{J}} - \int_\Gamma \partial_{\boldsymbol{n}} v_{\mathcal{J}}\, \varphi_{\mathcal{J}} &= \int_{\mathcal{J}} f_{\mathcal{J}}\, \varphi_{\mathcal{J}} + \int_\beta \partial_{\boldsymbol{n}}^{\mathcal{J}} v_{\mathcal{J}}\, \varphi_{\mathcal{J}},\\ \int_{\mathring{\mathcal{B}}} \nabla v_{\mathcal{B}} \cdot \nabla \varphi_{\mathcal{B}} + v_{\mathcal{B}}\, \varphi_{\mathcal{B}} &= \int_{\mathring{\mathcal{B}}} f_{\mathcal{B}}\, \varphi_{\mathcal{B}} + \int_\beta \partial_{\boldsymbol{n}}^{\mathcal{B}} v_{\mathcal{B}}\, \varphi_{\mathcal{B}}.\end{aligned}$$

Adding these equations and using the conditions (B.1b), (B.1f) and $\varphi_{\mathcal{J}}|_\beta = \varphi_{\mathcal{B}}|_\beta$ yields (B.3). □

As already mentioned above, Proposition 8 is not general enough to be apply in the proof of Theorem 9, since the kink inducing function σ is assumed to be in $H^{-1/2}(\beta)$; however, $H^{-3/2}$-regularity is required. Nevertheless, it seems reasonable to assume that an assertion analog to the proposition still holds true in the weaker context. This conjecture is based on [6, Thm. 7.1-2] that says, that the operators

$$\begin{aligned}(-\Delta + \mathrm{Id}, \tau_G) &: L^2(G,\Delta) \to L^2(G) \times H^{-\frac{1}{2}}(\partial G)\\ (-\Delta + \mathrm{Id}, \partial_{\boldsymbol{n}}) &: L^2(G,\Delta) \to L^2(G) \times H^{-\frac{3}{2}}(\partial G)\end{aligned}$$

are isomorphisms, and based on [69, Thm. 1.5.3.4 with comment on p. 55] which ensures that the trace mapping

$$(\tau_G, \partial_{\boldsymbol{n}}) : L^2(G,\Delta) \to H^{-\frac{1}{2}}(\partial G) \times H^{-\frac{3}{2}}(\partial G)$$

is continuous and onto for a bounded domain $G \subset \mathbb{R}^2$ of class $C^{1,1}$.

Conjecture 3 (Unique solvability of an elliptic BVP with a kink in $H^{-3/2}$)**:**
Let $\mathcal{B} \in \mathcal{O}$, where $\mathcal{O}$ is given by Definition 4 and use the notations of Definition 5. Moreover, let $\sigma \in H^{-3/2}(\beta)$, let $f_{\mathcal{J}} \in L^2(\mathcal{J})$ and let $f_{\mathcal{B}} \in L^2(\mathring{\mathcal{B}})$ be arbitrary.
Then the boundary value problem

$$\begin{aligned}
-\Delta v_{\mathcal{J}} + v_{\mathcal{J}} &= f_{\mathcal{J}} \quad \text{a. e. in } \mathcal{J}, && \text{(B.5a)} & \qquad -\Delta v_{\mathcal{B}} + v_{\mathcal{B}} &= f_{\mathcal{B}} \quad \text{a. e. in } \mathring{\mathcal{B}}, && \text{(B.5e)}\\
\partial_{\boldsymbol{n}} v_{\mathcal{J}} &= 0 \quad \text{a. e. on } \Gamma, && \text{(B.5b)} & & & &\\
v_{\mathcal{J}}|_\beta - v_{\mathcal{B}}|_\beta &= 0 \quad \text{a. e. on } \beta, && \text{(B.5c)} & \qquad \partial_{\boldsymbol{n}}^{\mathcal{J}} v_{\mathcal{J}} + \partial_{\boldsymbol{n}}^{\mathcal{B}} v_{\mathcal{B}} &= \sigma \quad \text{a. e. on } \beta, && \text{(B.5f)}\\
v_{\mathcal{J}} &\in L^2(\mathcal{J},\Delta), && \text{(B.5d)} & \qquad v_{\mathcal{B}} &\in L^2(\mathring{\mathcal{B}},\Delta) && \text{(B.5g)}
\end{aligned}$$

is uniquely solvable.

Equipped with Conjecture 3, it is possible to prove existence of Lagrange multipliers and adjoint states for the parametrized inner optimization problem (2.37), without reformulating the constraints to independent blocks on $\mathcal{J}$ and $\mathring{\mathcal{B}}$, as it is done in the proof of Theorem 5.

Theorem 9 (Existence of Lagrange multipliers for the inner optimization problem)**:**
Let the family of admissible sets $\mathcal{O}$ be given by Definition 4, let $\mathcal{B} \in \mathcal{O}$ be arbitrarily chosen, and let $(\bar{u}_{\mathcal{J}}, \bar{u}_{\mathcal{B}}, \bar{y}_{\mathcal{J}}, \bar{y}_{\mathcal{B}})$ be the optimal solution to the inner optimization problem (2.37) for the fixed parameter $\mathcal{B}$.
Then there exist multipliers $\hat{q}_{\mathcal{B}} \in L^2(\mathring{\mathcal{B}})$, $\hat{\sigma}_{\mathcal{J}} \in H^{-3/2}(\beta)$ associated with the BDD reformulated state constraints (2.37d) and (2.37e).
Furthermore, assume that there exists $\hat{p}_{\mathcal{J}} \in L^2(\mathcal{J},\Delta)$ and $\hat{p}_{\mathcal{B}} \in L^2(\mathring{\mathcal{B}},\Delta)$ (see Conjecture 3), such that

$$\begin{aligned}
-\Delta \hat{p}_{\mathcal{J}} + \hat{p}_{\mathcal{J}} &= \bar{y}_{\mathcal{J}} - y_d \quad \text{a. e. in } \mathcal{J}, && \text{(B.6a)} & \qquad -\Delta \hat{p}_{\mathcal{B}} + \hat{p}_{\mathcal{B}} &= \bar{y}_{\mathcal{B}} - y_d \quad \text{a. e. in } \mathring{\mathcal{B}}, && \text{(B.6c)}\\
\partial_{\boldsymbol{n}} \hat{p}_{\mathcal{J}} &= 0 \quad \text{a. e. on } \Gamma, && \text{(B.6b)} & \qquad \hat{p}_{\mathcal{J}}|_\beta - \hat{p}_{\mathcal{B}}|_\beta &= 0 \quad \text{a. e. on } \beta, && \text{(B.6d)}\\
& & & & \qquad \partial_{\boldsymbol{n}}^{\mathcal{J}} \hat{p}_{\mathcal{J}} + \partial_{\boldsymbol{n}}^{\mathcal{B}} \hat{p}_{\mathcal{B}} &= \hat{\sigma}_{\mathcal{J}} \quad \text{a. e. on } \beta, && \text{(B.6e)}
\end{aligned}$$

then there holds

$$\begin{aligned}
\lambda\,(\bar{u}_{\mathcal{J}} - u_d) + \hat{p}_{\mathcal{J}} &= 0 \quad \text{a. e. in } \mathcal{J}, && \text{(B.7a)}\\
\lambda\,(\bar{u}_{\mathcal{B}} - u_d) + \hat{p}_{\mathcal{B}} + \hat{q}_{\mathcal{B}} &= 0 \quad \text{a. e. in } \mathring{\mathcal{B}}. && \text{(B.7b)}
\end{aligned}$$

Proof. The proof consists of two parts. The first one provides existence of $\hat{q}_{\mathcal{B}}$ and $\hat{\sigma}_{\mathcal{J}}$ as Lagrange multipliers to a reduced problem. The second part shows, that the relations (B.7) hold with the assumed functions $\hat{p}_{\mathcal{J}}$ and $\hat{p}_{\mathcal{B}}$.

1) Let $\mathcal{B} \in \mathcal{O}$ be arbitrarily chosen, but fix. Consider the linear control-to-state operator $S = (S_{\mathcal{J}}, S_{\mathcal{B}})$ of the split boundary value problem (2.37f)–(2.37j),

$$S : L^2(\mathcal{J}) \times L^2(\mathring{\mathcal{B}}) \to H^2(\mathcal{J}) \times H^2(\mathring{\mathcal{B}}), \quad (u_{\mathcal{J}}, u_{\mathcal{B}}) \mapsto (y_{\mathcal{J}}, y_{\mathcal{B}}), \text{ where } \begin{cases} -\Delta y_{\mathcal{J}} + y_{\mathcal{J}} = u_{\mathcal{J}} & \text{in } \mathcal{J}\,, \\ -\Delta y_{\mathcal{B}} + y_{\mathcal{B}} = u_{\mathcal{B}} & \text{in } \mathring{\mathcal{B}}\,, \\ \partial_{\boldsymbol{n}} y_{\mathcal{J}} = 0 & \text{on } \Gamma\,, \\ y_{\mathcal{J}}|_{\beta} - y_{\mathcal{B}}|_{\beta} = 0 & \text{on } \beta\,, \\ \partial_{\boldsymbol{n}}^{\mathcal{J}} y_{\mathcal{J}} + \partial_{\boldsymbol{n}}^{\mathcal{B}} y_{\mathcal{B}} = 0 & \text{on } \beta\,. \end{cases}$$

S is known to be continuous (cf. [69, Thm. 2.3.3.2]). With use of the Dirichlet trace operators on the interface (cf. Definition 2 and Lemma 1)

$$\tau_{\mathcal{J}} : H^2(\mathcal{J}) \to H^{\frac{3}{2}}(\beta),$$
$$\tau_{\mathcal{B}} : H^2(\mathring{\mathcal{B}}) \to H^{\frac{3}{2}}(\beta),$$

the inner optimization problem (2.37) can be reduced to

$$\begin{aligned} &\text{minimize} \quad f(u_{\mathcal{J}}, u_{\mathcal{A}}) := \mathfrak{J}(\mathcal{B}; u_{\mathcal{J}}, u_{\mathcal{B}}, S_{\mathcal{J}}(u_{\mathcal{J}}, u_{\mathcal{B}}), S_{\mathcal{B}}(u_{\mathcal{J}}, u_{\mathcal{B}})) \\ &\text{subject to} \quad T(u_{\mathcal{J}}, u_{\mathcal{B}}) := \begin{pmatrix} \Delta y_{\min}^{\max} - y_{\min}^{\max} + u_{\mathcal{B}} \\ \tau_{\mathcal{B}} S_{\mathcal{B}}(u_{\mathcal{J}}, u_{\mathcal{B}}) - \tau_{\mathcal{B}} y_{\min}^{\max} \end{pmatrix} = 0, \end{aligned} \tag{B.8}$$

where $T : L^2(\mathcal{J}) \times L^2(\mathring{\mathcal{B}}) \to L^2(\mathring{\mathcal{B}}) \times H^{\frac{3}{2}}(\beta)$. This reduced problem fits into the usual framework of nonlinear optimization in Banach spaces.

In order to prove existence of multipliers, one has to show that a constraint qualification is valid. In current context the Zowe-Kurcyusz constraint qualification (cf. [164] and [159, p. 330]) is suitable and its validity for the operator T in $(\bar{u}_{\mathcal{J}}, \bar{u}_{\mathcal{B}})$ will be proven next.

Thus, for each arbitrary $z_1 \in L^2(\mathring{\mathcal{B}})$ and $z_2 \in H^{\frac{3}{2}}(\beta)$ one needs to find $(h_{\mathcal{J}}, h_{\mathcal{B}}) \in L^2(\mathcal{J}) \times L^2(\mathring{\mathcal{B}})$ such that

$$(\mathrm{D}T)(h_{\mathcal{J}}, h_{\mathcal{B}}) = \begin{pmatrix} h_{\mathcal{B}} \\ \tau_{\mathcal{B}} S_{\mathcal{B}}(h_{\mathcal{J}}, h_{\mathcal{B}}) \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}.$$

(Note that T is continuously Fréchet differentiable, since $S_{\mathcal{B}}$ is continuous and affine.) As a start, this defines $h_{\mathcal{B}} := z_1$. Now let $(v_{\mathcal{J}}^1, v_{\mathcal{B}}^1) := S(0, h_{\mathcal{B}})$. Consequently, $v_{\mathcal{B}}^1 \in H^2(\mathring{\mathcal{B}})$ and $\tau_{\mathcal{B}} v_{\mathcal{B}}^1 \in H^{\frac{3}{2}}(\beta)$. Next, let $v_{\mathcal{B}}^2 \in H^2(\mathring{\mathcal{B}})$ solve

$$\begin{aligned} -\Delta v_{\mathcal{B}}^2 + v_{\mathcal{B}}^2 &= 0 && \text{a. e. in } \mathring{\mathcal{B}}\,, \\ \tau_{\mathcal{B}} v_{\mathcal{B}}^2 &= z_2 - \tau_{\mathcal{B}} v_{\mathcal{B}}^1 && \text{a. e. on } \beta\,. \end{aligned}$$

Due to the extension operator of Lemma 1, there exists $v_{\mathcal{J}}^2 \in H^2(\mathcal{J})$ which suffices

$$\begin{aligned} \partial_{\boldsymbol{n}} v_{\mathcal{J}}^2 &= 0 && \text{a. e. on } \Gamma, \\ \partial_{\boldsymbol{n}}^{\mathcal{J}} v_{\mathcal{I}}^2 &= -\partial_{\boldsymbol{n}}^{\mathcal{B}} v_{\mathcal{B}}^2 && \text{a. e. on } \beta, \\ \tau_{\mathcal{J}} v_{\mathcal{J}}^2 &= \tau_{\mathcal{B}} v_{\mathcal{B}}^2 && \text{a. e. on } \beta. \end{aligned}$$

Defining $h_{\mathcal{J}} := -\Delta v_{\mathcal{J}}^2 + v_{\mathcal{J}}^2 \in L^2(\mathcal{J})$, $v_{\mathcal{J}} := v_{\mathcal{J}}^1 + v_{\mathcal{J}}^2$, and $v_{\mathcal{B}} := v_{\mathcal{B}}^1 + v_{\mathcal{B}}^2$, it follows

$$\begin{aligned} -\Delta v_{\mathcal{J}} + v_{\mathcal{J}} &= -\Delta v_{\mathcal{J}}^1 + v_{\mathcal{J}}^1 - \Delta v_{\mathcal{J}}^2 + v_{\mathcal{J}}^2 && = 0 + h_{\mathcal{J}} && \text{a. e. in } \mathcal{J}, \\ -\Delta v_{\mathcal{B}} + v_{\mathcal{B}} &= -\Delta v_{\mathcal{B}}^1 + v_{\mathcal{B}}^1 - \Delta v_{\mathcal{B}}^2 + v_{\mathcal{B}}^2 && = z_1 + 0 && \text{a. e. in } \mathring{\mathcal{B}}, \\ \partial_{\boldsymbol{n}} v_{\mathcal{J}} &= \partial_{\boldsymbol{n}} v_{\mathcal{J}}^1 + \partial_{\boldsymbol{n}} v_{\mathcal{J}}^2 && = 0 + 0 && \text{a. e. on } \Gamma, \\ \tau_{\mathcal{J}} v_{\mathcal{J}} - \tau_{\mathcal{B}} v_{\mathcal{B}} &= (\tau_{\mathcal{J}} v_{\mathcal{J}}^1 - \tau_{\mathcal{B}} v_{\mathcal{B}}^1) + (\tau_{\mathcal{J}} v_{\mathcal{J}}^2 - \tau_{\mathcal{B}} v_{\mathcal{B}}^2) && = 0 + 0 && \text{a. e. on } \beta, \\ \partial_{\boldsymbol{n}}^{\mathcal{J}} v_{\mathcal{J}} + \partial_{\boldsymbol{n}}^{\mathcal{B}} v_{\mathcal{B}} &= (\partial_{\boldsymbol{n}}^{\mathcal{J}} v_{\mathcal{J}}^1 + \partial_{\boldsymbol{n}}^{\mathcal{B}} v_{\mathcal{B}}^1) + (\partial_{\boldsymbol{n}}^{\mathcal{J}} v_{\mathcal{I}}^2 + \partial_{\boldsymbol{n}}^{\mathcal{B}} v_{\mathcal{B}}^2) && = 0 + 0 && \text{a. e. on } \beta, \end{aligned}$$

or, in other words, $S(h_\mathcal{J}, z_1) = (v_\mathcal{J}, v_\mathcal{B})$. Furthermore, there holds

$$(\mathrm{DT})(h_\mathcal{J}, h_\mathcal{B}) = \begin{pmatrix} h_\mathcal{B} \\ \tau_\mathcal{B} S_\mathcal{B}(h_\mathcal{J}, h_\mathcal{B}) \end{pmatrix} = \begin{pmatrix} z_1 \\ \tau_\mathcal{B} v_\mathcal{B}^1 + (z_2 - \tau_\mathcal{B} v_\mathcal{B}^1) \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}.$$

Consequently, the Zowe-Kurcyusz constraint qualification is fulfilled and there exist Lagrange multipliers $\hat{q}_\mathcal{B} \in L^2(\mathring{\mathcal{B}})$ and $\hat{\sigma}_\mathcal{J} \in H^{3/2}(\beta)^* = H^{-3/2}(\beta)$.

