

Eingereicht von  
Dipl.-Ing.  
Peter Gangl,  
Bakk. Techn.

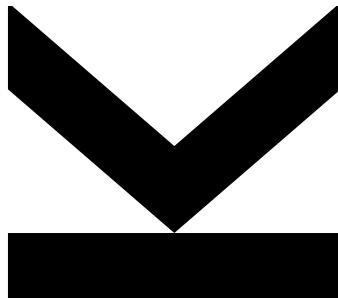
Angefertigt am  
Doktoratskolleg  
“Computational  
Mathematics”

Betreuer und  
Erstbeurteiler  
O. Univ.-Prof.  
Dipl.-Ing. Dr.  
Ulrich Langer

Zweitbeurteiler  
Univ.-Prof. Dr.  
Fredi Tröltzsch

Dezember 2016

# Sensitivity-Based Topology and Shape Optimization with Ap- plication to Electrical Machines



Dissertation  
zur Erlangung des akademischen Grades  
Doktor der technischen Wissenschaften  
im Doktoratsstudium  
Technische Wissenschaften



## Abstract

This thesis deals with topology and shape optimization methods for finding optimal geometries of devices from electrical engineering. As a model problem, we consider the design optimization of an electric motor. Here, the performance of the motor depends on the electromagnetic fields in its interior, which, among other factors, also depend on the geometry of the motor via the solution to Maxwell's equations. In our model, we use a special regime of Maxwell's equations, namely the partial differential equation (PDE) of nonlinear magnetostatics, and consider a two-dimensional setting of the electric motor. Thus, we are facing a PDE-constrained optimization problem where the unknown is the geometry of a given part of the motor.

An important tool for solving shape optimization problems is the *shape derivative*, i.e., the sensitivity of the domain-dependent objective function with respect to a smooth variation of a boundary or material interface. We derive the shape derivative for the optimization problem at hand, which involves a nonlinear PDE constraint, by means of a Lagrangian approach. We employ the shape derivative to obtain an improved design of the electric motor. One shortcoming of the class of shape optimization methods is that they can only vary boundaries or interfaces of given designs and cannot alter their topology, i.e., they cannot introduce holes or new components.

Using topology optimization methods, also the connectivity of a domain can change during the optimization procedure. In this thesis, we focus on topology optimization approaches based on topological sensitivities. On the one hand, we consider the sensitivities of the objective function with respect to a local variation of the material. On the other hand we rigorously derive the *topological derivative*, i.e., the sensitivity of a domain-dependent objective function with respect to the introduction of a hole in the interior of the domain. The latter approach is particularly involved in this case due to the nonlinear PDE constraint. The information provided by these sensitivities can be used for determining optimal designs whose topology may be different from the topology of the initial design.

In both classes of methods, we start with an initial geometry consisting of several materials and successively update the material interfaces in the course of the optimization procedure. The update is based on topological or shape sensitivities, which depend on the solutions to two PDEs (the state equation and the adjoint equation of the optimization problem). These PDEs are approximately solved by means of the finite element method on a triangular grid in each iteration. In order to obtain accurate solutions to these PDEs, the evolving interface should be resolved by the finite element discretization. We introduce a *local mesh adaptation strategy* which modifies the mesh only in a neighborhood of the interface and show optimal order of convergence as the mesh size approaches zero.

Finally, we combine the three components mentioned above and apply it to the optimization of electric motors. In a first step, we perform topology optimization in order to find the optimal connectivity of the design. In a second step, we use shape optimization together with the proposed mesh adaptation strategy as a post-processing in order to get smoother designs.

## Zusammenfassung

Diese Arbeit behandelt Methoden der Topologie- und Formoptimierung zur Bestimmung von optimalen Geometrien in Anwendungen aus der Elektrotechnik. Als ein Modellproblem betrachten wir die Optimierung der Geometrie eines Elektromotors. Das Verhalten des Motors wird bestimmt von den elektromagnetischen Feldern im Inneren des Motors, welche wiederum, über die Lösung der Maxwell-Gleichungen, auch von der Geometrie des Motors abhängen. Wir verwenden einen Spezialfall der Maxwell-Gleichungen, nämlich die partielle Differentialgleichung der nichtlinearen Magnetostatik. Des Weiteren betrachten wir ein zweidimensionales Modell des Elektromotors. Das Optimierungsproblem besteht also darin, die Geometrie eines gewissen Teils eines Elektromotors zu identifizieren, welche unter der Nebenbedingung einer nichtlinearen partiellen Differentialgleichung das bestmögliche Verhalten des Motors zur Folge hat.

Ein wichtiges Werkzeug zur Behandlung von Formoptimierungs-Problemen ist die *Formableitung*, also die Sensitivität eines Funktionals, das von der Form eines Gebietes abhängt, bezüglich einer glatten Variation des Randes dieses Gebietes. Wir berechnen die Formableitung für das beschriebene Formoptimierungs-Problem, welches eine nichtlineare partielle Differentialgleichung beinhaltet, mittels eines Lagrange'schen Zuganges und verwenden die Formableitung um eine verbesserte Geometrie des Elektromotors zu erhalten. Ein Nachteil der Klasse der Formoptimierungs-Verfahren ist, dass diese nur den Rand eines Gebietes variieren können, nicht aber seine Topologie. Es können also keine Löcher oder neuen Komponenten eingeführt werden.

Mittels Verfahren der Topologieoptimierung kann auch die Anzahl der zusammenhängenden Komponenten eines Gebietes im Laufe des Optimierungsverfahrens verändert werden. In dieser Arbeit behandeln wir Zugänge zur Topologieoptimierung, die auf topologischen Sensitivitäten beruhen. Einerseits betrachten wir die Sensitivität des Zielfunktionals bezüglich einer lokalen Variation des Materials. Andererseits berechnen wir rigoros die *topologische Ableitung*, also die Sensitivität eines Funktionals, welches von einem Gebiet abhängt, bezüglich der Einführung eines Loches im Inneren des Gebietes. Aufgrund der Nebenbedingung in Form einer nichtlinearen partiellen Differentialgleichung ist letzterer Zugang besonders aufwändig. Die Information aus diesen Sensitivitäten kann verwendet werden, um optimale Geometrien zu erhalten, deren Topologie von jener des ursprünglichen Designs abweicht.

In beiden Klassen von Verfahren gehen wir von einer Anfangsgeometrie aus, welche aus verschiedenen Materialien besteht, und bewegen die Interfaces zwischen den verschiedenen Materialien in eine Richtung, die mithilfe der Sensitivitäten des Funktionals bezüglich der Form oder Topologie des Gebietes bestimmt wird. Um diese Sensitivitäten berechnen zu können, müssen jedoch zwei partielle Differentialgleichungen gelöst werden (die Zustandsgleichung und die adjungierte Gleichung des Optimierungsproblems), was wir näherungsweise

mittels des Verfahrens der Finiten Elemente auf einem Dreiecksgitter bewerkstelligen. Um möglichst genaue Näherungslösungen dieser Gleichungen zu erhalten, sollte das Interface immer durch die Diskretisierung aufgelöst werden. Zu diesem Zweck führen wir eine *lokale Gitter-Anpassungs-Strategie* ein, welche das Gitter nur in einer Umgebung des Interfaces modifiziert, und zeigen optimale Konvergenzordnungen bei immer feiner werdendem Gitter.

Schließlich kombinieren wir diese drei Komponenten und wenden sie auf das Problem der Optimierung von Elektromotoren an. In einem ersten Schritt wenden wir ein Topologieoptimierungs-Verfahren an, um die optimale Topologie des Gebietes zu finden. In einem zweiten Schritt verwenden wir Formoptimierung gemeinsam mit der Modifizierung des Gitters zur Nachbearbeitung, um glattere Geometrien zu erhalten.

# Acknowledgments

First of all, I would like to express my gratitude to my supervisor Prof. Ulrich Langer who introduced me to the exciting field of shape and topology optimization in connection with electrical machines, which turned out to be not only mathematically challenging, but also practically relevant and a very promising topic for the future. I would like to thank him for guiding this thesis and for all his support on the way. Furthermore, I thank him for encouraging and enabling me to visit international conferences and to spend some time abroad during my PhD.

I am very thankful for the perfect environment in which I could conduct my research as well as the financial support both of which were provided by the Doctoral Program DK1214 (project DK4) on Computational Mathematics under the speaker Prof. Peter Paule, which is an excellence program of the Austrian Science Fund (FWF).

Furthermore, I gratefully acknowledge the collaboration and many fruitful discussions with Gerd Bramerdorfer and Prof. Wolfgang Amrhein from the institute for electrical drives and power electronics (EAL) of JKU as well as Armin Fohler and Siegfried Silber from the Linz Center of Mechatronics (LCM) GmbH who provided the ideas and the data for the numerical experiments involving electrical machines throughout this thesis, in particular for the results of Chapter 8. I am also thankful for the partial financial support provided by the LCM GmbH.

I want to thank Prof. Fredi Tröltzsch for agreeing to review this thesis and also for enabling me to spend a total of five months in his working group at the Technical University of Berlin. In Berlin, I was very happy to meet and cooperate with Antoine Laurain, Kevin Sturm and Houcine Meftahi, which resulted in Chapter 6 of this thesis. I am also very grateful to Prof. Samuel Amstutz for the two months I could spend at the University of Avignon and the many discussions I had with him. Furthermore, I would like to thank Prof. Walter Zulehner for many discussions over the past years.

Finally, I want to express my sincerest thanks to my parents Franz and Gertraud Gangl as well as my siblings Magdalena, Susanne and Elias without whose support this thesis would not have been possible. Furthermore, I want to thank my colleague and friend Daniela Saxenhuber on behalf of many others who accompanied me along my way over the past nine years of studying mathematics.

Peter Gangl  
Linz, December 2016



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	State of the Art in Topology and Shape Optimization . . . . .	1
1.2.1	Topology Optimization . . . . .	3
1.2.2	Shape Optimization . . . . .	7
1.3	Finite Element Methods for Interface Problems . . . . .	10
1.4	Organization of the Thesis . . . . .	11
<b>2</b>	<b>Physical Background</b>	<b>13</b>
2.1	Introduction to Electrical Machines . . . . .	13
2.2	Physical Model . . . . .	14
2.3	Model Problem . . . . .	17
2.4	Physical Properties . . . . .	20
2.5	Analysis of Two-dimensional Nonlinear Magnetostatics . . . . .	22
2.5.1	Existence and Uniqueness for the State Equation . . . . .	22
2.5.2	Newton Operator and Adjoint Equation . . . . .	24
<b>I</b>	<b>Topology Optimization</b>	<b>27</b>
<b>3</b>	<b>Material Sensitivities in Takahashi's On/Off Method</b>	<b>29</b>
3.1	Discrete Sensitivity Analysis . . . . .	30
3.2	Generalization to Continuous Level . . . . .	31
3.3	Application to Model Problem . . . . .	35
<b>4</b>	<b>Topological Derivative for Magnetostatic Problem</b>	<b>39</b>
4.1	Preliminaries . . . . .	41
4.1.1	Simplified Model Problem . . . . .	41
4.1.2	Perturbed State Equation . . . . .	42
4.1.3	Expansion of Cost Functional . . . . .	43
4.2	Linear Case . . . . .	44
4.2.1	Application to the Model Problem . . . . .	45
4.3	Preliminaries for the Nonlinear Case . . . . .	48
4.3.1	Summary of the Procedure . . . . .	48
4.3.2	Requirements . . . . .	49
4.3.3	Properties . . . . .	51

4.3.4	Weighted Sobolev Spaces . . . . .	55
4.3.5	Relation to Previous Work . . . . .	56
4.4	Topological Asymptotic Expansion: Case I . . . . .	57
4.4.1	Variation of Direct State . . . . .	58
4.4.2	Variation of Adjoint State . . . . .	65
4.4.3	Topological Asymptotic Expansion . . . . .	68
4.4.4	Proofs . . . . .	72
4.5	Topological Asymptotic Expansion: Case II . . . . .	103
4.5.1	Simplified Model Problem . . . . .	104
4.5.2	Perturbed Nonlinear Equation . . . . .	104
4.5.3	Expansion of Cost Functional . . . . .	104
4.5.4	Variation of Direct State . . . . .	104
4.5.5	Variation of Adjoint State . . . . .	110
4.5.6	Topological Asymptotic Expansion . . . . .	113
4.6	Polarization Matrices . . . . .	117
4.6.1	Preliminaries . . . . .	117
4.6.2	Change of Variables . . . . .	119
4.6.3	Case I . . . . .	120
4.6.4	Case II . . . . .	124
4.7	Computational Aspects . . . . .	125
4.7.1	Numerical Experiments . . . . .	129
4.8	Application to Model Problem . . . . .	129
<b>5</b>	<b>Comparison</b>	<b>137</b>
5.1	Linear Case . . . . .	137
5.1.1	Continuous Setting . . . . .	137
5.1.2	Discrete Setting . . . . .	138
5.2	Nonlinear Case . . . . .	139
5.3	General Remarks . . . . .	140
<b>II</b>	<b>Shape Optimization</b>	<b>143</b>
<b>6</b>	<b>Shape Optimization</b>	<b>145</b>
6.1	Introduction . . . . .	145
6.1.1	Computation of Shape Derivative . . . . .	147
6.1.2	Representation of Shape Derivative . . . . .	148
6.2	Preliminaries . . . . .	149
6.3	Existence of Optimal Shapes . . . . .	151
6.4	Derivation of the Shape Derivative . . . . .	151
6.4.1	Preliminaries . . . . .	151
6.4.2	An Abstract Differentiability Result . . . . .	152
6.4.3	Adjoint Equation . . . . .	153
6.5	Shape Derivative of the Cost Function . . . . .	153
6.6	Application to Model Problem . . . . .	155
6.6.1	Numerical Method . . . . .	156
6.6.2	Numerical Results . . . . .	158

<b>III Interface Handling</b>	<b>161</b>
<b>7 A Local Mesh Modification Strategy</b>	<b>163</b>
7.1 Preliminaries . . . . .	164
7.2 Description of the Method . . . . .	165
7.3 Maximum Angle Condition . . . . .	166
7.4 A Priori Error Estimates . . . . .	170
7.4.1 Numerical Experiments . . . . .	176
7.5 Condition Number . . . . .	176
7.5.1 Scaling of the Basis Functions . . . . .	178
7.5.2 Numerical Experiments . . . . .	182
<b>IV Combined Topology and Shape Optimization with Interface Handling</b>	<b>187</b>
<b>8 Numerical Optimization Results</b>	<b>189</b>
8.1 Implementation . . . . .	189
8.2 Combined Topology and Shape Optimization with Interface Handling . . . . .	190
8.3 Minimizing Total Harmonic Distortion . . . . .	190
8.4 Maximizing Torque for Synchronous Reluctance Motor . . . . .	193
<b>List of Notation</b>	<b>201</b>
<b>List of Figures</b>	<b>203</b>
<b>Bibliography</b>	<b>205</b>



# Chapter 1

## Introduction

### 1.1 Motivation

Over the past decades, electrical machines have become an integral part of our everyday life. They appear in household appliances, industrial applications, and of course also in the context of electromobility, a sector which is more topical now than ever it has been. It is desirable that these machines are designed in such a way that they fulfill their purpose in the best possible way. While, in former times, this design process was mainly based on the intuition and experience of engineers, nowadays computer-aided engineering tools which use numerical simulation and optimization algorithms have become an indispensable component. So far, the most commonly used methods for obtaining good designs of electrical machines include evolutionary algorithms where the design variables are geometric parameters such as the dimensions of a certain part of the motor. This approach yields a restriction of the set of possible designs as only a certain number of parameters can be considered. Here, more general topology and shape optimization methods allow for a wider variety of optimal designs, possibly including designs which could not have been imagined beforehand. In particular, with the rise of 3D printing and additive manufacturing technologies, more and more complex structures can be produced and, therefore, these general design optimization techniques seem to be a promising tool for optimal design problems not only in the context of electrical machines, but in many areas of engineering.

### 1.2 State of the Art in Topology and Shape Optimization

In this section, we give an overview over the most common approaches to mathematical design optimization and highlight their advantages and challenges, as well as relations between different approaches. The general goal in these kind of optimization problems is always to find a layout of a given object, e.g., a mechanical structure or an electrical device, such that its performance is as good as possible. The performance is always measured by means of an objective functional which assigns a real number to any given admissible design. Throughout this thesis, we will denote this objective functional by  $\mathcal{J}$  and we will always be interested in the minimization of  $\mathcal{J}$ , keeping in mind that a maximization of  $\mathcal{J}$  can be achieved by a minimization of the functional  $-\mathcal{J}$ . Although, in a wide range of practical applications, designs should be as good as possible with respect to several criteria, we restrict ourselves

to the case of a single objective function in this thesis. We remark that the most common approach to multi-objective optimization is by forming Pareto fronts, see e.g. [73], where designs obtained by single-objective optimization algorithms can be compared with respect to their overall performance. We also mention that, in our framework of single-objective optimization, it is always possible to choose a weighted sum of several performance criteria as an objective functional. However, it is usually not clear how to choose the weighting factors in this weighted sum.

Throughout this thesis, the design variable is the geometry of a given device and we assume all data of the problem such as loads or boundary conditions to be given and fixed.

The field of design optimization has its origins in the optimization of mechanical structures where most often the goal is either to minimize the compliance (i.e., to maximize the stiffness) of the structure while satisfying a volume constraint on the material used, or to minimize the volume while keeping a certain stiffness. However, these methods have also been applied to a wide variety of other industrial applications over the past decades. Design optimization methods are often categorized into the following three classes of methods, see e.g. [67]:

1. **Sizing optimization:** The most common sizing optimization problems involve trusses, i.e., mechanical structures consisting of several bars, where the goal is to find the optimal thickness of each bar in order to minimize the compliance of the structure under a given load. In this thesis, we will not deal with this class of methods and refer the reader to the review paper [40] and the references therein.
2. **Shape optimization:** In shape optimization, one is interested in finding the optimal shape of a boundary of a domain or of a material interface within a given domain. Here, we distinguish between parametric and non-parametric shape optimization. As the term suggests, in parametric shape optimization, the design is described by a set of parameters such as the dimensions or orientation of an object. This class of methods results in optimization problems with a finite dimensional design space. We remark that also shapes represented by spline curves, which are defined by so-called control points, fall into this category. In this thesis, we will only deal with non-parametric shape optimization, where (before discretization) the design space is infinite-dimensional. This class of shape optimization methods is based on the concept of the shape derivative, see Sections 1.2.2 and 6.1.
3. **Topology optimization:** The class of topology optimization methods does not only allow for a variation of boundaries or interfaces of a design, but, in contrast to shape optimization methods, also allows for a change of the topology. This means that, when using topology optimization methods, the number of holes or connected parts of the structure is not fixed from the beginning. This is a serious advantage of this class of methods over shape optimization methods since, in many practical applications, the optimal topology of an object is not known a priori.

We remark that the border between shape and topology optimization methods is not always very sharp, as there exist methods such as the level set method which theoretically are capable of altering the topology, but are very unlikely to do that in practice, see Section 1.2.1. We give an overview over design optimization methods which allow for topological changes in Section 1.2.1, and discuss the main aspects of “classical” shape optimization, i.e., smooth variations of given boundaries or material interfaces, in Section 1.2.2.

### 1.2.1 Topology Optimization

The concept of topology optimization originates from mechanical engineering, but has been applied to a large variety of other applications ranging from fluid dynamics [79, 92] over acoustics [77] to electromagnetics [102, 129, 162, 173, 174]. This subsection is meant to give a brief overview over the most common methods of topology optimization. To a large amount, it is based on the review articles [60, 145, 182, 208]. For a more detailed introduction to the single approaches we refer to these papers and the references therein.

The early work by Michell from 1904 on optimality criteria for the optimal design of trusses [149] is considered to be the first paper on topology optimization [182]. Later, in the 1970s, Rozvany and collaborators dealt with exact analytical optimization of similar grid-like structures [171, 180, 181]. However, the starting point of numerical topology optimization is widely considered to be the seminal paper by Bendsøe and Kikuchi [41] introducing the *homogenization method* for topology optimization, followed by the paper [38], where Bendsøe introduced what is now known as the *Solid Isotropic Material with Penalization (SIMP)* method, giving rise to the large class of *density-based* methods.

#### 1.2.1.1 Homogenization Method

The idea of the homogenization method is to represent a domain as a periodic microstructure (usually consisting of rectangular cells like a regular quadrilateral finite element grid) and then to find the optimal layout for each cell. Each of these cells is considered to consist of material and void regions (often a rectangular hole surrounded by solid material) and the dimensions and orientations of these holes are the design variables with respect to which the optimization is performed. Finally, one ends up with a perforated design which can be interpreted as a microstructure. A black-and-white structure can be obtained by setting those cells which are mostly occupied with material to solid and the other cells to void [60]. The method uses several degrees of freedom for each of the cells, amounting in a large number of degrees of freedom, which is considered a significant drawback of this method. For more details on the homogenization method, we refer the reader to the research papers [41, 215, 216] as well as the monograph [2].

#### 1.2.1.2 Density Methods

In topology optimization, one is interested in finding the optimal distribution of a given material within a design domain. A possible design can be represented by a function  $\rho$  which takes the value 1 in areas of material and the value 0 elsewhere. We remark that, in applications of mechanical engineering, if  $\rho$  is 0, the elasticity tensor vanishes and the global stiffness matrix becomes singular. Therefore, it is common practice in density-based topology optimization of mechanical structures to replace the value of 0 by a small, but positive number  $\rho_{min} > 0$ . The idea of density-based approaches to topology optimization is to relax this strict 0–1 nature of the problem by allowing the function  $\rho$  to attain any value between 0 and 1. The function  $\rho$  is called a density variable. This procedure amounts to the variable thickness sheet problem introduced in 1973 in [179]. In order to enforce a 0–1 structure of the final design, Bendsøe combined this idea with a penalization of intermediate density values in [38], i.e., he replaced the density function  $\rho$  in the state equation (and only there) by a penalized version of the density,  $\tilde{\rho}(\rho) = \rho^p$  for some  $p > 1$ . In combination with a constraint on the volume of

the arising structure, the algorithm favors the use of “black” and “white” regions, i.e., regions where  $\rho = 1$  and  $\rho = 0$ , respectively, because intermediate values “give very little stiffness at an unreasonable cost” [38]. As remarked in [192], a constraint which limits the volume is important for this penalizing effect to appear. The method described here together with the choice  $\tilde{\rho}(\rho) = \rho^p$  for some  $p > 0$  became well-known as the SIMP method. We remark that the method is sensitive with respect to the value of  $p$  and that good results are usually obtained by using  $p = 3$  or by gradually increasing the parameter from  $p = 1$  to higher values in the course of the optimization procedure [192]. For a comparison of different material interpolation schemes, see [42]. The penalized topology optimization problem is usually solved by a gradient-based optimization algorithm such as the Optimality Criteria method (see, e.g. [104]) or the Method of Moving Asymptotes (MMA, [201]).

While the penalization of intermediate density values yields designs with a 0–1 structure, these problems usually lack existence of a solution, a fact which often results in a mesh dependence of the optimized designs. For a detailed survey on the numerical problems resulting from the ill-posedness of such problems we refer the reader to [193]. The most popular approach to regularizing these ill-posed problems is by applying a filter to the sensitivities. This means that one replaces the actual sensitivity at a point by an average over the sensitivities in a neighborhood of a certain radius  $r_{min}$ , called the filter radius. Other approaches include a filtering of the density variable, see e.g. [97], adding a bound on the perimeter of the arising structure or on the gradient of the density variable  $\rho$ , see [193].

A more detailed overview of density-based topology optimization methods can be found in the monographs [39, 43] as well as in the review papers [182, 192].

### 1.2.1.3 Phase-Field Method

The phase-field method for topology optimization is a density-based method using a linear material interpolation,  $\tilde{\rho}(\rho) = \rho$ . A regularization is achieved by adding a term to the cost functional which approximates the total variation of the density variable. This term is a Cahn-Hilliard type functional, see [59], which itself is a weighted sum of two terms. One of these two terms causes a regularizing effect whereas the other term penalizes intermediate density values. We mention that the choices of the weighting factor between these two parts, as well as the weight of the Cahn-Hilliard type functional relative to the objective function, are often crucial for obtaining good results. The phase field method has been applied to many topology optimization problems, see e.g. [55, 91, 92, 197, 220].

### 1.2.1.4 Level Set Methods

In [165], Osher and Sethian introduced a framework for describing interfaces which evolve along a given velocity field in an implicit way, as zero level sets of an evolving scalar function  $\psi = \psi(x, t)$ . This level set function attains positive values in one subdomain and negative values in the other, while the material interface is given by the zero level set of the function  $\psi$ ,  $\Gamma(t) = \{x | \psi(x, t) = 0\}$ . Here,  $t$  is a pseudo-time variable. The evolution of  $\psi$  is given by the solution to the Hamilton-Jacobi equation

$$\frac{\partial}{\partial t} \psi + V \cdot \nabla \psi = 0, \quad (1.1)$$

where  $V$  determines the direction of the evolution and  $t$  is a pseudo-time parameter. In [166] this framework was first applied to structural optimization with the velocity field  $V$  given according to shape sensitivities. This concept has been used by many researchers [5–8, 72, 209]. We make a few comments on level set methods as a tool for optimal design:

- While using this implicit representation of a shape by means of a level set function alleviates dealing with topological changes, the level set method as described above must be categorized as a shape optimization method since the optimization process is guided by shape sensitivities. Therefore, no information about the sensitivity with respect to topological changes is available. It is possible that components merge or disappear, but no holes or new components can be created. For this reason, it is a very common approach to start with a perforated initial design with many circular holes and to let them merge or disappear in the course of the optimization procedure, see e.g. [3, 8].
- The numerical treatment of the Hamilton-Jacobi equation is non-trivial since it is a convection-dominated problem. Possible remedies for this issue include the introduction of an artificial diffusive term or the stabilization using upwinding techniques, see e.g. [8, 209]. We remark that, when using unstructured grids, special upwinding techniques must be applied, see e.g. [52].
- A further issue which has to be taken care of in the level set method is the fact that the level set function is likely to become very flat in the course of the optimization procedure. This problem is usually avoided by repeatedly re-initializing the level set function to a signed distance function which satisfies  $|\nabla\psi| = 1$  everywhere. For more details, see [164, 191, 208].

For an overview over different aspects of the level set method for shape and topology optimization, we refer the reader to the review papers [54, 208] as well as the monographs [164, 191].

### 1.2.1.5 Topological Derivative

The concept of the topological derivative was introduced in [78] as a means to allow for changes of the topology in the course of a classical shape optimization method. The topological derivative of a domain-dependent functional at an interior point of the domain describes its sensitivity with respect to the introduction of a hole around that point. We will deal with the topological derivative in detail in Chapter 4 and also give a thorough introduction there.

### 1.2.1.6 On/Off-type methods

Here, we mention two sensitivity-based methods which do not introduce intermediate materials, but rather switch between only two possible states (on or off).

In [163], Okamoto and Takahashi propose the gradient-based *On/Off method* for determining the optimal design of a magnetic shield for a magnetic recording system. After discretization, for each element of the finite element mesh, the sensitivity of the objective functional with respect to a perturbation of the magnetic reluctivity in only this element is computed. Based on this information, elements are switched between the two states at the most effective positions.

The idea of the evolutionary structural optimization (ESO) method in structural mechanics [212] is to start out from an initial design where all of the design subdomain is occupied with material and then to gradually remove material at the most favorable positions according to some sensitivities like the strain energy density. Later, the method was extended to the bi-directional evolutionary structural optimization method (BESO) which also allowed for the re-introduction of previously eliminated material, see [175].

The biggest difference between these two methods is probably that ESO/BESO comes from mechanical engineering and the On/Off method by the group of Takahashi was introduced with regard to applications from electrical engineering. The main difference between these two areas is that, in electromagnetics, void can be regarded as just another material with a different positive material coefficient and does not have to be excluded from the computational domain like in mechanics.

In Chapter 3, we will analyze in detail the On/Off method. In Chapter 5 we will draw a comparison between the material sensitivities used in the On/Off method and the topological derivative in the framework of linear and nonlinear two-dimensional magnetostatics.

### 1.2.1.7 Derivative-Free Approaches

We also mention the class of nature-inspired, derivative-free stochastic optimization algorithms which are widely used among engineers. Their big advantage over derivative-based methods is that they are not as prone to getting stuck in local minima. However, this comes at the price of much larger computational costs. We mention the class of genetic algorithms [94], the particle swarm optimization method [210] and the firefly algorithm [1]. An overview over this class of methods can be found in [214].

### 1.2.1.8 Interrelations

Finally, we mention some connections between some of the approaches mentioned above, which have been studied in the literature.

A very widely used approach is to couple the level set method, which is guided by shape sensitivity information, with the topological derivative in order to allow for a change of the topology in the course of the optimization procedure. We mention the papers [5, 6, 53, 105].

A different level set approach is given in [16, 19]. In contrast to the level set method introduced above, in these papers, the evolution of a level set function is guided by the topological derivative rather than by the shape derivative. This way, large topological changes can easily be achieved. We will make use of this algorithm in the numerical examples of Section 4.8 and Chapter 8.

We also mention the results from [17], where it is shown that, for a density method with a certain choice of the material interpolation function  $\tilde{\rho}(\rho)$ , the sensitivity with respect to the density variable coincides with the topological derivative, and the recent extension of this result [21]. In [21], the authors additionally establish a connection between the material sensitivities and the shape derivative on the material interface. Both results are only shown for the case of a linear PDE constraint.

Finally, we mention the approach presented in [213], which is a combination of a level set method and the phase field method. One feature of the approach presented there, which is

very useful in practice, is that the complexity of the arising final design can be adjusted by a parameter.