2) In addition, $(\bar{u}_\mathcal{J}, \bar{u}_\mathcal{B}, \hat{q}_\mathcal{B}, \hat{\sigma}_\mathcal{J})$ is a saddle point of the Lagrange function (cf. [159, Thm. 6.3])

$$\mathcal{L} : L^2(\mathcal{J}) \times L^2(\mathring{\mathcal{B}}) \times L^2(\mathring{\mathcal{B}}) \times H^{-\frac{3}{2}}(\beta) \to \mathbb{R},$$
$$\mathcal{L}(u_\mathcal{J}, u_\mathcal{B}, q_\mathcal{B}, \sigma_\mathcal{J}) := f(u_\mathcal{J}, u_\mathcal{B}) + \int_{\mathring{\mathcal{B}}} q_\mathcal{B} \left(u_\mathcal{B} + \Delta y_{\min}^{\max} - y_{\min}^{\max}\right) + \left\langle \sigma_\mathcal{J} \,,\, \tau_\mathcal{B} S_\mathcal{B}(u_\mathcal{J}, u_\mathcal{B}) - \tau_\mathcal{B} y_{\min}^{\max} \right\rangle_{H^{-\frac{3}{2}}(\beta), H^{\frac{3}{2}}(\beta)}.$$

Since the constraints (B.8) solely consist of equations, there holds

$$\begin{aligned}
0 &= \partial_{u_\mathcal{B}} \mathcal{L}(\bar{u}_\mathcal{J}, \bar{u}_\mathcal{B}, \hat{q}_\mathcal{B}, \hat{\sigma}_\mathcal{J}) h \\
&= \int_\mathcal{J} (S_\mathcal{J}(\bar{u}_\mathcal{J}, \bar{u}_\mathcal{B}) - y_d)\, S_\mathcal{J}(0, h) + \int_{\mathring{\mathcal{B}}} (S_\mathcal{B}(\bar{u}_\mathcal{J}, \bar{u}_\mathcal{B}) - y_d)\, S_\mathcal{B}(0, h) + \int_{\mathring{\mathcal{B}}} \lambda(\bar{u}_\mathcal{B} - u_d) h \\
&\quad + \int_{\mathring{\mathcal{B}}} \hat{q}_\mathcal{B}\, h + \left\langle \hat{\sigma}_\mathcal{J} \,,\, \tau_\mathcal{B} S_\mathcal{B}(0, h) \right\rangle_{H^{-\frac{3}{2}}(\beta), H^{\frac{3}{2}}(\beta)}, \quad \forall h \in L^2(\mathring{\mathcal{B}}).
\end{aligned}$$

Now let $(\hat{p}_\mathcal{J}, \hat{p}_\mathcal{B}) \in L^2(\mathcal{J}, \Delta) \times L^2(\mathring{\mathcal{B}}, \Delta)$ be introduced as the solution to (B.6). Using the suitable Green's formula (cf. [69, Thm. 1.5.3.6] in the special case without corners S_j), one can proceed

$$\begin{aligned}
0 &= \int_\mathcal{J} (-\Delta \hat{p}_\mathcal{J} + \hat{p}_\mathcal{J})\, S_\mathcal{J}(0, h) + \int_{\mathring{\mathcal{B}}} (-\Delta \hat{p}_\mathcal{B} + \hat{p}_\mathcal{B})\, S_\mathcal{B}(0, h) \\
&\quad + \int_{\mathring{\mathcal{B}}} \lambda\, (\bar{u}_\mathcal{B} - u_d) h + \int_{\mathring{\mathcal{B}}} \hat{q}_\mathcal{B}\, h + \left\langle \hat{\sigma}_\mathcal{J} \,,\, \tau_\mathcal{B} S_\mathcal{B}(0, h) \right\rangle_{H^{-\frac{3}{2}}(\beta), H^{\frac{3}{2}}(\beta)} \\
&= \int_\mathcal{J} \hat{p}_\mathcal{J} \underbrace{\left(-\Delta S_\mathcal{J}(0, h) + S_\mathcal{J}(0, h)\right)}_{=0} + \int_{\mathring{\mathcal{B}}} \hat{p}_\mathcal{B} \underbrace{\left(-\Delta S_\mathcal{B}(0, h) + S_\mathcal{B}(0, h)\right)}_{=h} \\
&\quad + \int_{\mathring{\mathcal{B}}} \lambda\, (\bar{u}_\mathcal{B} - u_d) h + \int_{\mathring{\mathcal{B}}} \hat{q}_\mathcal{B}\, h + \left\langle \hat{\sigma}_\mathcal{J} \,,\, \tau_\mathcal{B} S_\mathcal{B}(0, h) \right\rangle_{H^{-\frac{3}{2}}(\beta), H^{\frac{3}{2}}(\beta)} \\
&\quad - \left\langle \partial_{\boldsymbol{n}}^\mathcal{J} \hat{p}_\mathcal{J} \,,\, \underbrace{\tau_\mathcal{J} S_\mathcal{J}(0, h)}_{=\tau_\mathcal{B} S_\mathcal{B}(0, h)} \right\rangle_{H^{-\frac{3}{2}}(\beta), H^{\frac{3}{2}}(\beta)} + \left\langle \underbrace{\tau_\mathcal{J} \hat{p}_\mathcal{J}}_{=\tau_\mathcal{B} \hat{p}_\mathcal{B}} \,,\, \partial_{\boldsymbol{n}}^\mathcal{J} S_\mathcal{J}(0, h) \right\rangle_{H^{-\frac{1}{2}}(\beta), H^{\frac{1}{2}}(\beta)} \\
&\quad - \left\langle \partial_{\boldsymbol{n}}^\mathcal{B} \hat{p}_\mathcal{B} \,,\, \tau_\mathcal{B} S_\mathcal{B}(0, h) \right\rangle_{H^{-\frac{3}{2}}(\beta), H^{\frac{3}{2}}(\beta)} + \left\langle \tau_\mathcal{B} \hat{p}_\mathcal{B} \,,\, \partial_{\boldsymbol{n}}^\mathcal{B} S_\mathcal{B}(0, h) \right\rangle_{H^{-\frac{1}{2}}(\beta), H^{\frac{1}{2}}(\beta)} \\
&\quad - \left\langle \underbrace{\partial_{\boldsymbol{n}} \hat{p}_\mathcal{J}}_{=0} \,,\, \tau S_\mathcal{J}(0, h) \right\rangle_{H^{-\frac{3}{2}}(\beta), H^{\frac{3}{2}}(\beta)} + \left\langle \tau \hat{p}_\mathcal{J} \,,\, \underbrace{\partial_{\boldsymbol{n}} S_\mathcal{J}(0, h)}_{=0} \right\rangle_{H^{-\frac{1}{2}}(\beta), H^{\frac{1}{2}}(\beta)} \\
&= \int_{\mathring{\mathcal{B}}} \hat{p}_\mathcal{B}\, h + \lambda\, (\bar{u}_\mathcal{B} - u_d)\, h + \hat{q}_\mathcal{B}\, h \\
&\quad + \left\langle \underbrace{\hat{\sigma}_\mathcal{J} - \partial_{\boldsymbol{n}}^\mathcal{J} \hat{p}_\mathcal{J} - \partial_{\boldsymbol{n}}^\mathcal{B} \hat{p}_\mathcal{B}}_{=0} \,,\, \tau_\mathcal{B} S_\mathcal{B}(0, h) \right\rangle_{H^{-\frac{3}{2}}(\beta), H^{\frac{3}{2}}(\beta)} \\
&\quad + \left\langle \tau_\mathcal{B} \hat{p}_\mathcal{B} \,,\, \underbrace{\partial_{\boldsymbol{n}}^\mathcal{J} S_\mathcal{J}(0, h) + \partial_{\boldsymbol{n}}^\mathcal{B} S_\mathcal{B}(0, h)}_{=0} \right\rangle_{H^{-\frac{1}{2}}(\beta), H^{\frac{1}{2}}(\beta)}, \quad \forall h \in L^2(\mathring{\mathcal{B}}).
\end{aligned}$$

Finally, one obtains $\lambda\, (\bar{u}_\mathcal{B} - u_d) + \hat{p}_\mathcal{B} + \hat{q}_\mathcal{B} = 0$ in $L^2(\mathring{\mathcal{B}})$; this is (B.7b). Analogously, (B.7a) can be derived by investigation of $0 = \partial_{u_\mathcal{J}} \mathcal{L}(\bar{u}_\mathcal{J}, \bar{u}_\mathcal{B}, \hat{q}_\mathcal{B}, \hat{\sigma}_\mathcal{J}) h$. This completes the derivation of the claimed conditions.

□

After having provided the multipliers and the adjoint equations in the mentioned setting, where the adjoint states are connected via interface conditions, it is valuable to compare the results of Theorem 5 and Theorem 9. For convenience some of the properties discussed here are illustrated in Figure B.1.

Corollary 6 (Comparison of multipliers)**:**
Let the family of admissible sets $\mathcal{O}$ be given by Definition 4, let $\mathcal{B} \in \mathcal{O}$ be arbitrarily chosen, and let $(\bar{u}_\mathcal{J}, \bar{u}_\mathcal{B}, \bar{y}_\mathcal{J}, \bar{y}_\mathcal{B})$ be the optimal solution to the inner optimization problem (2.37) for the fixed parameter $\mathcal{B}$. Then the multipliers given by Proposition 3 and theorems 5 and 9 are connected in the following unique way

$$\hat{p}_\mathcal{J} = \bar{p}_\mathcal{J} \qquad \text{a. e. in } \mathcal{J}, \tag{B.9a}$$
$$\hat{p}_\mathcal{B}|_\beta = \bar{p}_\mathcal{J}|_\beta \qquad \text{a. e. on } \beta, \tag{B.9b}$$
$$\hat{\sigma}_\mathcal{J} - \partial_{\boldsymbol{n}}^{\mathcal{B}} \hat{p}_\mathcal{B} = \bar{\sigma}_\mathcal{J} \qquad \text{a. e. on } \beta, \tag{B.9c}$$
$$\hat{q}_\mathcal{B} + \hat{p}_\mathcal{B} = \bar{q}_\mathcal{B} + \bar{p}_\mathcal{B} = p_{\min}^{\max} \qquad \text{a. e. in } \mathring{\mathcal{B}}. \tag{B.9d}$$

Moreover, at the optimal active set $\mathcal{A}$ there holds

$$\hat{q}_\mathcal{A}|_\gamma = 0 \qquad \text{on } \gamma, \tag{B.10a}$$
$$-\Delta\hat{q}_\mathcal{A} + \hat{q}_\mathcal{A} = -\Delta\bar{q}_\mathcal{A} + \bar{q}_\mathcal{A} = \mu_{\mathring{\mathcal{A}}}^{\max} \qquad \text{a. e. in } \mathring{\mathcal{A}}_{\max}, \tag{B.10b}$$
$$-\Delta\hat{q}_\mathcal{A} + \hat{q}_\mathcal{A} = -\Delta\bar{q}_\mathcal{A} + \bar{q}_\mathcal{A} = -\mu_{\mathring{\mathcal{A}}}^{\min} \qquad \text{a. e. in } \mathring{\mathcal{A}}_{\min}, \tag{B.10c}$$
$$\bar{\sigma}_\mathcal{I} + \partial_{\boldsymbol{n}}^{\mathcal{B}}(\bar{q}_\mathcal{A} + \bar{p}_\mathcal{A}) = \mu_\gamma \qquad \text{a. e. on } \gamma, \tag{B.10d}$$
$$\hat{\sigma}_\mathcal{I} + \partial_{\boldsymbol{n}}^{\mathcal{B}} \hat{q}_\mathcal{A} = \mu_\gamma \qquad \text{a. e. on } \gamma, \tag{B.10e}$$
$$\hat{q}_\mathcal{A}|_{\mathring{\mathcal{A}}_{\max}} \geq 0 \qquad \text{in } \mathring{\mathcal{A}}_{\max}, \tag{B.10f}$$
$$\hat{q}_\mathcal{A}|_{\mathring{\mathcal{A}}_{\min}} \leq 0 \qquad \text{in } \mathring{\mathcal{A}}_{\min}, \tag{B.10g}$$
$$\hat{\sigma}_\mathcal{A}|_{\gamma_{\max}} \geq 0 \qquad \text{a. e. on } \gamma_{\max}, \tag{B.10h}$$
$$\hat{\sigma}_\mathcal{A}|_{\gamma_{\min}} \leq 0 \qquad \text{a. e. on } \gamma_{\min}, \tag{B.10i}$$

and regularity of different entities improves to

$$\hat{p}_\mathcal{I},\ \bar{p}_\mathcal{I} \in H^2(\mathcal{I}), \tag{B.10j}$$
$$\hat{p}_\mathcal{A},\ \hat{q}_\mathcal{A} \in H^2(\mathring{\mathcal{A}}), \tag{B.10k}$$
$$\hat{\sigma}_\mathcal{I},\ \bar{\sigma}_\mathcal{I} \in H^{\frac{1}{2}}(\gamma). \tag{B.10l}$$

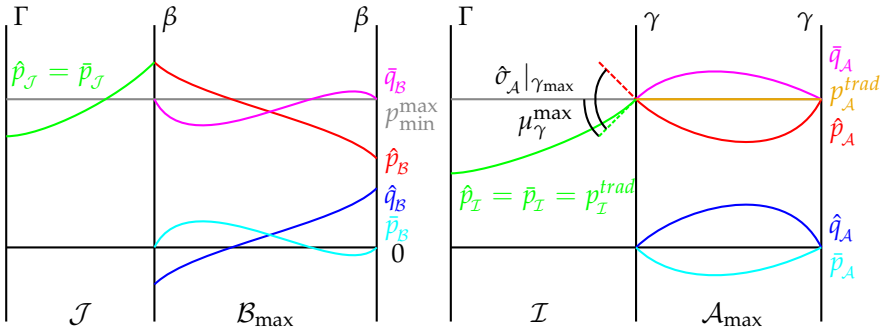


Figure B.1: Comparison of the Lagrange multipliers.

Proof. A comparison of the gradient equations (2.39d) and (B.7a), i. e.

$$\lambda\,(\bar{u}_\mathcal{J} - u_d) + \bar{p}_\mathcal{J} = 0 \quad \text{a. e. in } \mathcal{J},$$
$$\lambda\,(\bar{u}_\mathcal{J} - u_d) + \hat{p}_\mathcal{J} = 0 \quad \text{a. e. in } \mathcal{J},$$

yields (B.9a), since the optimal control $\bar{u}_\mathcal{J}$ is unique. Hence, (B.9b) directly results form the weak continuity of $\hat{p}$ (B.6d). Another direct consequence is (B.9c), since

$$\bar{\sigma}_\mathcal{J} \overset{(2.39c)}{=} \partial_{\boldsymbol{n}}^{\mathcal{J}} \bar{p}_\mathcal{J} = \partial_{\boldsymbol{n}}^{\mathcal{J}} \hat{p}_\mathcal{J} \overset{(B.6e)}{=} \hat{\sigma}_\mathcal{J} - \partial_{\boldsymbol{n}}^{\mathcal{B}} \hat{p}_\mathcal{B}.$$

Finally, a comparison of the gradient equations (2.39g) and (B.7b), i. e.

$$\begin{aligned} \lambda\,(\bar{u}_{\mathcal{B}} - u_d) + \bar{p}_{\mathcal{B}} + \bar{q}_{\mathcal{B}} &= 0 \quad \text{a. e. in } \mathring{\mathcal{B}}, \\ \lambda\,(\bar{u}_{\mathcal{B}} - u_d) + \hat{p}_{\mathcal{B}} + \hat{q}_{\mathcal{B}} &= 0 \quad \text{a. e. in } \mathring{\mathcal{B}}, \end{aligned} \tag{B.11}$$

yields the first equation of (B.9d), since the optimal control $\bar{u}_{\mathcal{B}}$ is unique. The second equation simply is due to the definition of $p_{\min}^{\max}$; see (2.46).