### 1.2.2 Shape Optimization

In contrast to topology optimization, in shape optimization the connectivity of a domain is assumed to be fixed. Here, one is interested in finding the shape of a domain or subdomain which is optimal with respect to a given criterion by means of smooth variations of the boundary or of a material interface. In this thesis, we will study transmission problems, i.e., we are looking for an optimal subdomain of a given hold-all domain. Therefore, we will describe all techniques for this case noting that all of these methods can be applied in a similar way to the case where the whole computational domain is varied, too.

Historically, the first contribution to shape optimization was the early work of Hadamard [96] who derived the shape derivative for the first eigenvalue of the clamped plate in 1907. The fields of shape calculus and shape optimization have become active areas of research since the 1970s, see [64, 142, 152, 170, 194]. For a comprehensive overview over the fields, see also the monographs [51, 75, 100, 101, 196].

In shape optimization, we make the distinction between parametric and non-parametric shape optimization. As mentioned above, in parametric shape optimization, the shape sensitivities are sensitivities with respect to certain parameters representing dimensions or orientations of an object. When a shape is represented by a spline curve or surface, the shape sensitivities may also be related to the position of the control points representing this curve or surface. In parametric shape optimization, the design space is represented by a certain, finite number of parameters, which results in a limitation of the possible designs. A comprehensive introduction to parametric shape optimization can be found in [67, 100].

In non-parametric shape optimization, on the contrary, one is interested in general shapes and the sensitivity of a shape function with respect to an arbitrary smooth perturbation of these shapes, called the shape derivative. In this thesis, we will only deal with non-parametric shape optimization.

For a (non-parametric) shape optimization problem, we consider a *shape functional*

$$\begin{aligned}\mathcal{J} : \mathcal{A} &\rightarrow \mathbb{R} \\ \Omega &\mapsto \mathcal{J}(\Omega),\end{aligned}$$

where  $\mathcal{A}$  is a set of admissible subsets of a hold-all domain  $D \in \mathbb{R}^d$ . For such shape functionals, one is interested in their sensitivity with respect to a perturbation of the domain  $\Omega$  which motivates the definition of the *shape derivative* of a shape function  $\mathcal{J}$ . However, it is important to note that the definition of the shape derivative is not straightforward since the domain of definition  $\mathcal{A}$  of the shape functional does not have a vector space structure. The shape functional  $\mathcal{J}$  is said to be shape differentiable if the limit

$$d\mathcal{J}(\Omega; V) = \lim_{t \searrow 0} \frac{\mathcal{J}(\Omega_t) - \mathcal{J}(\Omega)}{t}$$

exists and the mapping  $V \mapsto d\mathcal{J}(\Omega; V)$  is linear and continuous, see also Definition 6.1. Here,  $\Omega_t = T_t(\Omega)$  denotes the transformed domain under the flow  $T_t$  generated by a smooth vector

field  $V$ . We mention that there are two ways to define this flow given a smooth vector field  $V$ . In the *perturbation of identity* method, the transformation is given by  $T_t(X) = X + tV(X)$  for all  $X \in \mathbb{R}^d$  and  $t \geq 0$ , whereas in the *velocity or speed method*, it is given as  $T_t(X) = x(t, X)$  with  $x(t, X)$  the solution to the initial value problem

$$\begin{aligned}\frac{d}{dt}x(t, X) &= V(x(t, X)), \quad 0 < t < \tau, \\ x(0, X) &= X,\end{aligned}$$

which, for small  $\tau > 0$  has a unique solution, see [75, 196]. Note that, for simplicity, we assumed the vector field  $V$  to be autonomous. We remark that both approaches are equivalent for the derivation of first order shape derivatives, but differ by an acceleration term in the case of second order shape derivatives [75]. In Chapter 6, we will consider the latter approach.

A class of shape optimization problems, which is very important in engineering applications, is the class of PDE-constrained shape optimization problems where the shape function  $\mathcal{J}$  depends on the shape of a domain via the solution of a boundary value problem posed on this domain. In this thesis, we will only focus on this class of problems.

Among shape optimization methods there is not such a big variety of different approaches as in the case of topology optimization. The overall procedure in shape optimization could roughly be summarized in the following three steps:

1. Compute the shape sensitivity.
2. Extract a descent direction.
3. Move the shape in direction of this descent direction.

### 1.2.2.1 Computation of Shape Derivative

The idea of the shape derivative goes back to the early work of Hadamard [96] who showed for the first eigenvalue of a clamped plate that the shape sensitivity only depends on the normal component of the shape perturbation on the boundary. Later, Zolésio generalized this observation to general shape functions and showed that, for domains with smooth enough boundaries, the shape derivative can always be written as

$$d\mathcal{J}(\Omega; V) = \int_{\partial\Omega} g_\Gamma V \cdot n \, ds \tag{1.2}$$

with an integrable function  $g_\Gamma \in L^1(\Gamma)$ , see [75, Theorem 3.6]. This form is often called the Hadamard form of the shape derivative, but it is actually due to J.-P. Zolésio.

Besides this boundary integral form, the shape derivative can also be represented as a volume integral over the whole domain,

$$d\mathcal{J}(\Omega; V) = \int_{\Omega} g(V, DV) \, dx \tag{1.3}$$

for some function  $g$ . This representation has the advantage that it requires less regularity of the solutions to the PDE constraint as well as to the adjoint equation. We will give a comparison of these two possible representations in Section 6.1.2.

We will give a quick overview over different approaches of deriving the shape derivative using the velocity method in Section 6.1.1.

### 1.2.2.2 Extraction of Descent Direction

The shape derivative represents the sensitivity of a shape function with respect to a perturbation of the domain  $\Omega$  in the direction of a smooth vector field  $V$ . Given the analytic formula for the shape derivative of a cost function  $\mathcal{J}$ , it remains to extract such a vector field that yields a decrease of the objective function, i.e., such that  $d\mathcal{J}(\Omega; V) < 0$ . If the shape derivative is given in the Hadamard form (1.2), it can easily be seen that the vector field  $V = -g_\Gamma n$  is a descent direction. On the other hand, in the case of a volume formulation of the kind (1.3), one can extract a descent vector field  $V$  by solving the additional auxiliary boundary value

$$b(V, W) = -d\mathcal{J}(\Omega; V),$$

where  $b(\cdot, \cdot)$  denotes some positive definite bilinear form defined on a suitable Hilbert space, see also Section 6.6.1.2. This trick is also known as the “traction method”, see e.g. [27, 28], and has become a standard tool when dealing with the volume form of the shape derivative. Although this volume-based form seems to come at the costs of the additional solution of an auxiliary boundary value problem, there are several good reasons for choosing this representation of the shape derivative, which we will discuss in Section 6.1.2. A common and natural choice for  $b(\cdot, \cdot)$  is the bilinear form from linear elasticity, see [28, 153, 187], which gives the interpretation of  $V$  as a displacement field under loads represented by the shape derivative. Note that the solution  $V$  to the auxiliary boundary value problem above can be seen as the negative shape gradient in terms of the elliptic bilinear form  $b(\cdot, \cdot)$  by the Riesz representation theorem, see e.g. [9].

### 1.2.2.3 Evolution of the Shape

Once a descent vector field  $V$  is available, it remains to advect the shape in the direction of  $V$ . In principle, there are two different ways to represent a shape: explicitly or implicitly.

If the interface is given explicitly, e.g., by a spline or polygonal interface, which may or may not be aligned with an underlying finite element discretization, one simply moves the interface a certain distance into the direction given by  $V$ . If the interface is aligned with the mesh then it is advantageous if the vector field  $V$  is supported not only on the interface, but at least in a neighborhood of the interface, or on the whole computational domain. Thus, even though it is possible to extract a descent direction  $V$  from the boundary form of the shape derivative (1.2), this vector field is only defined on the interface and must be extended to a neighborhood in a suitable way.

In the level set method, the shape is represented in an implicit way, by means of a level set function  $\psi$  and its evolution is steered by the Hamilton-Jacobi equation (1.1) with  $V$  chosen as a descent vector field.

We also mention that, while the most widely used approach to numerical shape optimization is based on gradient descent methods, in recent years more and more authors have developed Newton-like methods for shape optimization which exploit second order shape sensitivity information. We mention the publications [160, 188, 199].

### 1.3 Finite Element Methods for Interface Problems

Throughout this thesis, we are concerned with finding the optimal distribution of ferromagnetic material within a given design subdomain. In the course of the optimization process, interfaces between different materials are moving in a direction determined by shape or topological sensitivities. In each iteration of the optimization algorithm, we have to solve the boundary value problem of two-dimensional magnetostatics as well as an adjoint equation in order to be able to evaluate these sensitivities for the next iteration. We solve these problems by means of a finite element method. When using standard finite element methods, the material interfaces where the solution exhibits lower regularity due to a jump of the material coefficient must be resolved by the finite element discretization in order to obtain accurate solutions [29]. Creating a new mesh which resolves the material interfaces at each iteration is very inefficient and should be avoided. A wide range of approaches dealing with this kind of interface problems is available in the literature. Of course, these methods are not restricted to applications from design optimization, but can be applied to a great variety of problems involving moving material interfaces such as, e.g., fluid-structure interaction or multi-phase flow problems. We give a brief overview over some of the most widely used methods.

The *extended finite element method (XFEM)* was introduced in [151] motivated by the problem of simulating crack propagation in structural mechanics without the need to resolve the small-scale features of the crack by the finite element mesh. Since then, the method has been applied to a wide variety of interface problems, see e.g. [37, 84, 157] and the references therein. The idea of the method is to enrich the finite element basis by additional basis functions which are modified or cut-off versions of the standard basis functions, see e.g. Figure 1.1(a) for the case of piecewise linear finite elements. The solution is sought in the enriched space  $V_h^\Gamma = V_h \oplus V_h^x$  where  $V_h$  is a standard finite element space, and  $V_h^x$  the space of standard finite element functions which are supported at the interface, multiplied with a so-called enrichment function [136]. Depending on the kind of discontinuity arising in the particular problem, this enrichment function is typically a distance function (for weak discontinuities, i.e., kinks) or a Heaviside function (for strong discontinuities, i.e., jumps), see e.g. [37, 84]. A thorough introduction to the XFEM from an application point of view can be found in [125]. The idea of enriching the basis is the same in the so-called *partition of unity method (PUM)*, also referred to as *generalized finite element method (GFEM)*, introduced in [31, 147], but here the enrichment functions are not necessarily local, but can have global support. Note that the interpretation of the term generalized finite element method is not always straightforward and often different authors mean different methods. We remark that XFEM, GFEM and PUM are closely related.

The *immersed interface method* was introduced in [138] in the framework of a finite difference method and adopted to what is now called the *immersed finite element method* in [139]. The idea of the method is similar to that of the XFEM. However, rather than adding basis functions to the basis, existing basis functions of the finite element space which are supported across the interface are modified in such a way that the interface jump conditions are satisfied, see Figures 1.1(b)–(c). Optimal order of convergence in the  $L^2(D)$  and  $H^1(D)$  norm have been established [119, 140].

In the *unfitted Nitsche method* introduced in [99], which is based on Nitsche's idea [155], a discontinuity of the solution across an interface is enforced in a weak sense, similar to

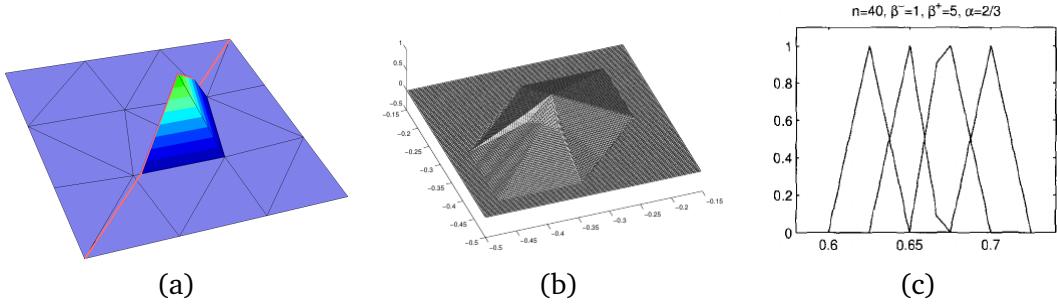


Figure 1.1: (a) Example of an enriched basis function in XFEM ([136]). (b)–(c) Examples of immersed basis functions satisfying a jump condition for the gradient in 1D and 2D ([140]).

discontinuous Galerkin methods, see e.g. [176]. This way of treating the interface conditions is often used in combination with XFEM, called the *Nitsche-XFEM*. In this method, just like in all other methods mentioned above, a crucial task is to establish stability of the method with respect to the location of the interface relative to the mesh. Generally, if an element of the underlying unfitted background mesh is cut by the interface very close to one of the vertices, the condition of the system becomes very bad. This issue is treated in the *CutFEM* [57], which is a stabilized version of the Nitsche-XFEM using a so-called ghost penalty term [56]. We also mention the recent work [58] where the CutFEM is combined with level-set based shape optimization. An alternative approach to getting stability of the Nitsche-XFEM by preconditioning was shown in [137].

All of the methods mentioned so far use a fixed background mesh and treat the interface by means of the finite element space. An alternative to these approaches is to modify the mesh and always work with a fitted discretization while still guaranteeing a certain quality of the mesh. We mention the mesh optimization approach of [34–36] and the *deformable simplicial complex (DSC) method* [68, 150].

In [83], an interface finite element method on a fixed mesh is introduced where the interface is resolved by locally modifying the finite element basis functions. Optimal order of convergence and also, when choosing a special hierarchical basis, optimal conditioning of the system matrix are shown. We note that this parametric approach can be equivalently interpreted as a fitted finite element method where some of the mesh nodes close to the interface are moved in such a way that the interface is resolved by the mesh. In Chapter 7, we will follow this approach and translate it to the case of triangular finite elements.

## 1.4 Organization of the Thesis

In this thesis, we combine the three concepts we introduced in Sections 1.2.1, 1.2.2 and 1.3 in the framework of electrical machines. The thesis consists of three theoretical parts dealing with topology optimization, shape optimization as well as a finite element method for resolving material interfaces, as well as a final part where the technologies we developed are combined and applied to some practical applications.

Here, we briefly highlight the main contributions of the author:

- Starting out from the On/Off method which has been applied to the optimization of electrical equipment and is defined on a discrete level, we generalized this method to the continuous level taking into account perturbations of the material coefficients in arbitrary smooth subdomains.
- The largest part of this thesis is concerned with the rigorous derivation of the topological derivative under the PDE constraint of two-dimensional nonlinear magnetostatics. Here, while the overall procedure is similar to the previous work done in [20, 46], the extension was by no means trivial. Furthermore, we put a strong emphasis on the applicability of the resulting formula. In particular, we derived an explicit form for the matrices  $\mathcal{M}$  and  $\mathcal{M}^{(2)}$ , which are related to the concept of polarization matrices, and found a way how the final formula can be computed efficiently in a numerical optimization algorithm. Moreover, we derived the topological derivative for the two different scenarios: introducing an inclusion of air inside ferromagnetic material, and introducing ferromagnetic material inside an air region. This is important for the usage of the formula in bi-directional optimization algorithms.
- We analyzed the relation between the material sensitivities used in the On/Off method, which are very similar to the sensitivities used in density-based topology optimization, and the topological derivative.
- We derived the shape derivative for the case of two-dimensional nonlinear magnetostatics in a rigorous way.
- We adapted the interface finite element method introduced in [83] to the case of triangular meshes.
- We combined topology and shape optimization with the interface finite element method and applied this combined algorithm to the optimization of electric motors.

The remainder of this thesis is organized as follows: In Chapter 2, we introduce the physical model used throughout this thesis and analyze its mathematical properties. Part I of the thesis deals with topology optimization and comprises Chapters 3–5. In Chapter 3 we consider the On/Off method, a topology optimization method based on sensitivities with respect to the material coefficient. The topological derivative for the case of two-dimensional nonlinear magnetostatics is investigated in great detail in Chapter 4. In Chapter 5, we compare these two kinds of sensitivities and point out their similarities and differences. Part II, which consists of Chapter 6, is concerned with shape optimization for the same magnetostatic problem. In Part III, consisting of Chapter 7, we introduce a finite element method which is suitable for tracking evolving interfaces. We combine all of these methods in Part IV of the thesis, and apply the developed techniques to two practical examples of the optimization of electric motors.

## Chapter 2

# Physical Background for Simulation of Electrical Machines

### 2.1 Introduction to Electrical Machines

Electrical machines convert electrical and mechanical energy into each other. An electric motor converts electrical energy into mechanical movement, whereas a generator produces electrical from mechanical energy. Electric motors can cause a linear or a rotational movement. Rotating electric motors generally consists of a fixed part, called the *stator*, and a rotating part, the *rotor*. Electric current is induced in coil areas in at least one of these two parts, which generates a magnetic field. The overall magnetic field in the electric motor, including a possible contribution from permanent magnets, is responsible for its rotation.

The most important classes of electric motors are *DC motors*, where direct current is induced, *induction motors* and *synchronous motors*, the latter two classes working on alternating current. As opposed to induction motors, which are of asynchronous type, the rotation of the rotor in a synchronous motor coincides with the rotating magnetic field caused by the alternating current in the coils. Throughout this thesis, we will be dealing with rotating synchronous electric motors. The magnetic field inside the motor may or may not be additionally influenced by the presence of permanent magnets, which can be placed in the interior of the motor or mounted on its surface. Figure 2.1 depicts an interior permanent magnet (IPM) synchronous electric motor which is of the same kind as the model problem introduced in Section 2.3 which we will consider throughout this thesis. In Section 8.4 we will, in addition, consider a synchronous reluctance motor without permanent magnets. For a comprehensive and thorough introduction to electrical machines, we refer the reader to [45].

Important criteria for the design of electric motors include high efficiency, high torque capability, high smoothness of the rotation and low production costs. In our model problem, we will consider a functional which is related to the smoothness of the rotation, and in Section 8.4 we will perform optimization to maximize the average torque of the motor. Most often, some of these criteria are conflicting and one has to find the best possible trade-off between different criteria, which is usually done by forming Pareto fronts, see e.g. [73, 218]. In this thesis, we will restrict ourselves to the case of single-objective optimization problems.



Figure 2.1: Interior permanent magnet synchronous motor as produced by Hanning Elektro-Werke GmbH & Co KG. Photo credit: LCM GmbH.

## 2.2 Physical Model

Electromagnetic phenomena are described by *Maxwell's equations* which were first published by James Clerk Maxwell in 1862, see [146]. The full set of equations reads

$$\operatorname{curl} \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}, \quad (2.1a)$$

$$\operatorname{curl} \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \quad (2.1b)$$

$$\operatorname{div} \mathbf{B} = 0, \quad (2.1c)$$

$$\operatorname{div} \mathbf{D} = \rho, \quad (2.1d)$$

where the unknown quantities are the electric field intensity  $\mathbf{E}$ , the electric flux density  $\mathbf{D}$ , the magnetic field intensity  $\mathbf{H}$  and the magnetic flux density  $\mathbf{B}$ . Note that all of these quantities are vector-valued functions from  $\mathbb{R}^3$  to  $\mathbb{R}^3$ . Furthermore, the equations involve the electric charge density  $\rho$  and the electric current density  $\mathbf{J}$ . This set of equations is complemented by the constitutive relations [120, 128]

$$\mathbf{B} = \mu \mathbf{H} + \mathbf{M}, \quad \mathbf{D} = \varepsilon \mathbf{E} + \mathbf{P}, \quad \mathbf{J} = \mathbf{J}_i + \sigma \mathbf{E}, \quad (2.2)$$

where  $\mathbf{M}$  is the magnetization which vanishes outside permanent magnets,  $\mathbf{P}$  denotes an electric polarization and  $\mathbf{J}_i$  an impressed current density (e.g., in the current-loaded coils). Furthermore,  $\mu$  denotes the magnetic permeability,  $\varepsilon$  the electric permittivity and  $\sigma$  the electric conductivity of a material. In general, these quantities can be matrices, but we will assume isotropic material throughout this thesis such that they become scalar quantities. Furthermore, we neglect possible effects of hysteresis. For the mathematical models used in the rest of this thesis, we will always be dealing with the reciprocal of the magnetic permeability which is called the magnetic reluctivity and is denoted by  $\nu$ . The magnetic reluctivity  $\nu$  and satisfies the relation

$$\mathbf{H} = \nu (\mathbf{B} - \mathbf{M}). \quad (2.3)$$

In the framework of electrical machines, we will be concerned with ferromagnetic materials where the magnetic reluctivity is a nonlinear function of the magnitude of the magnetic flux density,  $\nu = \nu(|\mathbf{B}|)$ . Note that the magnetic reluctivity  $\nu$  is a constant in the areas where the magnetization  $\mathbf{M}$  does not vanish [106]. For more details on the system of Maxwell's equations, we refer the reader to [116, 122, 128, 207].

For low-frequency applications like electrical machines, the displacement currents  $\frac{\partial \mathbf{D}}{\partial t}$  in (2.1a) can be neglected such that equation (2.1d) decouples from the other equations. The arising system (2.1a)–(2.1c) is called the *magnetoquasistatic problem* or *eddy current problem*. Since  $\mathbf{B}$  is divergence-free according to (2.1c), there exists a vector potential  $\mathbf{A}$ , which is unique up to a gradient field, such that

$$\mathbf{B} = \operatorname{curl} \mathbf{A}. \quad (2.4)$$

Substituting this relation into (2.3), (2.1b) and (2.1a) yields the vector potential formulation for the magnetoquasistatic problem [122, 128],

$$\sigma \frac{\partial \mathbf{A}}{\partial t} + \operatorname{curl} (\nu(|\operatorname{curl} \mathbf{A}|) \operatorname{curl} \mathbf{A}) = \mathbf{J}_i + \operatorname{curl} (\nu \mathbf{M}). \quad (2.5)$$

In the context of electric motors, this quasistatic model is used for the starting phase when the motor is accelerated from its resting position.

Once the rotor has reached a constant rotational speed, all involved electromagnetic quantities can be regarded as time-independent such that all time derivatives  $\frac{\partial}{\partial t}$  vanish. Under these assumptions, the *magnetostatic* model can be used to obtain the magnetic flux density  $\mathbf{B} = \operatorname{curl} \mathbf{A}$ ,

$$\operatorname{curl} (\nu(|\operatorname{curl} \mathbf{A}|) \operatorname{curl} \mathbf{A}) = \mathbf{J}_i + \operatorname{curl} (\nu \mathbf{M}). \quad (2.6a)$$

In order to be able to solve equation (2.6a) numerically by the finite element method, a bounded computational domain  $\hat{D}$  has to be introduced. This domain should be chosen in such a way that the magnetic fields are negligible outside the domain. Possible choices of boundary conditions for the magnetic fields are to force the tangential component of the magnetic field intensity  $\mathbf{H}$  to vanish, or to require the normal component of the magnetic flux density  $\mathbf{B}$  to disappear. Splitting the boundary of the computational domain  $\hat{D}$  into two parts  $\Gamma_H$  and  $\Gamma_B$  such that  $\partial \hat{D} = \overline{\Gamma}_H \cup \overline{\Gamma}_B$  with  $\Gamma_H \cap \Gamma_B = \emptyset$ , this amounts to the boundary conditions

$$\mathbf{H} \times n = 0 \text{ on } \Gamma_H, \quad (2.6b)$$

$$\mathbf{B} \cdot n = 0 \text{ on } \Gamma_B, \quad (2.6c)$$

where  $n$  denotes the outer unit normal vector to  $\hat{D}$ . Condition (2.6b) is called the perfect magnetic conductor (PMC) boundary condition and models materials with very high permeability, whereas the so-called induction boundary condition (2.6c) means that no magnetic flux leaves the computational domain, see [116, 128, 217]. The system of magnetostatic equations is completed by the interface conditions. Let  $\Gamma_I$  represent the material interfaces where the magnetic reluctivity  $\nu$  jumps, and let  $\llbracket v \rrbracket$  denote the jump of a function  $v$  across the interface  $\Gamma_I$ , i.e.,

$$\llbracket v \rrbracket = v^+|_{\Gamma_I} - v^-|_{\Gamma_I},$$

where  $v^+$  and  $v^-$  denote the restrictions of  $v$  to the respective subdomains. Then, the interface conditions for  $\mathbf{B}$  and  $\mathbf{H}$  read

$$[\![\mathbf{B} \cdot n]\!] = 0 \text{ on } \Gamma_I, \quad [\![\mathbf{H} \times n]\!] = 0 \text{ on } \Gamma_I. \quad (2.6d)$$

They follow from the integral forms of (2.1c) and (2.1a), respectively.

For the simulation of rotating electrical machines at a fixed rotational speed, a sequence of magnetostatic problems for a range of different rotor positions have to be solved. The current density  $\mathbf{J}_i$  is piecewise constant for each rotor position, but changes in the course of the rotation.

Under certain assumptions on the geometry of the computational domain  $\hat{D}$ , as well as on the sources  $\mathbf{J}_i$  and  $\mathbf{M}$  and the arising magnetic field, the model (2.6) can be approximated by a two-dimensional model. The assumptions are that

- one dimension of the computational domain is much larger than the other two,

$$\hat{D} = D \times (-l, l) \quad \text{with } l \gg \text{diam}(D),$$

- the data  $\mathbf{J}_i$  and  $\mathbf{M}$  as well as the magnetic field  $\mathbf{H}$  are constant with respect to  $x_3$  and of the form

$$\mathbf{J}_i = \begin{pmatrix} 0 \\ 0 \\ J_3(x_1, x_2) \end{pmatrix}, \quad \mathbf{M} = \begin{pmatrix} M_1(x_1, x_2) \\ M_2(x_1, x_2) \\ 0 \end{pmatrix}, \quad \mathbf{H} = \begin{pmatrix} H_1(x_1, x_2) \\ H_2(x_1, x_2) \\ 0 \end{pmatrix},$$

for  $(x_1, x_2) \in D$ .

Under these assumptions, it follows from the constitutive relation (2.2) that  $\mathbf{B}$  is of the same form as  $\mathbf{H}$  and  $\mathbf{M}$ . Regarding the equation (2.4), this form of  $\mathbf{B}$  is guaranteed by the ansatz

$$\mathbf{A} = \begin{pmatrix} 0 \\ 0 \\ u(x_1, x_2) \end{pmatrix},$$

which also yields that the so-called Coulomb gauging condition  $\text{div}\mathbf{A} = 0$  is satisfied and that  $|\mathbf{B}| = |\nabla u|$ . It can also be seen that, in this setting,  $\mathbf{H} \times n = \nu(|\nabla u|)\nabla u \cdot n$  and  $\mathbf{B} \cdot n = \nabla u \cdot \tau$  with  $\tau$  the tangential unit vector. This yields the boundary value problem of two-dimensional magnetostatics in the domain  $D \subset \mathbb{R}^2$ ,

Find  $u$  such that

$$-\text{div}(\nu(x, |\nabla u|)\nabla u) = J_3 - \nu \text{div}M^\perp, \quad x \in D, \quad (2.7a)$$

$$u = 0, \quad x \in \Gamma_D, \quad (2.7b)$$

$$\nu(x, |\nabla u|)\nabla u \cdot n = 0, \quad x \in \Gamma_N, \quad (2.7c)$$

$$[\![u]\!] = 0 \quad x \in \Gamma_I, \quad (2.7d)$$

$$[\![\nu(x, |\nabla u|)\nabla u \cdot n]\!] = 0 \quad x \in \Gamma_I, \quad (2.7e)$$

where  $M^\perp = (-M_2, M_1)^\top$  is the perpendicular of the first two components of the magnetization  $\mathbf{M}$  and  $\Gamma_D$  and  $\Gamma_N$  denote the part of the boundary  $\partial D$  where we require  $\mathbf{B} \cdot n = 0$  and  $\mathbf{H} \times n = 0$ , respectively.

**Remark 2.1.** *The quantity  $\mathbf{M}$  vanishes in materials without permanent magnetization. In permanent magnets, it holds the relation  $\mathbf{M} = \mu_{\text{mag}}\mathbf{H}_0$  where  $-\mathbf{H}_0$  denotes the magnetic field intensity such that the magnetic flux density vanishes [120].*

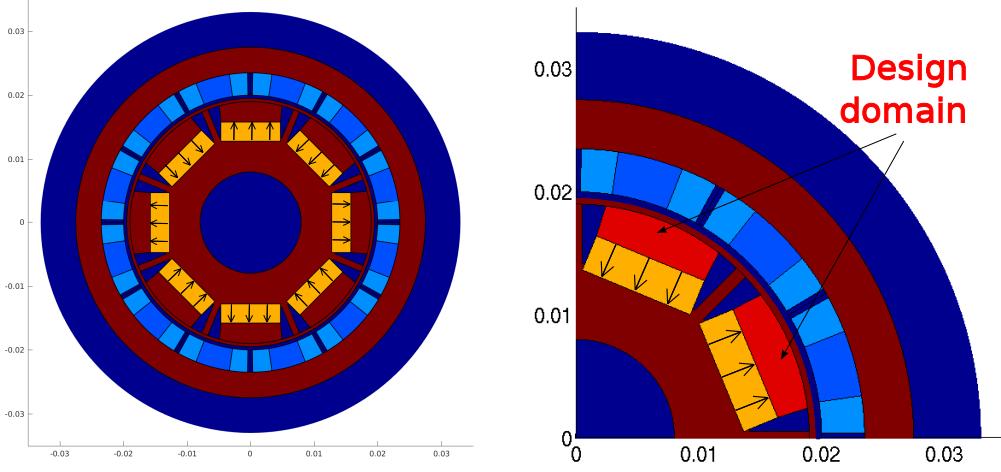


Figure 2.2: Left: Computational domain representing electric motor with different subdomains. Right: Zoom on upper left quarter with design region  $\Omega^d$  highlighted (for a different rotor-to-stator constellation).