Since the adjoint state $\hat{p}$ and the optimal control $\bar{u}$ are weakly continuous across the optimal interface γ (see (B.6d) and (2.5)), there holds (B.10a)

$$\hat{q}_{\mathcal{A}}|_{\gamma} = -\lambda\,(\bar{u}_{\mathcal{A}} - u_d)|_{\gamma} - \hat{p}_{\mathcal{A}}|_{\gamma} = -\lambda\,(\bar{u}_{\mathcal{I}} - u_d)|_{\gamma} - \hat{p}_{\mathcal{I}}|_{\gamma} = 0.$$

Since there is $p_{\mathcal{A}}^{trad} = p_{\min}^{\max}$ (compare (2.4a), (2.4b) and (2.46)) equation (B.9d) ensures, that

$$\hat{p}_{\mathcal{A}} + \hat{q}_{\mathcal{A}} = p_{\mathcal{A}}^{trad}.$$

Consequently, superposition of the adjoint equations for $\hat{p}_{\mathcal{A}}$ and $p_{\mathcal{A}}^{trad}$ yields

$$-\Delta\hat{q}_{\mathcal{A}} + \hat{q}_{\mathcal{A}} = \mu_{\mathring{\mathcal{A}}} = \mu_{\mathring{\mathcal{A}}}^{\max} - \mu_{\mathring{\mathcal{A}}}^{\min} \quad \text{a. e. in } \mathring{\mathcal{A}}.$$

In other words, there hold (B.10b) and (B.10c). Furthermore, a comparison of the definition of $p_{\min}^{\max}$ (see (2.46)) and the necessary conditions of Bergounioux and Kunisch for the adjoint state $p_{\mathcal{A}}^{trad}$ (see (2.4a), (2.4b)) yields

$$p_{\min}^{\max} = p_{\mathcal{A}}^{trad} \quad \text{in } \mathring{\mathcal{A}}.$$

Henceforth, there hold (B.10d) and (B.10e)

$$\begin{aligned} \mu_{\gamma} &= \partial_{\boldsymbol{n}}^{\mathcal{I}} p_{\mathcal{I}}^{trad} + \partial_{\boldsymbol{n}}^{\mathcal{A}} p_{\mathcal{A}}^{trad} = \partial_{\boldsymbol{n}}^{\mathcal{I}} \bar{p}_{\mathcal{I}} + \partial_{\boldsymbol{n}}^{\mathcal{A}} (\bar{p}_{\mathcal{A}} + \bar{q}_{\mathcal{A}}) = \bar{\sigma}_{\mathcal{I}} + \partial_{\boldsymbol{n}}^{\mathcal{A}} (\bar{p}_{\mathcal{A}} + \bar{q}_{\mathcal{A}}) \quad \text{a. e. on } \gamma, \\ \mu_{\gamma} &= \partial_{\boldsymbol{n}}^{\mathcal{I}} p_{\mathcal{I}}^{trad} + \partial_{\boldsymbol{n}}^{\mathcal{A}} p_{\mathcal{A}}^{trad} = \partial_{\boldsymbol{n}}^{\mathcal{I}} \hat{p}_{\mathcal{I}} + \partial_{\boldsymbol{n}}^{\mathcal{A}} (\hat{p}_{\mathcal{A}} + \hat{q}_{\mathcal{A}}) = \hat{\sigma}_{\mathcal{I}} + \partial_{\boldsymbol{n}}^{\mathcal{A}} \hat{q}_{\mathcal{A}} \quad \text{a. e. on } \gamma. \end{aligned}$$

The sign conditions for the multiplier $\hat{q}_{\mathcal{A}}$ are a consequence of nonnegativity of $\mu^{\max}$ and $\mu^{\min}$ and the weak maximum principle applied to the elliptic BVP (B.10b), (B.10a) and respectively (B.10c), (B.10a). Moreover, since $\hat{q}_{\mathcal{A}}|_{\mathring{\mathcal{A}}_{\max}} \geq 0$ and $\hat{q}_{\mathcal{A}}|_{\gamma_{\max}} = 0$ there must hold

$$\partial_{\boldsymbol{n}}^{\mathcal{A}} \hat{q}_{\mathcal{A}} \leq 0 \quad \text{on } \gamma_{\max}.$$

By means of (B.10e), this yields the sign condition (B.10h)

$$\hat{\sigma}_{\mathcal{A}}|_{\gamma_{\max}} = \mu_{\gamma}^{\max} - \partial_{\boldsymbol{n}}^{\mathcal{A}} \hat{q}_{\mathcal{A}}|_{\gamma_{\max}} \geq 0.$$

Vice versa, one obtains

$$\partial_{\boldsymbol{n}}^{\mathcal{A}} \hat{q}_{\mathcal{A}} \geq 0 \quad \text{on } \gamma_{\min},$$

due to $\hat{q}_{\mathcal{A}}|_{\mathring{\mathcal{A}}_{\min}} \leq 0$ and $\hat{q}_{\mathcal{A}}|_{\gamma_{\min}}$. Hence, the sign condition (B.10i) holds true

$$\hat{\sigma}_{\mathcal{A}}|_{\gamma_{\min}} = -\mu_{\gamma}^{\min} - \partial_{\boldsymbol{n}}^{\mathcal{A}} \hat{q}_{\mathcal{A}}|_{\gamma_{\min}} \leq 0.$$

The enhancement of regularity at the optimal configuration $\mathcal{A}$ can be recognized as follows. Since the Dirichlet trace $\bar{p}_{\mathcal{I}}|_{\gamma}$ (and respectively $\hat{p}_{\mathcal{I}}|_{\gamma}$, due to (B.9a)) is equal to $p_{\min}^{\max}|_{\gamma} \in H^{3/2}(\gamma)$ (cf. (2.57)), elliptic regularity yields $\hat{p}_{\mathcal{I}} = \bar{p}_{\mathcal{I}} \in H^2(\mathcal{I})$. Weak continuity of $\hat{p}$ at the optimal interface γ yields a $H^{3/2}$-regular Dirichlet trace for $\hat{p}_{\mathcal{A}}$ then. Hence, elliptic regularity ensures $\hat{p}_{\mathcal{A}}$ to be in $H^2(\mathring{\mathcal{A}})$. Consequently, the properties of the Neumann trace operator (see Definition 9) ensure $\hat{\sigma}_{\mathcal{A}}$ and $\bar{\sigma}_{\mathcal{A}}$ to be elements of $H^{1/2}(\gamma)$. In addition, $\bar{u}_{\mathcal{A}} \in H^2(\mathring{\mathcal{A}})$, such that the H^2-regularity transfers to $\hat{q}_{\mathcal{A}}$ by means of the gradient equation (B.11). Finally, the adjoint state $\bar{p}_{\mathcal{A}}$ and the Lagrange multiplier $\bar{q}_{\mathcal{A}}$ are known to be H^2-regular; cf. Theorem 5. □

Remarks:

Corollary 6 shows the direct connection between the multipliers in different versions.

1. The comparison of the dual variables in Corollary 6 and the results presented in Appendix A clarifies that there are plenty of different ways to use the BDD approach in the context of the model problem (2.1). However all those different ideas have in common, that the adjoint state $p_{\mathcal{A}}^{trad}$ is decomposed into a sum of a new adjoint state and a Lagrange multiplier $q_{\mathcal{A}}$. At this, the influence

of the state equation and the state constraint can be treated separately in the optimality system, whereas this is not possible by means of the necessary conditions of Bergounioux and Kunisch; cf. Proposition 3.

2. It should be emphasized that the multipliers $\hat{q}_{\mathcal{A}}$ and $\hat{\sigma}_{\mathcal{A}}$ inherit sign conditions from their counterparts $\mu_{\hat{\mathcal{A}}}$ and μ_γ, although they belong to equality constraints. This fact is not too surprising, since the constraints can be regarded as active inequality constraints.
3. However, the multipliers $\bar{q}_{\mathcal{A}}$ and $\bar{\sigma}_{\mathcal{A}}$ do not exhibit this property, which indicates, that the reformulation of the state equation in the proof of Theorem 5 yields a loss of sharpness.
4. The improvement of regularity of $\hat{q}_{\mathcal{A}}$ and $\bar{q}_{\mathcal{A}}$ vs. $\mu_{\hat{\mathcal{A}}}$ is linked to the treatment of the state constraint by means of the BDD approach. The equation was differentiated twice (cf. Paragraph 2.2.2) and the regularity improves from $L^2(\mathring{\mathcal{A}})$ to $H^2(\mathring{\mathcal{A}})$. Analogously, neither $\hat{\sigma}_{\mathcal{A}}$ nor $\bar{\sigma}_{\mathcal{A}}$ gains any improvement in regularity compared to μ_γ, which is due to the fact that the interface condition of the BDD reformulation (i. e. (2.27)) does not contain differential operations. However, improvements can be achieved by applying Neumann boundary conditions; cf. Appendix A.
5. Results, which are comparable to the 2nd and 4th item, are known from optimal control of ODE; cf. Maurer [122, (5.9), (5.10)].

C Remarks on Shape differentiability of the constraints

Essential parts of the derivation of the full first order necessary conditions in Section 2.3 rely on Lemma 8, which provides the local shape derivative of the equality constraints of the reduced shape optimization problem (2.45). Unfortunately, different attempts to prove shape differentiability of this nonstandard elliptic boundary value problem have not been successful. The crucial point is, that there is no direct access available to prove existence and higher regularity of solutions of a boundary value problem with an asymmetrical distribution of boundary conditions, to the best of the author's knowledge. The BVP is repeated here for convenience.

$$
\begin{aligned}
-\Delta \bar{y}_{\mathcal{J}} + \bar{y}_{\mathcal{J}} + \frac{1}{\lambda}\bar{p}_{\mathcal{J}} &= u_d && \text{in } \mathcal{J}, \\
\partial_n \bar{y}_{\mathcal{J}} &= 0 && \text{on } \Gamma, \\
\bar{y}_{\mathcal{J}}|_\beta &= y_{\min}^{\max}|_\beta && \text{on } \beta, \\
\partial_n^{\mathcal{J}} \bar{y}_{\mathcal{J}} &= \partial_n^{\mathcal{B}} y_{\min}^{\max} && \text{on } \beta,
\end{aligned}
\qquad
\begin{aligned}
-\Delta \bar{p}_{\mathcal{J}} + \bar{p}_{\mathcal{J}} - \bar{y}_{\mathcal{J}} &= -y_d && \text{in } \mathcal{J}, \\
\partial_n \bar{p}_{\mathcal{J}} &= 0 && \text{on } \Gamma, \\
& && \\
\bar{p}_{\mathcal{J}} \in L^2(\mathcal{J}, \Delta),\ \bar{y}_{\mathcal{J}} &\in H^2(\mathcal{J}). &&
\end{aligned}
$$

On the first glance such kind of BVP seem to be very artificial and this may give the impression that the associated problems can be avoided if the approach, which is pursued in Section 2.3, is modified in a suitable way. However, there are at least three reasons, which motivate a comprehensive investigation of PDEs that are equipped with asymmetrical boundary conditions, and which show that associated questions arise from different aspects of optimal control.

- A second order elliptic BVP which is stated on a bounded domain that possesses two distinct boundary components – an annulus, for example – and which has a Neumann- and a Dirichlet boundary condition on *one* of these boundary components, seems to be the justified generalization of a second order ODE initial value problem stated on a bounded interval.
- First order necessary conditions for time optimal control problems (of ODE and PDE) typically possess such an asymmetrical distribution of "boundary" conditions: one aims at controlling a state variable, which has to fulfill some sort of initial condition to a final state in minimal time. Here, the initial condition is usually formulated such that a unique simulation of the state is possible for (any) control. Henceforth, the end-time condition, which is required to determine the minimal time, appears as additional condition in the KKT system. This additional condition is compensated by a loss of an end-time condition for the adjoint state, see [112]. All in all, the optimality system is some sort of (P)DAE, whose kernel is the coupled state-adjoint-system, which features an asymmetrical distribution of "boundary"-, i. e. initial and end-time conditions.

Theses considerations draw a parallel between time optimal control and the approach of this thesis: the outer boundary Γ and the interface β correspond to initial time and free end-time, respectively. Consequently, the unknown optimal inactive set parallels the time interval. Moreover, the Neumann boundary condition on Γ together with either the Dirichlet- or the Neumann boundary condition at the interface can be regarded as the substitute for an initial condition, whereas the second interface condition corresponds to the end-time condition.

- It is illustrated in Section 2.7, that OC-PDE and the theory of PDAE are inextricably linked with each other. Moreover, it cannot be expected, that practical applications of OC-PDE are as easy as the model problem considered here. To be more precise, it is likely that the PDE constraint is accompanied by algebraic conditions or something else; see, for instance, [144]. Thus, it is reasonable to embed OC-PDE into the more general framework of OC-PDAE.

  Questions of solvability and regularity of the asymmetrical BVP is a typical question of theory of PDAE, where the investigation of compatibility of algebraic conditions is fundamental.

From this perspective, the asymmetrical distribution of boundary conditions in problem (2.45) does not appear to be artificial, not to say inevitable. In order to get a better understanding of the difficulties one is confronted with when proving shape differentiability, some aspects shall be discussed in the following.

Since the considered BVP is non-standard, there is no shape differentiability result available in literature so far. Hence, differentiability might be proven on an elementary basis by means of convergence of the difference quotient or by means of application of a theorem of Correa and Seeger (cf. [44, Chp. 10 Thm. 5.1]), which is concerned with differentiability of a saddle points with respect to a parameter.

In both cases let $V \in \mathcal{V}$ be a velocity field and for $t \in [0;\tau]$ (and suitable $\tau > 0$) let $T_t := T_t(V)$ be the associated transformation, cf. the 2nd item of the discussion on page 72. Furthermore, define $\mathcal{J}_t := T_t(\mathcal{J})$, $\beta_t := T_t(\beta)$, $\bar{y}_t := \bar{y}_{\mathcal{J}_t}$ and $\bar{p}_t := \bar{p}_{\mathcal{J}_t}$ and consider

$$\begin{aligned} -\Delta\bar{y}_t + \bar{y}_t + \frac{1}{\lambda}\bar{p}_t &= u_d && \text{in } \mathcal{J}_t, \\ \partial_{\boldsymbol{n}}\bar{y}_t &= 0 && \text{on } \Gamma, \\ \bar{y}_t|_{\beta_t} &= y_{\min}^{\max}|_{\beta_t} && \text{on } \beta_t, \\ \partial_{\boldsymbol{n}_t}^{\mathcal{J}}\bar{y}_t &= \partial_{\boldsymbol{n}_t}^{\mathcal{J}} y_{\min}^{\max} && \text{on } \beta_t, \end{aligned} \qquad \begin{aligned} -\Delta\bar{p}_t + \bar{p}_t - \bar{y}_t &= -y_d && \text{in } \mathcal{J}_t, \\ \partial_{\boldsymbol{n}}\bar{p}_t &= 0 && \text{on } \Gamma. \end{aligned}$$

The construction of $y_{\min}^{\max}$ in Lemma 4 ensures, that $y_{\min}^{\max}$ does not have to be redefined, if β_t remains in the sets $B_{\max}$ and $B_{\min}$, respectively. This is guaranteed, as long as $\tau > 0$ is chosen small enough.

Now the question arises, whether $\bar{y}_t$ and $\bar{p}_t$ converge to $\bar{y}_0 := \bar{y}_{\mathcal{J}}$ and $\bar{y}_0 := \bar{p}_{\mathcal{J}}$, respectively, such that the corresponding difference quotients converge. However, in order to be able to obtain a well defined difference "$\bar{y}_t - \bar{y}_0$" one has to ensure, that both constituents are elements of the same vector space. This topic is discussed in detail in Paragraph 2.6.2; in particular, in the 16th and 18th item on page 78f. Hence, the next step is to transport $\bar{y}_t$ and $\bar{p}_t$ back to the domain $\mathcal{J}$:

$$\bar{y}^t := \bar{y}_t \circ T_t, \qquad \bar{p}^t := \bar{p}_t \circ T_t.$$

The difference quotients $(\bar{y}^t - \bar{y}_0)/t$ and respectively $(\bar{p}^t - \bar{p}_0)/t$ are well-defined now, but in order to see whether they converge properly a variational form of the defining coupled PDE is required. Such a weak formulation can be obtained more easily when the boundary conditions are homogenized. By means of setting $y_t := \bar{y}_t - y_{\min}^{\max}$ and $p_t := \bar{p}_t - \theta_t$, where $\theta_t \in H^2(\mathcal{J}_t)$ is defined as a solution to

$$\begin{aligned} -\Delta\theta_t + \theta_t &= y_{\min}^{\max} - y_d && \text{in } \mathcal{J}_t, \\ \partial_{\boldsymbol{n}}\theta_t &= 0 && \text{on } \Gamma, \end{aligned}$$

the above coupled PDE can be homogenized

$$\begin{aligned} -\Delta y_t + y_t + \frac{1}{\lambda}p_t &= u_d + \Delta y_{\min}^{\max} - y_{\min}^{\max} - \theta_t =: F && \text{in } \mathcal{J}_t, \\ \partial_{\boldsymbol{n}} y_t &= 0 && \text{on } \Gamma, \\ y_t|_{\beta_t} &= 0 && \text{on } \beta_t, \\ \partial_{\boldsymbol{n}_t}^{\mathcal{J}} y_t &= 0 && \text{on } \beta_t, \end{aligned} \qquad \begin{aligned} -\Delta p_t + p_t - y_t &= 0 && \text{in } \mathcal{J}_t, \\ \partial_{\boldsymbol{n}} p_t &= 0 && \text{on } \Gamma. \end{aligned}$$

It is important to notice here, that θ_t can be chosen independently of $t \in [0;\tau]$, since $y_{\min}^{\max}$ is independent

of t and since θ_t can be defined as the restriction of the unique function $\Theta \in H^2(\Omega)$ given by

$$\begin{aligned} -\Delta\Theta + \Theta &= y_{\min}^{\max} - y_d && \text{in } \Omega, \\ \partial_{\boldsymbol{n}}\Theta &= 0 && \text{on } \Gamma. \end{aligned}$$

Now, there are two different approaches to obtain a weak formulation of the homogenized coupled system. At this point, the two abovementioned ideas (elementary approach vs. saddle point formulation and theorem of Correa and Seeger) split up. On the one hand an informal, asymmetrical variational formulation reads

$$\begin{aligned} \int_{\mathcal{J}_t} \nabla y_t \cdot \nabla \varphi_t + (y_t + \frac{1}{\lambda} p_t)\, \varphi_t \,\mathrm{d}x_t &= \int_{\mathcal{J}_t} F\, \varphi_t \,\mathrm{d}x_t, && \forall \varphi_t \in H^1(\mathcal{J}_t), \\ \int_{\mathcal{J}_t} \nabla p_t \cdot \nabla \phi_t + (p_t - y_t) \,\mathrm{d}x_t &= 0, && \forall \phi_t \in H^1_{\beta_t}(\mathcal{J}_t) := \{\phi_t \in H^1(\mathcal{J}_t) \,|\, \phi_t|_{\beta_t} = 0\}, \end{aligned}$$

where $y_t \in H^1_{\beta_t}(\mathcal{J}_t)$. On the other hand the system may be characterized informally as the saddle point of the functional

$$L_t : L^2(\mathcal{J}_t) \times H^1_{\beta_t}(\mathcal{J}_t) \times H^1(\mathcal{J}_t) \to \mathbb{R}, \quad (\varphi_t, \phi_t, \psi_t) \mapsto \int_{\mathcal{J}_t} \frac{1}{2}\phi_t^2 + \frac{\lambda}{2}\varphi_t^2 - \nabla\phi_t \cdot \nabla\psi_t - (\phi_t - \varphi_t - F)\,\psi_t \,\mathrm{d}x_t.$$

L_t is Fréchet differentiable (i. e. continuous in particular), convex with respect to (φ_t, ϕ_t) and concave with respect to ψ_t. Furthermore, $L^2(\mathcal{J}_t) \times H^1_{\beta_t}(\mathcal{J}_t)$ and $H^1(\mathcal{J}_t)$ are convex and non-empty sets. Hence, the functional L_t has saddle points (u_t, y_t, p_t), i. e.

$$(u_t, y_t, p_t) = \arg \min_{\substack{\varphi \in L^2(\mathcal{J}_t) \\ \phi \in H^1_{\beta_t}(\mathcal{J}_t)}} \max_{\psi \in H^1(\mathcal{J}_t)} L(\varphi_t, \phi_t, \psi_t).$$

They are equivalently characterized by (cf. [52, Prop. 1.6, p. 170])

$$\begin{aligned} 0 = \partial_{\varphi_t} L_t(u_t, y_t, p_t)\, \varphi_t &= \int_{\mathcal{J}_t} (\lambda u_t + p_t)\, \varphi =: f_{1,t}[u_t, y_t, p_t](\varphi_t), && \forall \varphi_t \in L^2(\mathcal{J}_t), \\ 0 = \partial_{\phi_t} L_t(u_t, y_t, p_t)\, \phi_t &= \int_{\mathcal{J}_t} y_t\, \phi_t - \nabla\phi_t \cdot \nabla p_t - \phi_t\, p_t =: f_{2,t}[u_t, y_t, p_t](\phi_t), && \forall \phi_t \in H^1_{\beta_t}(\mathcal{J}_t), \\ 0 = \partial_{\psi_t} L_t(u_t, y_t, p_t)\, \psi_t &= \int_{\mathcal{J}_t} -\nabla y_t \cdot \nabla\psi_t - (p_t - u_t - F)\, \psi_t =: f_{3,t}[u_t, y_t, p_t](\psi_t), && \forall \psi_t \in H^1(\mathcal{J}_t). \end{aligned}$$

However, both approaches require H^1-regularity of the adjoint p_t, which cannot be ensured. Thus, neither a proof in the style of [90, Appendix A], nor in the style of [44, Chp. 10 Sec. 6] is applicable here.