### 2.3 Model Problem

As a model problem, we consider an interior permanent magnet (IPM) synchronous electric motor. We are interested in the magnetic flux density  $\mathbf{B}$  in the motor due to the permanent magnets and the electric currents after the motor has reached a constant rotational speed. In this setting, considering one fixed rotor position, we can assume that the sources and, therefore, the electric and magnetic fields are time-independent and we can use the magnetostatic model (2.6a). Furthermore, since the axial dimension of the motor is large compared to its diameter, the assumptions for the reduction to a two-dimensional model are satisfied very well. This allows us to compute the magnetic field via the boundary value problem (2.7a)–(2.7e). This model is commonly used for the simulation of electrical machines [26, 45] and allows for a significant reduction of the computational costs, which is particularly useful in the context of design optimization. For a comparison between two- and three-dimensional models of electric motors, see [127, 206].

The geometry of the two-dimensional model is depicted in Figure 2.2. Let the hold-all domain  $D \subset \mathbb{R}^2$ , which we assume to have a boundary of class  $C^2$ , denote the computational domain which comprises all of the components of the motor as well as the air regions. The motor consists of a rotating inner part, called the rotor, and a fixed outer part, the stator. These two parts are separated by a thin air gap, which we denote by  $\Omega_g$ . We will be particularly interested in the magnetic field inside this air gap, because it has a big influence on the behavior of the electric motor. Both the rotor and the stator have an iron core consisting of ferromagnetic material, see the brown area in the left picture of Figure 2.2. We denote this ferromagnetic reference domain by  $\Omega_f^{ref}$  and its (open) complement by  $\Omega_{air}^{ref}$ , i.e.,  $\Omega_{air}^{ref} = D \setminus \overline{\Omega_f^{ref}}$ . The subdomain  $\Omega_{air}^{ref}$  represents all the materials that are not ferromagnetic and also contains the coil areas  $\Omega_c$ , the magnet areas  $\Omega_{mag}$  as well as the air gap  $\Omega_g$ . Let  $\Omega^d \subset$

$\Omega_f^{ref}$  denote the open design subdomain, which consists of the highlighted regions in the right picture of Figure 2.2. Note that,  $\Omega^d$  consists of eight parts. We are interested in the optimal distribution of ferromagnetic material and air regions in  $\Omega^d$ . Note the circular ring of ferromagnetic material between the design region and the air gap, which is fixed. This design restriction was chosen for mechanical reasons. For a given design we denote the subdomain of  $\Omega^d$  that is currently occupied with ferromagnetic material by  $\Omega$ . For any given configuration of ferromagnetic material inside  $\Omega^d$ , the set of all points of the motor that are occupied with ferromagnetic material is then given by

$$\Omega_f := \left( \Omega_f^{ref} \setminus \Omega^d \right) \cup \Omega. \quad (2.8)$$

Then, introducing  $\Omega_{air} = D \setminus \overline{\Omega}_f$ , we always have that  $\overline{D} = \overline{\Omega}_f \cup \overline{\Omega}_{air}$ . The magnetic reluctivity function  $\nu = \nu(x, |\nabla u|)$  introduced in (2.3) attains different values in the different subdomains of the motor. More precisely, we have

$$\nu(x, |\nabla u|) = \begin{cases} \hat{\nu}(|\nabla u|) & x \in \Omega_f, \\ \nu_0 & x \in \Omega_{air} \setminus \Omega_{mag}, \\ \nu_{mag} & x \in \Omega_{mag}, \end{cases} \quad (2.9)$$

where  $\hat{\nu}$  is a nonlinear function that is defined via the **B–H** relation which will be discussed in Section 2.4. Here, the constant

$$\nu_0 = 10^7 / (4\pi) \quad (2.10)$$

is the magnetic reluctivity of vacuum (expressed in the unit  $A^2 s^2 kg^{-1} m^{-1}$ ), which is practically the same as that of air, and  $\nu_{mag}$  denotes the material-dependent, but constant magnetic reluctivity of the permanent magnets. The value of  $\nu_{mag}$  is usually close to that of  $\nu_0$ . For sake of more compact presentation, we will set  $\nu_{mag} = \nu_0$  for the rest of this thesis and remark that in all numerical examples the realistic value  $\nu_{mag} = \nu_0/1.086$  was chosen. Note that also the ferromagnetic behavior of the coils, which consist of copper, are the same as that of air and we have  $\nu = \nu_0$  also in  $\Omega_c$ . Since the magnetic reluctivity in the electric motor also depends on the current design  $\Omega$ , we write  $\nu = \nu_\Omega$ . Then we have

$$\begin{aligned} \nu_\Omega(x, |\nabla u|) &= \chi_{\Omega_f}(x) \hat{\nu}(|\nabla u|) + \chi_{D \setminus \Omega_f}(x) \nu_0 \\ &= \left( \chi_{\Omega_f^{ref} \setminus \Omega^d}(x) + \chi_\Omega(x) \right) \hat{\nu}(|\nabla u|) + \left( \chi_{D \setminus \Omega_f^{ref}}(x) + \chi_{\Omega^d \setminus \Omega}(x) \right) \nu_0, \end{aligned} \quad (2.11)$$

where  $\chi_S$  denotes the characteristic function of a given set  $S$ . Note that the expression above is meaningful since  $\Omega \subset \Omega^d \subset \Omega_f^{ref}$ .

The typical boundary condition for the simulation of electrical machines in the setting of two-dimensional magnetostatics is to prescribe the homogeneous Dirichlet boundary conditions (2.7b) on the entire boundary  $\partial D$ , which implies  $\mathbf{B} \cdot \mathbf{n}|_{\partial D} = 0$ , i.e., that no magnetic flux leaves the domain. Note that the computational domain includes a layer of air outside the actual motor, such that this boundary condition is actually realistic. Therefore, both our ansatz and test functions will be from the space  $H_0^1(D)$ .

We introduce the nonlinear operator  $A_\Omega : H_0^1(D) \rightarrow H^{-1}(D)$  representing the left hand side of equation (2.7a) defined by

$$\langle A_\Omega(u), \eta \rangle = \int_D \nu_\Omega(x, |\nabla u|) \nabla u \cdot \nabla \eta \, dx, \quad (2.12)$$

for all  $u, \eta \in H_0^1(D)$ . The right hand side of equation (2.7a) comprises two different sources, the impressed currents and the magnetization. For a fixed rotor position, the current density  $J_3$  is piecewise constant in the coil areas  $\Omega_c$  and vanishes outside  $\Omega_c$ . Likewise, the magnetization is piecewise constant and vanishes outside the magnet areas  $\Omega_{mag}$ . Given a test function  $\eta \in H_0^1(D)$ , the weak form of the right hand side of equation (2.7a) reads

$$\langle F, \eta \rangle = \int_D J_3 \eta + \nu_{mag} M^\perp \cdot \nabla \eta \, dx, \quad (2.13)$$

where  $\langle \cdot, \cdot \rangle$  is to be understood as a duality product between  $H^{-1}(D)$  and  $H_0^1(D)$ . The interface conditions (2.7d) and (2.7e) are automatically included in the weak formulation of the problem. Now, the weak formulation of boundary value problem (2.7) reads

$$\text{Find } u \in H_0^1(D) \text{ such that } \langle A_\Omega(u), \eta \rangle = \langle F, \eta \rangle \quad \text{for all } \eta \in H_0^1(D). \quad (2.14)$$

Our goal is to find a set  $\Omega$  such that a given domain-dependent shape functional  $\mathcal{J}$  is minimized. In the case of electric motors, this objective function  $\mathcal{J}$  generally is supported only in the air gap  $\Omega_g \subset \Omega_{air}^{ref}$ . Therefore, a perturbation of the material coefficient inside the design domain  $\Omega^d$  will not directly affect the functional and the functional depends on the configuration of the design subdomain  $\Omega^d$  only via the solution  $u$  of the state equation,  $\mathcal{J} = \mathcal{J}(u)$ . In our model optimization problem, we consider one fixed rotor position and no electric currents, i.e.,  $J_3 = 0$ . Thus, the magnetic field is generated solely by the permanent magnets. We are interested in the radial component of the magnetic flux density  $\mathbf{B}$  along a circular curve in the air gap, see the blue curve in Figure 2.3. Note that the radial component of  $\mathbf{B}$  in the air gap is positive in regions close to magnets whose magnetization is pointing outwards, and negative close to the other magnets. The smoothness of this curve is related to the smoothness of the rotation of the motor at nominal speed. Therefore, our goal is to find the ferromagnetic set  $\Omega \subset \Omega^d$  such that the radial component of  $\mathbf{B}$  in the air gap comes as close as possible to a given smooth sine curve. Noting that  $\mathbf{B} = \mathbf{B}(u) = \operatorname{curl}((0, 0, u)^\top)$ , this amounts to minimizing the functional

$$\mathcal{J}(u) = \int_{\Omega_g} |\mathbf{B}(u) \cdot n_g - B_d|^2 \, dx = \int_{\Omega_g} |\nabla u \cdot \tau_g - B_d|^2 \, dx, \quad (2.15)$$

where, for  $x_m = (x_{m,1}, x_{m,2})^\top$  denoting the center of the electric motor and  $x$  a point in the air gap  $\Omega_g$ , the vector fields

$$n_g(x) = \frac{1}{|x - x_m|} \begin{pmatrix} x_1 - x_{m,1} \\ x_2 - x_{m,2} \end{pmatrix} \quad \text{and} \quad \tau_g(x) = \frac{1}{|x - x_m|} \begin{pmatrix} -(x_2 - x_{m,2}) \\ x_1 - x_{m,1} \end{pmatrix} \quad (2.16)$$

denote the unit vectors pointing in radial and tangential directions, respectively. The desired radial component  $B_d$  is here given by  $B_d(x) = \frac{1}{2} \sin(4\theta(x))$  where  $\theta(x)$  denotes the angular coordinate of a point  $x \in \Omega_g$  in a polar coordinate system centered at  $x_m$ . Minimizing this

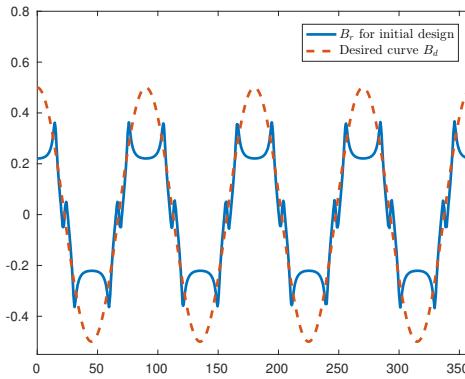


Figure 2.3: Radial component of the magnetic flux density  $B_r$  along a circular curve inside the air gap for initial design (blue) compared with desired curve  $B_d$  (green)

functional leads to a reduction of the total harmonic distortion (THD; see [45, 66]) of the flux density which causes the rotor to rotate more smoothly. We will denote the radial component of the magnetic flux density in the air gap by  $B_r$ ,

$$B_r(u)(x) := \mathbf{B}(u(x)) \cdot n_g(x) = \nabla u(x) \cdot \tau_g(x), \quad x \in \Omega_g.$$

Summarizing, we consider the PDE-constrained optimization problem

$$\inf_{\Omega \in \mathcal{A}} \mathcal{J}(u) \tag{2.17a}$$

$$\text{subject to } \langle A_\Omega(u), \eta \rangle = \langle F, \eta \rangle \quad \forall \eta \in H_0^1(D), \tag{2.17b}$$

where  $\mathcal{J}$  is defined in (2.15),  $A_\Omega$  and  $F$  are defined in (2.12) and (2.13), respectively, and the state variable  $u = u(\Omega)$  is from  $H_0^1(D)$ . Here, the set  $\mathcal{A}$  denotes a set of admissible shapes which will be specified later. In Chapter 8 we will consider more complicated objective functions including the case of rotating electrical machines.

## 2.4 Physical Properties

In this section, we have a closer look at the relation between the magnetic field intensity  $\mathbf{H}$  and the magnetic flux density  $\mathbf{B}$ . Recall that we are only considering the case of isotropic materials such that these two fields are parallel, and that we neglect possible effects of hysteresis.

The magnetic field intensity  $\mathbf{H}$  is also called the magnetizing force because it creates a magnetic flux  $\mathbf{B}$  and thus induces magnetic behavior of a body. The magnitude of the generated magnetic field  $B := |\mathbf{B}|$  depends on the magnitude of the magnetizing force  $H := |\mathbf{H}|$  and on the material properties [169]. In many materials, this relation is linear and, assuming the absence of permanent magnetization, we have that  $B = \mu H$  with a constant magnetic permeability  $\mu$ . In ferromagnetic materials on the contrary, the magnetic flux density  $B$  is

amplified and the relation between  $B$  and  $H$  is a nonlinear one, described by the so-called  $B$ - $H$ -curve

$$f : \mathbb{R}_0^+ \rightarrow \mathbb{R}_0^+ : H \mapsto B = f(H),$$

where  $\mathbb{R}_0^+$  denotes the non-negative real numbers. For a given material, the magnetic permeability  $\mu$  and the magnetic reluctivity  $\nu$  introduced in (2.2) and (2.3), respectively, are defined based on this notion,

$$\mu(s) := f(s)/s, \quad \text{and} \quad \nu(s) := f^{-1}(s)/s, \quad (2.18)$$

such that we have

$$\mathbf{B} = \mu(|\mathbf{H}|)\mathbf{H}, \quad \text{and} \quad \mathbf{H} = \nu(|\mathbf{B}|)\mathbf{B}.$$

**Remark 2.2.** *In the presence of permanent magnetization, the amount of magnetization  $\mathbf{B}$  induced in the body is the sum of the magnetizing force  $\mathbf{H}$  and the permanent magnetization  $\mathbf{M}$ , see (2.2). Recall that the relation between  $B$  and  $H$  in permanent magnets is in general linear [106]. In view of Remark 2.1, we have that  $\mathbf{B} = \mu\mathbf{H} + \mathbf{M} = \mu(\mathbf{H} + \mathbf{H}_0)$ .*

Figure 2.4 depicts the  $B$ - $H$ -curve used in the ferromagnetic subdomains  $\Omega_f$  of the electric motors considered in this thesis, as well as the corresponding magnetic permeability and the magnetic reluctivity, which we will denote by  $\hat{\mu}$  and  $\hat{\nu}$ , respectively. The following natural assumptions on the  $B$ - $H$ -curve  $f$  follows from physical properties (cf. [169]):

**Assumption 1.** Let  $f : \mathbb{R}_0^+ \rightarrow \mathbb{R}_0^+$  be a  $B$ - $H$ -curve. Then the following holds:

1.  $f$  is continuously differentiable on  $\mathbb{R}_0^+$ ,
2.  $f(0) = 0$ ,
3.  $f'(s) \geq \mu_0$  for all  $s \geq 0$ ,
4.  $\lim_{s \rightarrow \infty} f'(s) = \mu_0$ .

Here,  $\mu_0$  denotes the magnetic permeability in vacuum and is the reciprocal of the magnetic reluctivity of vacuum defined in (2.10),  $\mu_0 = 1/\nu_0$ . The following properties are immediate consequences of Assumption 1:

**Lemma 2.3.** *Let Assumption 1 hold. Then, the function  $\hat{\nu}$  is continuously differentiable on  $[0, \infty)$  and there exists  $\underline{\nu} > 0$  such that, for all  $s \in \mathbb{R}_0^+$ , we have*

$$\underline{\nu} \leq \hat{\nu}(s) \leq \nu_0, \quad (2.19a)$$

$$\underline{\nu} \leq (\hat{\nu}(s)s)' \leq \nu_0. \quad (2.19b)$$

*Proof.* A proof can be found in, e.g. [168]. □

Note that (2.19b) yields that the mapping  $s \mapsto \hat{\nu}(s)s$  is strongly monotone with monotonicity constant  $\underline{\nu}$ , i.e.,

$$(\hat{\nu}(s)s - \hat{\nu}(t)t)(s - t) \geq \underline{\nu}(s - t)^2 \quad \forall s, t \in \mathbb{R}_0^+, \quad (2.20)$$

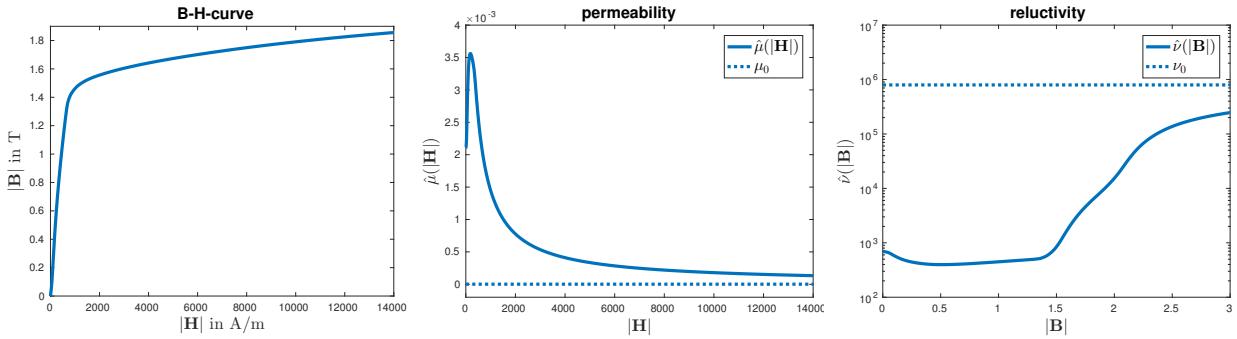


Figure 2.4: Left:  $B$ - $H$ -curve  $f$  of ferromagnetic material used in simulation of electric motor. Center: magnetic permeability  $\hat{\mu}$ . Right: magnetic reluctivity  $\hat{\nu}$  in semi-logarithmic plot.

and Lipschitz continuous with Lipschitz constant  $\nu_0$ , i.e.,

$$|\hat{\nu}(s)s - \hat{\nu}(t)t| \leq \nu_0 |s - t| \quad \forall s, t \in \mathbb{R}^+. \quad (2.21)$$

We will see in Section 2.5 that properties (2.20) and (2.21) are important for the well-posedness of the boundary value problem (2.14).

**Remark 2.4.** Since (2.20) and (2.21) obviously also hold for the constant magnetic reluctivity function  $\nu(s) = \nu_0$ , we get these relations also for the global magnetic reluctivity function of the motor defined in (2.11). This means that also the mapping  $s \mapsto \nu_\Omega(x, s)s$  is strongly monotone and Lipschitz continuous with constants the same constants  $\underline{\nu}$  and  $\nu_0$ , independent of the spatial position  $x \in D$ .

For ferromagnetic materials, the  $B$ - $H$  relations are generally not known analytically. In practice, they are usually interpolated or approximated from measured values. It is an essential task in the numerical computation of electromagnetic fields in the presence of nonlinear material to approximate this  $B$ - $H$ -curve suitably. In particular, it must be ensured that properties (2.20) and (2.21) remain fulfilled, even in the presence of uncertainties on the given data due to, e.g., measurement errors. This issue was addressed in [121] from the perspective of inverse problems, and in [169] by a monotonicity-preserving interproximation of the given data points. We remark that in all numerical computations of this thesis, the  $B$ - $H$ -curve was computed by a software library which implements the approach presented in [169] that was kindly provided by Dr. Clemens Pechstein.

## 2.5 Analysis of Two-dimensional Nonlinear Magnetostatics

### 2.5.1 Existence and Uniqueness of a Solution to the State Equation

The physical properties of  $B$ - $H$ -curves presented in Section 2.4 allow us to show existence and uniqueness of the boundary value problem (2.14). The result follows from the Theorem of Zarantonello [219, pp. 503], which can also be regarded as a nonlinear extension of the lemma of Lax-Milgram.

**Theorem 2.5** (Zarantonello, 1960). *Let  $X$  be a real Hilbert space and  $A : X \rightarrow X^*$  a nonlinear operator satisfying the following conditions:*

1.  *$A$  is strongly monotone on  $X$ , i.e., there exists a constant  $c > 0$  such that, for all  $u, v \in X$ , it holds*

$$\langle A(u) - A(v), u - v \rangle \geq c \|u - v\|^2, \quad (2.22)$$

2.  *$A$  is Lipschitz continuous on  $X$ , i.e., there exists a constant  $L > 0$  such that, for all  $u, v \in X$ , it holds*

$$\|A(u) - A(v)\| \leq L \|u - v\|. \quad (2.23)$$

*Then, for each  $b \in X^*$ , the operator equation  $A(u) = b$ ,  $u \in X$ , has a unique solution which depends continuously on  $b$ . More precisely, it follows from  $A(u_j) = b_j$ ,  $j = 1, 2$ , that*

$$\|u_1 - u_2\| \leq c^{-1} \|b_1 - b_2\|.$$

In order to apply this theorem to the two-dimensional magnetostatic boundary value problem (2.14), we need the strong monotonicity and the Lipschitz continuity of the operator  $A_\Omega$  defined in (2.12) on the real Hilbert space  $H_0^1(D)$ . Both results were shown in [106] under the (rather restrictive) assumption that the magnetic reluctivity  $\hat{\nu}$  is monotonically increasing, and in, e.g., [168] in the more general setting (2.20)–(2.21) where the mapping  $s \mapsto \hat{\nu}(s)s$  is monotone, but not necessarily  $\hat{\nu}$  itself. Note that this is the case for a large class of ferromagnetic materials including the one used in the numerical experiments, see also Figure 2.4 and Figure 4.1.

**Lemma 2.6** ([168, 222]). *Assume that the mapping  $s \mapsto \nu_\Omega(s)s$  is strongly monotone and Lipschitz continuous from  $\mathbb{R}_0^+$  to  $\mathbb{R}_0^+$  with monotonicity constant  $\underline{\nu}$  and Lipschitz constant  $\nu_0$ . Then the operator  $A_\Omega$  defined in (2.12) is strongly monotone with monotonicity constant  $\underline{\nu}$  and Lipschitz constant  $\nu_0$ .*

Since the right hand  $F$  defined in (2.13) is an element of the dual space  $H_0^1(D)^* = H^{-1}(D)$ , we get the well-posedness of boundary value problem (2.14):

**Theorem 2.7.** *Let  $\hat{\nu}$  be the magnetic reluctivity according to a B–H-curve  $f$  which satisfies Assumption 1. Then, the two-dimensional magnetostatic boundary value problem defined by (2.14), (2.12), (2.13) has a unique solution  $u \in H_0^1(D)$  and there exists a constant  $c > 0$  such that*

$$\|u\|_{H_0^1(D)} \leq c \|F\|_{H^{-1}(D)}.$$

*Proof.* Since the underlying B–H-curve  $f$  satisfies Assumption 1, we have the strong monotonicity and the Lipschitz continuity of the reluctivity function  $\hat{\nu}$  in the ferromagnetic subdomain  $\Omega_f$  and also of the global reluctivity function  $\nu_\Omega$  as stated in Remark 2.4. Therefore, Lemma 2.6 implies that Theorem 2.5 is applicable to the operator  $A_\Omega$  introduced in (2.12) along with the right hand side  $F$  (2.13) for the real Hilbert space  $X = H_0^1(D)$ , which yields the statement with  $c = \underline{\nu}$ .  $\square$

A solution to the nonlinear variational equation (2.14) can be computed by Newton's method (see e.g. [156]). In Newton's method, one chooses an initial guess  $u^0$ , and iteratively solves a linearized version of (2.14) to obtain an update for the solution  $u$ . The method converges quadratically provided that the initial guess is chosen sufficiently close to the true solution.

### 2.5.2 Newton Operator and Adjoint Equation

In order to apply Newton's method to (2.14), we need the Fréchet derivative  $A'_\Omega$  of the operator  $A_\Omega$ . The same operator  $A'_\Omega$  will also show up later in the boundary value problem defining the adjoint state of the PDE-constrained optimization problem (2.17).

For sake of better readability we introduce the operator  $T: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  given by

$$T(W) := \hat{\nu}(|W|)W \quad (2.24)$$

with its Jacobian

$$DT(W) = \begin{cases} \hat{\nu}(|W|)I + \frac{\hat{\nu}'(|W|)}{|W|}W \otimes W, & W \neq (0, 0)^\top \\ \hat{\nu}(|W|)I & W = (0, 0)^\top, \end{cases} \quad (2.25)$$

where  $W \in \mathbb{R}^2$ ,  $I$  denotes the identity matrix in  $\mathbb{R}^2$  and  $\otimes$  the outer product between two column vectors,

$$a \otimes b := a b^\top,$$

for  $a, b \in \mathbb{R}^2$ . Note that  $DT$  is continuous also in  $W = 0$ . Let further

$$\begin{aligned} T_\Omega(x, W) &:= \nu_\Omega(x, W)W = \chi_{\Omega_f}(x)T(W) + \chi_{D \setminus \Omega_f}(x)\nu_0W \\ &= \left( \chi_{\Omega_f^{ref} \setminus \Omega^d}(x) + \chi_\Omega(x) \right)T(W) + \left( \chi_{D \setminus \Omega_f^{ref}}(x) + \chi_{\Omega^d \setminus \Omega}(x) \right)\nu_0W, \end{aligned} \quad (2.26)$$

such that  $\langle A_\Omega(u), \eta \rangle = \int_D T_\Omega(x, \nabla u) \cdot \nabla \eta \, dx$ , and note that

$$DT_\Omega(x, W) = \chi_{\Omega_f}(x)DT(W) + \chi_{D \setminus \Omega_f}(x)\nu_0I.$$

The Fréchet derivative of the operator  $A_\Omega : H_0^1(D) \rightarrow H^{-1}(D)$  is then given by

$$\begin{aligned} A'_\Omega &: H_0^1(D) \rightarrow \mathcal{L}(H_0^1(D), H^{-1}(D)), \\ \langle A'_\Omega(u)w, \eta \rangle &= \int_D DT_\Omega(x, \nabla u) \nabla w \cdot \nabla \eta \, dx, \end{aligned} \quad (2.27)$$

where  $u, w, \eta \in H_0^1(D)$ . For  $W = (w_1, w_2)^\top \in \mathbb{R}^2$ , let  $\theta_W$  the angle between  $W$  and the  $x_1$ -axis such that  $\cos \theta_W = \langle W/|W|, e_1 \rangle$  with  $e_1 = (1, 0)^\top$ , and denote  $R_{\theta_W}$  the counter-clockwise rotation matrix around an angle  $\theta_W$ , i.e.,

$$R_{\theta_W} = \begin{pmatrix} \cos \theta_W & -\sin \theta_W \\ \sin \theta_W & \cos \theta_W \end{pmatrix}.$$

Note that

$$DT(W) = R_{\theta_W} \begin{pmatrix} \hat{\nu}(|W|) + \hat{\nu}'(|W|)|W| & 0 \\ 0 & \hat{\nu}(|W|) \end{pmatrix} R_{\theta_W}^\top \quad (2.28)$$

for all  $W \in \mathbb{R}^2$ . From this, it is easy to see that the eigenvalues and corresponding eigenvectors of  $DT(W)$  are

$$\begin{aligned} \lambda_1 &= \lambda_1(|W|) = \hat{\nu}(|W|), & v_1 &= \begin{pmatrix} -w_2 \\ w_1 \end{pmatrix}, \\ \lambda_2 &= \lambda_2(|W|) = \hat{\nu}(|W|) + \hat{\nu}'(|W|)|W|, & v_2 &= \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}. \end{aligned} \quad (2.29)$$

Note that  $DT(W)$  is symmetric and, due to (2.19a) and (2.19b), positive definite for all  $W \in \mathbb{R}^2$ . Now we can show the existence of a unique solution to a problem that represents a linearization of the boundary value problem (2.14).

**Lemma 2.8.** *Let  $u \in H_0^1(D)$  fixed and define the bilinear form  $a_u(w, \eta) := \langle A'_\Omega(u)w, \eta \rangle$  with  $A'_\Omega : H_0^1(D) \rightarrow \mathcal{L}(H_0^1(D), H^{-1}(D))$  given in (2.27). Let furthermore  $b \in H^{-1}(D)$ . Then the variational problem*

$$\text{Find } w \in H_0^1(D) \text{ such that } a_u(w, \eta) = \langle b, \eta \rangle \quad \forall \eta \in H_0^1(D) \quad (2.30)$$

has a unique solution.

*Proof.* It can easily be seen from (2.19) that the bilinear form  $a_u(\cdot, \cdot)$  is elliptic,

$$\begin{aligned} a_u(w, w) &= \langle A'_\Omega(u)w, w \rangle = \int_D DT_\Omega(x, \nabla u) \nabla w \cdot \nabla w \, dx \\ &\geq \int_D \min\{\lambda_1(|\nabla u|), \lambda_2(|\nabla u|)\} |\nabla w|^2 \, dx \\ &\geq \underline{\nu} \|\nabla w\|_{L^2(D)}^2 \geq C \underline{\nu} \|w\|_{H^1(D)}^2, \end{aligned}$$

and bounded,

$$\begin{aligned} a_u(w, \eta) &= \langle A'_\Omega(u)w, \eta \rangle = \int_D DT_\Omega(x, \nabla u) \nabla w \cdot \nabla \eta \, dx \\ &\leq \int_D \max\{\lambda_1(|\nabla u|), \lambda_2(|\nabla u|)\} |\nabla w| |\nabla \eta| \, dx \\ &\leq \nu_0 \|\nabla w\|_{L^2(D)} \|\nabla \eta\|_{L^2(D)} \leq C \nu_0 \|w\|_{H^1(D)} \|\eta\|_{H^1(D)}, \end{aligned}$$

in  $H^1(D)$  for any given  $u \in H_0^1(D)$ . Here, we exploited the equivalence of the norm in  $H^1(D)$  and the  $L^2(D)$  norm of the gradient due to the Friedrichs inequality, see e.g. [47], since  $w, \eta \in H_0^1(D)$ . Thus, the statement follows by the lemma of Lax-Milgram, see e. g. [47].  $\square$

**Remark 2.9** (Newton's method). *Boundary value problem (2.14) can be solved by applying Newton's method to the operator equation  $R_\Omega(u) := A_\Omega(u) - F = 0$  in  $H^{-1}(D)$ . For that purpose, we start with an initial guess  $u^0 \in H_0^1(D)$  and, for  $k = 0, 1, 2, \dots$ , compute the next iterate  $u^{k+1}$  as the sum of the current iterate  $u^k$  and an update  $w^k$ . Here, the update  $w^k \in H_0^1(D)$  is the solution to the problem*

$$\text{Find } w \in H_0^1(D) \text{ such that } \langle A'_\Omega(u^k)w, \eta \rangle = -\langle R_\Omega(u^k), \eta \rangle \quad \forall \eta \in H_0^1(D),$$

which is well-posed due to Lemma 2.8. The fact that Newton's method converges only locally can be overcome by using the damped version of Newton's method, i.e., by making the update step  $u^{k+1} = u^k + \tau^k w^k$  with a damping parameter  $\tau^k \in (0, 1]$  which is chosen small enough such that  $\|R_\Omega(u^k + \tau^k w^k)\| < \|R_\Omega(u^k)\|$ . It can be shown that such a choice of  $\tau^k$  is always possible [156, 222].

**Remark 2.10** (Adjoint equation). *We will also encounter equation (2.30) as the equation defining the adjoint state  $p$  of the PDE-constrained optimization problem (2.17). The adjoint problem reads*

$$\text{Find } p \in H_0^1(D) \text{ such that } \langle A'_\Omega(u)\eta, p \rangle = -\langle \frac{\partial \mathcal{J}}{\partial u}(u), \eta \rangle \quad \forall \eta \in H_0^1(D), \quad (2.31)$$

where  $u$  is the solution to the primal problem (2.14). Note that  $A'_\Omega(u)$  is symmetric for all  $u \in H_0^1(D)$ , i.e., we have  $\langle A'_\Omega(u)\eta, w \rangle = \langle A'_\Omega(u)w, \eta \rangle$  for all  $u, \eta, w \in H_0^1(D)$ , and thus the above equation is equivalent to (2.30).

Because of these contexts, we call the operator  $A'_\Omega$  defined in (2.27) the adjoint operator or the Newton operator.

# **Part I**

# **Topology Optimization**



## Chapter 3

# Material Sensitivities in Takahashi's On/Off Method

In this chapter, we investigate a first sensitivity-based design optimization method for the optimization of electrical equipment. The gradient-based *On/Off method* was introduced by the group around Norio Takahashi in [163] and, since then, has been successfully applied to a wide range of problems in the electrical engineering community, see, e.g., [161, 162, 202–205]. The method uses sensitivities of the objective functional with respect to a local perturbation of the magnetic reluctivity and distributes ferromagnetic material according to this sensitivity information.

The On/Off method acts on the discrete level, i.e., after a discretization of the computational domain. It is based on the fact that the difference between having ferromagnetic material or air in an element of the finite element (FE) mesh is only reflected in the value of the magnetic reluctivity. Recall that the magnetic reluctivity of air is the constant  $\nu_0$ , see (2.10), whereas it is a nonlinear function  $\hat{\nu} = \hat{\nu}(|\nabla u|)$  depending on the absolute value of the magnetic flux density,  $|\mathbf{B}| = |\nabla u|$ , in the ferromagnetic material, see Section 2.4. Note that, for ferromagnetic materials, this value is usually much smaller than the reluctivity of air,  $\hat{\nu}(|\nabla u|) \ll \nu_0$ . Then, for each element of the mesh inside the design area, the sensitivity of the objective function with respect to a perturbation of the material coefficient only in this one element is calculated. If this sensitivity is negative, a larger value of the magnetic reluctivity  $\nu$  is favorable for reducing the value of the objective function, which is realized by assigning the reluctivity value of air to this element. The element is said to be switched “off”. On the other hand, if the sensitivity is positive, it is favorable to have ferromagnetic material in this element, i.e., the element is switched “on”. Note that this procedure is based on a monotonicity assumption which cannot be rigorously guaranteed. Therefore, the method is heuristic.

Thus, the principle of the method is very similar to classical density-based topology optimization methods where a density variable  $\rho$  which can attain values between 0 and 1 is used to interpolate between two materials. In density-based methods, the optimization is driven by the sensitivity of the objective function with respect to this density variable, see, e.g. [43]. In the context of structural mechanics, this sensitivity is called the strain energy density. The On/Off method does not introduce a density variable and does not allow for intermediate values of the material coefficient, but still uses the same kind of sensitivities. In particular, the On/Off method is very similar to the evolutionary structural optimization (ESO) method,

see [50, 212]. Also the ESO is based on an implicit monotonicity assumption which is not always satisfied, see also the discussion in [50, Section 7.6].

Later, in Chapter 5, we will draw a comparison between the sensitivities with respect to the material coefficient used in the On/Off method (henceforth referred to as the *On/Off sensitivities*), and the topological derivative, which is derived in Chapter 4, for both the case of linear and nonlinear material behavior in the ferromagnetic subdomain.

In Section 3.1, we present the sensitivity analysis method proposed in [163] where the sensitivities are calculated for each element of the FE mesh inside the design area. In Section 3.2, we generalize this idea to the continuous level where we consider perturbations of coefficients in arbitrary, smooth subdomains  $\omega$  of the design domain  $\Omega^d$ . We conclude the chapter with numerical results for problem (2.17) obtained by the On/Off method.

This chapter is based on [88].

### 3.1 Discrete Sensitivity Analysis

In this section, we present the On/Off method introduced in [163]. The On/Off sensitivities of the objective function have to be computed for every element of the finite element mesh inside the design area. By introducing and adjoint variable, only one linear problem has to be solved in order to determine the On/Off sensitivities for all elements.

In this section, we will consider problem (2.17) in the case of the nonlinear state equation with  $\nu$  given in (2.9). Note that the case of a linear state equation where  $\hat{\nu}$  is replaced by a constant  $\nu_1 < \nu_0$  is a special case. The discretization of the state equation (2.17b) by means of linear triangular finite elements yields a system of nonlinear finite element equations of the form

$$\mathbf{K}(\mathbf{u})\mathbf{u} = \mathbf{F}, \quad (3.1)$$

where  $\mathbf{u}$  denotes the nodal parameter vector that we have to determine, see, e.g., [106]. Given an objective function  $\mathcal{J} = \mathcal{J}(\nu_k, \mathbf{u})$ , we are interested in the sensitivities

$$\frac{d\mathcal{J}}{d\nu_k} = \frac{\partial \mathcal{J}}{\partial \nu_k} + \frac{\partial \mathcal{J}}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial \nu_k}, \quad (3.2)$$

where the design parameter  $\nu_k$  is the magnetic reluctivity in a triangular element  $T_k$  in the FE mesh inside the design area. Since we are using linear triangular elements, the gradient of the finite element function is constant in every finite element. Thus, for the finite element solution, the reluctivity is constant in every finite element as well. In our model problem, the objective functional  $\mathcal{J}$  defined in (2.15) does not depend explicitly on the reluctivity inside the design area, therefore  $\frac{\partial \mathcal{J}}{\partial \nu_k} = 0$ . In order to determine the sensitivities  $\frac{\partial \mathbf{u}}{\partial \nu_k}$ , we consider the residual identity

$$\mathbf{r}(\nu_k, \nu(\mathbf{u}(\nu_k)), \mathbf{u}(\nu_k)) := \mathbf{K}(\nu_k, \nu(\mathbf{u}(\nu_k)))\mathbf{u}(\nu_k) - \mathbf{F} \equiv 0 \quad (3.3)$$

at the solution, where the dependencies on  $\nu_k$  are now explicitly specified. Differentiating

both sides of (3.3) with respect to  $\nu_k$ , we obtain

$$\begin{aligned} 0 &= \frac{d\mathbf{r}}{d\nu_k} = \frac{\partial\mathbf{r}}{\partial\nu_k} + \frac{\partial\mathbf{r}}{\partial\nu} \frac{\partial\nu}{\partial\mathbf{u}} \frac{\partial\mathbf{u}}{\partial\nu_k} + \frac{\partial\mathbf{r}}{\partial\mathbf{u}} \frac{\partial\mathbf{u}}{\partial\nu_k} \\ &= \frac{\partial\mathbf{K}}{\partial\nu_k} \mathbf{u} + (\mathbf{N} + \mathbf{K}) \frac{\partial\mathbf{u}}{\partial\nu_k}, \end{aligned}$$

where we introduced

$$\mathbf{N} := \frac{\partial\mathbf{r}}{\partial\nu} \frac{\partial\nu}{\partial\mathbf{u}} = \mathbf{u}^\top \frac{\partial\mathbf{K}}{\partial\nu} \frac{\partial\nu}{\partial\mathbf{u}} = \mathbf{u}^\top \frac{d\mathbf{K}}{d\mathbf{u}}.$$

From this, it can be seen that the sensitivity  $\frac{\partial\mathbf{u}}{\partial\nu_k}$  is given by

$$\frac{\partial\mathbf{u}}{\partial\nu_k} = -(\mathbf{N} + \mathbf{K})^{-1} \frac{\partial\mathbf{K}}{\partial\nu_k} \mathbf{u}. \quad (3.4)$$

Here we used the fact that, for our model problem,  $\frac{\partial\mathbf{F}}{\partial\nu_k} = 0$  since the right hand side  $\mathbf{F}$ , which is the finite element discretized version of the linear functional (2.13), does not explicitly depend on the reluctivity  $\nu_k$  in elements  $T_k$  located in the design area. Inserting (3.4) into (3.2) yields the formula for the On/Off sensitivities,

$$\frac{d\mathcal{J}}{d\nu_k} = \mathbf{p}^\top \left( \frac{\partial\mathbf{K}}{\partial\nu_k} \mathbf{u} \right) \quad (3.5)$$

where the adjoint state  $\mathbf{p}$  solves the adjoint equation

$$(\mathbf{K} + \mathbf{N})^\top \mathbf{p} = -\frac{\partial\mathcal{J}}{\partial\mathbf{u}}. \quad (3.6)$$

**Remark 3.1.** Note that the matrix  $\mathbf{K} + \mathbf{N}$  depends on the solution vector  $\mathbf{u}$  and is the finite element matrix corresponding to the adjoint operator  $A'_\Omega(u)$  introduced in (2.27).

**Remark 3.2.** In the case of a linear state equation, the On/Off sensitivities are of the same form (3.5). The only differences lie in the computation of the direct state  $\mathbf{u}$  from (3.1) as the nonlinear operator  $\mathbf{K}(\cdot)$  degenerates to the linear operator  $\mathbf{K}$ , the stiffness matrix of the boundary value problem (2.17b), and in the computation of the adjoint state  $\mathbf{p}$  from (3.6) since the matrix  $\mathbf{N}$  vanishes.

## 3.2 Generalization to Continuous Level

In this section, we will generalize the idea of [163], which is based on a FE discretization, to the continuous level. We will consider perturbations of the material parameter on arbitrary, smooth subdomains  $\omega$  of the design domain  $\Omega^d$  rather than only on the single elements of the FE mesh, and we will derive the formula for the sensitivities in terms of operators rather than FE vectors.

Let  $\omega \subset \Omega^d$  be fixed and let  $g : H_0^1(D) \rightarrow L^2(D)$  be defined by

$$g(u) := |\nabla u(\cdot)|,$$

with the Fréchet derivative  $g' : H_0^1(D) \rightarrow \mathcal{L}(H_0^1(D), L^2(D))$  given by

$$g'(u) = \frac{1}{|\nabla u|} \nabla u \cdot \nabla(\cdot).$$

Moreover, we define  $\tilde{\nu} : H_0^1(D) \rightarrow L^\infty(D)$ ,

$$\tilde{\nu}(u) := \hat{\nu}(g(u)), \quad (3.7)$$

where  $\hat{\nu} : \mathbb{R} \rightarrow \mathbb{R}$  is a nonlinear function given via a *B–H*-curve which fulfills Assumption 1, see Section 2.4. Then we have  $\tilde{\nu}' : H_0^1(D) \rightarrow \mathcal{L}(H_0^1(D), L^\infty(D))$  given by

$$\tilde{\nu}'(u) = \hat{\nu}'(g(u))g'(u) = \frac{\hat{\nu}'(|\nabla u|)}{|\nabla u|} \nabla u \cdot \nabla(\cdot). \quad (3.8)$$

Recall the notation introduced in Section 2.3. For the rest of this chapter, let  $\Omega = \Omega^d$ , i.e., all of the design subdomain  $\Omega^d$  is occupied by ferromagnetic material. Then, it holds that  $\omega \subset \Omega^d \subset \Omega_f$  and the global reluctivity function (2.11) can be rewritten as  $\nu : D \times H_0^1(D) \rightarrow L^\infty(D)$  defined by

$$\begin{aligned} \nu(x, u) &:= \chi_{\Omega_f}(x) \tilde{\nu}(u) + \chi_{\Omega_{air}}(x) \nu_0 \\ &= \chi_\omega(x) \tilde{\nu}_\omega(u) + \chi_{\Omega_f \setminus \omega}(x) \tilde{\nu}(u) + \chi_{\Omega_{air}}(x) \nu_0 \quad \forall x \in D \text{ a.e.}, \end{aligned} \quad (3.9)$$

where  $\chi_A$  denotes the characteristic function of a set  $A$ , and  $\tilde{\nu}_\omega$  denotes the restriction of  $\tilde{\nu}$  onto  $\omega$ . Here, for a more compact presentation, we used the magnetic reluctivity of air  $\nu_0$  also inside the magnet areas  $\Omega_{mag} \subset \Omega_{air}$ .

We will derive the sensitivity of a functional  $\mathcal{J} = \mathcal{J}(\nu, u(\nu))$  with respect to a perturbation of the magnetic reluctivity inside  $\omega$ , i.e., with respect to a perturbation of  $\tilde{\nu}_\omega$ . Application of the chain rule yields

$$\frac{d\mathcal{J}}{d\tilde{\nu}_\omega} = \frac{\partial \mathcal{J}}{\partial \nu} \frac{\partial \nu}{\partial \tilde{\nu}_\omega} + \frac{\partial \mathcal{J}}{\partial u} \frac{\partial u}{\partial \tilde{\nu}_\omega}. \quad (3.10)$$

Again the sensitivity  $\frac{\partial u}{\partial \tilde{\nu}_\omega}$  is obtained by setting the residual operator to zero and forming the Fréchet derivative of both sides. For that purpose, we introduce the operators

$$\begin{aligned} R_1(\tilde{\nu}_\omega) &:= \int_{\omega} \tilde{\nu}(u(\tilde{\nu}_\omega)) \nabla u(\tilde{\nu}_\omega) \cdot \nabla(\cdot) dx, \\ R_2(\tilde{\nu}_\omega) &:= \int_{\Omega_f \setminus \omega} \tilde{\nu}(u(\tilde{\nu}_\omega)) \nabla u(\tilde{\nu}_\omega) \cdot \nabla(\cdot) dx + \int_{\Omega_{air}} \nu_0 \nabla u(\tilde{\nu}_\omega) \cdot \nabla(\cdot) dx, \end{aligned}$$

and define the residual operator  $R(\tilde{\nu}_\omega) := R_1(\tilde{\nu}_\omega) + R_2(\tilde{\nu}_\omega) - F$ , i.e.,

$$\begin{aligned} R(\tilde{\nu}_\omega) &= \int_{\omega} \tilde{\nu}(u(\tilde{\nu}_\omega)) \nabla u(\tilde{\nu}_\omega) \cdot \nabla(\cdot) dx \\ &\quad + \int_{\Omega_f \setminus \omega} \tilde{\nu}(u(\tilde{\nu}_\omega)) \nabla u(\tilde{\nu}_\omega) \cdot \nabla(\cdot) dx + \int_{\Omega_{air}} \nu_0 \nabla u(\tilde{\nu}_\omega) \cdot \nabla(\cdot) dx - F, \end{aligned}$$

where  $F$  defined in (2.13). Note that, for the solution  $u$  of the boundary value problem (2.17b), the residual  $R$  vanishes, i.e.,

$$R(\tilde{\nu}_\omega) \equiv 0. \quad (3.11)$$

Also note that, in our case, the right hand side  $F$  as defined in (2.13) is independent of the magnetic reluctivity  $\nu$  in the design subdomain  $\Omega^d$ . We differentiate both sides of (3.11) with respect to  $\tilde{\nu}_\omega$ . We begin with  $R_1$ :

$$\begin{aligned} \frac{dR_1}{d\tilde{\nu}_\omega} h_\omega &= \lim_{t \rightarrow 0} \frac{1}{t} (R_1(\tilde{\nu}_\omega + th_\omega) - R_1(\tilde{\nu}_\omega)) \\ &= \lim_{t \rightarrow 0} \frac{1}{t} \left\{ \int_{\omega} \tilde{\nu}(u(\tilde{\nu}_\omega + th_\omega)) \nabla u(\tilde{\nu}_\omega + th_\omega) \cdot \nabla(\cdot) dx \right. \\ &\quad \left. - \int_{\omega} \tilde{\nu}(u(\tilde{\nu}_\omega)) \nabla u(\tilde{\nu}_\omega) \cdot \nabla(\cdot) dx + t \int_{\omega} h_\omega(u(\tilde{\nu}_\omega + th_\omega)) \nabla u(\tilde{\nu}_\omega + th_\omega) \cdot \nabla(\cdot) dx \right\}. \end{aligned}$$

Using the expansions

$$u(\tilde{\nu}_\omega + th_\omega) = u(\tilde{\nu}_\omega) + t \frac{\partial u}{\partial \tilde{\nu}_\omega}(h_\omega) + \mathcal{O}(t^2), \quad (3.12)$$

$$\tilde{\nu} \left( u(\tilde{\nu}_\omega) + t \frac{\partial u}{\partial \tilde{\nu}_\omega}(h_\omega) \right) = \tilde{\nu}(u(\tilde{\nu}_\omega)) + t \tilde{\nu}'(u(\tilde{\nu}_\omega)) \frac{\partial u}{\partial \tilde{\nu}_\omega}(h_\omega) + \mathcal{O}(t^2), \quad (3.13)$$

we obtain

$$\begin{aligned} \frac{dR_1}{d\tilde{\nu}_\omega} h_\omega &= \int_{\omega} \tilde{\nu}(u(\tilde{\nu}_\omega)) \nabla \frac{\partial u}{\partial \tilde{\nu}_\omega}(h_\omega) \cdot \nabla(\cdot) dx + \int_{\omega} \tilde{\nu}'(u(\tilde{\nu}_\omega)) \frac{\partial u}{\partial \tilde{\nu}_\omega}(h_\omega) \nabla u(\tilde{\nu}_\omega) \cdot \nabla(\cdot) dx \\ &\quad + \int_{\omega} h_\omega(u(\tilde{\nu}_\omega)) \nabla u(\tilde{\nu}_\omega) \cdot \nabla(\cdot) dx. \end{aligned} \quad (3.14)$$

Note that, here,  $\frac{\partial u}{\partial \tilde{\nu}_\omega}(h_\omega)$  denotes the sensitivity of the state variable  $u$  when the material coefficient  $\tilde{\nu}_\omega$  inside  $\omega$  is perturbed by  $h_\omega \in L^\infty(\omega)$ . For  $R_2$ , we get

$$\begin{aligned} \frac{dR_2}{d\tilde{\nu}_\omega} h_\omega &= \lim_{t \rightarrow 0} \frac{1}{t} (R_2(\tilde{\nu}_\omega + th_\omega) - R_2(\tilde{\nu}_\omega)) \\ &= \lim_{t \rightarrow 0} \frac{1}{t} \left\{ \int_{\Omega_f \setminus \omega} \tilde{\nu}(u(\tilde{\nu}_\omega + th_\omega)) \nabla u(\tilde{\nu}_\omega + th_\omega) \cdot \nabla(\cdot) dx \right. \\ &\quad \left. - \int_{\Omega_f \setminus \omega} \tilde{\nu}(u(\tilde{\nu}_\omega)) \nabla u(\tilde{\nu}_\omega) \cdot \nabla(\cdot) dx + \int_{\Omega_{air}} \nu_0 (\nabla u(\tilde{\nu}_\omega + th_\omega) - \nabla u(\tilde{\nu}_\omega)) \cdot \nabla(\cdot) dx \right\}. \end{aligned}$$

Again, using expansions (3.12) and (3.13), we get

$$\begin{aligned} \frac{dR_2}{d\tilde{\nu}_\omega} h_\omega &= \int_{\Omega_f \setminus \omega} \tilde{\nu}(u(\tilde{\nu}_\omega)) \nabla \frac{\partial u}{\partial \tilde{\nu}_\omega}(h_\omega) \cdot \nabla(\cdot) dx \\ &\quad + \int_{\Omega_f \setminus \omega} \tilde{\nu}'(u(\tilde{\nu}_\omega)) \frac{\partial u}{\partial \tilde{\nu}_\omega}(h_\omega) \nabla u(\tilde{\nu}_\omega) \cdot \nabla(\cdot) dx + \int_{\Omega_{air}} \nu_0 \nabla \frac{\partial u}{\partial \tilde{\nu}_\omega}(h_\omega) \cdot \nabla(\cdot) dx. \end{aligned} \quad (3.15)$$

Combining (3.11), (3.14) and (3.15), as well as using the notation (3.9) yields

$$\begin{aligned} 0 &= \frac{dR}{d\tilde{\nu}_\omega} h_\omega = \frac{dR_1}{d\tilde{\nu}_\omega} h_\omega + \frac{dR_2}{d\tilde{\nu}_\omega} h_\omega \\ &= \int_D \nu(x, u(\tilde{\nu}_\omega)) \nabla \frac{\partial u}{\partial \tilde{\nu}_\omega}(h_\omega) \cdot \nabla(\cdot) dx + \int_D \nu'(x, u(\tilde{\nu}_\omega)) \frac{\partial u}{\partial \tilde{\nu}_\omega}(h_\omega) \nabla u(\tilde{\nu}_\omega) \cdot \nabla(\cdot) dx \\ &\quad + \int_\omega h_\omega(u(\tilde{\nu}_\omega)) \nabla(u(\tilde{\nu}_\omega)) \cdot \nabla(\cdot) dx, \end{aligned}$$

where the differentiation in  $\nu'(x, u(\tilde{\nu}_\omega))$  is to be understood with respect to  $u$  as in (3.8). Here,  $h_\omega \in L^\infty(\omega)$  is the direction of the perturbation we are considering. For our purposes, it is sufficient to consider constant perturbations of  $\tilde{\nu}_\omega$ , therefore we set

$$h_\omega \equiv 1. \quad (3.16)$$

Note that by using general  $h_\omega \in L^\infty(\omega)$ , a weighted perturbation of  $\tilde{\nu}_\omega$  can be simulated. Plugging in (3.16) as well as (3.9), (3.7) and (3.8), we get the equality

$$\begin{aligned} 0 &= \int_{\Omega_f} \hat{\nu}(|\nabla u|) \nabla \frac{\partial u}{\partial \tilde{\nu}_\omega} \cdot \nabla(\cdot) dx + \int_{\Omega_f} \frac{\hat{\nu}'(|\nabla u|)}{|\nabla u|} (\nabla u \cdot \nabla \frac{\partial u}{\partial \tilde{\nu}_\omega}) (\nabla u \cdot \nabla(\cdot)) dx \\ &\quad + \int_{\Omega_{air}} \nu_0 \nabla \frac{\partial u}{\partial \tilde{\nu}_\omega} \cdot \nabla(\cdot) dx + \int_\omega \nabla u \cdot \nabla(\cdot) dx, \end{aligned}$$

from which we can obtain  $\frac{\partial u}{\partial \tilde{\nu}_\omega}$ . Introducing the linear operators

$$\begin{aligned} K_u : H_0^1(D) &\rightarrow H^{-1}(D), \\ K_u w &= \int_{\Omega_f} \hat{\nu}(|\nabla u|) \nabla w \cdot \nabla(\cdot) dx + \int_{\Omega_{air}} \nu_0 \nabla w \cdot \nabla(\cdot) dx, \\ N_u : H_0^1(D) &\rightarrow H^{-1}(D), \\ N_u w &= \int_{\Omega_f} \frac{\hat{\nu}'(|\nabla u|)}{|\nabla u|} (\nabla u \cdot \nabla w) (\nabla u \cdot \nabla(\cdot)) dx, \end{aligned}$$

for fixed  $u \in H_0^1(D)$ , we can formally write

$$\frac{\partial u}{\partial \tilde{\nu}_\omega} = -(K_u + N_u)^{-1} M_\omega u, \quad (3.17)$$

with

$$\begin{aligned} M_\omega : H_0^1(D) &\rightarrow H^{-1}(D), \\ M_\omega u &= \int_\omega \nabla u \cdot \nabla(\cdot) dx. \end{aligned}$$

Note that the operator  $K_u + N_u$  is nothing but the operator  $A'_\Omega(u)$  defined in (2.27), which is invertible due to Lemma 2.8. Combining (3.10) and (3.17) gives

$$\begin{aligned} \frac{d\mathcal{J}}{d\nu_\omega} &= \frac{\partial \mathcal{J}}{\partial \nu} \frac{\partial \nu}{\partial \nu_\omega} - \frac{\partial \mathcal{J}}{\partial u} (K_u + N_u)^{-1} M_\omega u \\ &= \frac{\partial \mathcal{J}}{\partial \nu} \frac{\partial \nu}{\partial \nu_\omega} + p^* M_\omega u \end{aligned} \quad (3.18)$$

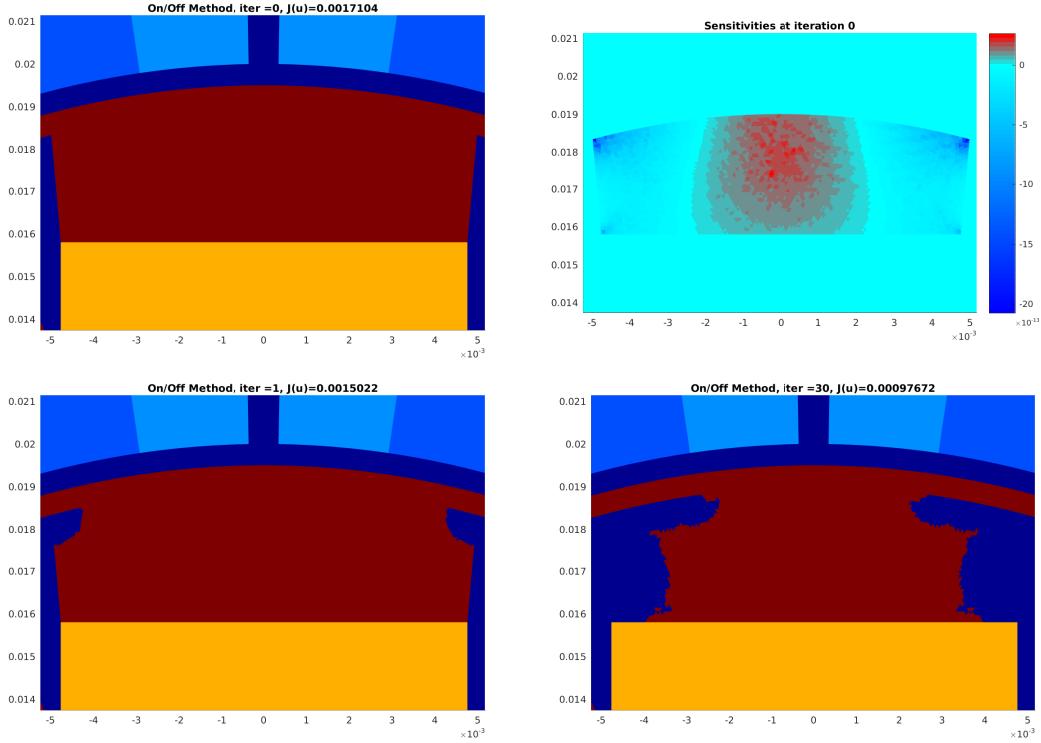


Figure 3.1: Top left: Initial design (all elements on). Top right: On/Off sensitivities. Bottom left: Design after first iteration. Bottom right: Final design after 30 iterations of Algorithm 1.

where the adjoint state  $p \in H_0^1(D)$  is given by the adjoint equation

$$(K_u + N_u)^* p = -\frac{\partial \mathcal{J}}{\partial u}, \quad (3.19)$$

which is the same as (2.31). Again, if  $\mathcal{J}$  does not depend on  $\nu$  explicitly as it is the case in our model problem with  $\mathcal{J}$  given by (2.15), (3.18) simplifies to

$$\frac{d\mathcal{J}}{d\nu_\omega} = p^* M_\omega u = \int_\omega \nabla u \cdot \nabla p \, dx,$$

with  $p$  defined by (3.19).

**Remark 3.3.** Again, note that this formula is the same for the case of a linear or a nonlinear state equation. Like in the case of Section 3.1, the only difference lies in the computation of the direct and adjoint states  $u$  and  $p$ , respectively, as problem (2.17b) becomes linear and the operator  $N_u$  vanishes.

### 3.3 Application to Model Problem

In this section, we apply the On/Off method described in Section 3.1 to the minimization problem (2.17) in the case of a nonlinear state equation (2.17b). The idea of the method

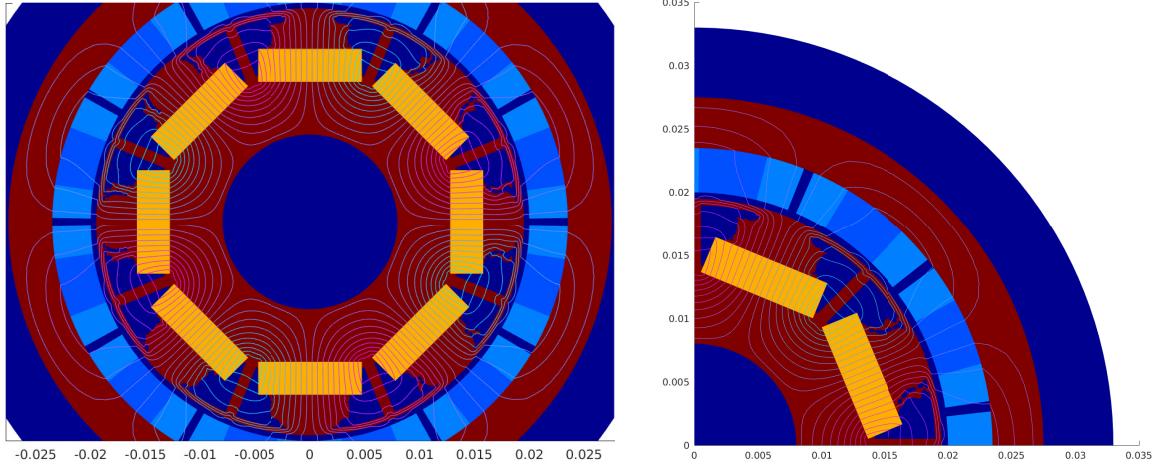


Figure 3.2: Final design after 30 iterations of Algorithm 1 with magnetic field.

is that, for elements occupied with ferromagnetic material where the sensitivity is negative, switching these elements to air (i.e., switching them off) will (most likely) yield a decrease of the objective function. Likewise, switching elements with positive sensitivity from air to ferromagnetic material (i.e., switching them on) should also decrease the objective function. However, it is important to note that the sensitivity in a triangle indicates how the objective function is affected if the material coefficient in this and only this element is modified. Thus, switching all elements to the state indicated by their sensitivity at the same time might even yield a deterioration of the objective value.