D Some notions from group theory

The different notions introduced in this section are required to understand the group theoretic perspective on the considerations of Paragraph 2.6.1 concerning the structure of the set $\mathcal{O}$, which was defined in Definition 4. These notions are elementary and can be looked up in any textbook on algebra, for instance [114, §5] or [83, Sec. 10.1]. Nonetheless, they are given for convenience.

Definition 21:
Let M be a set and let $(G, \circ)$ be a group, with *unit element* $\mathbb{1}$. Then one can define the following notions:

- *G operates on the set M* (also *G acts on the set M*), if there is a mapping (*group operation*)

$$M \times G \to M, \quad (x, g) \mapsto g(x)$$

with the properties

$$\forall x \in M,\, g, h \in G: \quad \begin{cases} h(g(x)) = (h \circ g)(x), \\ \mathbb{1}(x) = x. \end{cases}$$

Let G operate on M. Then

- the *stabilizer* (or *isotropy group*) of $x \in M$ is defined as

$$G_x := \{g \in G \,|\, g(x) = x\};$$

- the group operation is *faithful*, if
$$\forall g \in G \setminus \{\mathbb{1}\} \quad \exists x \in M: \quad g(x) \neq x;$$
- the *orbit of* $x \in M$ *under* G is defined as the set of images of x
$$G(x) := \{g(x) \in M \,|\, g \in G\};$$
the orbits are the equivalence classes of the equivalence relation $x \sim y :\Leftrightarrow \exists g \in G: \, g(x) = y$;
- the group operation is *transitive*, if M consist of a single orbit only; that is to say $M = G(x)$.

E Derivation of second order derivatives of the Lagrangian

In order to establish a Lagrange-Newton method for finding critical points of the Lagrangian – see Paragraph 3.3.3 – one has to derive its second order (partial) derivatives, which is the goal of this section.[2] The statement of the Lagrangian and its first order derivatives is repeated here, for convenience. Let
$$o := (\mathcal{B}; f) := (\mathcal{B}; u_{\mathcal{J}}, u_{\mathcal{B}}, y_{\mathcal{J}}, y_{\mathcal{B}}; p_{\mathcal{J}}, p_{\mathcal{B}}, q_{\mathcal{B}}, \sigma_{\mathcal{J}}, \sigma_{\mathcal{B}})$$
be the tuple of its variables, then the Lagrangian is defined as
$$\begin{aligned}\mathcal{L}(o) = \mathfrak{J}(\mathcal{B}; u_{\mathcal{J}}, u_{\mathcal{B}}, y_{\mathcal{J}}, y_{\mathcal{B}}) \\ - \int_{\mathcal{J}} (-\Delta y_{\mathcal{J}} + y_{\mathcal{J}} - u_{\mathcal{J}})\, p_{\mathcal{J}} - \langle p_{\mathcal{J}} \,,\, \partial_{\boldsymbol{n}} y_{\mathcal{J}} \rangle_{\Gamma} - \left\langle p_{\mathcal{J}} \,,\, \partial_{\boldsymbol{n}}^{\mathcal{J}} y_{\mathcal{J}} - \partial_{\boldsymbol{n}}^{\mathcal{J}} y_{\min}^{\max} \right\rangle_{\beta} \\ - \int_{\mathring{\mathcal{B}}} (-\Delta y_{\mathcal{B}} + y_{\mathcal{B}} - u_{\mathcal{B}})\, p_{\mathcal{B}} \\ + \int_{\mathring{\mathcal{B}}} (\Delta y_{\min}^{\max} - y_{\min}^{\max} + u_{\mathcal{B}})\, q_{\mathcal{B}} + \langle\!\langle \sigma_{\mathcal{B}} \,,\, y_{\mathcal{B}} - y_{\min}^{\max} \rangle\!\rangle_{\beta} + \langle\!\langle \sigma_{\mathcal{J}} \,,\, y_{\mathcal{J}} - y_{\min}^{\max} \rangle\!\rangle_{\beta}.\end{aligned}$$
As already derived in Paragraph 2.4.3 the first order derivatives are as follows. In contrast to the former derivation, the derivatives are not simplified by means of known first order necessary conditions, since they are the starting point for the derivation of second order derivatives and thus are not evaluated at the optimal configuration. In abuse of notation, the duality pairings for boundary expressions are substituted by a formal integral notation. The derivatives are evaluated with respect to the direction
$$h := (V; v_{\mathcal{J}}, v_{\mathcal{B}}, z_{\mathcal{J}}, z_{\mathcal{B}}; s_{\mathcal{J}}, s_{\mathcal{B}}, Q_{\mathcal{B}}, \Sigma_{\mathcal{J}}, \Sigma_{\mathcal{B}}).$$
This yields
$$\begin{aligned}
(\partial_{u_{\mathcal{J}}} \mathcal{L}(o)) v_{\mathcal{J}} &= \int_{\mathcal{J}} \lambda (u_{\mathcal{J}} - u_d)\, v_{\mathcal{J}} + v_{\mathcal{J}}\, p_{\mathcal{J}}, \\
(\partial_{u_{\mathcal{B}}} \mathcal{L}(\bar{o})) v_{\mathcal{B}} &= \int_{\mathring{\mathcal{B}}} \lambda (u_{\mathcal{B}} - u_d)\, v_{\mathcal{B}} + v_{\mathcal{B}}\, p_{\mathcal{B}} + v_{\mathcal{B}}\, q_{\mathcal{B}}, \\
(\partial_{y_{\mathcal{J}}} \mathcal{L}(o)) z_{\mathcal{J}} &= \int_{\mathcal{J}} (y_{\mathcal{J}} - y_d)\, z_{\mathcal{J}} + (\Delta z_{\mathcal{J}} - z_{\mathcal{J}})\, p_{\mathcal{J}} - \int_{\Gamma} p_{\mathcal{J}}\, \partial_{\boldsymbol{n}} z_{\mathcal{J}} + \int_{\beta} z_{\mathcal{J}}\, \sigma_{\mathcal{J}} - p_{\mathcal{J}}\, \partial_{\boldsymbol{n}}^{\mathcal{J}} z_{\mathcal{J}}, \\
(\partial_{y_{\mathcal{B}}} \mathcal{L}(o)) z_{\mathcal{B}} &= \int_{\mathring{\mathcal{B}}} (y_{\mathcal{B}} - y_d)\, z_{\mathcal{B}} + (\Delta z_{\mathcal{B}} - z_{\mathcal{B}})\, p_{\mathcal{B}} + \int_{\beta} z_{\mathcal{B}}\, \sigma_{\mathcal{B}}, \\
(\partial_{p_{\mathcal{J}}} \mathcal{L}(o)) s_{\mathcal{J}} &= \int_{\mathcal{J}} -(-\Delta y_{\mathcal{J}} + y_{\mathcal{J}} - u_{\mathcal{J}})\, s_{\mathcal{J}} - \int_{\Gamma} s_{\mathcal{J}}\, \partial_{\boldsymbol{n}} y_{\mathcal{J}} - \int_{\beta} s_{\mathcal{J}}\, \partial_{\boldsymbol{n}}^{\mathcal{J}} (y_{\mathcal{J}} - y_{\min}^{\max}), \\
(\partial_{p_{\mathcal{B}}} \mathcal{L}(o)) s_{\mathcal{B}} &= \int_{\mathring{\mathcal{B}}} -(-\Delta y_{\mathcal{B}} + y_{\mathcal{B}} - u_{\mathcal{B}})\, s_{\mathcal{B}}, \\
(\partial_{q_{\mathcal{B}}} \mathcal{L}(o))\, Q_{\mathcal{B}} &= \int_{\mathring{\mathcal{B}}} (\Delta y_{\min}^{\max} - y_{\min}^{\max} + u_{\mathcal{B}})\, Q_{\mathcal{B}}, \\
(\partial_{\sigma_{\mathcal{J}}} \mathcal{L}(o)) \Sigma_{\mathcal{J}} &= \int_{\beta} (y_{\mathcal{J}} - y_{\min}^{\max})\, \Sigma_{\mathcal{J}}, \\
(\partial_{\sigma_{\mathcal{B}}} \mathcal{L}(o)) \Sigma_{\mathcal{B}} &= \int_{\beta} (y_{\mathcal{B}} - y_{\min}^{\max})\, \Sigma_{\mathcal{B}},
\end{aligned}$$

[2] In particular, it deals with the Lagrangian $\mathcal{L}$ given by Definition 7, but the derivations could also be performed for the Lagrangian of the original model problem (2.1).

$$
\begin{aligned}
\partial_{\mathcal{B}}(\mathcal{L}(o);V) &= \int_\beta \left(\frac{1}{2}(y_{\mathcal{J}}-y_d)^2-\frac{1}{2}(y_{\mathcal{B}}-y_d)^2+\frac{\lambda}{2}(u_{\mathcal{J}}-u_d)^2-\frac{\lambda}{2}(u_{\mathcal{B}}-u_d)^2\right) V\cdot\boldsymbol{n}_{\mathcal{J}}\\
&+\int_\beta \Big(-\left(-\Delta y_{\mathcal{J}}+y_{\mathcal{J}}-u_{\mathcal{J}}\right)p_{\mathcal{J}}+\left(-\Delta y_{\mathcal{B}}+y_{\mathcal{B}}-u_{\mathcal{B}}\right)p_{\mathcal{B}}-\left(\Delta y^{\max}_{\min}-y^{\max}_{\min}+u_{\mathcal{B}}\right)q_{\mathcal{B}}\Big)V\cdot\boldsymbol{n}_{\mathcal{J}}\\
&-\int_\beta \left(\partial_{\boldsymbol{n}}^{\mathcal{J}}p_{\mathcal{J}}\,\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{J}}-y^{\max}_{\min})+p_{\mathcal{J}}\,\partial_{\boldsymbol{nn}}(y_{\mathcal{J}}-y^{\max}_{\min})+p_{\mathcal{J}}\,\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{J}}-y^{\max}_{\min})\,\kappa_{\mathcal{J}}\right)\,V\cdot\boldsymbol{n}_{\mathcal{J}}\\
&+\int_\beta \Big(\Big(\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{B}}-y^{\max}_{\min})+(y_{\mathcal{B}}-y^{\max}_{\min})\,\kappa_{\mathcal{J}}\Big)\sigma_{\mathcal{B}}+\Big(\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{J}}-y^{\max}_{\min})+(y_{\mathcal{J}}-y^{\max}_{\min})\,\kappa_{\mathcal{J}}\Big)\sigma_{\mathcal{J}}\Big)V\cdot\boldsymbol{n}_{\mathcal{J}}.
\end{aligned}
$$

The next goal is to derive the derivative of each of those semiderivatives with respect to the direction

$$\delta := (W;\delta_f) = (W;\delta_{u_{\mathcal{J}}},\delta_{u_{\mathcal{B}}},\delta_{y_{\mathcal{J}}},\delta_{y_{\mathcal{B}}};\delta_{p_{\mathcal{J}}},\delta_{p_{\mathcal{B}}},\delta_{q_{\mathcal{B}}},\delta_{\sigma_{\mathcal{J}}},\delta_{\sigma_{\mathcal{B}}}).$$

The constituents of the direction h are treated as shape independent test functions (and in particular V is autonomous) what considerably simplifies calculations (see Paragraph 2.4.2) and which should yield a result, which is comparable to a second covariant derivative; cf. the 21st item of the discussion on page 81. Furthermore, it should be mentioned that the derivation is only formally and disregards any question concerning regularity.

$$
\begin{aligned}
d\left(\partial_{u_{\mathcal{J}}}\mathcal{L}(o)v_{\mathcal{J}}\right)\delta &= \int_{\mathcal{J}}\left(\lambda\,\delta_{u_{\mathcal{J}}}+\delta_{p_{\mathcal{J}}}\right)v_{\mathcal{J}}+\int_\beta\left(\lambda\,(u_{\mathcal{J}}-u_d)+p_{\mathcal{J}}\right)\,v_{\mathcal{J}}\,W\cdot\boldsymbol{n}_{\mathcal{J}},\\
d\left(\partial_{u_{\mathcal{B}}}\mathcal{L}(o)v_{\mathcal{B}}\right)\delta &= \int_{\mathring{\mathcal{B}}}\left(\lambda\,\delta_{u_{\mathcal{B}}}+\delta_{p_{\mathcal{B}}}+\delta_{q_{\mathcal{B}}}\right)v_{\mathcal{B}}-\int_\beta\left(\lambda\,(u_{\mathcal{B}}-u_d)+p_{\mathcal{B}}+q_{\mathcal{B}}\right)\,v_{\mathcal{B}}\,W\cdot\boldsymbol{n}_{\mathcal{J}},\\
d\left(\partial_{y_{\mathcal{J}}}\mathcal{L}(o)z_{\mathcal{J}}\right)\delta &= \int_{\mathcal{J}}\delta_{y_{\mathcal{J}}}\,z_{\mathcal{J}}+(\Delta z_{\mathcal{J}}-z_{\mathcal{J}})\,\delta_{p_{\mathcal{J}}}-\int_\Gamma\delta_{p_{\mathcal{J}}}\partial_{\boldsymbol{n}}z_{\mathcal{J}}+\int_\beta z_{\mathcal{J}}\,\delta_{\sigma_{\mathcal{J}}}-\delta_{p_{\mathcal{J}}}\,\partial_{\boldsymbol{n}}^{\mathcal{J}}z_{\mathcal{J}}\\
&+\int_\beta\left((y_{\mathcal{J}}-y_d)\,z_{\mathcal{J}}+(\Delta z_{\mathcal{J}}-z_{\mathcal{J}})\,p_{\mathcal{J}}\right)W\cdot\boldsymbol{n}_{\mathcal{J}}\\
&+\int_\beta\left(\partial_{\boldsymbol{n}}^{\mathcal{J}}z_{\mathcal{J}}\,\sigma_{\mathcal{J}}+z_{\mathcal{J}}\,\sigma_{\mathcal{J}}\,\kappa_{\mathcal{J}}-\partial_{\boldsymbol{n}}^{\mathcal{J}}p_{\mathcal{J}}\,\partial_{\boldsymbol{n}}^{\mathcal{J}}z_{\mathcal{J}}-p_{\mathcal{J}}\,\partial_{\boldsymbol{nn}}z_{\mathcal{J}}-p_{\mathcal{J}}\,\partial_{\boldsymbol{n}}^{\mathcal{J}}z_{\mathcal{J}}\,\kappa_{\mathcal{J}}\right)W\cdot\boldsymbol{n}_{\mathcal{J}},\\
d\left(\partial_{y_{\mathcal{B}}}\mathcal{L}(o)z_{\mathcal{B}})\right)\delta &= \int_{\mathcal{B}}\delta_{y_{\mathcal{B}}}\,z_{\mathcal{B}}+(\Delta z_{\mathcal{B}}-z_{\mathcal{B}})\,\delta_{p_{\mathcal{B}}}+\int_\beta z_{\mathcal{B}}\,\delta_{\sigma_{\mathcal{B}}}\\
&+\int_\beta\left(-(y_{\mathcal{B}}-y_d)\,z_{\mathcal{B}}-(\Delta z_{\mathcal{B}}-z_{\mathcal{B}})\,p_{\mathcal{B}}+\partial_{\boldsymbol{n}}^{\mathcal{J}}z_{\mathcal{B}}\,\sigma_{\mathcal{B}}+z_{\mathcal{B}}\,\sigma_{\mathcal{B}}\,\kappa_{\mathcal{J}}\right)W\cdot\boldsymbol{n}_{\mathcal{J}},\\
d\left(\partial_{p_{\mathcal{J}}}\mathcal{L}(o)s_{\mathcal{J}}\right)\delta &= \int_{\mathcal{J}}-(-\Delta\delta_{y_{\mathcal{J}}}+\delta_{y_{\mathcal{J}}}-\delta_{u_{\mathcal{J}}})\,s_{\mathcal{J}}-\int_\Gamma s_{\mathcal{J}}\,\partial_{\boldsymbol{n}}\delta_{y_{\mathcal{J}}}-\int_\beta s_{\mathcal{J}}\,\partial_{\boldsymbol{n}}^{\mathcal{J}}\delta_{y_{\mathcal{J}}}\\
&+\int_\beta\left(-(-\Delta y_{\mathcal{J}}+y_{\mathcal{J}}-u_{\mathcal{J}})\,s_{\mathcal{J}}\right)W\cdot\boldsymbol{n}_{\mathcal{J}}\\
&-\int_\beta\left(\partial_{\boldsymbol{n}}^{\mathcal{J}}s_{\mathcal{J}}\,\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{J}}-y^{\max}_{\min})+s_{\mathcal{J}}\,\partial_{\boldsymbol{nn}}(y_{\mathcal{J}}-y^{\max}_{\min})+s_{\mathcal{J}}\,\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{J}}-y^{\max}_{\min})\,\kappa_{\mathcal{J}}\right)W\cdot\boldsymbol{n}_{\mathcal{J}},\\
d\left(\partial_{p_{\mathcal{B}}}\mathcal{L}(o)s_{\mathcal{B}}\right)\delta &= \int_{\mathring{\mathcal{B}}}-(-\Delta\delta_{y_{\mathcal{B}}}+\delta_{y_{\mathcal{B}}}-\delta_{u_{\mathcal{B}}})\,s_{\mathcal{B}}-\int_\beta-(-\Delta y_{\mathcal{B}}+y_{\mathcal{B}}-u_{\mathcal{B}})\,s_{\mathcal{B}}\,W\cdot\boldsymbol{n}_{\mathcal{J}},\\
d\left(\partial_{q_{\mathcal{B}}}\mathcal{L}(o)Q\right)\delta &= \int_{\mathring{\mathcal{B}}}\delta_{u_{\mathcal{B}}}\,Q-\int_\beta(\Delta y^{\max}_{\min}-y^{\max}_{\min}+u_{\mathcal{B}})\,Q\,W\cdot\boldsymbol{n}_{\mathcal{J}},\\
d\left(\partial_{\sigma_{\mathcal{J}}}\mathcal{L}(o)\Sigma_{\mathcal{J}}\right)\delta &= \int_\beta\delta_{y_{\mathcal{J}}}\,\Sigma_{\mathcal{J}}+\left(\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{J}}-y^{\max}_{\min})\,\Sigma_{\mathcal{J}}+(y_{\mathcal{J}}-y^{\max}_{\min})\,\Sigma_{\mathcal{J}}\,\kappa_{\mathcal{J}}\right)W\cdot\boldsymbol{n}_{\mathcal{J}},\\
d\left(\partial_{\sigma_{\mathcal{B}}}\mathcal{L}(o)\Sigma_{\mathcal{B}}\right)\delta &= \int_\beta\delta_{y_{\mathcal{B}}}\,\Sigma_{\mathcal{B}}+\left(\partial_{\boldsymbol{n}}^{\mathcal{J}}(y_{\mathcal{B}}-y^{\max}_{\min})\,\Sigma_{\mathcal{B}}+(y_{\mathcal{B}}-y^{\max}_{\min})\,\Sigma_{\mathcal{B}}\,\kappa_{\mathcal{J}}\right)W\cdot\boldsymbol{n}_{\mathcal{J}},\\
d\left(\partial_{\mathring{\mathcal{B}}}(\mathcal{L}(o);V)\right)\delta &= \int_\beta\left((y_{\mathcal{J}}-y_d)\,\delta_{y_{\mathcal{J}}}-(y_{\mathcal{B}}-y_d)\,\delta_{y_{\mathcal{B}}}+\lambda\,(u_{\mathcal{J}}-u_d)\,\delta_{u_{\mathcal{J}}}-\lambda\,(u_{\mathcal{B}}-u_d)\,\delta_{u_{\mathcal{B}}}\right)V\cdot\boldsymbol{n}_{\mathcal{J}}\\
&+\int_\beta\Big(-(-\Delta\delta_{y_{\mathcal{J}}}+\delta_{y_{\mathcal{J}}}-\delta_{u_{\mathcal{J}}})\,p_{\mathcal{J}}-(-\Delta y_{\mathcal{J}}+y_{\mathcal{J}}-u_{\mathcal{J}})\,\delta_{p_{\mathcal{J}}}\\
&\qquad+(-\Delta\delta_{y_{\mathcal{B}}}+\delta_{y_{\mathcal{B}}}-\delta_{u_{\mathcal{B}}})\,p_{\mathcal{B}}+(-\Delta y_{\mathcal{B}}+y_{\mathcal{B}}-u_{\mathcal{B}})\,\delta_{p_{\mathcal{B}}}\\
&\qquad-\delta_{u_{\mathcal{B}}}\,q_{\mathcal{B}}-(\Delta y^{\max}_{\min}-y^{\max}_{\min}+u_{\mathcal{B}})\,\delta_{q_{\mathcal{B}}}\Big)V\cdot\boldsymbol{n}_{\mathcal{J}}
\end{aligned}
$$