There are basically two ways to use the On/Off Sensitivities in an iterative optimization procedure. One approach, which is used in [162, 163] is to consider the absolute value of the sensitivities and to switch all elements whose sensitivity exceeds a certain threshold value (e.g.  $(1 - \gamma)$  times the maximum sensitivity for some  $\gamma \in (0, 1)$ ). This approach is also used in a similar way in [63, 80] using the topological derivative. Another approach, which we will follow in this Section, is to consider the element of the design region with the largest (in terms of absolute value) sensitivity and to switch all elements around this extremal element within a certain radius  $r$  which contain the same material as the extremal element. Both approaches involve a parameter ( $\gamma$  and  $r$ ) which influences the evolution of the design and should be chosen small enough such that an improvement of the objective function is achieved. Thus, we recommend to use a line search in this parameter.

The optimization procedure using this second approach is summarized in Algorithm 1:

**Algorithm 1.** Initialization: Set  $j = 0$ , choose initial design  $\Omega_0$  compute  $\mathcal{J}(\Omega_0)$  and On/Off Sensitivities (3.5). Set positive sensitivities in  $\Omega_f$  and negative sensitivities in  $\Omega_{air}$  to zero. Choose the parameter  $\bar{r}$ .

- (i) Determine element  $T_{ext}$  with extremal sensitivity
- (ii) Switch state of all elements within a radius  $r_j$  of the element  $T_{ext}$  which have the same state as  $T_{ext}$  and set  $\Omega_{j+1}$  the updated geometry, where  $r_j = \bar{r} \max\{1, 1/2, 1/4, \dots\}$  such that  $\mathcal{J}(\Omega_{j+1}) < \mathcal{J}(\Omega_j)$ .

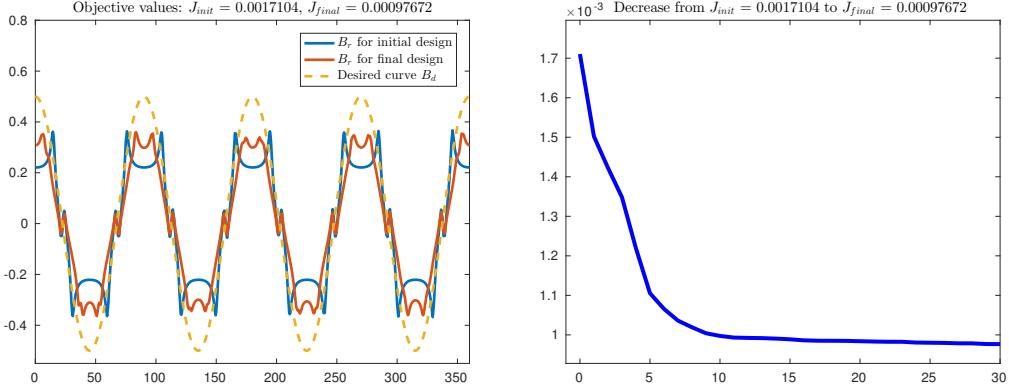


Figure 3.3: Left: Radial component of magnetic flux density for initial and final design together with desired curve. Right: Evolution of objective function (2.15) during the optimization procedure.

- (iii) Compute On/Off Sensitivities (3.5) for updated geometry  $\Omega_j$  and set positive sensitivities in  $\Omega_f$  and negative sensitivities in  $\Omega_{air}$  to zero.
- (iv) If On/Off Sensitivities are zero everywhere, stop, else set  $j \leftarrow j + 1$  and go to (i).

Note that for every evaluation of the objective function  $\mathcal{J}$ , the state equation (3.1) has to be solved, and for every evaluation of the On/Off sensitivities (3.5), in addition the adjoint state, i.e., the solution to (3.6), is needed. The parameter  $\bar{r}$  determines the maximum radius of the holes that are introduced and must be set by the user.

In our experiments, we chose  $\bar{r}$  to be four times the length of the smallest edge in the mesh of the design area. In order to decide whether an element is in a neighborhood of radius  $r_j$  of the extremal element, we considered the distance between the centroids of these triangles. The results of Algorithm 1 applied to our model problem (2.17) can be seen in Figures 3.1–3.2. Figure 3.1 shows the initial configuration of one of the eight parts of the design subdomain  $\Omega^d$ , as well as the On/Off sensitivity and the design after one iteration of the algorithm. In each iteration, we modified each of the eight subdomains in two positions due to symmetry. The final design after 33 iterations of Algorithm 1 can be seen in the last picture of Figure 3.1 as well as in Figure 3.2. We treated the eight parts of  $\Omega^d$  individually. The difference in the designs in the eight parts is probably due to the mesh which was not chosen as periodic. In Figure 3.2, we can see the final design along with the field lines of the magnetic flux density. Figure 3.3 shows the radial component of the B-field of the improved design compared to the initial design and the desired curve, as well as the decrease of the objective function in the course of the optimization procedure. Note that the desired curve  $B_d$  is not reachable, i.e., functional (2.15) cannot become zero, due to the definition of the design area  $\Omega^d$ .



## Chapter 4

# Topological Derivative for Magnetostatic Problem

The goal of this chapter is the rigorous derivation of the topological derivative for the design optimization problem (2.17) constrained by the nonlinear partial differential equation (PDE) of two-dimensional magnetostatics introduced in Section 2.3.

The *topological derivative* of a domain-dependent shape functional  $\mathcal{J} = \mathcal{J}(\Omega)$  indicates whether a perturbation of the domain (i.e., an introduction of a hole) around a spatial point  $x_0$  would lead to an increase or decrease of the objective functional. The idea of the topological derivative was first introduced for the compliance functional in linear elasticity in [78, 189] in the framework of the bubble method, where classical shape optimization methods are combined with the repeated introduction of holes (so-called “bubbles”) at optimal positions. The mathematical concept of the topological derivative was rigorously introduced in [195]. Given an open set  $\Omega \subset \mathbb{R}^d$  with  $d$  the space dimension, and a fixed bounded, smooth domain  $\omega$  containing the origin, the topological derivative of a shape functional  $\mathcal{J} = \mathcal{J}(\Omega)$  at a spatial point  $x_0$  is defined as the quantity  $G(x_0)$  satisfying a topological asymptotic expansion of the form

$$\mathcal{J}(\Omega_\varepsilon) - \mathcal{J}(\Omega) = f(\varepsilon) G(x_0) + o(f(\varepsilon)), \quad (4.1)$$

where  $\Omega_\varepsilon = \Omega \setminus \bar{\omega}_\varepsilon$  with  $\omega_\varepsilon = x_0 + \varepsilon \omega$  denotes the perturbed domain, and  $f$  is a positive first order correction function that vanishes with  $\varepsilon \rightarrow 0$ . We remark that, in the case where  $\mathcal{J}$  depends on the domain  $\Omega_\varepsilon$  via the solution of a boundary value problem on  $\Omega_\varepsilon$ , boundary conditions also have to be specified on the boundary of the hole, i.e., on  $\partial\omega_\varepsilon$ . Then, the choice of the boundary conditions on this boundary has a great influence on the resulting formula for the topological derivative  $G$ . In the original paper [195], the authors introduced the topological derivative concept with  $f(\varepsilon)$  being the volume of the ball of radius  $\varepsilon$  in  $\mathbb{R}^d$ . Later, in [143], the concept was generalized to the form (4.1) which also allowed to deal with Dirichlet boundary conditions on the boundary of the hole, see also [63, 159].

Topological asymptotic expansions of the form (4.1) have been derived for many different problems constrained by linear PDEs. We refer the interested reader to [13, 14, 22–24, 61, 65, 80, 144] as well as the monograph [159]. Besides the field of shape and topology optimization, topological derivatives are also used in applications from mathematical imaging, such as image segmentation [108] or electric impedance tomography [109, 133], or other geometric

inverse problems such as the detection of obstacles, of cracks or of impurities of a materials, see e.g. [65, 95] and the references therein.

In the context of magnetostatics, introducing a hole into a domain does not correspond to excluding this hole from the computational domain, but rather corresponds to the presence of an inclusion of a different material, namely air. Thus, in this scenario, both the perturbed and the unperturbed configuration live on the same domain  $\Omega$ , and only the material coefficient of the underlying PDE constraint is perturbed. Let  $u_\varepsilon$  and  $u_0$  denote the solutions to the perturbed and unperturbed state equation and  $\mathcal{J}_\varepsilon$  and  $\mathcal{J}_0$  the objective functionals defined on the perturbed and unperturbed configurations, respectively. Then, the asymptotic expansion corresponding to (4.1) reads

$$\mathcal{J}_\varepsilon(u_\varepsilon) - \mathcal{J}_0(u_0) = f(\varepsilon) G(x_0) + o(f(\varepsilon)), \quad (4.2)$$

where, again, the function  $f$  is positive and tends to zero with  $\varepsilon$ . The quantity  $G(x_0)$  is then often referred to as the *configurational derivative* of the shape functional  $\mathcal{J}$  at point  $x_0$ , see [159]. This sensitivity is analyzed for a class of linear PDE constraints in [14]. We remark that, in the limit case where the material coefficient inside the inclusion tends to zero, the classical topological derivative defined by (4.1) with homogeneous Neumann boundary conditions on the boundary of the hole is recovered, see [159, Remark 5.3]. In our case, the function  $f$  in (4.2) will be given by  $f(\varepsilon) = \varepsilon^d$  with  $d = 2$  the space dimension. Under a slight abuse of notation, we will refer to the configurational derivative defined by (4.2) as the topological derivative.

In this thesis, we derive the topological derivative for a design optimization problem that is constrained by the quasilinear equation of two-dimensional magnetostatics. As opposed to the linear case, only a few problems constrained by nonlinear PDEs have been studied in the literature. We mention the paper [158] where the topological derivative is estimated for the  $p$ -Poisson problem and the papers [15] and [117] for the topological asymptotic expansion in the case of a semilinear elliptic PDE constraint. In the recent work [20] which is based on [46], the authors considered a class of quasilinear PDEs and rigorously derived the topological derivative according to (4.2), which consists of two terms: a first term that resembles the topological derivative in the linear case, and a second term which accounts for the nonlinearity of the problem.

The quasilinear PDE we consider in this thesis does not exactly fit the framework of [20, 46]. However, it is similar and we will follow the steps taken there in order to derive the topological derivative for the electromagnetic design optimization problem (2.17) introduced in Section 2.3.

The rest of this chapter is organized as follows: We collect some mathematical preliminaries of our problem in Section 4.1 and show the main steps for the derivation of the topological derivative in the case of a linear PDE constraint in Section 4.2. In Section 4.3, we collect some assumptions and estimates on the nonlinearity of the problem and introduce suitable spaces which we will need later on. The rigorous derivation of the topological asymptotic expansion for two different cases is performed in Sections 4.4 and 4.5. In Section 4.4, we consider the case where an inclusion of air is introduced into a domain of ferromagnetic material, whereas Section 4.5 deals with the reverse scenario. We derive explicit formulas for the matrices occurring in these formulas, which are related to the concept of polarization

matrices, in Section 4.6, before considering computational aspects of the derived formulas, which ensure the practical applicability of the result, in Section 4.7. Finally, we apply the derived formulas to the optimization of a model electric motor in Section 4.8 by means of a level set algorithm.

## 4.1 Preliminaries

We aim at solving problem (2.17) by exploiting topological sensitivity information, or, more precisely, by means of the topological derivative introduced in (4.2). It is important to note that the topological derivative for introducing air in ferromagnetic material is different from that for introducing ferromagnetic material in a domain of air. Therefore, we distinguish between the following two cases:

1. **Case I:** An *inclusion of air* is introduced inside an area that is occupied with *ferromagnetic material*, see Figure 4.2 on page 58.
2. **Case II:** An *inclusion of ferromagnetic material* is introduced inside an area that is occupied with *air*, see Figure 4.3 on page 105.

In order to distinguish these two sensitivities, we denote the topological derivative in Case I by  $G^{f \rightarrow \text{air}}$  and in Case II by  $G^{\text{air} \rightarrow f}$ . It is important to have access to both these sensitivities for employing bidirectional optimization algorithms which are capable of both introducing and removing material at the most favorable positions. In Section 4.2, we will see that, in the case of a linear state equation (2.17b), the two sensitivities  $G^{f \rightarrow \text{air}}$  and  $G^{\text{air} \rightarrow f}$  differ only by a constant factor. In the more realistic case introduced in Section 2.3 however, where the nonlinear material behavior of ferromagnetic material is accounted for, the two topological derivatives must be derived individually. We will rigorously derive  $G^{f \rightarrow \text{air}}$  for Case I in Section 4.4. Most of the steps for deriving  $G^{\text{air} \rightarrow f}$  in Case II are similar to Case I. In Case II however, we have nonlinear material inside the inclusion and linear material outside the inclusion, which simplifies some of the steps. We will show the main steps for the derivation of  $G^{\text{air} \rightarrow f}$  in Section 4.5.

Throughout this chapter, for sake of more compact presentation, we will drop the differential  $dx$  in the volume integrals whenever there is no danger of confusion.

### 4.1.1 Simplified Model Problem

In order to alleviate some calculations, we introduce a simplified model of the PDE constraint (2.17b). The model we introduce here, is meant for Case I. The analogous simplified model for Case II will be introduced in the beginning of Section 4.5.

The simplification consists in the fact that, in the unperturbed configuration, we assume the material coefficient  $\nu$  to be homogeneous in the entire bounded domain  $D$ . In the notation of Section 2.3, we assume that  $\Omega_f^{ref} = D$  and, in the unperturbed case,  $\Omega = \Omega^d$ . Then, the unperturbed state equation (2.17b) simplifies to

$$\text{Find } u_0 \in H_0^1(D) \text{ such that } \int_D T(\nabla u_0) \cdot \nabla \eta = \langle F, \eta \rangle \quad \forall \eta \in H_0^1(D), \quad (4.3)$$

as can be seen from the definitions of the operator  $A_\Omega$  (2.12) and the reluctivity function  $\nu_\Omega$  (2.11), as well as the definition of the operator  $T$  (2.24). Here,  $F \in H^{-1}(D)$  is as in (2.13) and represents the sources due to the electric currents in the coil areas of the motor and the permanent magnetization in the magnets.

We will assume this simplified setting for the rest of this chapter and derive the formula for the topological derivative under these assumptions.

**Remark 4.1.** *The reason why we have to make this simplification will come clear in the proofs of Proposition 4.27 and Lemma 4.35. We remark that the topological derivative denotes the sensitivity of the objective function with respect to a perturbation inside an inclusion whose radius tends to zero. Therefore, the material coefficients “far away” from the point of perturbation, e.g., outside the design subdomain  $\Omega^d$  when the point of perturbation is inside  $\Omega^d$ , should not influence the formula for the sensitivity and it is justified to use the same formula also for the realistic setting introduced in Section 2.3. Note that, for all numerical computations, the realistic state equation (2.17b) and adjoint equation (2.31) were solved.*

#### 4.1.2 Perturbed State Equation

We are interested in the sensitivity of the objective functional  $\mathcal{J}$  with respect to a local perturbation of the material coefficient around a fixed point  $x_0$  in the design subdomain  $\Omega^d$ . For that purpose, we introduce a perturbed version of the simplified state equation (4.3).

Let  $\text{supp}(F)$  denote the support of the distribution  $F \in H^{-1}(D)$ , i.e., the complement in  $D$  of the largest open set where  $F$  vanishes, see [113]. We assume that  $\text{supp}(F)$  is compactly contained in  $D$ ,  $\text{supp}(F) \subset\subset D$ , and that the design subdomain  $\Omega^d$  is open and compactly contained in  $D \setminus \text{supp}(F)$ ,

$$\Omega^d \subset\subset D \setminus \text{supp}(F).$$

Let  $x_0$  be a fixed point in  $\Omega^d$ . Let furthermore  $\omega \subset \mathbb{R}^2$  be a bounded open domain with  $C^2$  boundary which contains the origin, and let  $\omega_\varepsilon = x_0 + \varepsilon \omega$  represent the inclusion of different material in the physical domain. For simplicity and without loss of generality, we assume that  $x_0 = (0, 0)^\top$ . Furthermore, let  $0 < \rho < R$  and  $\lambda \in (0, 1]$  such that

$$\lambda \omega \subset\subset B(0, \rho) \subset B(0, R) \subset\subset D \setminus \text{supp}(F). \quad (4.4)$$

Note that such a choice of  $\lambda, \rho, R$  is always possible if  $x_0 \in \Omega^d$ .

Recall the notation introduced in Section 2.3. In the perturbed configuration, the inclusion  $\omega_\varepsilon$  of radius  $\varepsilon$  is occupied by air. Therefore, we have  $\Omega = \Omega^d \setminus \omega_\varepsilon$ , and, according to (2.8),  $\Omega_f = \Omega_f^{ref} \setminus \Omega^d \cup \Omega = D \setminus \omega_\varepsilon$  the set of points occupied by ferromagnetic material, see Figure 4.2. Here we used that, in the simplified setting introduced in Section 4.1.1,  $\Omega_f^{ref} = D$ . We define the operator

$$T_\varepsilon(x, W) := \chi_{D \setminus \omega_\varepsilon}(x) T(W) + \chi_{\omega_\varepsilon}(x) \nu_0 W, \quad (4.5)$$

for  $\varepsilon > 0$ ,  $x \in D$  and  $W \in \mathbb{R}^2$  with its Jacobian given by

$$DT_\varepsilon(x, W) = \chi_{D \setminus \omega_\varepsilon}(x) DT(W) + \chi_{\omega_\varepsilon}(x) \nu_0 I.$$

Note that, given the special setting introduced above,  $T_\varepsilon(x, W) = T_{\Omega^d \setminus \omega_\varepsilon}(x, W)$  according to (2.26).

Thus, in the simplified setting introduced in Section 4.1.1, the perturbed state equation reads

$$\text{Find } u_\varepsilon \in H_0^1(D) \text{ such that } \int_D T_\varepsilon(x, \nabla u_\varepsilon) \cdot \nabla \eta = \langle F, \eta \rangle \quad \forall \eta \in H_0^1(D). \quad (4.6)$$

**Remark 4.2.** *In this thesis, the rigorous derivation of the topological derivative will be carried out for the case where  $\omega$  is the unit disk in  $\mathbb{R}^2$ ,  $\omega = B(0, 1)$ . In the linear case of Section 4.2, a generalization to ellipse-shaped inclusions is possible, see Remark 4.73 in Section 4.6. Concerning the nonlinear case, which is treated in the rest of this chapter, the possibility of an extension of the results to more general shapes of  $\omega$  is discussed in Remark 4.45.*

#### 4.1.3 Expansion of Cost Functional

We assume that the functional to be minimized is of the following form:

**Assumption 2.** For  $\varepsilon \geq 0$  small enough, let  $\mathcal{J}_\varepsilon : H_0^1(D) \rightarrow \mathbb{R}$  such that

$$\mathcal{J}_\varepsilon(u_\varepsilon) = \mathcal{J}_0(u_0) + \langle \tilde{G}, u_\varepsilon - u_0 \rangle + \delta_J \varepsilon^2 + R(\varepsilon) \quad (4.7)$$

where

1.  $\tilde{G}$  denotes a bounded linear form on  $H_0^1(D)$
2.  $\delta_J \in \mathbb{R}$
3. the remainder  $R(\varepsilon)$  is of the form

$$R(\varepsilon) = O \left( \int_{D \setminus B(0, \hat{\alpha} \varepsilon \tilde{r})} |\nabla(u_\varepsilon - u_0)|^2 \right) \quad (4.8)$$

for a given  $\hat{\alpha} > 0$  and a given  $\tilde{r} \in (0, 1)$ .

In the case of electric motors, the objective function  $\mathcal{J}$  is generally supported only in the air gap  $\Omega_g \subset \Omega_{air}^{ref}$ . Therefore, a perturbation of the material coefficient inside the design domain  $\Omega^d$  will only affect the functional via the solution  $u$  of the PDE constraint, but not in a direct way. For that reason, we assume that the functional for the perturbed and the unperturbed configuration coincide, i.e.,  $\mathcal{J}_\varepsilon = \mathcal{J}_0 = \mathcal{J}$  in the expansion (4.2), for the rest of this chapter.

**Remark 4.3.** *Functional (2.15) in our model problem does not depend explicitly on  $\varepsilon$ , i.e.,  $\mathcal{J}_\varepsilon = \mathcal{J}_0 = \mathcal{J}$ , and fulfills an expansion of the form (4.7) with  $\tilde{G} = D\mathcal{J}(u_0)$ ,  $\delta_J = 0$ :*

$$\begin{aligned} \mathcal{J}(u_\varepsilon) - \mathcal{J}(u_0) &= 2 \int_{\Omega_g} (\nabla u_\varepsilon \cdot \tau_g - \nabla u_0 \cdot \tau_g)(\nabla u_0 \cdot \tau_g - B^d) + \int_{\Omega_g} |\nabla u_\varepsilon \cdot \tau_g - \nabla u_0 \cdot \tau_g|^2 \\ &= \langle D\mathcal{J}(u_0), \nabla u_\varepsilon - \nabla u_0 \rangle + \int_{\Omega_g} |\nabla u_\varepsilon \cdot \tau_g - \nabla u_0 \cdot \tau_g|^2. \end{aligned}$$

Since  $\mathcal{J}$  is integrated over the air gap region  $\Omega_g$  which is not part of the design subdomains  $\Omega^d$ , see Figure 2.2, it can be seen that the remainder  $R(\varepsilon) = \int_{\Omega_g} |\nabla u_\varepsilon \cdot \tau_g - \nabla u_0 \cdot \tau_g|^2$  is of the form (4.8).

## 4.2 Linear Case

This section is based on [14, 46, 88]. In this section, we illustrate the procedure of deriving the topological derivative according to (4.2) for the simpler case of a linear state equation. For this purpose, we consider linear material behavior in both materials, i.e., we replace the nonlinear reluctivity function  $\hat{\nu}$  by a constant reluctivity value  $\nu_1$  satisfying  $0 < \nu_1 < \nu_0$ . Note that also ferromagnetic materials exhibit linear behavior in regions of low magnetic flux density, see Figure 2.4. However, we remark that, in most applications involving electrical machines, this simplified linear model is not a realistic model. For  $\varepsilon \geq 0$  small enough, we introduce the piecewise constant reluctivity function  $\nu_\varepsilon : D \rightarrow \mathbb{R}$  defined by

$$\nu_\varepsilon(x) = \chi_{D \setminus \omega_\varepsilon}(x) \nu_1 + \chi_{\omega_\varepsilon}(x) \nu_0,$$

and, for  $u, \eta \in H_0^1(D)$ , the bilinear form

$$a_\varepsilon(u, \eta) = \int_D \nu_\varepsilon \nabla u \cdot \nabla \eta, \quad x \in D.$$

Then, the state equation in the perturbed ( $\varepsilon > 0$ ) or unperturbed setting ( $\varepsilon = 0$ ) reads,

$$\text{Find } u_\varepsilon \in H_0^1(D) \text{ such that } a_\varepsilon(u_\varepsilon, \eta) = \langle F, \eta \rangle \quad \forall \eta \in H_0^1(D),$$

with  $F \in H^{-1}(D)$  representing the currents and the magnetization. With this setting, it is possible to apply the following result from [14] in order to obtain the topological derivative of a given cost functional.

**Proposition 4.4** ([14]). *Let  $V$  be a real Hilbert space. For all parameters  $\varepsilon \in [0, \varepsilon_0)$ ,  $\varepsilon_0 > 0$ , consider a function  $u_\varepsilon \in V$  solving a variational problem of the form*

$$a_\varepsilon(u_\varepsilon, \eta) = l_\varepsilon(\eta) \quad \forall \eta \in V, \tag{4.9}$$

where  $a_\varepsilon$  and  $l_\varepsilon$  are a bilinear and a linear form on  $V$ , respectively. Consider a cost function

$$j(\varepsilon) = J_\varepsilon(u_\varepsilon).$$

Suppose that the following hypotheses hold:

1. There exists  $\tilde{G} \in V^*$  and a number  $\delta_J$  such that

$$J_\varepsilon(u_\varepsilon) = J_0(u_0) + \langle \tilde{G}, u_\varepsilon - u_0 \rangle + f(\varepsilon) \delta_J + o(f(\varepsilon)), \tag{4.10}$$

2. There exist two numbers  $\delta a$  and  $\delta l$  and a function  $f(\varepsilon) \geq 0$  such that, when  $\varepsilon$  goes to zero,

$$(a_\varepsilon - a_0)(u_0, p_\varepsilon) = f(\varepsilon) \delta a + o(f(\varepsilon)), \tag{4.11}$$

$$(l_\varepsilon - l_0)(p_\varepsilon) = f(\varepsilon) \delta l + o(f(\varepsilon)), \tag{4.12}$$

$$\lim_{\varepsilon \rightarrow 0} f(\varepsilon) = 0,$$

where  $p_\varepsilon \in V$  is an adjoint state satisfying

$$a_\varepsilon(\varphi, p_\varepsilon) = -\langle \tilde{G}, \varphi \rangle \quad \forall \varphi \in V. \tag{4.13}$$

Then the first variation of the cost function with respect to  $\varepsilon$  is given by

$$j(\varepsilon) - j(0) = f(\varepsilon)(\delta a - \delta l + \delta_J) + o(f(\varepsilon)).$$

*Proof.* Due to (4.9), we have

$$j(\varepsilon) - j(0) = [J_\varepsilon(u_\varepsilon) - J_0(u_0)] + [a_\varepsilon(u_\varepsilon, p_\varepsilon) - a_0(u_0, p_\varepsilon)] - [l_\varepsilon(p_\varepsilon) - l_0(p_\varepsilon)].$$

Using (4.11) and (4.12), we get

$$j(\varepsilon) - j(0) = J_\varepsilon(u_\varepsilon) - J_0(u_0) + a_\varepsilon(u_\varepsilon - u_0, p_\varepsilon) + f(\varepsilon)(\delta a - \delta l) + o(f(\varepsilon)).$$

It follows from (4.10) that

$$j(\varepsilon) - j(0) = \langle \tilde{G}, u_\varepsilon - u_0 \rangle + a_\varepsilon(u_\varepsilon - u_0, p_\varepsilon) + f(\varepsilon)(\delta a - \delta l + \delta_J) + o(f(\varepsilon)).$$

The adjoint equation (4.13) yields the claimed result.  $\square$

### 4.2.1 Application to the Model Problem

In this section, we apply Proposition 4.4 to the simplified model problem introduced in Section 4.1.1 in the linear setting introduced in the beginning of Section 4.2. We show the main steps for deriving the variations  $\delta a$ ,  $\delta l$ ,  $\delta J_1$ ,  $\delta J_2$ . The derivation of  $\delta a$  will not be done in detail, but we rather just outline the main steps. Starting in Section 4.3, we will perform the rigorous derivation for the topological derivative in the case of a nonlinear reluctivity function  $\hat{\nu}$  in one of the subdomains, which, of course, also covers the linear case by setting  $\hat{\nu} = \nu_1 = \text{const} > 0$ . The rigorous derivation of the variation of the bilinear form  $\delta a$  specifically for the linear case can be found in [14].

#### 4.2.1.1 Variation of the Linear Form

In Section 4.1.2, we assumed for the point  $x_0$  around which we perturb the material that  $x_0 \in \Omega^d \subset\subset D \setminus \text{supp}(F)$ , such that, for all  $\eta \in H_0^1(D)$ ,

$$l_\varepsilon(\eta) = l_0(\eta) = \langle F, \eta \rangle,$$

and therefore relation (4.12) trivially holds with  $\delta l = 0$ .

#### 4.2.1.2 Variation of the Cost Function

By Assumption 2, we have

$$\mathcal{J}_\varepsilon(u_\varepsilon) = \mathcal{J}_0(u_0) + \langle \tilde{G}, u_\varepsilon - u_0 \rangle + \delta_J \varepsilon^2 + O\left(\int_{D \setminus B(0, \hat{\alpha} \varepsilon \tilde{r})} |\nabla(u_\varepsilon - u_0)|^2\right),$$

for some  $\hat{\alpha} > 0$  and  $\tilde{r} \in (0, 1)$ . As we will see later by combining (4.68) and (4.71), it holds that  $\|u_\varepsilon - u_0\|_{H^1(D \setminus B(0, \hat{\alpha} \varepsilon \tilde{r}))}^2 = o(\varepsilon^2)$  for all  $\hat{\alpha} > 0$  and  $\tilde{r} \in (0, 1)$ . Thus we get (4.10) with  $\delta_J$  as in (4.7) and  $f(\varepsilon) = \varepsilon^2$ . In particular, for a functional of type (2.15) which only depends on the values of the state variable outside the design region,

$$\mathcal{J}_\varepsilon(u) = \mathcal{J}_0(u) = \mathcal{J}(u|_{D \setminus \Omega^d}),$$

we get that  $\delta_J = 0$ , see Remark 4.3.

### 4.2.1.3 Variation of the Bilinear Form

The rigorous derivation of a variation  $\delta a$  of the bilinear form is more involved and we will only sketch the main steps in this section.

First, we introduce the boundary value problem defining the variation of the direct state  $\tilde{u}_\varepsilon := u_\varepsilon - u_0$ . The problem reads

Find  $\tilde{u}_\varepsilon \in H_0^1(D)$  such that

$$\int_D \nu_\varepsilon \nabla \tilde{u}_\varepsilon \cdot \nabla \eta + (\nu_0 - \nu_1) \int_{\omega_\varepsilon} \nabla u_0 \cdot \nabla \eta = 0 \quad \forall \eta \in H_0^1(D). \quad (4.14)$$

This problem is derived from (4.9) by subtracting the variational equation for  $\varepsilon = 0$  from the same equation with  $\varepsilon > 0$ , noting that  $l_\varepsilon = F$  is independent of  $\varepsilon$ .

Analogously, we deduce the boundary value problem defining the variation of the adjoint state,  $\tilde{p}_\varepsilon := p_\varepsilon - p_0$ . From (4.13), we get

Find  $\tilde{p}_\varepsilon \in H_0^1(D)$  such that

$$\int_D \nu_\varepsilon \nabla \eta \cdot \nabla \tilde{p}_\varepsilon + (\nu_0 - \nu_1) \int_{\omega_\varepsilon} \nabla p_0 \cdot \nabla \eta = 0 \quad \forall \eta \in H_0^1(D). \quad (4.15)$$

Then, by plugging in the definition and using (4.14) with  $\eta = \tilde{p}_\varepsilon$ , we get

$$\begin{aligned} (a_\varepsilon - a_0)(u_0, p_\varepsilon) &= (\nu_0 - \nu_1) \int_{\omega_\varepsilon} \nabla u_0 \cdot \nabla p_\varepsilon \\ &= (\nu_0 - \nu_1) \int_{\omega_\varepsilon} \nabla u_0 \cdot \nabla \tilde{p}_\varepsilon + (\nu_0 - \nu_1) \int_{\omega_\varepsilon} \nabla u_0 \cdot \nabla p_0 \\ &= - \int_D \nu_\varepsilon \nabla \tilde{u}_\varepsilon \cdot \nabla \tilde{p}_\varepsilon + (\nu_0 - \nu_1) \int_{\omega_\varepsilon} \nabla u_0 \cdot \nabla p_0. \end{aligned} \quad (4.16)$$

Assuming enough regularity for the unperturbed direct and adjoint state, it can be seen that, for the second term in (4.16), we have

$$(\nu_0 - \nu_1) \int_{\omega_\varepsilon} \nabla u_0 \cdot \nabla p_0 = |\omega| \varepsilon^2 (\nu_0 - \nu_1) \nabla u_0(x_0) \cdot \nabla p_0(x_0) + o(\varepsilon^2) \quad (4.17)$$

as  $\varepsilon$  approaches zero.

In order to treat the first term in (4.16), we define  $\tilde{\nu}(x) = \chi_{\mathbb{R}^2 \setminus \omega}(x) \nu_1 + \chi_\omega \nu_0$  for  $x \in \mathbb{R}^2$ , and introduce  $\varepsilon$ -independent approximations to boundary value problems (4.14) and (4.15). After a change of scale, we get the transmission problem defining the variation of the direct state at scale 1,

Find  $H \in \mathcal{H}$  such that

$$\int_{\mathbb{R}^2} \tilde{\nu} \nabla H \cdot \nabla \eta + (\nu_0 - \nu_1) \int_{\omega} \nabla u_0(x_0) \cdot \nabla \eta = 0 \quad \forall \eta \in \mathcal{H}, \quad (4.18)$$

approximating (4.14), and the problem defining the variation of the adjoint state at scale 1,

Find  $K \in \mathcal{H}$  such that

$$\int_{\mathbb{R}^2} \tilde{\nu} \nabla \eta \cdot \nabla K + (\nu_0 - \nu_1) \int_{\omega} \nabla p_0(x_0) \cdot \nabla \eta = 0 \quad \forall \eta \in \mathcal{H}, \quad (4.19)$$

as an approximation of (4.15), where  $\mathcal{H}$  is a suitable Hilbert space which will be defined in Section 4.3.4. The solutions  $H, K$  are approximations to  $\tilde{u}_\varepsilon$  and  $\tilde{p}_\varepsilon$ , respectively, at scale 1 and it holds

$$\tilde{u}_\varepsilon(x) \approx \varepsilon H(\varepsilon^{-1}x) \quad \text{and} \quad \tilde{p}_\varepsilon(x) \approx \varepsilon K(\varepsilon^{-1}x),$$

for almost every  $x \in D$ . An important ingredient for deriving the variation of the bilinear form (4.11) is to show that these  $\varepsilon$ -independent approximations of  $\tilde{u}_\varepsilon$  and  $\tilde{p}_\varepsilon$  have a sufficiently fast decay as  $|x|$  approaches infinity. This would imply that the impact of the local variation of the material is small “far away” from the inclusion. In the case of a linear state equation, this sufficiently fast decay can be established by convolution of the right hand side of problems (4.18) and (4.19) with the fundamental solution of the Laplace equation. Exploiting these sufficiently fast decays will allow us to show that

$$\int_D \nu_\varepsilon \nabla \tilde{u}_\varepsilon \cdot \nabla \tilde{p}_\varepsilon = \varepsilon^2 \int_{\mathbb{R}^2} \tilde{\nu} \nabla H \cdot \nabla K + o(\varepsilon^2),$$

which, by means of (4.18) tested with  $\eta = K$ , together with the term (4.17), yields (4.11) with

$$\delta a = (\nu_0 - \nu_1) \int_\omega \nabla u_0(x_0) \cdot (\nabla K + \nabla p_0(x_0)).$$

It can be seen from (4.19) that  $K$  depends linearly on  $\nabla p_0(x_0)$  and, therefore,  $\delta a$  can be represented by means of a matrix  $\mathcal{M}$ . Finally, we get

$$\begin{aligned} \delta a &= \nabla u_0(x_0)^\top \mathcal{M} \nabla p_0(x_0), \\ f(\varepsilon) &= \varepsilon^2. \end{aligned} \tag{4.20}$$

Here,  $\mathcal{M} = \nu_1 \mathcal{P}(\omega, \nu_0/\nu_1)$  where the matrix  $\mathcal{P}(\omega, \nu_0/\nu_1)$  only depends on the shape of the inclusion  $\omega$  and the contrast  $\nu_0/\nu_1$  and is called a polarization matrix, see, e.g., [11]. Explicit formulas for these matrices are available if  $\omega$  is a disk or ellipse in two space dimensions, or a ball or ellipsoid in three space dimensions, see also [11, 14, 65] as well as Section 4.6 of this thesis. We mention that in the case where  $\omega$  is the unit disk in  $\mathbb{R}^2$ , the polarization matrix in the linear setting reads

$$\mathcal{P}_{\omega, \nu_0/\nu_1} = 2 \frac{\nu_0/\nu_1 - 1}{\nu_0/\nu_1 + 1} |\omega| I, \tag{4.21}$$

where  $I$  is the identity matrix.

#### 4.2.1.4 Summary

Summarizing, by applying Proposition 4.4, we have found the topological asymptotic expansion

$$\mathcal{J}_\varepsilon(u_\varepsilon) - \mathcal{J}_0(u_0) = \varepsilon^2 \left( \nu_1 \nabla u_0(x_0)^\top \mathcal{P}_{\omega, \nu_0/\nu_1} \nabla p_0(x_0) \right) + o(\varepsilon^2).$$

Choosing  $\omega = B(0, 1)$  as the unit disk, it follows from (4.21) that the topological derivative at a point  $x_0$  for introducing air with material coefficient  $\nu_0$  inside another material with coefficient  $\nu_1$  reads

$$G^{f \rightarrow \text{air}}(x_0) = 2\nu_1 \frac{\nu_0/\nu_1 - 1}{\nu_0/\nu_1 + 1} \pi \nabla u_0(x_0) \cdot \nabla p_0(x_0). \tag{4.22}$$

**Remark 4.5.** Of course, in the reverse scenario of introducing material with coefficient  $\nu_1$  inside an area of air with coefficient  $\nu_0$ , the topological derivative is obtained in the exact same way and we get the corresponding formula by just exchanging the values of  $\nu_0$  and  $\nu_1$  in (4.22):

$$G^{air \rightarrow f}(x_0) = 2\nu_0 \frac{\nu_1/\nu_0 - 1}{\nu_1/\nu_0 + 1} \pi \nabla u_0(x_0) \cdot \nabla p_0(x_0). \quad (4.23)$$

Note that this equivalence is not given in the case of nonlinear material behavior in one of the two subdomains. Then, the two cases have to be considered individually, see Section 4.4 for the case of introducing air inside ferromagnetic material and Section 4.5 for the reverse scenario, as well as Chapter 5.

### 4.3 Preliminaries for the Nonlinear Case

This section serves as a preparation for Section 4.4 where the rigorous derivation of the topological derivative in the case of nonlinear ferromagnetic material in one of the subdomains is performed. Here, we collect some requirements and properties of the nonlinear function  $\hat{\nu}$ , i.e., the magnetic reluctivity function in the ferromagnetic subdomain, which was replaced by the constant  $\nu_1$  in the simplified setting of Section 4.2. We remark that the nonlinear case of Section 4.4 is an extension of the linear case, for which we sketched the procedure in Section 4.2. This means that all steps taken in Section 4.2 also occur in Section 4.4. In addition, we will have to deal with another term which accounts for the nonlinearity of the function  $\hat{\nu}$ .

Recall the operator  $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2 : W \mapsto \hat{\nu}(|W|)W$  introduced in (2.24) and its Jacobian  $DT$  (2.25). For  $V, W \in \mathbb{R}^2$ , we introduce the operator

$$S_W(V) := T(W + V) - T(W) - DT(W)V. \quad (4.24)$$

For  $\varepsilon > 0$  and  $\omega_\varepsilon = x_0 + \varepsilon\omega$  as in Section 4.1.2, define

$$S_W^\varepsilon(x, V) := \chi_{D \setminus \omega_\varepsilon}(x) S_W(V), \quad (4.25)$$

Note that, if  $\hat{\nu}$  is a constant as it was the case in Section 4.2, the operator  $T$  is linear and thus  $S_W(V)$  and  $S_W^\varepsilon(x, V)$  vanish for all  $V, W \in \mathbb{R}^2$ . The operator defined in (4.24) can be seen as a characterization of the nonlinearity of the problem.

Furthermore, for  $\varepsilon > 0$ , we define the scaled version of the domain  $D$  as

$$D/\varepsilon = \{y = x/\varepsilon | x \in D\}.$$

#### 4.3.1 Summary of the Procedure

We give a brief overview over the main steps taken in the rest of this chapter.

We are interested in a topological asymptotic expansion of the form (4.2). By Assumption 2, this reduces to showing that

$$\langle \tilde{G}, u_\varepsilon - u_0 \rangle = \varepsilon^2 G(x_0) + o(\varepsilon^2) \quad \text{and} \quad R(\varepsilon) = o(\varepsilon^2).$$

with the remainder  $R(\varepsilon)$  of the form (4.8), where we chose  $f(\varepsilon) = \varepsilon^2$ . In order to show these relations, we investigate in detail the difference  $u_\varepsilon - u_0$ , called the variation of the direct

state. We introduce an approximation of this variation which is independent of the small parameter  $\varepsilon$ . This approximation, which we will denote by  $H$ , is the solution to a transmission problem on the entire plane  $\mathbb{R}^2$  and is an element of a weighted Sobolev space as introduced in Section 4.3.4. We establish relations between this approximation  $H$  and the variation  $u_\varepsilon - u_0$  on the domain  $D$ . A very important ingredient for this is to show that  $H$  satisfies a sufficiently fast decay towards infinity, meaning that this approximation to the difference between the perturbed and unperturbed state is small “far away” from the inclusion. This result, which is rather technical, is obtained in Theorem 4.24. All of these steps are shown in detail in Section 4.4.1.

Similar results are needed for the variation of the adjoint state  $p_\varepsilon - p_0$  which is approximated by the function  $K$ . Again, a sufficiently fast decay of  $K$  towards infinity is important. We remark that, also in the case of a nonlinear state equation, the boundary value problem defining the adjoint state is always linear. Therefore, the treatment of the variation of the adjoint state is less technical. These steps are carried out in Section 4.4.2.

Given the relations of Sections 4.4.1 and 4.4.2, a topological asymptotic expansion of the form (4.2) is shown in Section 4.4.3. The topological derivative  $G$  consists of two terms. The first term is very similar to (4.20) involving a matrix  $\mathcal{M}$  the explicit form of which is derived in Section 4.6. The second term is specific for the nonlinear case and its numerical computation is treated in Section 4.7.

Finally we apply the obtained formula to the model problem of Section 2.3 in Section 4.8 using a level set algorithm which is based on topological sensitivity information.

### 4.3.2 Requirements

In addition to Assumption 1, which yields the physical properties (2.19a) and (2.19b), we have to make further assumptions on the nonlinear function  $\hat{\nu}$  representing the magnetic reluctivity in the ferromagnetic subdomains.

**Assumption 3.** We assume that the nonlinear magnetic reluctivity function  $\hat{\nu}$  satisfies the following:

1.  $\hat{\nu} \in C^3(\mathbb{R}_0^+)$ .
2. There exists  $\tilde{c} > 0$  such that  $\frac{|\hat{\nu}'(s)|}{s} \leq \tilde{c}$  for all  $s \geq 0$ .
3. There exist non-negative constants  $\tilde{c}_4, \tilde{c}_5, \tilde{c}', \tilde{c}''$  such that it holds

$$\left| (f^{-1})''(s) \right| \leq \tilde{c}_4, \quad \forall s \geq 0, \quad (4.26)$$

$$\left| (f^{-1})'''(s) \right| \leq \tilde{c}_5, \quad \forall s \geq 0, \quad (4.27)$$

$$|\hat{\nu}'(s)| \leq \tilde{c}', \quad \forall s \geq 0, \quad (4.28)$$

$$|\hat{\nu}''(s)| \leq \tilde{c}'', \quad \forall s \geq 0, \quad (4.29)$$

where  $(f^{-1})''(s) = 2\hat{\nu}'(s) + \hat{\nu}''(s)s$  and  $(f^{-1})'''(s) = 3\hat{\nu}''(s) + \hat{\nu}'''(s)s$  denote the second and third derivative of the inverse of the  $B$ - $H$ -curve,  $f^{-1}(s) = \hat{\nu}(s)s$ , see (2.18), respectively.

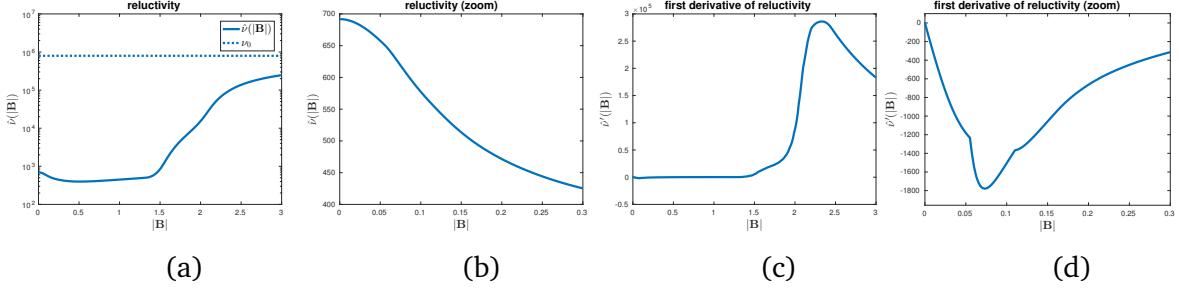


Figure 4.1: (a) Magnetic reluctivity of ferromagnetic subdomain  $\hat{\nu}$  in semilogarithmic scale. (b) Zoom of reluctivity  $\hat{\nu}$ . (c) First derivative of magnetic reluctivity,  $\hat{\nu}'$ . (d) Zoom of first derivative of magnetic reluctivity,  $\hat{\nu}'$ .

**Assumption 4.** Let  $\delta_{\hat{\nu}} := \inf_{s>0} (\hat{\nu}'(s)s)/\hat{\nu}(s)$ . We assume that

$$\delta_{\hat{\nu}} > \max \left\{ \delta_{\hat{\nu}}^{R_1}, \delta_{\hat{\nu}}^{R_2} \right\}$$

where  $\delta_{\hat{\nu}}^{R_1} = -1/3$  and  $\delta_{\hat{\nu}}^{R_2} = -\frac{(1+k_1)^2}{(1+k_1)^2+2}$  with  $k_1 = (\underline{\nu} - \nu_0)/\nu_0$ .

Note that the first assumption implies that  $\hat{\nu}$  is Lipschitz continuous on  $[0, \infty)$  and we denote the Lipschitz constant by  $L_{\hat{\nu}}$ . According to Lemma 2.3, the reluctivity function  $\hat{\nu}$  is once continuously differentiable. However, in our asymptotic analysis, we will make use of derivatives of order up to three and thus assume  $\hat{\nu} \in C^3(\mathbb{R}_0^+)$ . This assumption is realistic in practice, since the function  $\hat{\nu}$  is not known explicitly but only approximated from measured data by smooth functions, see [169]. Assumption 3.2 does not automatically follow from physical properties, but it is satisfied for the (realistic) set of data we used for all of the numerical computations, see Figure 4.1. Note that Assumption 3.2 implies that  $\hat{\nu}'(0) = 0$ .

Assumption 4 is needed to show Propositions 4.21 and 4.22 which will then yield the asymptotic behavior of the variation of the direct state at scale 1, see Theorem 4.24. Due to the big contrast between maximum and minimum value of the magnetic reluctivity, see Figure 4.1(a), the value of  $\delta_{\hat{\nu}}^{R_2} < 0$  is very close to zero such that, in order for  $\hat{\nu}$  to fulfill Assumption 4, the function  $\hat{\nu}$  would have to be almost monotone. This would rule out a big class of reluctivity functions including the one used throughout this thesis where  $\delta_{\hat{\nu}} = -0.3091$ . However, we remark that Assumption 4 is only a sufficient condition for the result of Theorem 4.24 and it may very well be possible to show the result with weaker assumptions on  $\delta_{\hat{\nu}}$ . The relaxation of Assumption 4 is subject of future investigation.

In the Section 4.3.3, we will make use of the following estimates:

**Lemma 4.6.** *Let Assumption 3 hold. Then there exist constants  $c_4, c_5$  such that, for all  $\varphi \in \mathbb{R}^2$  the following estimates hold:*

$$4|\hat{\nu}'(|\varphi|)| + |\hat{\nu}''(|\varphi|)| |\varphi| \leq c_4, \quad (4.30)$$

$$|\hat{\nu}'''(|\varphi|)| |\varphi| + 9|\hat{\nu}''(|\varphi|)| + 12 \frac{|\hat{\nu}'(|\varphi|)|}{|\varphi|} \leq c_5. \quad (4.31)$$

*Proof.* Estimate (4.30) can be easily seen using (4.26) and (4.28):

$$\begin{aligned} 4|\hat{\nu}'(|\varphi|)| + |\hat{\nu}''(|\varphi|)| |\varphi| &= 4|\hat{\nu}'(|\varphi|)| + |\hat{\nu}''(|\varphi|)| |\varphi| \\ &= 4|\hat{\nu}'(|\varphi|)| + |\hat{\nu}''(|\varphi|)| |\varphi| + 2\hat{\nu}'(|\varphi|) - 2\hat{\nu}'(|\varphi|)| \\ &\leq 4|\hat{\nu}'(|\varphi|)| + |\hat{\nu}''(|\varphi|)| |\varphi| + 2\hat{\nu}'(|\varphi|)| + 2|\hat{\nu}'(|\varphi|)| \\ &\leq \tilde{c}_4 + 6\tilde{c}' =: c_4. \end{aligned}$$

Similarly, from (4.27) and (4.29), it follows that

$$\begin{aligned} |\hat{\nu}'''(|\varphi|)| |\varphi| &= |\hat{\nu}'''(|\varphi|)| |\varphi| \\ &= |\hat{\nu}'''(|\varphi|)| |\varphi| + 3\hat{\nu}''(|\varphi|) - 3\hat{\nu}''(|\varphi|)| \\ &\leq |\hat{\nu}'''(|\varphi|)| |\varphi| + 3\hat{\nu}''(|\varphi|)| + 3|\hat{\nu}''(|\varphi|)| \\ &\leq \tilde{c}_5 + 3\tilde{c}'', \end{aligned}$$

which yields estimate (4.31) with  $c_5 := \tilde{c}_5 + 12\tilde{c}'' + 12\tilde{c}$  by means of estimate (4.29) and Assumption 3.2.  $\square$

### 4.3.3 Properties

Given the physical properties of Section 2.4 as well as the additional requirements of Section 4.3.2, we can show the relations of Lemmas 4.7 and 4.8, which we will make use of throughout this chapter.

**Lemma 4.7.** *Let Assumption 1 hold. Then, for the mapping  $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  given in (2.24), the following properties hold:*

1. *For all  $\varphi \in \mathbb{R}^2$ , we have that*

$$\underline{\nu}|\varphi| \leq |T(\varphi)| \leq \nu_0|\varphi|.$$

2. *There exist  $0 < \underline{c}_1 \leq \bar{c}_1$  such that*

$$\underline{c}_1|\psi|^2 \leq \psi^\top DT(\varphi)\psi \leq \bar{c}_1|\psi|^2 \quad \forall \varphi, \psi \in \mathbb{R}^2. \quad (4.32)$$

3. *There exists  $c_2 > 0$  such that*

$$(T(\varphi + \psi) - T(\varphi)) \cdot \psi \geq c_2|\psi|^2 \quad \forall \varphi, \psi \in \mathbb{R}^2. \quad (4.33)$$

4. *There exists a Lipschitz constant  $c_3 > 0$  such that*

$$|T(\varphi + \psi) - T(\varphi)| \leq c_3|\psi| \quad \forall \varphi, \psi \in \mathbb{R}^2. \quad (4.34)$$

*Proof.* 1. By definition of the operator  $T$  and property (2.19a) which follows from Assumption 1, we have

$$\underline{\nu}|\varphi| \leq |T(\varphi)| = \hat{\nu}(|\varphi|)|\varphi| \leq \nu_0|\varphi|.$$

2. Let  $\varphi = (\varphi_1, \varphi_2)^\top \in \mathbb{R}^2$ . According to (2.29), the eigenvalues of the  $2 \times 2$  matrix  $DT(\varphi)$  are given by  $\lambda_1 = \hat{\nu}(|\varphi|)$ ,  $\lambda_2 = \hat{\nu}(|\varphi|) + \hat{\nu}'(|\varphi|)|\varphi|$ . Properties (2.19a) and (2.19b) yield the claimed result with  $\underline{c}_1 = \underline{\nu}$  and  $\bar{c}_1 = \nu_0$ .

3. Property (4.33) can be seen in the same way. It holds that

$$(T(\varphi + \psi) - T(\varphi)) \cdot \psi = \int_0^1 \psi^\top DT(\varphi + t\psi) \psi dt \geq \underline{c}_1 |\psi|^2,$$

which yields (4.33) holds with  $c_2 = \underline{c}_1$ .

4. In a similar way, we obtain (4.34) with  $c_3 = \nu_0$ ,

$$\begin{aligned} |T(\varphi + \psi) - T(\varphi)| &= \left| \int_0^1 DT(\varphi + t\psi) \psi dt \right| \leq \int_0^1 |DT(\varphi + t\psi) \psi| dt \\ &\leq \int_0^1 \max \{ \lambda_1(\varphi + t\psi), \lambda_2(\varphi + t\psi) \} |\psi| dt \leq \nu_0 |\psi|. \end{aligned}$$

Here, we used (2.29) as well as (2.19a) and (2.19b).  $\square$

**Lemma 4.8.** *Let Assumption 3 hold. Then, the mapping  $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2, W \mapsto \hat{\nu}(|W|)W$  given in (2.24) has the following properties:*

1.  $T \in C^3(\mathbb{R}^2)$ .

2. For the constant  $c_4 \geq 0$  in (4.30), it holds

$$|S_\varphi(\psi_2) - S_\varphi(\psi_1)| \leq c_4 |\psi_2 - \psi_1|(|\psi_1| + |\psi_2|) \quad \forall \varphi, \psi_1, \psi_2 \in \mathbb{R}^2. \quad (4.35)$$

3. For the constant  $c_5 \geq 0$  in (4.31), it holds

$$|S_{\varphi_2}(\psi) - S_{\varphi_1}(\psi)| \leq c_5 |\varphi_2 - \varphi_1| |\psi|^2 \quad \forall \varphi_1, \varphi_2, \psi \in \mathbb{R}^2. \quad (4.36)$$

*Proof.* 1. We consider the first, second and third derivative of  $T$ :

- Recall the Jacobian of the mapping  $T$  given in (2.25),

$$DT(\varphi)(\psi) = \begin{cases} \hat{\nu}(|\varphi|)\psi + \frac{\hat{\nu}'(|\varphi|)}{|\varphi|}(\varphi \otimes \varphi)\psi, & \varphi \neq (0, 0)^\top \\ \hat{\nu}(|\varphi|)\psi & \varphi = (0, 0)^\top, \end{cases}$$

for  $\psi \in \mathbb{R}^2$ , and note that  $DT$  is continuous also in  $\varphi = 0$  due to

$$\begin{aligned} \lim_{|\varphi| \rightarrow 0} |DT(\varphi)(\psi) - DT(0)(\psi)| &\leq \lim_{|\varphi| \rightarrow 0} \left( |(\hat{\nu}(|\varphi|) - \hat{\nu}(0))\psi| + \left| \frac{\hat{\nu}'(|\varphi|)}{|\varphi|}(\varphi \otimes \varphi)\psi \right| \right) \\ &\leq \lim_{|\varphi| \rightarrow 0} (|(\hat{\nu}(|\varphi|) - \hat{\nu}(0))| |\psi| + |\hat{\nu}'(|\varphi|)| |\varphi| |\psi|) = 0. \end{aligned}$$

because of (4.28) and the continuity of  $\hat{\nu}$ .

- For  $D^2T$ , we get for all  $\varphi \neq 0$  and all  $\psi, \eta \in \mathbb{R}^2$ ,

$$\begin{aligned} D^2T(\varphi)(\psi, \eta) &= \frac{\hat{\nu}'(|\varphi|)}{|\varphi|} (\varphi \otimes \eta + (\varphi \cdot \eta)I + \eta \otimes \varphi) \psi \\ &\quad + \left( \frac{\hat{\nu}''(|\varphi|)}{|\varphi|^2} - \frac{\hat{\nu}'(|\varphi|)}{|\varphi|^3} \right) (\varphi \cdot \eta)(\varphi \otimes \varphi)\psi. \end{aligned} \quad (4.37)$$

In the point  $\varphi = 0$ , the Fréchet derivative of  $D^2T$  is given by  $D^2T(0)(\psi, \eta) = 0$ , which can be seen as follows:

$$\begin{aligned} \lim_{|\eta| \rightarrow 0} \frac{|DT(\eta)(\psi) - DT(0)(\psi) - 0|}{|\eta|} &= \lim_{|\eta| \rightarrow 0} \left| \frac{\hat{\nu}'(|\eta|) - \hat{\nu}'(0)}{|\eta|} \psi + \hat{\nu}'(|\eta|) \left( \frac{\eta}{|\eta|} \otimes \frac{\eta}{|\eta|} \right) \psi \right| \\ &\leq 2 |\hat{\nu}'(0)| |\psi| = 0, \end{aligned}$$

since  $\hat{\nu}'(0) = 0$  due to Assumption 3.2. Thus, we have,  $D^2T(0)(\psi, \eta) = 0$ . Also here, we can see the continuity in  $\varphi = 0$  as, for any  $\psi, \eta \in \mathbb{R}^2$ , it holds

$$\begin{aligned} \lim_{|\varphi| \rightarrow 0} |D^2T(\varphi)(\psi, \eta) - D^2T(0)(\psi, \eta)| &= \lim_{|\varphi| \rightarrow 0} \left[ \hat{\nu}'(|\varphi|) \left( \frac{\varphi}{|\varphi|} \otimes \eta + \left( \frac{\varphi}{|\varphi|} \cdot \eta \right) I + \eta \otimes \frac{\varphi}{|\varphi|} - \left( \frac{\varphi}{|\varphi|} \cdot \eta \right) \left( \frac{\varphi}{|\varphi|} \otimes \frac{\varphi}{|\varphi|} \right) \right) \right. \\ &\quad \left. + \hat{\nu}''(|\varphi|)(\varphi \cdot \eta) \left( \frac{\varphi}{|\varphi|} \otimes \frac{\varphi}{|\varphi|} \right) \right] \psi = 0, \end{aligned}$$

where, we used that  $\varphi/|\varphi|$  is a unit vector and, again, that  $\hat{\nu}'(0) = 0$  due to Assumption 3.2, as well as (4.29).