$$
\begin{aligned}
&- \int_\beta \Big(\partial_{\boldsymbol{n}}^{\mathcal{J}} \delta_{p_{\mathcal{J}}} \, \partial_{\boldsymbol{n}}^{\mathcal{J}} (y_{\mathcal{J}} - y_{\min}^{\max}) + \delta_{p_{\mathcal{J}}} \, \partial_{\boldsymbol{nn}} (y_{\mathcal{J}} - y_{\min}^{\max}) + \delta_{p_{\mathcal{J}}} \, \partial_{\boldsymbol{n}}^{\mathcal{J}} (y_{\mathcal{J}} - y_{\min}^{\max}) \, \kappa_{\mathcal{J}} \\
&\qquad + \partial_{\boldsymbol{n}}^{\mathcal{J}} p_{\mathcal{J}} \, \partial_{\boldsymbol{n}}^{\mathcal{J}} \delta_{y_{\mathcal{J}}} + p_{\mathcal{J}} \, \partial_{\boldsymbol{nn}} \delta_{y_{\mathcal{J}}} + p_{\mathcal{J}} \, \partial_{\boldsymbol{n}}^{\mathcal{J}} \delta_{y_{\mathcal{J}}} \, \kappa_{\mathcal{J}} \Big) \, V \cdot \boldsymbol{n}_{\mathcal{J}} \\
&+ \int_\beta \Big(\Big(\partial_{\boldsymbol{n}}^{\mathcal{J}} \delta_{y_{\mathcal{B}}} + \delta_{y_{\mathcal{B}}} \, \kappa_{\mathcal{J}} \Big) \, \sigma_{\mathcal{B}} + \Big(\partial_{\boldsymbol{n}}^{\mathcal{J}} (y_{\mathcal{B}} - y_{\min}^{\max}) + (y_{\mathcal{B}} - y_{\min}^{\max}) \, \kappa_{\mathcal{J}} \Big) \, \delta_{\sigma_{\mathcal{B}}} \\
&\qquad \Big(\partial_{\boldsymbol{n}}^{\mathcal{J}} \delta_{y_{\mathcal{J}}} + \delta_{y_{\mathcal{J}}} \, \kappa_{\mathcal{J}} \Big) \, \sigma_{\mathcal{J}} + \Big(\partial_{\boldsymbol{n}}^{\mathcal{J}} (y_{\mathcal{J}} - y_{\min}^{\max}) + (y_{\mathcal{J}} - y_{\min}^{\max}) \, \kappa_{\mathcal{J}} \Big) \, \delta_{\sigma_{\mathcal{J}}} \Big) \, V \cdot \boldsymbol{n}_{\mathcal{J}} \\
&+ \int_\beta \left((y_{\mathcal{J}} - y_d) \, \partial_{\boldsymbol{n}}^{\mathcal{J}} (y_{\mathcal{J}} - y_d) + \frac{1}{2} (y_{\mathcal{J}} - y_d)^2 \, \kappa_{\mathcal{J}} \right) V \cdot \boldsymbol{n}_{\mathcal{J}} W \cdot \boldsymbol{n}_{\mathcal{J}} \\
&- \int_\beta \left((y_{\mathcal{B}} - y_d) \, \partial_{\boldsymbol{n}}^{\mathcal{J}} (y_{\mathcal{B}} - y_d) + \frac{1}{2} (y_{\mathcal{B}} - y_d)^2 \, \kappa_{\mathcal{J}} \right) V \cdot \boldsymbol{n}_{\mathcal{J}} W \cdot \boldsymbol{n}_{\mathcal{J}} \\
&+ \int_\beta \left(\lambda \, (u_{\mathcal{J}} - u_d) \, \partial_{\boldsymbol{n}}^{\mathcal{J}} (u_{\mathcal{J}} - u_d) + \frac{\lambda}{2} (u_{\mathcal{J}} - u_d)^2 \, \kappa_{\mathcal{J}} \right) V \cdot \boldsymbol{n}_{\mathcal{J}} W \cdot \boldsymbol{n}_{\mathcal{J}} \\
&- \int_\beta \left(\lambda \, (u_{\mathcal{B}} - u_d) \, \partial_{\boldsymbol{n}}^{\mathcal{J}} (u_{\mathcal{B}} - u_d) + \frac{\lambda}{2} (u_{\mathcal{B}} - u_d)^2 \, \kappa_{\mathcal{J}} \right) V \cdot \boldsymbol{n}_{\mathcal{J}} W \cdot \boldsymbol{n}_{\mathcal{J}} \\
&- \int_\beta \Big(\partial_{\boldsymbol{n}}^{\mathcal{J}} (-\Delta y_{\mathcal{J}} + y_{\mathcal{J}} - u_{\mathcal{J}}) \, p_{\mathcal{J}} + (-\Delta y_{\mathcal{J}} + y_{\mathcal{J}} - u_{\mathcal{J}}) \, (\partial_{\boldsymbol{n}}^{\mathcal{J}} p_{\mathcal{J}} + p_{\mathcal{J}} \, \kappa_{\mathcal{J}}) \Big) \, V \cdot \boldsymbol{n}_{\mathcal{J}} W \cdot \boldsymbol{n}_{\mathcal{J}} \\
&+ \int_\beta \Big(\partial_{\boldsymbol{n}}^{\mathcal{J}} (-\Delta y_{\mathcal{B}} + y_{\mathcal{B}} - u_{\mathcal{B}}) \, p_{\mathcal{B}} + (-\Delta y_{\mathcal{B}} + y_{\mathcal{B}} - u_{\mathcal{B}}) \, (\partial_{\boldsymbol{n}}^{\mathcal{J}} p_{\mathcal{B}} + p_{\mathcal{B}} \, \kappa_{\mathcal{J}}) \Big) \, V \cdot \boldsymbol{n}_{\mathcal{J}} W \cdot \boldsymbol{n}_{\mathcal{J}} \\
&- \int_\beta \Big(\partial_{\boldsymbol{n}}^{\mathcal{J}} (\Delta y_{\min}^{\max} - y_{\min}^{\max} + u_{\mathcal{B}}) \, q_{\mathcal{B}} + (\Delta y_{\min}^{\max} - y_{\min}^{\max} + u_{\mathcal{B}}) \, (\partial_{\boldsymbol{n}}^{\mathcal{J}} q_{\mathcal{B}} + q_{\mathcal{B}} \, \kappa_{\mathcal{J}}) \Big) \, V \cdot \boldsymbol{n}_{\mathcal{J}} W \cdot \boldsymbol{n}_{\mathcal{J}} \\
&- \int_\beta \Big(\partial_{\boldsymbol{n}}^{\mathcal{J}} p_{\mathcal{J}} \, \partial_{\boldsymbol{nn}} (y_{\mathcal{J}} - y_{\min}^{\max}) + (\partial_{\boldsymbol{nn}} p_{\mathcal{J}} + \partial_{\boldsymbol{n}}^{\mathcal{J}} p_{\mathcal{J}} \, \kappa_{\mathcal{J}}) \partial_{\boldsymbol{n}}^{\mathcal{J}} (y_{\mathcal{J}} - y_{\min}^{\max}) \Big) \, V \cdot \boldsymbol{n}_{\mathcal{J}} W \cdot \boldsymbol{n}_{\mathcal{J}} \\
&- \int_\beta \Big(p_{\mathcal{J}} \, \partial_{\boldsymbol{nnn}}^{\mathcal{J}} (y_{\mathcal{J}} - y_{\min}^{\max}) + (\partial_{\boldsymbol{n}}^{\mathcal{J}} p_{\mathcal{J}} + p_{\mathcal{J}} \, \kappa_{\mathcal{J}}) \, \partial_{\boldsymbol{nn}} (y_{\mathcal{J}} - y_{\min}^{\max}) \Big) \, V \cdot \boldsymbol{n}_{\mathcal{J}} W \cdot \boldsymbol{n}_{\mathcal{J}} \\
&- \int_\beta \Big(p_{\mathcal{J}} \, \partial_{\boldsymbol{nn}} (y_{\mathcal{J}} - y_{\min}^{\max}) \, \kappa_{\mathcal{J}} + (\partial_{\boldsymbol{n}}^{\mathcal{J}} p_{\mathcal{J}} \, \kappa_{\mathcal{J}} + p_{\mathcal{J}} \, \kappa_{\mathcal{J}}^2) \, \partial_{\boldsymbol{n}}^{\mathcal{J}} (y_{\mathcal{J}} - y_{\min}^{\max}) \Big) \, V \cdot \boldsymbol{n}_{\mathcal{J}} W \cdot \boldsymbol{n}_{\mathcal{J}} \\
&+ \int_\beta \Big(\partial_{\boldsymbol{nn}} (y_{\mathcal{B}} - y_{\min}^{\max}) \, \sigma_{\mathcal{B}} + \partial_{\boldsymbol{n}}^{\mathcal{J}} (y_{\mathcal{B}} - y_{\min}^{\max}) \, \sigma_{\mathcal{B}} \, \kappa_{\mathcal{J}} \Big) \, V \cdot \boldsymbol{n}_{\mathcal{J}} W \cdot \boldsymbol{n}_{\mathcal{J}} \\
&+ \int_\beta \Big(\partial_{\boldsymbol{n}}^{\mathcal{J}} (y_{\mathcal{B}} - y_{\min}^{\max}) \, \kappa_{\mathcal{J}} \, \sigma_{\mathcal{B}} + (y_{\mathcal{B}} - y_{\min}^{\max}) \, \kappa_{\mathcal{J}}^2 \, \sigma_{\mathcal{B}} \Big) \, V \cdot \boldsymbol{n}_{\mathcal{J}} W \cdot \boldsymbol{n}_{\mathcal{J}} \\
&+ \int_\beta \Big(\partial_{\boldsymbol{nn}} (y_{\mathcal{J}} - y_{\min}^{\max}) \, \sigma_{\mathcal{J}} + \partial_{\boldsymbol{n}}^{\mathcal{J}} (y_{\mathcal{J}} - y_{\min}^{\max}) \, \sigma_{\mathcal{J}} \, \kappa_{\mathcal{J}} \Big) \, V \cdot \boldsymbol{n}_{\mathcal{J}} W \cdot \boldsymbol{n}_{\mathcal{J}} \\
&+ \int_\beta \Big(\partial_{\boldsymbol{n}}^{\mathcal{J}} (y_{\mathcal{J}} - y_{\min}^{\max}) \, \kappa_{\mathcal{J}} \, \sigma_{\mathcal{J}} + (y_{\mathcal{J}} - y_{\min}^{\max}) \, \kappa_{\mathcal{J}}^2 \, \sigma_{\mathcal{J}} \Big) \, V \cdot \boldsymbol{n}_{\mathcal{J}} W \cdot \boldsymbol{n}_{\mathcal{J}} .
\end{aligned}
$$

Bibliography

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, 2008. http://press.princeton.edu/books/absil/. 67, 68, 70, 71, 78, 91, 92, 93, 95, 96, 99, 109

[2] R. A. Adams and J. J. F. Fournier. *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics*. Elsevier, Amsterdam, 2nd edition, 2003. 10.1016/S0079-8169(03)80001-6. 6

[3] J.-J. Alibert and J.-P. Raymond. Boundary control of semilinear elliptic equations with discontinuous leading coefficients and unbounded controls. *Numerical Functional Analysis and Optimization*, 3&4:235–250, 1997. 10.1080/01630569708816758. 7

[4] H. W. Alt. *Lineare Funktionalanalysis – Eine anwendungsorientierte Einführung*. Springer, Berlin, 2006. 10.1007/3-540-34187-0. 7, 12, 13

[5] H. Amann. *Ordinary differential equations*, volume 13 of *De Gruyter Studies in Mathematics*. Walter de Gruyter, Berlin, 1990. 10.1515/9783110853698. 95

[6] J.-P. Aubin. *Approximation of Elliptic Boundary-Value Problems*, volume XXVI of *Pure and Applied Mathematics*. Wiley-Interscience, New York, 1972. 16, 36, 37, 148

[7] B. Aulbach. *Gewöhnliche Differenzialgleichungen*. Elsevier, Heidelberg, 2004. 95

[8] J. W. Barrett and C. M. Elliott. A finite-element method for solving elliptic equations with Neumann data on a curved boundary using unfitted meshes. *IMA Journal of Numerical Analysis*, 4(3):309–325, 1984. 10.1093/imanum/4.3.309. 125

[9] J. W. Barrett and C. M. Elliott. Fitted and unfitted finite-element methods for elliptic equations with smooth interfaces. *IMA Journal of Numerical Analysis*, 7(3):283–300, 1987. 10.1093/imanum/7.3.283. 125

[10] S. Bechmann and M. Frey. *Regularisierungsmethoden für Optimalsteuerungsprobleme*, volume 80 of *Bayreuther Mathematische Schriften*. Mathematisches Institut der Universität Bayreuth, 2008. 6, 116, 117, 138

[11] M. Bergounioux, M. Haddou, M. Hintermüller, and K. Kunisch. A comparison of a Moreau-Yosida based active set strategy and interior point methods for constrained optimal control problems. *SIAM Journal on Optimization*, 11:495–521, 2000. 10.1137/S1052623498343131. 115, 116, 117