- By differentiating (4.37), we obtain for all  $\varphi, \psi, \eta, \xi \in \mathbb{R}^2$  with  $\varphi \neq 0$  that

$$\begin{aligned} D^3T(\varphi)(\psi, \eta, \xi) &= \frac{1}{|\varphi|^2} \left( \hat{\nu}''(|\varphi|) - \frac{\hat{\nu}'(|\varphi|)}{|\varphi|} \right) \varphi \cdot \xi [(\varphi \cdot \eta)\psi + (\varphi \cdot \psi)\eta + (\psi \cdot \eta)\varphi] \\ &\quad + \frac{\hat{\nu}'(|\varphi|)}{|\varphi|} [(\xi \cdot \eta)\psi + (\xi \cdot \psi)\eta + (\psi \cdot \eta)\xi] \\ &\quad + \frac{1}{|\varphi|^2} \left( \hat{\nu}''(|\varphi|) - \frac{\hat{\nu}'(|\varphi|)}{|\varphi|} \right) [(\xi \cdot \eta)(\varphi \cdot \psi)\varphi + (\varphi \cdot \eta)(\xi \cdot \psi)\varphi + (\varphi \cdot \eta)(\varphi \cdot \psi)\xi] \\ &\quad + \left( \frac{\hat{\nu}'''(|\varphi|)}{|\varphi|^3} - 3 \frac{\hat{\nu}''(|\varphi|)}{|\varphi|^4} + 3 \frac{\hat{\nu}'(|\varphi|)}{|\varphi|^5} \right) (\varphi \cdot \xi)(\varphi \cdot \eta)(\varphi \cdot \psi)\varphi, \end{aligned} \tag{4.38}$$

where we used that

$$\frac{d}{d\varphi} \left( \frac{\hat{\nu}'(|\varphi|)}{|\varphi|} \right) \eta = \frac{1}{|\varphi|^2} \left( \hat{\nu}''(|\varphi|) - \frac{\hat{\nu}'(|\varphi|)}{|\varphi|} \right) \varphi \cdot \eta.$$

We show that, under Assumption 3.2, the Fréchet derivative of  $D^2T$  at the point  $\varphi = 0$  is given by

$$D^3T(0)(\psi, \eta, \xi) = \nu''(0) (\xi \otimes \eta + (\xi \cdot \eta)I + \eta \otimes \xi) \psi. \tag{4.39}$$

Exploiting that  $D^2T(0)(\psi, \eta) = 0$ , we get

$$\begin{aligned} \lim_{|\xi| \rightarrow 0} \frac{|D^2T(\xi)(\psi, \eta) - D^2T(0)(\psi, \eta) - D^3T(0)(\psi, \eta, \xi)|}{|\xi|} &= \lim_{|\xi| \rightarrow 0} \left| \frac{\hat{\nu}'(|\xi|)}{|\xi|} \left( \frac{\xi}{|\xi|} \otimes \eta + \left( \frac{\xi}{|\xi|} \cdot \eta \right) I + \eta \otimes \frac{\xi}{|\xi|} \right) \psi \right. \\ &\quad \left. + \left( \hat{\nu}''(|\xi|) - \frac{\hat{\nu}'(|\xi|)}{|\xi|} \right) \left( \frac{\xi}{|\xi|} \cdot \eta \right) \left( \frac{\xi}{|\xi|} \otimes \frac{\xi}{|\xi|} \right) \psi \right. \\ &\quad \left. - \nu''(0) \left( \frac{\xi}{|\xi|} \otimes \eta + \left( \frac{\xi}{|\xi|} \cdot \eta \right) I + \eta \otimes \frac{\xi}{|\xi|} \right) \psi \right|. \end{aligned}$$

Noting that, under Assumption 3.2, we have  $\hat{\nu}'(0) = 0$  and thus

$$\lim_{t \rightarrow 0} \nu'(t)/t = \lim_{t \rightarrow 0} (\nu'(t) - \nu'(0))/(t - 0) = \nu''(0),$$

we see that the above expression vanishes, which proves the form (4.39). The continuity of  $D^3T(\varphi)(\psi, \eta, \xi)$  is clear for  $\varphi \neq 0$  and can be seen for the point  $\varphi = 0$  noting that  $\lim_{t \rightarrow 0} \nu'(t)/t = \nu''(0)$  which finishes the proof of statement 1 of Lemma 4.8.

2. We follow the lines of the proof of Proposition 4.1.3(7) on page 73 of [46]: Since  $T \in C^3(\mathbb{R}^2)$ , we can apply the fundamental theorem of calculus and get

$$\begin{aligned} S_\varphi(\psi_2) - S_\varphi(\psi_1) &= T(\varphi + \psi_2) - T(\varphi + \psi_1) - DT(\varphi)(\psi_2 - \psi_1) \\ &= \int_0^1 [DT(\varphi + \psi_1 + t(\psi_2 - \psi_1)) - DT(\varphi)] (\psi_2 - \psi_1) dt \\ &= \int_0^1 \int_0^1 D^2T(\varphi + s[(1-t)\psi_1 + t\psi_2])((1-t)\psi_1 + t\psi_2, \psi_2 - \psi_1) dt ds. \end{aligned}$$

From (4.37), it follows that

$$\begin{aligned} |D^2T(\varphi)(\psi, \eta)| &\leq 3 \frac{|\hat{\nu}'(|\varphi|)|}{|\varphi|} |\varphi| |\psi| |\eta| + \left( \frac{|\hat{\nu}''(|\varphi|)|}{|\varphi|^2} + \frac{|\hat{\nu}'(|\varphi|)|}{|\varphi|^3} \right) |\varphi|^3 |\psi| |\eta| \\ &= (4|\hat{\nu}'(|\varphi|)| + |\hat{\nu}''(|\varphi|)| |\varphi|) |\psi| |\eta|. \end{aligned} \quad (4.40)$$

Estimate (4.40), together with requirement (4.30), yields that

$$\begin{aligned} |S_\varphi(\psi_2) - S_\varphi(\psi_1)| &\leq \int_0^1 \int_0^1 |D^2T(\varphi + s[(1-t)\psi_1 + t\psi_2])| |(1-t)\psi_1 + t\psi_2| |\psi_2 - \psi_1| dt ds \\ &\leq c_4 |(1-t)\psi_1 + t\psi_2| |\psi_2 - \psi_1| \\ &\leq c_4 (|\psi_1| + |\psi_2|) |\psi_2 - \psi_1|. \end{aligned}$$

3. Since  $T$  is three times continuously differentiable, we get by the fundamental theorem of calculus that

$$\begin{aligned} |S_{\varphi_2}(\psi) - S_{\varphi_1}(\psi)| &= |T(\varphi_2 + \psi) - T(\varphi_2) - DT(\varphi_2)\psi - (T(\varphi_1 + \psi) - T(\varphi_1) - DT(\varphi_1)\psi)| \\ &= \left| \int_0^1 (DT(\varphi_2 + t\psi) - DT(\varphi_2)) \psi dt - \int_0^1 (DT(\varphi_1 + t\psi) - DT(\varphi_1)) \psi dt \right| \\ &= \left| \int_0^1 \int_0^1 D^2T(\varphi_2 + st\psi)(\psi, t\psi) ds dt - \int_0^1 \int_0^1 D^2T(\varphi_1 + st\psi)(\psi, t\psi) ds dt \right| \\ &\leq \int_0^1 \int_0^1 |D^2T(\varphi_2 + st\psi)(\psi, t\psi) - D^2T(\varphi_1 + st\psi)(\psi, t\psi)| ds dt \\ &\leq \int_0^1 \int_0^1 \int_0^1 t |D^3T(\varphi_1 + st\psi + r(\varphi_2 - \varphi_1))(\psi, \psi, \varphi_2 - \varphi_1)| dr ds dt. \end{aligned}$$

From (4.38), it is seen that

$$\begin{aligned} |D^3T(\varphi)(\psi, \psi, \xi)| &\leq 6 \left| \left( \hat{\nu}''(|\varphi|) - \frac{\hat{\nu}'(|\varphi|)}{|\varphi|} \right) \right| |\psi|^2 |\xi| + 3 \frac{|\hat{\nu}'(|\varphi|)|}{|\varphi|} |\psi|^2 |\xi| \\ &\quad + \left| \left( \hat{\nu}'''(|\varphi|) |\varphi| - 3\hat{\nu}''(|\varphi|) + 3 \frac{\hat{\nu}'(|\varphi|)}{|\varphi|} \right) \right| |\psi|^2 |\xi| \\ &\leq \left( |\hat{\nu}'''(|\varphi|)| |\varphi| + 9|\hat{\nu}''(|\varphi|)| + 12 \frac{|\hat{\nu}'(|\varphi|)|}{|\varphi|} \right) |\psi|^2 |\xi|. \end{aligned}$$

Denoting  $\varphi_{r,s}^t = \varphi_1 + s t \psi + r(\varphi_2 - \varphi_1)$  for  $0 \leq r, s, t \leq 1$ , we get

$$\begin{aligned} |S_{\varphi_2}(\psi) - S_{\varphi_1}(\psi)| &\leq \int_0^1 \int_0^1 \int_0^1 |D^3T(\varphi_{r,s}^t)(\psi, \psi, \varphi_2 - \varphi_1)| dr ds dt \\ &\leq \int_0^1 \int_0^1 \int_0^1 \left| |\hat{\nu}'''(|\varphi_{r,s}^t|)| |\varphi_{r,s}^t| + 9|\hat{\nu}''(|\varphi_{r,s}^t|)| + 12 \frac{|\hat{\nu}'(|\varphi_{r,s}^t|)|}{|\varphi_{r,s}^t|} \right| |\psi|^2 |\varphi_2 - \varphi_1| dr ds dt \\ &\leq c_5 |\psi|^2 |\varphi_2 - \varphi_1|, \end{aligned}$$

where we used (4.31), which holds under Assumption 3.

This concludes the proof of Lemma 4.8  $\square$

**Remark 4.9.** It is easy to see that the statements of Lemma 4.7 and Lemma 4.8 also hold for the  $x$ -dependent operators  $T_\varepsilon$ ,  $DT_\varepsilon$  introduced in (4.5) and the operator  $S^\varepsilon$  defined in (4.25) for each point in their domain of definition with the same constants.

**Remark 4.10.** Note that relation (4.35) implies that

$$|S_\varphi(\psi)| \leq c_4 |\psi|^2. \quad (4.41)$$

#### 4.3.4 Weighted Sobolev Spaces

In order to analyze the asymptotic behavior of the variation of the direct and adjoint state at scale 1 in Sections 4.4.1 and 4.4.2 as well as in Sections 4.5.4 and 4.5.5, we need to define an appropriate function space. We follow the presentation given in [46]. For more details on weighted Sobolev spaces, we refer the reader to [12]. Let the weight function  $w : \mathbb{R}^2 \rightarrow \mathbb{R}$  be defined as

$$w(x) = \frac{1}{(1 + |x|^2)^{1/2} (\log(2 + |x|))}. \quad (4.42)$$

Note that  $w \in L^2(\mathbb{R}^2)$  and  $w(x) > 0$  for all  $x \in \mathbb{R}^2$  with

$$\inf_{x \in \mathbb{R}^2} w(x) = 0 \quad \text{and} \quad \sup_{x \in \mathbb{R}^2} w(x) < \infty.$$

For all open  $\mathcal{O} \subset \mathbb{R}^2$ , the space of distributions in  $\mathcal{O}$  is denoted by  $\mathcal{D}'(\mathcal{O})$ . We define the weighted Sobolev space

$$\mathcal{H}^w(\mathcal{O}) := \{u \in \mathcal{D}'(\mathcal{O}) : w u \in L^2(\mathcal{O}), \nabla u \in L^2(\mathcal{O})\},$$

together with the inner product

$$\langle u, v \rangle_{\mathcal{H}^w(\mathcal{O})} := \langle w u, w v \rangle_{L^2(\mathcal{O})} + \langle \nabla u, \nabla v \rangle_{L^2(\mathcal{O})}, \quad u, v \in \mathcal{H}^w(\mathcal{O}),$$

and the norm

$$\|u\|_{\mathcal{H}^w(\mathcal{O})} := \langle u, u \rangle_{\mathcal{H}^w(\mathcal{O})}^{\frac{1}{2}}, \quad u \in \mathcal{H}^w(\mathcal{O}).$$

The following result is shown in [46].

**Lemma 4.11.** *The space  $\mathcal{H}^w(\mathcal{O})$  endowed with the inner product  $\langle \cdot, \cdot \rangle_{\mathcal{H}^w(\mathcal{O})}$  is a separable Hilbert space.*

We define the weighted quotient Sobolev space

$$\mathcal{H}(\mathbb{R}^2) := \mathcal{H}^w(\mathbb{R}^2)/\mathbb{R}$$

where we factor out the constants, and equip it with the quotient norm

$$\|[u]\|_{\mathcal{H}(\mathbb{R}^2)} := \inf_{m \in \mathbb{R}} \|\tilde{u} + m\|_{\mathcal{H}^w(\mathbb{R}^2)}, \quad [u] \in \mathcal{H}(\mathbb{R}^2),$$

where  $\tilde{u} \in \mathcal{H}^w(\mathbb{R}^2)$  is any element of the class  $[u]$ . We note that  $\mathcal{H}(\mathbb{R}^2)$  is a Hilbert space because  $\mathcal{H}^w(\mathbb{R}^2)$  is a Hilbert space and  $\mathbb{R}$  is a closed subspace. For the space  $\mathcal{H}(\mathbb{R}^2)$ , we can state the following Poincaré inequality, which is proved in [46]:

**Lemma 4.12.** *There exists  $c_P > 0$  such that*

$$\|[u]\|_{\mathcal{H}(\mathbb{R}^2)} \leq c_P \|\nabla \tilde{u}\|_{L^2(\mathbb{R}^2)}, \quad \forall [u] \in \mathcal{H}(\mathbb{R}^2),$$

where  $\tilde{u} \in \mathcal{H}^w(\mathbb{R}^2)$  is any element of the class  $[u]$ .

For all  $[u] \in \mathcal{H}(\mathbb{R}^2)$ , let  $\tilde{u} \in \mathcal{H}(\mathbb{R}^2)$  denote any element of the class  $[u]$ . We endow  $\mathcal{H}(\mathbb{R}^2)$  with the semi-norm

$$|[u]|_{\mathcal{H}(\mathbb{R}^2)} := \|\nabla \tilde{u}\|_{L^2(\mathbb{R}^2)}.$$

The following corollary follows directly from Lemma 4.12.

**Corollary 4.13.** *The semi-norm  $|\cdot|_{\mathcal{H}(\mathbb{R}^2)}$  is a norm and is equivalent to the norm  $\|\cdot\|_{\mathcal{H}(\mathbb{R}^2)}$  in  $\mathcal{H}(\mathbb{R}^2)$ .*

### 4.3.5 Relation to Previous Work

As we mentioned in the introduction of Chapter 4, large parts of this chapter are following the lines of [20, 46]. Here, we want to give a brief overview over the main differences to the results obtained there.

The main technical difference of the considered problems can be seen from the definition of the perturbed operator  $T_\varepsilon$  given in Section 4.1.2. In this thesis, we consider the perturbation of a nonlinear subdomain by an inclusion of linear material or the other way around,

$$T_\varepsilon(x, W) = \begin{cases} \nu_0 W & \text{in } \omega_\varepsilon, \\ T(W) & \text{in } D \setminus \omega_\varepsilon, \end{cases} \quad T_\varepsilon^{(2)}(x, W) = \begin{cases} T(W) & \text{in } \omega_\varepsilon, \\ \nu_0 W & \text{in } D \setminus \omega_\varepsilon, \end{cases}$$

with  $T(W) = \hat{\nu}(|W|)W$  for  $W \in \mathbb{R}^2$ , whereas in [20, 46], the authors consider the same nonlinear function multiplied by a different constant factor inside and outside the inclusion. In our notation, this would correspond to

$$T_\varepsilon(x, W) = \begin{cases} \gamma_1 T_a(W) & \text{in } \omega_\varepsilon, \\ \gamma_0 T_a(W) & \text{in } D \setminus \omega_\varepsilon. \end{cases}$$

There, the operator  $T_a(W)$  represents a regularized version of the  $p$ -Laplace operator,  $T_a(W) = (a^2 + |W|^2)^{(p-2)/2}W$  for some  $a > 0$  and  $p \geq 2$ . On the one hand, many of the steps taken for the derivation of the topological derivative can be used analogously in our context. The function space setting and the majority of the estimations even simplify here since all involved quantities are defined in the Hilbert spaces  $H^1(D)$  and  $\mathcal{H}(\mathbb{R}^2)$  rather than in the Sobolev space  $W^{1,p}(D)$  or the corresponding weighted Sobolev space over  $\mathbb{R}^2$ . On the other hand, especially the proof of Theorem 4.24, which is based on Propositions 4.21 and 4.22, required some additional effort.

On the other hand, the work presented in this thesis extends the results of [20, 46] in several directions. Our work is motivated by a concrete application from electrical engineering. Therefore, our focus is not only on the rigorous theoretical derivation of the correct formula for the topological derivative, but also on the practical applicability of this formula. In order to be able to use the derived formula for computational shape and topology optimization, we had to consider the following additional aspects:

- We compute both sensitivities  $G^{f \rightarrow air}$  (see Section 4.4) and  $G^{air \rightarrow f}$  (see Section 4.5) in order to be able to apply a bi-directional optimization algorithms which is capable of both introducing and removing ferromagnetic material. Note that, in [20, 46], the derivation of the topological derivative in the reverse scenario cannot be achieved by simply exchanging the values for  $\gamma_0$  and  $\gamma_1$  since the result that corresponds to our Theorem 4.24 assumes that  $\gamma_1 < \gamma_0$ .
- We derive explicit formulas for the matrices  $\mathcal{M}$ ,  $\mathcal{M}^{(2)}$ , see (4.233) and (4.234).
- It is a priori not clear, how the new terms  $J_2$ ,  $J_2^{(2)}$  defined in (4.104) and (4.211), respectively, which account for the nonlinearity of the problem, can be computed numerically in an efficient way. In Section 4.7, we show a way to efficiently evaluate these terms by precomputing some values in an off-line stage and looking them up during the optimization process.

## 4.4 Topological Asymptotic Expansion: Case I

In this section, we derive the topological asymptotic expansion (4.2) for the introduction of an inclusion of air, which has linear material behavior, inside ferromagnetic material, which behaves nonlinearly. For the reader's convenience, we moved all longer, technical proofs of this section to Section 4.4.4.

We begin the topological asymptotic expansion by the expansion of our cost function. Due to Assumption 2, we have the expansion (4.7), i.e.,

$$\mathcal{J}_\varepsilon(u_\varepsilon) = \mathcal{J}_0(u_0) + \langle \tilde{G}, u_\varepsilon - u_0 \rangle + \delta_J \varepsilon^2 + R(\varepsilon)$$

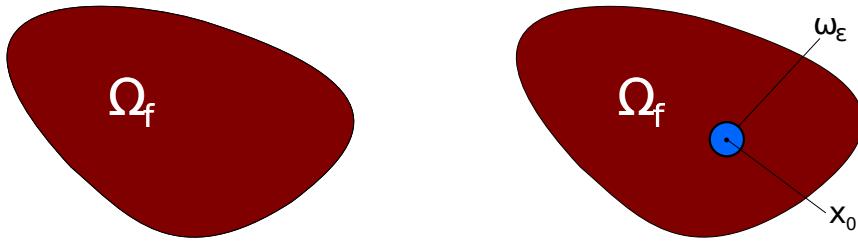


Figure 4.2: Left: Unperturbed configuration for Case I,  $D = \Omega_f$ . Right: Perturbed configuration for Case I,  $\bar{D} = \overline{\Omega_f} \cup \overline{\omega_\varepsilon}$ .

with  $\delta_J \in \mathbb{R}$ ,  $\tilde{G} \in H^{-1}(D)$  and the remainder  $R$  of the form (4.8). Estimate (4.73) in Section 4.4.1.6 will allow us to show that, under condition (4.8), it holds  $R(\varepsilon) = o(\varepsilon^2)$ . This is important for showing a topological asymptotic expansion of the form (4.2) with the first order correction function  $f(\varepsilon) = \varepsilon^2$ . The goal of the rest of this section is to identify the topological derivative  $G(x_0)$  at the point  $x_0$  such that

$$\langle \tilde{G}, u_\varepsilon - u_0 \rangle = \varepsilon^2 G(x_0) + o(\varepsilon^2). \quad (4.43)$$

For dealing with this term, we will first collect some properties about the behavior of the difference between the solution of the perturbed and the unperturbed problem,  $u_\varepsilon - u_0$ , in Section 4.4.1. We will refer to this difference as the *variation of the direct state*. Then, in Section 4.4.2, we will introduce an adjoint boundary value problem with the bounded linear form  $\tilde{G}$  on the right hand side, again in the unperturbed ( $\varepsilon = 0$ ) and perturbed ( $\varepsilon > 0$ ) setting, and investigate the difference between the solutions to these problems,  $p_\varepsilon - p_0$ , the *variation of the adjoint state*. In Section 4.4.3, we will combine our findings to derive the final formula for the topological derivative  $G(x_0)$ .

#### 4.4.1 Variation of Direct State

In this section, we first define a boundary value problem whose solution is the variation of the direct state,  $u_\varepsilon - u_0$  (Section 4.4.1.2), and introduce an approximation to this problem (Section 4.4.1.3). Note that the left hand side of (4.43) depends on  $\varepsilon$  only via the solution to the perturbed equation  $u_\varepsilon$ , whereas, on the right hand side, the dependence of  $\varepsilon$  is explicit and the quantity  $G(x_0)$  is independent of  $\varepsilon$ . The key to achieve a relation like (4.43) is to define a transmission problem at scale 1 which is defined on the entire plane  $\mathbb{R}^2$  and independent of  $\varepsilon$  (see Section 4.4.1.4) and show subsequently the relation between the solution  $H$  of this problem and the variation  $u_\varepsilon - u_0$  in terms of  $\varepsilon$ , see Section 4.4.1.6. An important ingredient for the estimates in Section 4.4.1.6 will be a sufficiently fast decay of  $H$  as  $|x|$  tends to infinity, which is provided in Section 4.4.1.5.

##### 4.4.1.1 Regularity Assumptions

In order to perform the asymptotic analysis for the derivation of the topological derivative, we need some regularity of the solution to the unperturbed state problem (4.3) in a neighborhood of the point of the perturbation  $x_0 \in \Omega^d$ . Henceforth, we make the following assumption:

**Assumption 5.** There exists  $\beta > 0$  such that

$$u_0|_{\Omega^d} \in C^{1,\beta}(\Omega^d).$$

**Remark 4.14.** If the right hand side of the quasilinear boundary value problem (4.3) is given by a smooth function  $f \in C^{0,\gamma}(\overline{D})$ , Assumption 5 follows from [141, Theorem 3.20] (assuming that  $u_0 \in L^\infty(D)$ ). In the case of the model problem introduced in Section 2.3, the right hand side is a distribution  $F \in H^{-1}(D)$ , which is, however, only supported outside the design area. Therefore, the assumption that the solution  $u_0$  is smooth in the design area is reasonable.

Assumption 5 immediately implies the following regularity properties:

$$\nabla u_0|_{\Omega^d} \in C^{0,\beta}(\Omega^d), \quad (4.44)$$

$$\nabla u_0|_{\Omega^d} \in L^\infty(\Omega^d). \quad (4.45)$$

#### 4.4.1.2 Step 1: Variation $u_\varepsilon - u_0$

Subtracting the perturbed problem (4.3) from the unperturbed problem (4.6) and noting that the right hand sides coincide, we get

$$\begin{aligned} 0 &= \int_D T_\varepsilon(x, \nabla u_\varepsilon) \cdot \nabla \eta - \int_D T(\nabla u_0) \cdot \nabla \eta \\ &= \int_D (T_\varepsilon(x, \nabla u_\varepsilon) - T_\varepsilon(x, \nabla u_0)) \cdot \nabla \eta + \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|\nabla u_0|)) \nabla u_0 \cdot \nabla \eta. \end{aligned}$$

The boundary value problem defining the variation of the direct state at scale  $\varepsilon$ ,  $\tilde{u}_\varepsilon := u_\varepsilon - u_0$ , is therefore given by

$$\begin{aligned} &\text{Find } \tilde{u}_\varepsilon \in H_0^1(D) \text{ such that} \\ &\int_D (T_\varepsilon(x, \nabla u_0 + \nabla \tilde{u}_\varepsilon) - T_\varepsilon(x, \nabla u_0)) \cdot \nabla \eta \\ &= - \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|\nabla u_0|)) \nabla u_0 \cdot \nabla \eta \quad \forall \eta \in H_0^1(D). \end{aligned} \quad (4.46)$$

#### 4.4.1.3 Step 2: Approximation of Variation $u_\varepsilon - u_0$

We approximate problem (4.46) by the same boundary value problem where we replace the function  $\nabla u_0$  by its value at the point of interest  $x_0 = (0, 0)^\top$ , i.e., we replace  $\nabla u_0$  by the constant  $U_0 := \nabla u_0(x_0)$ . Note that this point evaluation makes sense due to Assumptions 5. Denoting the solution to the arising boundary value problem by  $h_\varepsilon$ , we get

$$\begin{aligned} &\text{Find } h_\varepsilon \in H_0^1(D) \text{ such that} \\ &\int_D (T_\varepsilon(x, U_0 + \nabla h_\varepsilon) - T_\varepsilon(x, U_0)) \cdot \nabla \eta \\ &= - \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|U_0|)) U_0 \cdot \nabla \eta \quad \forall \eta \in H_0^1(D). \end{aligned} \quad (4.47)$$

The relation between the solutions to boundary value problems (4.46) and (4.47) will be investigated in Proposition 4.28.

#### 4.4.1.4 Step 3: Change of Scale

Next, we make another approximation to boundary value problem (4.47). First, we perform a change of scale, i.e., we go over from the domain  $D$  with the inclusion  $\omega_\varepsilon$  of size  $\varepsilon$  to the much larger domain  $D/\varepsilon$  with the inclusion  $\omega$  of unit size, e.g.,  $\omega = B(0, 1)$ . In a second step we approximate this scaled version of (4.47) by sending the boundary of the “very large” domain  $D/\varepsilon$  to infinity. This yields a transmission problem on the plane  $\mathbb{R}^2$  which is independent of  $\varepsilon$ . We introduce the  $\varepsilon$ -independent operators corresponding to (4.5) and (4.25) at scale 1,

$$\begin{aligned}\tilde{T}(x, W) &:= \chi_{\mathbb{R}^2 \setminus \omega}(x)T(W) + \chi_\omega(x)\nu_0 W, \\ \tilde{S}_W(x, V) &:= \chi_{\mathbb{R}^2 \setminus \omega}(x)S_W(V),\end{aligned}\tag{4.48}$$

for  $x \in \mathbb{R}^2$  and  $V, W \in \mathbb{R}^2$  with  $T$  and  $S$  given in (2.24) and (4.24), respectively, and note that

$$D\tilde{T}(x, W) = \chi_{\mathbb{R}^2 \setminus \omega}(x)DT(W) + \chi_\omega(x)\nu_0 I.$$

**Remark 4.15.** Again, the statements of Lemma 4.7 and Lemma 4.8 also hold for the  $x$ -dependent operators  $\tilde{T}, D\tilde{T}, \tilde{S}$  for each point in their domain of definition with the same constants.

With this notation, we arrive at the nonlinear transmission problem on  $\mathbb{R}^2$  defining  $H$ , the variation of the direct state at scale 1:

Find  $H \in \mathcal{H}(\mathbb{R}^2)$  such that

$$\begin{aligned}&\int_{\mathbb{R}^2} (\tilde{T}(x, U_0 + \nabla H) - \tilde{T}(x, U_0)) \cdot \nabla \eta \\ &= - \int_{\omega} (\nu_0 - \hat{\nu}(|U_0|)) U_0 \cdot \nabla \eta \quad \forall \eta \in \mathcal{H}(\mathbb{R}^2).\end{aligned}\tag{4.49}$$

**Remark 4.16.** We remark that, for  $U_0 = (0, 0)^\top$ , problems (4.47) and (4.49) only admit the trivial solution which yields that  $H, \nabla H, h_\varepsilon, \nabla h_\varepsilon$  are identical zero and also  $S_{U_0}(\nabla H)$  and  $S_{U_0}(\nabla h_\varepsilon)$  vanish. In this case, many computations simplify significantly. For the rest of this chapter, we exclude the trivial case and assume that  $U_0 \neq (0, 0)^\top$ .

Next, we show existence and uniqueness of a solution to (4.49) using Theorem 2.5 (Zarantonello).

**Proposition 4.17.** Let Assumption 1 hold. Then there exists a unique solution  $H \in \mathcal{H}(\mathbb{R}^2)$  to problem (4.49).

*Proof.* We apply Theorem 2.5 to problem (4.49) rewritten in the form

$$\text{Find } H \in \mathcal{H}(\mathbb{R}^2) \text{ such that } A H = L,$$

where the operator  $A : \mathcal{H}(\mathbb{R}^2) \rightarrow \mathcal{H}(\mathbb{R}^2)^*$  and the right hand side  $L \in \mathcal{H}(\mathbb{R}^2)^*$  are defined by

$$\begin{aligned}\langle A\eta_1, \eta_2 \rangle &= \int_{\mathbb{R}^2} (\tilde{T}(x, U_0 + \nabla \eta_1) - \tilde{T}(x, U_0)) \cdot \nabla \eta_2, \\ \langle L, \eta \rangle &= \int_{\omega} (\hat{\nu}(|U_0|) - \nu_0) U_0 \cdot \nabla \eta,\end{aligned}$$

for  $\eta_1, \eta_2, \eta \in \mathcal{H}(\mathbb{R}^2)$ . We verify the strong monotonicity (2.22) and Lipschitz continuity (2.23) of the operator A.

Property (4.33) together with Remark 4.15 gives

$$\begin{aligned} \langle A\eta_1 - A\eta_2, \eta_1 - \eta_2 \rangle &= \int_{\mathbb{R}^2} \left( \tilde{T}(x, U_0 + \nabla \eta_1) - \tilde{T}(x, U_0 + \nabla \eta_2) \right) (\nabla \eta_1 - \nabla \eta_2) \\ &\geq c_2 \int_{\mathbb{R}^2} |\nabla \eta_1 - \nabla \eta_2|^2 = c_2 \|\nabla \eta_1 - \nabla \eta_2\|_{L^2(\mathbb{R}^2)}^2. \end{aligned}$$

The Poincaré inequality of Lemma 4.12 yields the strong monotonicity property (2.22) in  $\mathcal{H}(\mathbb{R}^2)$ .