[12] M. Bergounioux, K. Ito, and K. Kunisch. Primal-dual strategy for constrained optimal control problems. *SIAM Journal on Control and Optimization*, 37:1176–1194, 1999. 10.1137/S0363012997328609. 115, 117

[13] M. Bergounioux and K. Kunisch. Primal-dual strategy for state-constrained optimal control problems. *Computational Optimization and Applications*, 22:193–224, 2002. 10.1023/A:1015489608037. 115

[14] M. Bergounioux and K. Kunisch. On the structure of Lagrange multipliers for state-constrained optimal control problems. *Systems & Control Letters*, 48:169–176, 2003. 10.1016/S0167-6911(02)00262-1. 7, 49, 50, 141

[15] A. Borzì and V. Schulz. *Computational Optimization of Systems Governed by Partial Differential Equations*. Computational Science & Engineering. SIAM, Philadelphia, 2012. 10.1137/1.9781611972054. 92

[16] K. E. Brenan, S. L. Campbell, and L. R. Petzold. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, volume 14 of *Classics in Applied Mathematics*. SIAM, Philadelphia, 2nd edition, 1996. Reprint of the 1989 original, 10.1137/1.9781611971224. 84, 85, 86, 87

[17] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer, New York, 1991. 19

[18] A. E. Bryson, Jr., W. F. Denham, and S. E. Dreyfus. Optimal programming problems with inequality constraints I: Necessary conditions for extremal solutions. *AIAA Journal*, 1(11):2544–2550, 1963. 10.2514/3.2107. 2, 9, 23

[19] D. Bucur and G. Buttazzo. *Variational Methods in Shape Optimization Problems*, volume 65 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser, Boston, 2005. 10.1007/b137163. 8, 30

[20] R. Z. Bulirsch, F. Montrone, and H. J. Pesch. Abort landing in the presence of windshear as a minimax optimal control problem. Part 1: Necessary conditions. *Journal of Optimization Theory and Applications*, 70:1–23, 1991. 10.1007/BF00940502. 9, 84

[21] R. Z. Bulirsch, F. Montrone, and H. J. Pesch. Abort landing in the presence of windshear as a minimax optimal control problem. Part 2: Multiple shooting and homotopy. *Journal of Optimization Theory and Applications*, 70:223–254, 1991. 10.1007/BF00940625. 9, 84

[22] S. L. Campbell and W. Marszalek. ODE/DAE integrators and MOL problems. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 76(S1):251–254, 1996. 10.1002/zamm.19960761107. 88

[23] S. L. Campbell and W. Marszalek. The index of an infinite dimensional implicit system. *Mathematical and Computer Modelling of Dynamical Systems*, 5(1):18–42, 1999. 10.1076/mcmd.5.1.18.3625. 86

[24] E. Casas. *Análysis numérico de algunos problemas de optimización estructural*. PhD thesis, Univ. Santiago de Compostela (Spain), 1982. 7

[25] E. Casas. Control of an elliptic problem with pointwise state constraints. *SIAM Journal on Control and Optimization*, 4:1309–1322, 1986. 10.1137/0324078. 6, 7, 8

[26] E. Casas. Boundary control of semilinear elliptic equations with pointwise state constraints. *SIAM Journal on Control and Optimization*, 31:993–1006, 1993. 10.1137/0331044. 7

[27] J. Céa, A. Gioan, and J. Michel. Quelques résultats sur l'identification de domaines. *Calcolo*, 10(3–4):207–232, 1973. 10.1007/BF02575843. 8

[28] K. W. Cheng and Th.-P. Fries. Higher-order XFEM for curved strong and weak discontinuities. *International Journal for Numerical Methods in Engineering*, 82(5):564–590, 2010. 10.1002/nme.2768. 125

[29] S. Cherednichenko and A. Rösch. Error estimates for the regularization of optimal control problems with pointwise control and state constraints. *Zeitschrift für Analysis und ihre Anwendungen*, 27:195–212, 2008. 10.4171/ZAA/1351. 116

[30] J. Chessa, P. Smolinski, and T. Belytschko. The extended finite element method (XFEM) for solidification problems. *International Journal for Numerical Methods in Engineering*, 53:1959–1977, 2002. 10.1002/nme.386. 125

[31] K. Chudej. Index analysis for singular PDE models of fuel cells. In H.-G. Bock, F. Hoog, A. Friedman, A. Gupta, H. Neunzert, W. R. Pulleyblank, T. Rusten, F. Santosa, A.-K. Tornberg, V. Capasso, R. Mattheij, H. Neunzert, O. Scherzer, A. Bucchianico, R. Mattheij, and M. Peletier, editors, *Progress in Industrial Mathematics at ECMI 2004*, volume 8 of *Mathematics in Industry*, pages 212–216. Springer, Berlin, 2006. 10.1007/3-540-28073-1_30. 1

[32] K. Chudej, P. Heidebrecht, V. Petzet, S. Scherdel, K. Schittkowski, H. J. Pesch, and K. Sundmacher. Index analysis and numerical solution of a large scale nonlinear PDAE system describing the dynamical behaviour of molten carbonate fuel cells. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 85(2):132–140, 2005. 10.1002/zamm.200310166. 1

[33] K. Chudej, H. J. Pesch, and K. Sternberg. Optimal control of load changes for molten carbonate fuel cell systems: A challenge in PDE constrained optimization. *SIAM Journal on Applied Mathematics*, 70(2):621–639, 2009. 10.1137/080722102. 1, 101

[34] K. Chudej, V. Petzet, S. Scherdel, H. J. Pesch, K. Schittkowski, P. Heidebrecht, and K. Sundmacher. Index analysis of a nonlinear PDAE system describing a molten carbonatefuel cell. *PAMM - Proceedings in Applied Mathematics and Mechanics*, 3(1):563–564, 2003. 10.1002/pamm.200310549. 1

[35] D. Clever and J. Lang. Optimal control of radiative heat transfer in glass cooling with restrictions on the temperature gradient. *Optimal Control Applications and Methods*, 2011. 10.1002/oca.984. 1

[36] D. D. Hömberg and S. Volkwein. Control of laser surface hardening by a reduced-order approach using proper orthogonal decomposition. *Mathematical and Computer Modelling*, 38(10):1003–1028, 2003. 10.1016/S0895-7177(03)90102-6. 1

[37] M. Dambrine. On variations of the shape Hessian and sufficient conditions for stability of critical shapes. *Real Academica de Ciencias Serie a Matemáticas (RACSAM)*, 96(1):95–121, 2002. http://www.rac.es/ficheros/doc/00073.pdf. 58

[38] M. Dambrine and M. Pierre. About stability of equilibrium shapes. *ESAIM: Mathematical Modelling and Numerical Analysis*, 34(4):811–834, 2000. 10.1051/m2an:2000105. 58

[39] M. Dambrine, J. Sokołowski, and A. Żochowski. On stability analysis in shape optimisation: critical shapes for Neumann problem. *Control and Cybernetics*, 32(3):503–528, 2003. http://matwbn.icm.edu.pl/ksiazki/cc/cc32/cc3235.pdf. 58

[40] M. Delfour, G. Payre, and J.-P. Zolésio. An optimal triangulation for second-order elliptic problems. *Computer Methods in Applied Mechanics and Engineering*, 50(3):231–261, 1985. 10.1016/0045-7825(85)90095-7. 119

[41] M. C. Delfour and J.-P. Zolésio. Anatomy of the Shape Hessian. *Annali di Matematica pura ed applicata*, 159(1):315–339, 1991. 10.1007/BF01766307. 80

[42] M. C. Delfour and J.-P. Zolésio. Velocity method and Lagrangian formulation for the computation of the Shape Hessian. *SIAM Journal on Control and Optimization*, 29(6):1414–1442, 1991. 10.1137/0329072. 72, 80

[43] M. C. Delfour and J.-P. Zolésio. Structure of shape derivatives for nonsmooth domains. *Journal of Functional Analysis*, 104(1):1–33, 1992. 10.1016/0022-1236(92)90087-Y. 73

[44] M. C. Delfour and J.-P. Zolésio. *Shape and Geometries*, volume 22 of *Advances in Design and Control*. SIAM, Philadelphia, 2nd edition, 2011. 10.1137/1.9780898719826. 5, 8, 18, 35, 42, 43, 44, 53, 54, 55, 58, 61, 63, 64, 65, 67, 68, 69, 72, 73, 75, 76, 77, 78, 80, 81, 93, 119, 154, 155

[45] J. Donea, A. Huerta, J. Ponthot, and A. Rodríguez-Ferran. Arbitrary Lagrangian-Eulerian methods. In *Encyclopedia of Computational Mechanics*. John Wiley & Sons, Ltd., 2004. 10.1002/0470091355.ecm009. 125

[46] J. C. Dunn. Local attractors for gradient-related descent iterations. In Ch. A. Floudas and P. M. Pardalos, editors, *Encyclopedia of Optimization*, pages 1911–1919. Springer US, 2009. 10.1007/978-0-387-74759-0_344. 91

[47] B. Düring, A. Jüngel, and S. Volkwein. Sequential quadratic programming method for volatility estimation in option pricinig. *Journal of Optimization Theory and Applications*, 139:515–540, 2008. 10.1007/s10957-008-9404-4. 1

[48] R. P. Dwight. Robust mesh deformation using the linear elasticity equations. In H. Deconinck and E. Dick, editors, *Computational Fluid Dynamics 2006*, pages 401–406, Berlin, 2009. Springer. 10.1007/978-3-540-92779-2_62. 125

[49] G. Dziuk. Finite elements for the Beltrami operator on arbitrary surfaces. In S. Hildebrandt and R. Leis, editors, *Partial Differential Equations and Calculus of Variations*, volume 1357 of *Lecture Notes in Mathematics*, pages 142–155. Springer, Berlin, 1988. 10.1007/BFb0082865. 46

[50] Ch. Eck, H. Garcke, and P. Knabner. *Mathematische Modellierung*. Springer, Heidelberg, 2011. 10.1007/978-3-642-18424-6. 111

[51] C. Eichler-Liebenow. *Zur numerischen Behandlung räumlich mehrdimensionaler parabolischer Differentialgleichungen mit linear-impliziten Splitting-Methoden und linearer partieller differentiell-algebraischer Systeme*. PhD thesis, Martin-Luther-Universität Halle-Wittenberg, 1999. http://sundoc.bibliothek.uni-halle.de/diss-online/99/99H128/. 87

[52] I. Ekeland and R. Témam. *Convex Analysis and Variational Problems*, volume 28 of *Classics in Applied Mathematics*. SIAM, Philadelphia, 1999. Reprint of the 1976 original, 10.1137/1.9781611971088. 55, 155

[53] K. Eppler and H. Harbrecht. Efficient treatment of stationary free boundary problems. *Applied Numerical Mathematics*, 56(10–11):1326–1339, 2006. Selected Papers from the First

Chilean Workshop on Numerical Analysis of Partial Differential Equations (WONAPDE 2004), 10.1016/j.apnum.2006.03.017. 101

[54] K. Eppler and H. Harbrecht. Tracking Neumann data for stationary free boundary problems. *SIAM Journal on Control and Optimization*, 48(5):2901–2916, 2009. 10.1137/080733760. 101, 103

[55] K. Eppler and H. Harbrecht. Tracking the Dirichlet data in L^2 is an ill-posed problem. *Journal of Optimization Theory and Applications*, 145:17–35, 2010. 10.1007/s10957-009-9630-4. 101

[56] 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*, 141:513–531, 2009. 10.1007/s10957-008-9507-y. 46

[57] K. Eppler and F. Tröltzsch. Fast optimization methods in the selective cooling of steel. In M. Grötschel, S. O. Krumke, and J. Rambau, editors, *Online Optimization of Large Scale Systems*, pages 185–204. Springer, Berlin, 2001. http://www.springer.com/mathematics/book/978-3-540-42459-8. 1

[58] R. A. Feijóo, A. A. Novotny, E. Taroco, and C. Padra. The topological derivative for the Poisson's problem. *Mathematical Models and Methods in Applied Sciences*, 13(12):1825–1844, 2003. 10.1142/S0218202503003136. 52

[59] J. Fischer. *Optimal Control Problems Governed by Nonlinear Partial Differential Equations and Inclusions*. PhD thesis, Universität Bayreuth, Bayreuth, 2010. http://opus.ub.uni-bayreuth.de/volltexte/2010/709. 30

[60] K. Fister and S. Lenhart. Optimal control of a competitive system with age-structure. *Journal of Mathematical Analysis and Applications*, 291(2):526–537, 2004. 10.1016/j.jmaa.2003.11.031. 1

[61] W. Forst and D. Hoffmann. *Optimization – Theory and Practice*. Springer Undergraduate Texts in Mathematics and Technology. Springer, New York, 2010. 10.1007/978-0-387-78977-4. 31

[62] A. Friedman. *Variational principles and free-boundary problems*. Pure and Applied Mathematics. Wiley-Interscience, New York, 1982. 101

[63] Th.-P. Fries. A corrected XFEM approximation without problems in blending elements. *International Journal for Numerical Methods in Engineering*, 75:503–532, 2008. 10.1002/nme.2259. 125

[64] Ch. Goulaouic and P. Grisvard. Existence de traces pour les éléments d'espaces de distributions définis comme domaines d'opérateurs différentiels maximaux. *Inventiones Mathematicae*, 9:308–317, 1970. 10.1007/BF01425485. 36

[65] E. Griepentrog, M. Hanke, and R. März. Toward a better understanding of differential algebraic equations (introductory survey). Number 2 in Berliner Seminar on Differential-Algebraic Equations. Humboldt-Universität zu Berlin, Institut für Mathematik, 1992. http://edoc.hu-berlin.de/docviews/abstract.php?id=25605. 84, 85

[66] E. Griepentrog and R. März. *Differential-algebraic equations and their numerical treatment*, volume 88 of *Teubner-Texte zur Mathematik*. Teubner, Leipzig, 1986. 84

[67] A. Griewank. On solving nonlinear equations with simple singularities or nearly singular solutions. *SIAM Review*, 27(4):537–563, 1985. 10.1137/1027141. 135

[68] A. Griewank and M. R. Osborne. Newton's method for singular problems when the dimension of the null space is > 1. *SIAM Journal on Numerical Analysis*, 18(1):145–149, 1981. 10.1137/0718011. 135

[69] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*, volume 24 of *Monographs and studies in mathematics*. Pitman Advanced Publishing Program, Boston, 1985. 6, 11, 16, 18, 34, 36, 37, 38, 43, 46, 97, 119, 124, 148, 149, 150

[70] A. Günther and M. Hinze. Elliptic control problems with gradient constraints – variational discrete versus piecewise constant controls. *Computational Optimization and Applications*, 49:549–566, 2011. 10.1007/s10589-009-9308-8. 88

[71] E. Hairer, Ch. Lubich, and M. Roche. *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods*, volume 1409 of *Lecture Notes in Mathematics*. Springer, Berlin, 1989. 10.1007/BFb0093947. 84

[72] E. Hairer and G. Wanner. *Solving ordinary differential equations II: Stiff and differential-algebraic problems*, volume 14 of *Series in Computational Mathematics*. Springer, Berlin, 2nd edition, 2010. 10.1007/978-3-642-05221-7. 84, 85, 87, 88

[73] W. Hamilton, Jr. On nonexistence of boundary arcs in control problems with bounded state variables. *Automatic Control, IEEE Transactions on*, 17(3):338–343, 1972. 10.1109/TAC.1972.1099982. 84

[74] A. Hansbo and P. Hansbo. An unfitted finite element method, based on Nitsche's method, for elliptic interface problems. *Computer Methods in Applied Mechanics and Engineering*, 191(47–48):5537–5552, 2002. 10.1016/S0045-7825(02)00524-8. 125

[75] R. F. Hartl, S. P. Sethi, and R. G. Vickson. A survey of the maximum principles for optimal control problems with state constraints. *SIAM Review*, 37(2):181–218, 1995. 10.1137/1037043. 9, 57, 84, 88

[76] J. Haslinger, K. Ito, T. Kozubek, K. Kunisch, and G. Peichl. On the shape derivative for problems of Bernoulli type. *Interfaces and Free Boundaries*, 11(2):317–330, 2009. 10.4171/IFB/213. 52

[77] J. Haslinger, T. Kozubek, K. Kunisch, and G. Peichl. Shape optimization and fictitious domain approach for solving free boundary problems of Bernoulli type. *Computational Optimization and Applications*, 26(3):231–251, 2003. 10.1023/A:1026095405906. 52

[78] J. Haslinger and P. Neittaanmäki. *Finite Element Approximation for Optimal Shape, Material and Topology Design*. John Wiley & Sons, Chichester, 2nd edition, 1996. 8, 119

[79] D. W. Henderson. Infinite-dimensional manifolds are open susets of hilbert space. *Bulletin of the American Mathematical Society*, 75:759–762, 1969. 10.1090/S0002-9904-1969-12276-7. 77

[80] H. Hermes and J. P. Lasalle. *Functional Analysis and Time Optimal Control*, volume 56 of *Mathematics in Science and Engineering*. Academic Press, New York, 1969. 10.1016/S0076-5392(08)60049-1. 142

[81] R. Herzog and K. Kunisch. Algorithms for PDE-constrained optimization. *GAMM-Mitteilungen*, 33(2):163–176, 2010. 10.1002/gamm.201010013. 1

[82] M. R. Hestenes. *Calculus of Variation and Optimal Control Theory*. Applied Mathematics Series. John Wiley & Sons, New York, 1966. 32

[83] J. Hilgert and K.-H. Neeb. *Structure and Geometry of Lie Groups*. Springer Monographs in Mathematics. Springer, New York, 2012. 10.1007/978-0-387-84794-8. 67, 72, 155

[84] M. Hintermüller and M. Hinze. Moreau-Yosida regularization in state constrained elliptic control problems: Error estimates and parameter adjustment. *SIAM Journal on Numerical Analysis*, 47(3):1666–1683, 2009. 10.1137/080718735. 139

[85] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set method as a semi-smooth Newton method. *SIAM Journal on Optimization*, 13(3):865–888, 2003. 10.1137/S1052623401383558. 116

[86] 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. 10.1137/050637480. 116, 138

[87] 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. 10.1137/040611598. 116, 138

[88] M. Hintermüller and K. Kunisch. Stationary optimal control problems with pointwise state constraints. In *Numerical PDE Constrained Optimization*, volume 72 of *Lecture Notes in Computational Science and Engineering*. Springer, Berlin, 2009. 6, 116

[89] M. Hintermüller and A. Laurain. A shape and topology optimization technique for solving a class of linear complementarity problems in function space. *Computational Optimization and Applications*, 46:535–569, 2010. 10.1007/s10589-008-9201-x. 91

[90] M. Hintermüller and W. Ring. A level set approach for the solution of a state constrained optimal control problem. *Numerische Mathematik*, 98:135–166, 2004. 10.1007/s00211-004-0531-z. 88, 89, 98, 104, 155

[91] M. Hintermüller, F. Tröltzsch, and I. Yousept. Mesh-independence of semismooth Newton methods for Lavrentiev-regularized state constrained nonlinear optimal control problems. *Numerische Mathematik*, 108(4):571–603, 2008. 10.1007/s00211-007-0134-6. 7, 116

[92] M. Hinze and Ch. Meyer. Variational discretization of Lavrentiev-regularized state constrained elliptic optimal control problems. *Computational Optimization and Applications*, 46:487–510, 2010. 10.1007/s10589-008-9198-1. 116

[93] S. -Y. Hsu and Ch. -L. Chang. Mesh deformation based on fully stressed design: the method and 2-d examples. *International Journal for Numerical Methods in Engineering*, 72(5):606–629, 2007. 10.1002/nme.2027. 125

[94] K. Ito and K. Kunisch. Augmented Lagrangian methods for nonsmooth, convex optimization in Hilbert spaces. *Nonlinear Analysis: Theory, Methods & Applications*, 41(5&6):591–616, 2000. 10.1016/S0362-546X(98)00299-5. 116

[95] K. Ito and K. Kunisch. Semi-smooth Newton methods for state-constrained optimal control problems. *Systems & Control Letters*, 50(3):221–228, 2003. 10.1016/S0167-6911(03)00156-7. 116

[96] K. Ito and K. Kunisch. Semi-smooth Newton methods for variational inequalities of the first kind. *ESIAM: Mathematical Modelling and Numerical Analysis*, 37(1):41–62, 2003. 10.1051/m2an:2003021. 116

[97] K. Ito and K. Kunisch. The primal-dual active set method for nonlinear optimal control problems with bilateral constraints. *SIAM Journal on Control and Optimization*, 43(1):357–376, 2004. 10.1137/S0363012902411015. 116

[98] K. Ito and K. Kunisch. *Lagrange Multiplier Approach to Variational Problems and Applications*, volume 15 of *Advances in Design and Control*. SIAM, Philadelphia, 2008. 10.1137/1.9780898718614. 115

[99] K. Ito and K. Kunisch. Semismooth Newton methods for time-optimal control for a class of ODEs. *SIAM Journal on Control and Optimization*, 48(6):3997–4013, 2010. 10.1137/090753905. 142

[100] K. Ito, K. Kunisch, and G. Peichl. Variational approach to shape derivatives. *ESAIM: COCV*, 14(3):517–539, 2008. 10.1051/cocv:2008002. 31, 52

[101] J. J. Sokołowski and A. Żochowski. On the topological derivative in shape optimization. *SIAM Journal on Control and Optimization*, 37(4):1251–1272, 1999. 10.1137/S0363012997323230. 8

[102] D. H. Jacobson, M. M. Lele, and J. L. Speyer. New necessary conditions of optimality for control problems with state-variable inequality constraints. *Journal of Mathematical Analysis and Applications*, 35(2):255–284, 1971. 10.1016/0022-247X(71)90219-8. 9

[103] J. Jahn. *Introduction to the Theory of Nonlinear Optimization*. Spinger, Berlin, 3rd edition, 2007. 10.1007/978-3-540-49379-2. 26

[104] K. Kärkkäinen. *Shape Sensitivity Analysis for Numerical Solution of Free Boundary Problems*. Jyväskylä studies in computing, 58, University of Jyväskylä, 2005. http://urn.fi/URN:ISBN:951-39-2395-9. 105, 106, 112, 120

[105] K. Kärkkäinen and T. Tiihonen. Free surfaces: shape sensitivity analysis and numerical methods. *International Journal for Numerical Methods in Engineering*, 44(8):1079–1098, 1999. 10.1002/(SICI)1097-0207(19990320)44:8<1079::AID-NME543>3.0.CO;2-I. 105

[106] B. Khesin and R. Wendt. *The Geometriy of Infinite-Dimensional Groups*, volume 51 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics*. Springer, 2009. 10.1007/978-3-540-77263-7. 67

[107] D. Kinderlehrer and G. Stampacchia. *An Introduction to Variational Inequalities and Their Applications*, volume 31 of *Classics in Applied Mathematics*. SIAM, Philadelphia, 2000. Reprint of the 1980 original, 10.1137/1.9780898719451. 101

[108] S. Kobayashi and K. Nomizu. *Foundations of Differential Geometry, Vol. I*, volume 15 of *Interscience Tracts in Pure and Applied Mathemtics*. Interscience Publishers (a division of John Wiley & Sons), New York, 1963. 73

[109] I. Kolář, P. W. Michor, and J. Slovák. *Natural Operations in Differential Geometry*. Springer, Berlin, 1993. corrected electronic edition available http://www.emis.de/monographs/KSM/kmsbookh.pdf. 78, 79

[110] K. Krumbiegel and A. Rösch. A virtual control concept for state constrained optimal control problems. *Computational Optimization and Applications*, 43:213–233, 2009. 10.1007/s10589-007-9130-0. 24, 141

[111] N. P. Kruyt, C. Cuvelier, A. Segal, and J. van der Zanden. A total linearization method for solving viscous free boundary flow problems by the finite element method. *International Journal for Numerical Methods in Fluids*, 8(3):351–363, 1988. 10.1002/fld.1650080308. 105

[112] K. Kunisch and D. Wachsmuth. Time optimal control of the wave equation, its regularization and numerical realization. *Accepted for publication in ESAIM: COCV*, 2012. 10.1051/cocv/2011105. 142, 153

[113] S. Lang. *Differential and Riemannian Manifolds*, volume 160 of *Graduate Texts in Mathematics*. Springer, New York, 3rd edition, 1995. http://www.springer.com/mathematics/analysis/book/978-0-387-94338-1. 67, 70, 72, 73, 74, 75, 78, 93, 94, 95

[114] S. Lang. *Algebra*, volume 211 of *Graduate Texts in Mathematics*. Springer, New York, 3rd edition, 2002. http://www.springer.com/mathematics/algebra/book/978-0-387-95385-4. 75, 155

[115] E. Laporte and P. Le Tallec. *Numerical Methods in Sensitivity Analysis and Shape Optimization*. Modeling and Simulation in Science, Engineering and Technology. Birkhäuser, Boston, 2002. http://www.springer.com/birkhauser/mathematics/book/978-0-8176-4322-5. 8

[116] P. -L. Lions. *Optimal Control of Systems Governed by Partial Differential Equations*, volume 170 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 1971. 142

[117] P.-L. Lions and E. Magenes. *Non-Homogeneous Boundary Value Problems and Applications, I*, volume 181 of *Die Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 1972. 18

[118] W. Lucht and K. Strehmel. Discretization based indices for semilinear partial differential algebraic equations. *Applied Numerical Mathematics*, 28(2–4):371–386, 1998. 10.1016/S0168-9274(98)00054-3. 88

[119] W. Lucht, K. Strehmel, and C. Eichler-Liebenow. Indexes and special discretization methods for linear partial differential algebraic equations. *BIT Numerical Mathematics*, 39:484–512, 1999. 10.1023/A:1022370703243. 88

[120] W. S. Martinson and P. I. Barton. A differentiation index for partial differential-algebraic equations. *SIAM Journal on Scientific Computing*, 21(6):2295–2315, 2000. 10.1137/S1064827598332229. 86, 88

[121] H. Maurer. *Optimale Steuerprozesse mit Zustandsbeschränkungen*. Habilitationsschrift, Universität Würzburg, 1976. 9

[122] H. Maurer. On the minimum principle for optimal control problems with state constraints. *Schriftenreihe des Rechenzentrums der Universität Münster*, 41, 1979. 9, 57, 84, 88, 153

[123] H. Maurer and H. J. Pesch. Direct optimization methods for solving a complex state-constrained optimal control problem in microeconomics. *Applied Mathematics and Computation*, 204(2):568–579, 2008. 10.1016/j.amc.2008.05.035. 9

[124] H. Maurer and J. Zowe. First and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems. *Mathematical Programming*, 16:98–110, 1979. 10.1007/BF01582096. 26

[125] Ch. Meyer, U. Prüfert, and F. Tröltzsch. On two numerical methods for state-constrained elliptic control problems. *Optimization Methods and Software*, 22(6):871–899, 2007. 10.1080/10556780701337929. 116

[126] Ch. Meyer, A. Rösch, and F. Tröltzsch. Optimal control problems of PDEs with regularized pointwise state constraints. *Computational Optimization and Applications*, 33(2-3):209–228, 2006. 10.1007/s10589-005-3056-1. 116

[127] Ch. Meyer and I. Yousept. Regularization of state-constrained elliptic optimal control problems with nonlocal radiation interface conditions. *Computational Optimization and Applications*, 44:183–212, 2009. 10.1007/s10589-007-9151-8. 1, 116

[128] J. Mossino. Approximation numérique de problèmes de contrôle optimal avec contrainte sur le contrôle et sur l'état. *Calcolo*, 13:21–62, 1976. 10.1007/BF02575950. 7

[129] F. Murat and J. Simon. Étude de problemès d'optimal design. In J. Céa, editor, *Optimization Techniques Modeling and Optimization in the Service of Man Part 2*, volume 41 of *Lecture Notes in Computer Science*, pages 54–62. Springer, Berlin, 1976. 10.1007/3-540-07623-9_279. 8

[130] J. W. Neuberger. *Sobolev Gradients and Differential Equations*, volume 1670 of *Lecture Notes in Mathematics*. Springer, Berlin, 2nd edition, 2010. 10.1007/978-3-642-04041-2. 46

[131] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2nd edition, 2006. 10.1007/978-0-387-40065-5. 91, 104, 105

[132] A. A. Novotny, R. A. Feijóo, E. Taroco, and C. Padra. Topological sensitivity analysis. *Computer Methods in Applied Mechanics and Engineering*, 192(7-8):803–829, 2003. 10.1016/S0045-7825(02)00599-6. 52

[133] A. A. Novotny, R. A. Feijóo, E. Taroco, and C. Padra. Topological-shape sensitivity method: Theory and applications. In G. M. L. Gladwell, M. P. Bendsøe, N. Olhoff, and O. Sigmund, editors, *IUTAM Symposium on Topological Design Optimization of Structures, Machines and Materials*, volume 137 of *Solid Mechanics and Its Applications*, pages 469–478. Springer, 2006. 10.1007/1-4020-4752-5_45. 52

[134] H. J. Pesch, V. A. Karkhin, A. S. Ilin, A. A. Prikhodovsky, V. V. Plochikhin, M. V. Makhutin, and H.-W. Zoch. Effects of latent heat of fusion on thermal processes in laser welding of alumium alloys. *Science and Technology of Welding and Joining*, 10(5):1–7, 2005. 10.1179/174329305X19286. 1

[135] V. Petzet, Ch. Büskens, H. J. Pesch, V. A. Karkhin, M. V. Makhutin, A. A. Prikhodovsky, and V. V. Ploshikhin. OPTILAS: Numerical optimization as a key tool for the improvement of advanced multi-beam laser welding techniques. In A. Bode and F. Durst, editors, *High Performance Computing in Science and Engineering, Garching 2004*, pages 153–166, Berlin, 2005. Springer. 10.1007/3-540-28555-5. 1

[136] V. Petzet, Ch. Büskens, H. J. Pesch, A. A. Prikhodovsky, V. A. Karkhin, and V. V. Ploshikhin. Elimination of hot cracking in laser beam welding. *PAMM - Proceedings in Applied Mathematics and Mechanics*, 4(1):580–581, 2004. 10.1002/pamm.200410271. 1

[137] V. Petzet, H. J. Pesch, A. A. Prikhodovsky, and V. V. Ploshikhin. Different optimization models for crack-free laser welding. *PAMM - Proceedings in Applied Mathematics and Mechanics*, 5(1):755–756, 2005. 10.1002/pamm.200510352. 1

[138] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko. *The Mathematical Theory of Optimal Processes*. Interscience Publishers (a division of John Wiley & Sons), 1962. 142

[139] O. Prionneau. *Optimal Shape Design of Elliptic Systems*. Springer Series in Computational Physics. Springer, New York, 1983. 8

[140] J. Rang and L. Angermann. Perturbation index of linear partial differential-algebraic equations. *Applied Numerical Mathematics*, 53(2–4):437–456, 2005. 10.1016/j.apnum.2004.08.017. 86, 88

[141] H. Robbins. Junction phenomena for optimal control with state-variable inequality constraints of third order. *Journal of Optimization Theory and Applications*, 31:85–99, 1980. 10.1007/BF00934790. 84

[142] A. Rösch and F. Tröltzsch. Existence of regular Lagrange multipliers for a nonlinear elliptic optimal control problem with pointwise control-state constraints. *SIAM Journal on Control and Optimization*, 45(2):548–564, 2006. 10.1137/050625114. 116

[143] A. Rösch and F. Tröltzsch. On regularity of solutions and Lagrange multipliers of optimal control problems for semilinear equations with mixed pointwise control-state constraints. *SIAM Journal on Control and Optimization*, 46(3):1098–1115, 2007. 10.1137/060671565. 116

[144] A. Rund. *Beiträge zur Optimalen Steuerung partiell-differential algebraischer Gleichungen*. PhD thesis, Fakultät für Mathematik, Physik und Informatik, Universität Bayreuth, 2012. http://opac.uni-bayreuth.de/query/bvb/BV039936043. 142, 154

[145] A. Rund and K. Chudej. Optimal control for a simplified 1D fuel cell model. *Mathematical and Computer Modelling of Dynamical Systems*, 2012. Accepted for publication in Mathematical and Computer Modelling of Dynamical Systems: Methods, Tools and Applications in Engineering and Related Sciences, 10.1080/13873954.2011.642389. 1, 101

[146] S. Schmidt. *Efficient Large Scale Aerodynamic Design Based on Shape Calculus*. PhD thesis, University of Trier, Germany, 2010. http://ubt.opus.hbz-nrw.de/volltexte/2010/569/. 8, 46, 58

[147] S. Schmidt and V. Schulz. Shape derivatives for general objective functions and the incompressible Navier-Stokes equations. *Control and Cybernetics*, 39(3):677–713, 2010. http://control.ibspan.waw.pl:3000/contents/export?filename=Szmidt-Schulz.pdf. 42

[148] W. M. Seiler. Index concepts for general systems of partial differential equations. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 81(S3):629–632, 2001. 10.1002/zamm.20010811591. 86

[149] J. A. Sethian. *Level Set Methods and Fast Marching Methods*, volume 3 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, 2nd edition, 1999. 10.2277/0521645573. 93, 94

[150] J. Simon. Differentiation with respect to the domain in boundary value problems. *Numerical Functional Analysis and Optimization*, 2(7–8):649–687, 1980. 10.1080/01630563.1980.10120631. 8

[151] J. Sokołowski and J.-P. Zolésio. *Introduction to Shape Optimization*, volume 16 of *Springer Series in Computational Mathematics*. Springer, Berlin, 1992. 8, 11, 18, 42, 43, 54, 58, 74, 80, 107, 119

[152] K. Sternberg, K. Chudej, and H. J. Pesch. Partial differential-algebraic dynamic model of a molten carbonate fuel cell. *PAMM - Proceedings in Applied Mathematics and Mechanics*, 4(1):584–585, 2004. 10.1002/pamm.200410273. 1

[153] K. Sternberg, K. Chudej, and H. J. Pesch. Suboptimal control of a 2D molten carbonate fuel cell PDAE model. *Mathematical and Computer Modelling of Dynamical Systems*, 13(5):471–485, 2007. 10.1080/13873950701377288. 1

[154] K. Sternberg, K. Chudej, H. J. Pesch, and A. Rund. Parametric sensitivity analysis of fast load changes of a dynamic MCFC model. *Journal of Fuel Cell Science and Technology*, 5(2):021002, 2008. 10.1115/1.2885400. 1

[155] J. Stoer and R. Z. Bulirsch. *Introduction to Numerical Analysis*, volume 12 of *Texts in applied mathematics*. Springer, New York, 2 edition, 1993. http://www.springer.com/mathematics/computational+science+26+engineering/book/978-0-387-95452-3. 9

[156] T. Tiihonen. Shape optimization and trial methods for free boundary problems. *RAIRO - Modélisation methématique et analyse numérique*, 31(7):805–825, 1997. http://www.numdam.org/item?id=M2AN_1997__31_7_805_0. 58, 112

[157] T. Tiihonen. Fixed point methods for internal free boundary problems. *Numerical Functional Analysis and Optimization*, 19(3–4):399–413, 1998. 10.1080/01630569808816835. 112

[158] F. Tröltzsch. Regular Lagrange multipliers for control problems with mixed pointwise control-state constraints. *SIAM Journal on Optimization*, 15:616–634, 2005. 10.1137/S1052623403426519. 116

[159] F. Tröltzsch. *Optimal Control of Partial Differential Equations*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, Rhode Island, 2010. 6, 7, 33, 34, 37, 38, 58, 149, 150

[160] A. Unger and F. Tröltzsch. Fast solution of optimal control problems in selective cooling of steel. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 81:447–456, 2001. 10.1002/1521-4001(200107)81:7<447::AID-ZAMM447>3.0.CO;2-U. 1

[161] L. Younes. *Shapes and Diffeomorphisms*, volume 171 of *Applied Mathematical Sciences*. Springer, Berlin, 2010. 10.1007/978-3-642-12055-8. 61, 68

[162] J. Zolésio. *Identification de domaines par déformation*. PhD thesis, Univeristé de Nice, France, 1979. 8

[163] J.-P. Zolésio. Weak shape formulation of free boundary problems. *Annali della Scuola Normale Superiore di Pisa, Classe di Scienze 4e sèrie*, 21(1):11–44, 1994. http://www.numdam.org/item?id=ASNSP_1994_4_21_1_11_0. 26

[164] J. Zowe and S. Kurcyusz. Regularity and stability for the mathematical programming problem in Banach spaces. *Applied Mathematics & Optimization*, 5:49–62, 1979. 10.1007/BF01442543. 37, 149

List of symbols and abbreviations

Abbreviations

ALE	arbitrary Lagrangian-Eulerian
BDD	Bryson-Denham-Dreyfus
BiOP	bilevel optimization problem
BVP	boundary value problem
DAE	differential-algebraic equation
FE	finite element
FEM	finite element method
FFT	fast Fourier transformation
iff	if and only if
iOP	inner optimization problem
KKT	Karush-Kuhn-Tucker (conditions)
NC	(first order) necessary conditions
NLP	nonlinear programming nonlinear optimization problem
OC	optimal control
OC-ODE	optimal control of ODEs
OC-PDE	optimal control of PDEs
OCP	optimal control problem
ODE	ordinary differential equation
oOP	outer optimization problem
OP	optimization problem
PDAE	partial differential-algebraic equation
PDAS	primal-dual active set strategy
PDAS-EPF	PDAS equipped with exact path-following
PDE	partial differential equation
set-OCP	set optimal control problem
SQP	sequential quadratic programming

Coefficients and function(-al)s

α	smooth curve in manifold $\mathcal{M}$
$\dot{\alpha}(0)$	tangent vector of a curve α at $t = 0$
λ	Tikhonov regularization parameter
$\mu := \mu^{\max} - \mu^{\min}$	
$\mu_M := \mu\|_M$	where $M \in \{\mathring{\mathcal{A}}, \gamma\}$
$\mu^{\max}$	Lagrange multiplier to the upper state constraint
$\mu^{\max}_{\mathcal{I}} := \mu^{\max}\|_{\mathcal{I}\cup\mathcal{A}_{\min}}$	
$\mu^{\max}_M := \mu^{\max}\|_M$	where $M \in \{\Omega, \Gamma, \mathring{\mathcal{A}}_{\max}, \gamma_{\max}\}$
$\mu^{\min}$	Lagrange multiplier to the lower state constraint
$\mu^{\min}_{\mathcal{I}} := \mu^{\min}\|_{\mathcal{I}\cup\mathcal{A}_{\max}}$	
$\mu^{\min}_M := \mu^{\min}\|_M$	where $M \in \{\Omega, \Gamma, \mathring{\mathcal{A}}_{\min}, \gamma_{\min}\}$
$\mu[.]$	covector field
$\mu_x[.]$	covector to a manifold at point x
$\bar{\sigma}_{\mathcal{J}}$	Lagrange multiplier associated with interface BDD reformulation
ξ	vector field on a manifold
ξ_x	tangent vector to a manifold at point x
$c_{\max}$	regular part of the multiplier $\mu^{\max}$
$c_{\min}$	regular part of the multiplier $\mu^{\min}$
$\mathcal{F}$	reduced objective functional
J	objective (functional)
$\mathfrak{J}$	split objective with active set as explicit variable
$\mathscr{J}$	split objective
$\mathfrak{K}_{\cdot}^{\cdot}$	various quadratic "variational" merit functionals
$\tilde{\mathfrak{K}}_{\cdot}^{\cdot}$	various linear "variational" merit functionals
$K_{\cdot}^{\cdot}$	various merit functionals
$\mathcal{L}$	Lagrangian of the set optimal control problem
$\boldsymbol{n}_M$	(extension of) the outer unit normal vector field of a set M
o	abbreviation for the tuple of variables of the Lagrangian
$\bar{p}_{\mathcal{B}}$	adjoint state associated with state equation in $\mathcal{B}$
$\bar{p}_{\mathcal{J}}$	adjoint state associated with state equation in $\mathcal{J}$
p^{trad}	adjoint state of Casas' necessary conditions
$p^{trad}_M := p^{trad}\|_M$	where $M \in \{\mathcal{I}, \mathcal{A}, \mathcal{A}_{\max}, \mathcal{A}_{\min}\}$
$p^{\max}_{\min}$	interpolation of known parts of the adjoint state
$P_{\mathcal{J}}$	shape adjoint state
$\bar{q}_{\mathcal{B}}$	Lagrange multiplier associated with distributed BDD reformulation

u	control (variable)
u_d	control shift
V	velocity field; sometimes used for a vector space
v	section of a vector bundle
v^α	section of a vector bundle over a curve α, which is transported to the standard fiber
v_α	index section of a vector bundle over a curve α
$y_{\min}^{\max}$	interpolation of the state constraining functions
y	state (variable)
$Y_{\mathcal{J}}$	shape adjoint state
y_d	desired state
$y_{\max}$	upper state constraint
$y_{\min}$	lower state constraint

Miscellaneous notations

$(.)'[V]$	local shape derivative with respect to the velocity field V
$(.)\|_M$	restriction of a function to a set M; frequently used as substitute for τ_M
$(.)^c$	complement of a set
$(.)^t := (.)_t \circ T_t$	variable that is transported back to the original set; confer v^α
$(.)_t$	variable on a transformed set M_t; confer v_α
$(.,.)_H$	inner product of a Hilbert space H
$(. \cdot .)$	scalar product in $\mathbb{R}^2$
$\bar{(.)}$	optimal variables
$\hat{(.)}$	Lagrange multipliers from Appendix B
$\mathring{(.)}$	interior of a set
$\overline{(.)}$	closure of a set (do not confuse with the shorter bar that denotes optimal variables)
$\underline{(.)}$	discretized entities
$. \circ .$	composition of functions
$. \sim .$	equivalence relation in $\mathcal{O}$
$. \sim_{\mathcal{B}} .$	equivalence relation in $\mathcal{H}(\Omega)$
$[.]_{\mathcal{B}}$	equivalence class in $\mathcal{H}(\Omega)$; often abbreviated by $[.]$
$\langle\!\langle . , . \rangle\!\rangle_M := \langle . , . \rangle_{H^{-\frac{3}{2}}(M), H^{\frac{3}{2}}(M)}$	duality pairing for $M \in \{\Gamma, \beta, \gamma\}$
$\langle . , . \rangle_M := \langle . , . \rangle_{H^{-\frac{1}{2}}(M), H^{\frac{1}{2}}(M)}$	duality pairing for $M \in \{\Gamma, \beta, \gamma\}$
$\langle . , . \rangle_{X^*,X}$	duality pairing of a Banach space X
$\|.\|_V$	norm of a normed vector space V
$\subset\subset$	compactly contained

Operators and other notations

$\dot{\alpha}(0)$	tangent vector of a curve α at $t = 0$
δ	unique operator due to Green's formula; typically associated with the Neumann trace operator $\partial_{\boldsymbol{n}}$ in the classical setting; sometimes δ is used for a generic positive constant or the direction in a semiderivative
$\varphi_\circ$	trivializing map of the tangent bundle $T\mathcal{H}(\Omega)$ by means of the composition of transformations
φ_{Id}	trivializing map of the tangent bundle $T\mathcal{H}(\Omega)$ by means of the identity
φ_U	trivializing map of a vector bundle in the neighborhood U
φ_U^x	isomorphism between a fiber $\pi^{-1}(x)$ and the standard fiber B of a vector bundle
κ_M	(mean) curvature of the boundary of a set M
Δ	Laplace operator $\sum_i \partial^2_{x_i}$
Δ_β	Laplace-Beltrami operator along the boundary β
Λ	formal operator associated with bilinear form
τ_M^m	trace operator of m-th order to a set M
τ_M	(Dirichlet) trace (operator) to a set M
ω_M^m	extension operator; right inverse of τ_M^m
$a(.,.)$	bilinear form
$A(\mathcal{B})$	quadratic penalization term at the set $\mathcal{B}$
b_M	oriented distance function to a set $M \subset \mathbb{R}^2$
$\partial_{\boldsymbol{n}}^M$	Neumann trace (operator) or normal derivative to a set M
$\partial_{\boldsymbol{nn}}$	binormal trace (operator) or binormal derivative
∂M	boundary of a set M
d_M	distance function to a set $M \in \mathbb{R}^2$
$d(.,.)$	metric on $\mathcal{G}(\Theta)$ and $\mathcal{X}(\mathcal{B})$, respectively
$d_0(.,.)$	semimetric on $\mathcal{G}(\Theta)$
$d_{\mathcal{K}}(.,.)$	Courant metric (right-invariant metric) on the quotient group $\mathcal{H}(\Omega)/\mathcal{K}(\mathcal{B})$
$d\mathcal{F}(\mathcal{B}; V)$	shape semiderivative of $\mathcal{F}$ at the set $\mathcal{B}$ with respect to the velocity field V
$d^2\mathcal{F}(\mathcal{B}; V, W)$	second order shape semiderivative of $\mathcal{F}$ at the set $\mathcal{B}$ with respect to the velocity fields V and W
D	covariant derivative (operator); "common" differential operator in $\mathbb{R}^N$
D	derivation on $\mathfrak{F}(\mathcal{M})$
D_x	derivation at a point $x \in \mathcal{M}$
grad f	gradient of $f \in \mathfrak{F}(\mathcal{M})$ (in order to distinguish it from an affine connection ∇)
$\nabla\mathcal{F}(.)$	(L^1-) shape gradient of $\mathcal{F}$
$\nabla^2 f[.,.]$	second covariant derivative of a function $f \in \mathfrak{F}(\mathcal{M})$
∇_β	tangential gradient along the boundary β
$\nabla_\eta \xi$	covariant derivative of a vector field $\xi \in \mathfrak{V}(\mathcal{M})$ with respect to $\eta \in \mathfrak{V}(\mathcal{M})$

$\nabla^{\circ}_{\eta}\xi$	covariant derivative of ξ with respect to η induced by the trivializing map $\varphi_{\circ}$
$\nabla^{\mathrm{Id}}_{\eta}\xi$	covariant derivative of ξ with respect to η induced by the trivializing map φ_{Id}
$\nabla_{\eta}\mu[.]$	covariant derivative of a covector field $\mu \in \mathfrak{V}^*(\mathcal{M})$ with respect to $\eta \in \mathfrak{V}(\mathcal{M})$
G	geometry-to-solution operator; sometimes G used for a set or an element of $\mathcal{H}(\Omega)$
g	Riemannian metric
$g_x(.,.)$	inner product of $T_x\mathcal{M}$ induced by a Riemannian metric on a manifold $\mathcal{M}$
Hess f	Hessian of $f \in \mathfrak{F}(\mathcal{M})$ (in order to distinguish it from a second covariant derivative ∇^2
$\mathrm{Id}_{(.)}$	identity operator on a set or space
$P^{s\leftarrow t}_{\alpha}$	parallel translation along a curve α
p_{β}	distance projection on a boundary β
R	retraction
S	typically a control-to-state operator
T	constraining operator
$T_t(f) := \mathrm{Id} + F + tf$	transformation induced by $f := F - \mathrm{Id} \in \Theta_0$ for $F \in \mathcal{H}(\Omega)$ and $t \in [0;\tau]$
$T_t(V)$	transformation induced by velocity field $V \in \mathcal{V}$ for $t \in [0;\tau]$

Spaces and other sets

$\mathcal{A} := \mathcal{A}_{\max} \cup \mathcal{A}_{\min}$	(optimal) active set
$\mathcal{A}_{\max}$	(optimal) upper active set
$\mathcal{A}_{\min}$	(optimal) lower active set
$\beta := \partial\mathcal{B}$	interface between (candidate) in- and active set
$\beta_{\max} := \partial\mathcal{B}_{\max}$	(candidate) upper interface
$\beta_{\min} := \partial\mathcal{B}_{\min}$	(candidate) lower interface
B	base space of a vector bundle (also called standard fiber); sometimes an interface node or a subset of $\mathbb{R}^2$
$\mathcal{B}$	(candidate) active set
$\mathcal{B}_{\max}$	(candidate) upper active set
$\mathcal{B}_{\min}$	(candidate) lower active set
$\mathcal{B}_{[F]} := F(\mathcal{B})$	image set of $\mathcal{B}$ with respect to the equivalence class $[F] \in \mathcal{H}(\Omega)/\mathcal{K}(\mathcal{B})$
B_{ε}	ε-ball around the identity in $\mathcal{H}(\Omega)$
$B_r(x)$	ball with radius r around a point x
$B_x := \pi^{-1}(x)$	fiber of a vector bundle (E,π) through x
$\mathcal{C}$	candidate set for the active set; obtained via coefficients only
$\mathcal{C}_{\max}$	candidate set for the upper active set
$\mathcal{C}_{\min}$	candidate set for the lower active set
$C^0(.)$	space of continuous functions

$C^1(\overline{\mathbb{R}^2}, \mathbb{R}^2)$	space of all uniformly continuously differentiable and bounded functions
$C^1(\mathbb{R}^2, \mathbb{R}^2)$	space of all continuously differentiable functions
$C_0^\infty(.)$	space of infinitely differentiable functions with compact support
$C^{1,1}(\overline{\mathbb{R}^2}, \mathbb{R}^2)$	space of all Lipschitz-continuous differentiable and bounded functions
(E, π)	vector bundle (with different meaning)
E	total space of a vector bundle
$\mathfrak{F}(\mathcal{M})$	space of real-valued, smooth functions defined on $\mathcal{M}$
$\mathfrak{F}_x(\mathcal{M})$	space of real-valued, smooth functions defined in a neighborhood of $x \in \mathcal{M}$
$\mathcal{G}(\Theta)$	space/group of perturbations of identity
$\mathcal{G}(\Theta_0) = \mathcal{H}(\Omega)$	space/group of perturbations of identity, which do not act on Ω^c
Γ	boundary of Ω
$\gamma := \gamma_{\max} \cup \gamma_{\min}$	(optimal) interface
$\gamma_{\max} := \partial\mathcal{A}_{\max}$	(optimal) upper interface
$\gamma_{\min} := \partial\mathcal{A}_{\min}$	(optimal) lower interface
Ω	spacial domain; sometimes called holdall
$\boldsymbol{H}$	horizontal subbundle
$\mathcal{H}(\Omega)$	subgroup of $\mathcal{G}(\Theta)$; stabilizer of Ω^c with respect to the pointwise group operation
$H^1(., \Delta)$	domain Hilbert space of the Laplacian
$H_s^1(.) := \{\varphi \in H^1(.) \mid \varphi\|_s = 0\}$	
$H^s(.) := W^{s,2}(.)$	L^2-based Sobolev space for $s > 0$
$H_0^s(.)$	closure of $C_0^\infty(.)$ in $H^s(.)$ for $s > 0$
$H^{-s}(.)$	dual space of $H_0^s(.)$
$\mathcal{I} := \Omega \subset \mathcal{A}$	(optimal) inactive set
$\mathcal{J} := \Omega \setminus \mathcal{B}$	(candidate) inactive set
$\mathcal{K}(\mathcal{B})$	subgroup of $\mathcal{H}(\Omega)$; stabilizer of $\mathcal{B}$ with respect to the setwise group operation
$L^2(.)$	space of square-integrable functions
$L^2(., \Delta)$	very weak Sobolev space
$\mathcal{M}$	(real) manifold
$\mathcal{M}(.)$	space of (signed) regular Borel measures
$M_t := T_t(M)$	Image of a set M with respect to a transformation $T_t := T_t(V)$ or $T_t := T_t(f)$
$\mathcal{O}$	family of feasible sets
$\mathcal{O}(\mathcal{B})$	set of all sets in $\mathcal{O}$ which are homeomorphically homotope to $\mathcal{B}$ in Ω
$\mathcal{T}(B_\varepsilon)$	family of paths in $\mathcal{H}(\Omega)$ associated with B_ε
$\Theta := C^{1,1}(\overline{\mathbb{R}^2}, \mathbb{R}^2)$	base space for the construction of $\mathcal{G}(\Theta)$
Θ_0	closed subspace of Θ
$T\mathcal{M}$	tangent bundle of a manifold $\mathcal{M}$

$T^*\mathcal{M}$	cotangent bundle of a manifold $\mathcal{M}$
$T_x\mathcal{M}$	tangent space to a manifold $\mathcal{M}$ at x
$T_x^*\mathcal{M}$	cotangent space to a manifold $\mathcal{M}$ at x
U	typically a neighborhood on a manifold; sometime an admissible set of controls
$\boldsymbol{V}$	vertical subbundle
$\mathcal{V}$	space of velocity fields
$\mathfrak{V}(\mathcal{M})$	set of all vector fields on $\mathcal{M}$
$\mathfrak{V}(\mathcal{M})^*$	set of all covector fields on $\mathcal{M}$
$V(\Lambda)$	domain Hilbert space of the formal operator Λ
$W^{s,p}(.)$	$L^p(.)$ based Sobolev for $s \in \mathbb{R}$
$\mathcal{X}(\mathcal{B})$	family of all images of $\mathcal{B}$ which can be obtained via transformations in $\mathcal{H}(\Omega)$
X^*	dual space of a Banach space X

Index