For the Lipschitz condition, we get by property (4.34) together with Remark 4.15, Cauchy's inequality and the norm equivalence of Corollary 4.13 that

$$\begin{aligned} \|A\eta_1 - A\eta_2\|_{\mathcal{H}(\mathbb{R}^2)^*} &= \sup_{\eta \neq 0} \frac{1}{\|\eta\|_{\mathcal{H}(\mathbb{R}^2)}} |\langle A\eta_1 - A\eta_2, \eta \rangle| \\ &= \sup_{\eta \neq 0} \frac{1}{\|\eta\|_{\mathcal{H}(\mathbb{R}^2)}} \left| \int_{\mathbb{R}^2} \left( \tilde{T}(x, U_0 + \nabla \eta_1) - \tilde{T}(x, U_0 + \nabla \eta_2) \right) \cdot \nabla \eta \right| \\ &\leq c_3 \sup_{\eta \neq 0} \frac{1}{\|\eta\|_{\mathcal{H}(\mathbb{R}^2)}} \int_{\mathbb{R}^2} |\nabla \eta_1 - \nabla \eta_2| |\nabla \eta| \\ &\leq c_3 \sup_{\eta \neq 0} \frac{1}{\|\eta\|_{\mathcal{H}(\mathbb{R}^2)}} \|\nabla \eta_1 - \nabla \eta_2\|_{\mathcal{H}(\mathbb{R}^2)} \|\eta\|_{\mathcal{H}(\mathbb{R}^2)} \\ &= c_3 \|\nabla \eta_1 - \nabla \eta_2\|_{\mathcal{H}(\mathbb{R}^2)}. \end{aligned}$$

Therefore, Theorem (2.5) yields the existence of a unique solution  $H \in \mathcal{H}(\mathbb{R}^2)$  to the variational problem (4.49) since  $L \in \mathcal{H}(\mathbb{R}^2)^*$ .  $\square$

#### 4.4.1.5 Step 4: Asymptotic Behavior of Variations of Direct State

In this section, we investigate the asymptotic behavior of the solution  $H$  to problem (4.49) as  $|x|$  goes to infinity. In order to show (4.43), we need a sufficiently fast decay of  $H$  “far away” from the inclusion.

For that purpose, we introduce the function  $H_\varepsilon : D \rightarrow \mathbb{R}$ , which brings back the solution  $H$  of problem (4.49) from the plane  $\mathbb{R}^2$  to the bounded domain  $D$ . For  $H \in \mathcal{H}(\mathbb{R}^2)$  the solution to (4.49),  $\hat{H} \in \mathcal{H}^w(\mathbb{R}^2)$  a given element of the class  $H$  and  $\varepsilon > 0$ ,  $H_\varepsilon$  is defined by

$$H_\varepsilon(x) := \varepsilon \hat{H}(\varepsilon^{-1}x), \quad x \in D. \quad (4.50)$$

Note that, when one is only interested in the gradient of  $H_\varepsilon$ , the specific choice of  $\hat{H}$  in the class  $H$  does not matter. Noting that

$$\underline{w} := \inf_{x \in D} w\left(\frac{x}{\varepsilon}\right) > 0,$$

it is easy to see that  $\hat{H} \in \mathcal{H}^w(\mathbb{R}^2)$  implies  $H_\varepsilon \in H^1(D)$ . We can show some first estimates which we will make use of in later estimations:

**Lemma 4.18.** *Let Assumption 1 and Assumption 5 hold. Then*

$$\|\nabla \tilde{u}_\varepsilon\|_{L^2(D)}^2 = \mathcal{O}(\varepsilon^2), \quad (4.51)$$

$$\|\nabla h_\varepsilon\|_{L^2(D)}^2 = \mathcal{O}(\varepsilon^2), \quad (4.52)$$

$$\|\nabla H_\varepsilon\|_{L^2(D)}^2 = \mathcal{O}(\varepsilon^2). \quad (4.53)$$

The proof of Lemma 4.18 can be found in Section 4.4.4.1 on page 72.

**Remark 4.19.** *By the triangle inequality, it follows immediately from estimates (4.51), (4.52) and (4.53) that*

$$\|\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon\|_{L^2(D)}^2 = \mathcal{O}(\varepsilon^2), \quad (4.54)$$

$$\|\nabla h_\varepsilon - \nabla H_\varepsilon\|_{L^2(D)}^2 = \mathcal{O}(\varepsilon^2). \quad (4.55)$$

In Section 4.4.1.6, we will show more estimates for the difference between the solution  $\tilde{u}_\varepsilon$  of (4.46) and  $h_\varepsilon$  of (4.47) on the one hand, and for the difference between  $h_\varepsilon$  and  $H_\varepsilon$  on the other hand, which will allow us to derive the topological derivative  $G(x_0)$  in Section 4.4.3. In particular, we will show stronger estimates of the type (4.54) in (4.71), and of (4.55) in (4.67).

In order to show estimate (4.67), we need that there exists a representative  $\tilde{H}$  of the solution  $H$  to (4.49) which satisfies a sufficiently fast decay for  $|x| \rightarrow \infty$ . For that purpose, let the nonlinear operator  $Q : \mathcal{H}(\mathbb{R}^2) \rightarrow \mathcal{H}(\mathbb{R}^2)^*$  be defined by

$$\langle Q\eta_1, \eta_2 \rangle := \int_{\mathbb{R}^2} [\tilde{T}(x, U_0 + \nabla \eta_1) - \tilde{T}(x, U_0)] \cdot \nabla \eta_2 + \int_{\omega} (\nu_0 - \hat{\nu}(|U_0|)) U_0 \cdot \nabla \eta_2. \quad (4.56)$$

Note that for  $H$  the solution of (4.49), we have that

$$\langle Q H, \eta \rangle = 0 \quad \text{for all } \eta \in \mathcal{H}(\mathbb{R}^2). \quad (4.57)$$

In the following, we show that there exist a supersolution  $R_1$  satisfying  $\langle Q R_1, \eta \rangle \geq 0$  and a subsolution  $R_2$  such that  $\langle Q R_2, \eta \rangle \leq 0$  for all test functions  $\eta$  in a subset of  $\mathcal{H}(\mathbb{R}^2)$ , both of which satisfy a sufficient decay at infinity. Then we make use of a comparison principle to show that there exists a representative  $\tilde{H}$  of the solution  $H$  of (4.57) which satisfies  $R_2(x) \leq \tilde{H}(x) \leq R_1(x)$  almost everywhere and conclude that  $\tilde{H}$  must have the same decay at infinity as  $R_1$  and  $R_2$ .

For that purpose, we first introduce a coordinate system that is aligned with the fixed vector  $U_0$ . Since we excluded the trivial case where  $U_0 = (0, 0)^\top$  (see Remark 4.16), we can introduce the unit vector  $e_1 = U_0/|U_0|$  and the orthonormal basis  $(e_1, e_2)$  of  $\mathbb{R}^2$ . We denote  $(x_1, x_2)$  the system of coordinates in this basis and introduce the half space  $\mathbb{R}_+^2 := \{x \in \mathbb{R}^2 : U_0 \cdot x \geq 0\}$ . We first show that there exists a representative  $\tilde{H}$  of the solution  $H$  to (4.49) that is odd with respect to the first coordinate.

**Lemma 4.20.** *Let  $H \in \mathcal{H}(\mathbb{R}^2)$  be the unique solution to the operator equation  $QH = 0$  with  $Q$  defined in (4.56) and assume that  $\omega$  is symmetric with respect to the line  $\{x \in \mathbb{R}^2 : U_0 \cdot x = 0\}$ . Then there exists an element  $\tilde{H}$  of the class  $H$  such that, for all  $(x_1, x_2) \in \mathbb{R}^2$ ,*

$$\tilde{H}(-x_1, x_2) = -\tilde{H}(x_1, x_2).$$

*In particular,  $\tilde{H}(0, x_2) = 0$  for all  $x_2 \in \mathbb{R}$ .*

A proof of Lemma 4.20 can be found in Section 4.4.4.1 on page 73.

Lemma 4.20 allows us to investigate the asymptotic behavior of  $\tilde{H}$  only in the half plane  $\mathbb{R}_+^2$ . However, Lemma 4.20 is based on the assumption that  $\omega$  is symmetric with respect to the line  $\{x \in \mathbb{R}^2 : U_0 \cdot x = 0\}$ . In order to fulfill this assumption for any given  $U_0 \in \mathbb{R}^2$ , we will from now on assume that  $\omega = B(0, 1)$ .

**Proposition 4.21.** *Let  $\omega = B(0, 1)$  and let Assumption 4 hold. Then there exists  $\sigma > 1$  and  $k \in (0, 1]$  such that the function  $R_1 \in \mathcal{H}(\mathbb{R}^2) \cap L^\infty(\mathbb{R}^2)$  defined by*

$$R_1(x) := \begin{cases} k(U_0 \cdot x)|x|^{-\sigma} & x \in \mathbb{R}^2 \setminus \omega, \\ k(U_0 \cdot x) & x \in \omega, \end{cases} \quad (4.58)$$

satisfies

$$\langle Q R_1, \eta \rangle \geq 0 \quad \forall \eta \in \mathcal{H}(\mathbb{R}^2) : \text{supp}(\eta) \subset \mathbb{R}_+^2, \eta \geq 0 \text{ a.e.} \quad (4.59)$$

for the operator  $Q$  defined in (4.56).

The proof of Proposition 4.21 is very technical and can be found in Section 4.4.4.1 on page 74.

Next, we provide a subsolution  $R_2$  satisfying  $\langle Q R_2, \eta \rangle \leq 0$  for all  $\eta$  from the same set of test functions.

**Proposition 4.22.** *Let  $\omega = B(0, 1)$  and let Assumption 4 hold. Then there exists  $\sigma > 1$  such that the function  $R_2 \in \mathcal{H}(\mathbb{R}^2) \cap L^\infty(\mathbb{R}^2)$  defined by*

$$R_2(x) := \begin{cases} k(U_0 \cdot x)|x|^{-\sigma} & x \in \mathbb{R}^2 \setminus \omega, \\ k(U_0 \cdot x) & x \in \omega, \end{cases} \quad (4.60)$$

with

$$k = \frac{\underline{\nu} - \nu_0}{\nu_0 + \underline{\nu}(\sigma - 1)} \in (-1, 0)$$

satisfies

$$\langle Q R_2, \eta \rangle \leq 0 \quad \forall \eta \in \mathcal{H}(\mathbb{R}^2) : \text{supp}(\eta) \subset \mathbb{R}_+^2, \eta \geq 0 \text{ a.e.} \quad (4.61)$$

for the operator  $Q$  defined in (4.56).

The technical proof of Proposition 4.22 can be found in Section 4.4.4.1 on page 80.

Next, we can show that there exists an element  $\tilde{H}$  of the class  $H$ , where  $H \in \mathcal{H}(\mathbb{R}^2)$  is the solution to (4.49), which has the same asymptotic behavior as  $R_1$  and  $R_2$  defined in (4.58) and (4.60), respectively, by means of a comparison principle.

**Proposition 4.23.** *Let  $\omega = B(0, 1)$  and let Assumption 4 hold. Let  $R_1$  the supersolution defined in Proposition 4.21,  $R_2$  the subsolution defined in Proposition 4.22 and  $H \in \mathcal{H}(\mathbb{R}^2)$  the unique solution to the operator equation  $QH = 0$  with  $Q$  defined in (4.56). Then there exists an element  $\tilde{H}$  of the class  $H$  such that*

$$R_2(x) \leq \tilde{H}(x) \leq R_1(x) \quad \forall x \in \mathbb{R}_+^2 \text{ a.e.} \quad (4.62)$$

The proof of Proposition 4.23 can be found in Section 4.4.4.1 on page 85.

Finally, collecting the results of Propositions 4.21, 4.22 and 4.23, we can state the following theorem:

**Theorem 4.24.** *Let  $\omega = B(0, 1)$ , let Assumption 4 hold and let  $H \in \mathcal{H}(\mathbb{R}^2)$  be the unique solution to the operator equation  $QH = 0$  with  $Q$  defined in (4.56). Then there exists an element  $\tilde{H}$  of the class  $H \in \mathcal{H}(\mathbb{R}^2)$  and  $\tau > 0$  such that*

$$\tilde{H}(y) = \mathcal{O}(|y|^{-\tau}) \text{ as } |y| \rightarrow \infty. \quad (4.63)$$

The proof of Theorem 4.24 can be found in Section 4.4.4.1 on page 86.

**Remark 4.25.** *As  $R_1$  and  $R_2$  are both in  $L^\infty(\mathbb{R}^2)$ , we also have that  $H \in L^\infty(\mathbb{R}^2)$ .*

From now on, the function  $H_\varepsilon$  is defined choosing  $\hat{H} = \tilde{H}$  in (4.50) where  $\tilde{H}$  is as in Theorem 4.24, i.e.,

$$H_\varepsilon(x) := \varepsilon \tilde{H}(\varepsilon^{-1}x), \quad x \in D.$$

#### 4.4.1.6 Estimates for the variations of the direct state

We show a technical lemma which we will need for the proof of Proposition 4.27. Let  $\theta : \mathbb{R}^2 \rightarrow \mathbb{R}$  be a smooth function such that

$$\theta(x) = 0, \quad x \in B(0, \rho), \quad \text{and} \quad \theta(x) = 1, \quad x \in \mathbb{R}^2 \setminus B(0, R), \quad (4.64)$$

where  $0 < \rho < R$  together with  $\lambda \in (0, 1]$  were defined in (4.4) in such a way that  $\lambda \omega \subset \subset B(0, \rho) \subset B(0, R) \subset \subset D \setminus \text{supp}(F)$ . Recall function  $H_\varepsilon$  defined by

$$H_\varepsilon(x) = \varepsilon \tilde{H}(\varepsilon^{-1}x), \quad x \in D.$$

Then let the function  $\kappa_\varepsilon : D \rightarrow \mathbb{R}$  be defined by

$$\kappa_\varepsilon(x) = \theta(x)H_\varepsilon(x). \quad (4.65)$$

**Lemma 4.26.** *It holds  $\kappa_\varepsilon \in H^1(D)$  and  $H_\varepsilon - \kappa_\varepsilon \in H_0^1(D)$ . Moreover, under the assumptions of Theorem 4.24, it holds*

$$\|\nabla \kappa_\varepsilon\|_{L^2(D)}^2 = o(\varepsilon^2). \quad (4.66)$$

The proof of Lemma 4.26 can be found in Section 4.4.4.1 on page 86.

**Proposition 4.27.** *Let  $\omega = B(0, 1)$  and let Assumptions 1, 4 and 5 hold. Then*

$$\|\nabla h_\varepsilon - \nabla H_\varepsilon\|_{L^2(D)}^2 = o(\varepsilon^2), \quad (4.67)$$

$$\forall \alpha > 0 \ \forall r \in (0, 1) : \int_{D \setminus B(0, \alpha \varepsilon^r)} |\nabla h_\varepsilon|^2 = o(\varepsilon^2), \quad (4.68)$$

$$\int_D |\nabla u_0 - U_0| |\nabla h_\varepsilon|^2 = o(\varepsilon^2), \quad (4.69)$$

$$\int_D |\nabla h_\varepsilon - \nabla H_\varepsilon| (|\nabla h_\varepsilon| + |\nabla H_\varepsilon|) = o(\varepsilon^2). \quad (4.70)$$

The proof of Proposition 4.27 can be found in Section 4.4.4.1 on page 87. Note that, in the proof of (4.67), we exploited the structure of the simplified setting introduced in Section 4.1.1.

**Proposition 4.28.** *Let  $\omega = B(0, 1)$  and let Assumptions 1, 4 and 5 hold. Then*

$$\|\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon\|_{L^2(D)}^2 = o(\varepsilon^2), \quad (4.71)$$

$$\int_D |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon|(|\nabla \tilde{u}_\varepsilon| + |\nabla h_\varepsilon|) = o(\varepsilon^2), \quad (4.72)$$

$$\forall \alpha > 0 \forall r \in (0, 1) : \int_{D \setminus B(0, \alpha \varepsilon^r)} |\nabla \tilde{u}_\varepsilon|^2 = o(\varepsilon^2). \quad (4.73)$$

The proof of Proposition 4.28 can be found in Section 4.4.4.1 on page 89.

#### 4.4.2 Variation of Adjoint State

For  $\varepsilon > 0$ , we introduce the perturbed adjoint equation to the PDE-constrained optimization problem (2.17) in the simplified setting of Section 4.1.1,

Find  $p_\varepsilon \in H_0^1(D)$  such that

$$\int_D DT_\varepsilon(x, \nabla u_0) \nabla p_\varepsilon \cdot \nabla \eta = -\langle \tilde{G}, \eta \rangle \quad \forall \eta \in H_0^1(D), \quad (4.74)$$

where  $T_\varepsilon$  is given in (4.5) and  $\tilde{G}$  fulfills Assumption 2 together with the objective function  $\mathcal{J}$ . Note that  $DT_\varepsilon$  is a symmetric matrix. For  $\varepsilon = 0$  we get the unperturbed adjoint equation,

Find  $p_0 \in H_0^1(D)$  such that

$$\int_D DT(\nabla u_0) \nabla p_0 \cdot \nabla \eta = -\langle \tilde{G}, \eta \rangle \quad \forall \eta \in H_0^1(D), \quad (4.75)$$

where we used that  $DT_0(x, \nabla u_0) = DT(\nabla u_0)$  according to the definition of  $T_\varepsilon$  (4.5). For  $\varepsilon > 0$ , we call  $p_\varepsilon$  the perturbed adjoint state, and  $p_0$  the unperturbed adjoint state. Note that we use the same right hand side  $\tilde{G}$ , independently of the parameter  $\varepsilon \geq 0$ .

##### 4.4.2.1 Regularity Assumptions

Similarly to Assumption 5, we also need the unperturbed adjoint state  $p_0$  to be sufficiently regular in a neighborhood of the point of the perturbation  $x_0 \in \Omega^d$ . We assume the following:

**Assumption 6.** There exists  $\tilde{\beta} > 0$  such that

$$p_0|_{\Omega^d} \in C^{1, \tilde{\beta}}(\Omega^d).$$

**Remark 4.29.** *If the right hand side of the adjoint equation (4.75) is a smooth function  $f \in C^{0, \gamma}(\overline{D})$  then the Hölder condition of Assumption 6 can be shown by [93, Theorem 8.34]. In our case, the right hand side is a distribution  $\tilde{G} \in H^{-1}(D)$  satisfying an expansion of the form (4.7). However, in the case of our model problem where we consider functionals  $\mathcal{J}$  of the type (2.15),  $\tilde{G}$  is not supported in  $\Omega^d$  and thus the assumption that  $p_0$  is smooth in  $\Omega^d$  is reasonable.*

Assumption 6 immediately yields the regularity properties

$$\nabla p_0|_{\Omega^d} \in C^{0, \tilde{\beta}}(\Omega^d), \quad (4.76)$$

$$\nabla p_0|_{\Omega^d} \in L^\infty(\Omega^d). \quad (4.77)$$

#### 4.4.2.2 Step 1: Variation $p_\varepsilon - p_0$

We proceed in an analogous way to Section 4.4.1.2. Subtracting (4.75) from (4.74), we get

$$\begin{aligned} 0 &= \int_D DT_\varepsilon(x, \nabla u_0) \nabla p_\varepsilon \cdot \nabla \eta - DT(\nabla u_0) \nabla p_0 \cdot \nabla \eta \\ &= \int_D DT_\varepsilon(x, \nabla u_0) \nabla \tilde{p}_\varepsilon \cdot \nabla \eta - DT(\nabla u_0) \nabla p_0 \cdot \nabla \eta + DT_\varepsilon(x, \nabla u_0) \nabla p_0 \cdot \nabla \eta. \end{aligned}$$

Thus we get the boundary value problem defining the variation of the adjoint state at scale  $\varepsilon$ ,  $\tilde{p}_\varepsilon := p_\varepsilon - p_0$ ,

Find  $\tilde{p}_\varepsilon \in H_0^1(D)$  such that

$$\int_D DT_\varepsilon(x, \nabla u_0) \nabla \tilde{p}_\varepsilon \cdot \nabla \eta = - \int_{\omega_\varepsilon} (\nu_0 I - DT(\nabla u_0)) \nabla p_0 \cdot \nabla \eta \quad \forall \eta \in H_0^1(D), \quad (4.78)$$

where we used that  $DT(\nabla u_0)$  and  $DT_\varepsilon(x, \nabla u_0)$  coincide in  $D \setminus \omega_\varepsilon$ .

#### 4.4.2.3 Step 2: Approximation of Variation $p_\varepsilon - p_0$

Analogously to Section 4.4.1.3, we approximate boundary value problem (4.78) by the same boundary value problem where the functions  $\nabla u_0, \nabla p_0$  are replaced by their values at the point  $x_0$ ,  $U_0 := \nabla u_0(x_0)$  and  $P_0 = \nabla p_0(x_0)$ , respectively. Again, note that this point evaluation makes sense due to Assumptions 5 and 6. We denote the solution to the arising boundary value problem by  $k_\varepsilon$  and get

Find  $k_\varepsilon \in H_0^1(D)$  such that

$$\int_D DT_\varepsilon(x, U_0) \nabla k_\varepsilon \cdot \nabla \eta = - \int_{\omega_\varepsilon} (\nu_0 I - DT(U_0)) P_0 \cdot \nabla \eta \quad \forall \eta \in H_0^1(D). \quad (4.79)$$

#### 4.4.2.4 Step 3: Change of Scale

Also here, we proceed analogously to the case of the variation of the direct state presented in 4.4.1.4. We perform a change of scale and then approximate boundary value problem (4.79) by sending the outer boundary to infinity, which yields the linear transmission problem

Find  $K \in \mathcal{H}(\mathbb{R}^2)$  such that

$$\int_{\mathbb{R}^2} D\tilde{T}(x, U_0) \nabla K \cdot \nabla \eta = - \int_{\omega} (\nu_0 I - DT(U_0)) P_0 \cdot \nabla \eta \quad \forall \eta \in \mathcal{H}(\mathbb{R}^2). \quad (4.80)$$

Note that (4.80) is independent of  $\varepsilon$ .

**Remark 4.30.** We remark that, for  $P_0 = (0, 0)^\top$ , problems (4.79) and (4.80) only admit the trivial solution such that  $K, \nabla K, k_\varepsilon, \nabla k_\varepsilon$  are identical zero. In this case, many computations simplify significantly. For the rest of this work, we exclude the trivial case and assume that  $P_0 \neq (0, 0)^\top$ .

It is straightforward to establish the well-posedness of problem (4.80):

**Lemma 4.31.** *Let Assumption 1 hold. Then, there exists a unique solution  $K \in \mathcal{H}(\mathbb{R}^2)$  to problem (4.80).*

*Proof.* We show existence and uniqueness of a solution  $H \in \mathcal{H}(\mathbb{R}^2)$  to (4.80) by the lemma of Lax-Milgram, see e.g. [47]. The coercivity and boundedness of the left hand side of (4.80) can be shown analogously to the proof of Lemma 2.8 exploiting (4.32) together with Remark 4.15, the physical properties (2.19) and the norm equivalence of Corollary 4.13. The right hand side of (4.80) is obviously a bounded linear functional on  $\mathcal{H}(\mathbb{R}^2)$ .  $\square$

#### 4.4.2.5 Step 4: Asymptotic Behavior of Variations of the Adjoint State

Let  $K \in \mathcal{H}(\mathbb{R}^2)$  be the unique solution in  $\mathcal{H}(\mathbb{R}^2)$  to (4.80) and let  $\hat{K} \in \mathcal{H}^w(\mathbb{R}^2)$  denote a given element of the class  $K$ . For  $\varepsilon > 0$ , let  $K_\varepsilon : D \rightarrow \mathbb{R}$  be defined by

$$K_\varepsilon(x) := \varepsilon \hat{K}(\varepsilon^{-1}x). \quad (4.81)$$

As in the case of the variation of the direct state, making the change of scale backwards, it follows from  $\hat{K} \in \mathcal{H}^w(\mathbb{R}^2)$  that  $K_\varepsilon \in H^1(D)$ , since

$$\inf_{x \in D} w\left(\frac{x}{\varepsilon}\right) > 0.$$

**Lemma 4.32.** *Let Assumptions 1, 3, 5 and 6. Then it holds*

$$\|\nabla \tilde{p}_\varepsilon\|_{L^2(D)}^2 = O(\varepsilon^2), \quad (4.82)$$

$$\|\nabla k_\varepsilon\|_{L^2(D)}^2 = O(\varepsilon^2), \quad (4.83)$$

$$\|\nabla K_\varepsilon\|_{L^2(D)}^2 = O(\varepsilon^2). \quad (4.84)$$

The proof of Lemma 4.32 can be found in Section 4.4.4.2 on page 93.

Next, we show an asymptotic behavior of an element of the class  $K \in \mathcal{H}(\mathbb{R}^2)$  similar to (4.63).

**Proposition 4.33.** *Let Assumption 1 hold and let  $K \in \mathcal{H}(\mathbb{R}^2)$  the unique solution to (4.80) according to Lemma 4.31. Then there exists an element  $\tilde{K}$  of the class  $K$  such that*

$$\tilde{K}(y) = \mathcal{O}(|y|^{-1}) \quad \text{as } |y| \rightarrow \infty. \quad (4.85)$$

The proof of Proposition 4.33 can be found in Section 4.4.4.2 on page 93.

Let, from now on, the function  $K_\varepsilon$  (4.81) be defined by choosing  $\hat{K} = \tilde{K}$  where  $\tilde{K}$  is the element of the class  $K \in \mathcal{H}(\mathbb{R}^2)$ , which satisfies the asymptotic behavior (4.85). Here,  $K \in \mathcal{H}(\mathbb{R}^2)$  is the unique solution to (4.80) according to Lemma 4.31. Recall the smooth function  $\theta$  defined in (4.64). Analogously to the function  $\kappa_\varepsilon$  in (4.65), let the function  $\kappa_{a\varepsilon} : D \rightarrow \mathbb{R}$  be defined by

$$\kappa_{a\varepsilon}(x) = \theta(x)K_\varepsilon(x). \quad (4.86)$$

**Lemma 4.34.** *Let Assumption 1 hold. It holds  $\kappa_{a\varepsilon} \in H^1(D)$  and  $K_\varepsilon - \kappa_{a\varepsilon} \in H_0^1(D)$ . Moreover, it holds*

$$\|\nabla \kappa_{a\varepsilon}\|_{L^2(D)}^2 = o(\varepsilon^2). \quad (4.87)$$

*Proof.* The proof is analogous to the proof of Lemma 4.26, exploiting the asymptotic behavior (4.85) of  $K$ . The only difference lies in the second step of the proof of Lemma 4.26 on page 87 as the asymptotic behavior of  $\tilde{H}$  (4.63), i.e.,  $\tilde{H} = \mathcal{O}(|y|^{-\tau})$  with  $\tau > 0$ , is different from that of  $\tilde{K}$  (4.85),  $\tilde{K} = \mathcal{O}(|y|^{-1})$ . Here, it suffices to replace the exponent  $\tau$  by 1 in the proof and the result (4.87) follows.  $\square$

**Lemma 4.35.** *Let Assumption 1 hold. Then, it holds*

$$\|\nabla k_\varepsilon - \nabla K_\varepsilon\|_{L^2(D)}^2 = o(\varepsilon^2), \quad (4.88)$$

and

$$\forall \alpha > 0 \forall r \in (0, 1) : \int_{D \setminus B(0, \alpha \varepsilon^r)} |\nabla k_\varepsilon|^2 = o(\varepsilon^2). \quad (4.89)$$

The proof of Lemma 4.35 can be found in Section 4.4.4.2 on page 94.

**Lemma 4.36.** *Let Assumptions 1, 3, 5 and 6 be satisfied. Then, it holds*

$$\|\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon\|_{L^2(D)}^2 = o(\varepsilon^2). \quad (4.90)$$

The proof of Lemma 4.36 can be found in Section 4.4.4.2 on page 96.

#### 4.4.3 Topological Asymptotic Expansion

Recall Assumption 2, i.e., that the cost function  $\mathcal{J}_\varepsilon : H_0^1(D) \rightarrow \mathbb{R}$  is of the form

$$\mathcal{J}_\varepsilon(u_\varepsilon) = \mathcal{J}_0(u_0) + \langle \tilde{G}, \tilde{u}_\varepsilon \rangle + \delta_J \varepsilon^2 + R(\varepsilon), \quad (4.91)$$

with  $\tilde{u}_\varepsilon = u_\varepsilon - u_0$ ,  $\tilde{G} \in H^{-1}(D)$  and where the remainder  $R(\varepsilon)$  is of the form (4.8). By estimate (4.73), it follows that

$$R(\varepsilon) = o(\varepsilon^2). \quad (4.92)$$

We have a closer look at the term  $\langle \tilde{G}, \tilde{u}_\varepsilon \rangle$ . Testing adjoint equation (4.74) for  $\varepsilon > 0$  with  $\eta = \tilde{u}_\varepsilon$  and exploiting the symmetry of  $DT_\varepsilon$ , we get

$$\begin{aligned} \langle \tilde{G}, \tilde{u}_\varepsilon \rangle &= - \int_D DT_\varepsilon(x, \nabla u_0) \nabla \tilde{u}_\varepsilon \cdot \nabla p_\varepsilon \\ &= - \int_D DT_\varepsilon(x, \nabla u_0) \nabla \tilde{u}_\varepsilon \cdot \nabla p_\varepsilon \\ &\quad + \int_D (T_\varepsilon(x, \nabla u_0 + \nabla \tilde{u}_\varepsilon) - T_\varepsilon(x, \nabla u_0)) \cdot \nabla p_\varepsilon + \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|\nabla u_0|)) \nabla u_0 \cdot \nabla p_\varepsilon, \end{aligned}$$

where we added the left and right hand side of (4.46) tested with  $\eta = p_\varepsilon$ . According to the definition of the operator  $S^\varepsilon$  (4.25), we get

$$\langle \tilde{G}, \tilde{u}_\varepsilon \rangle = \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|\nabla u_0|)) \nabla u_0 \cdot \nabla p_\varepsilon + \int_D S_{\nabla u_0}^\varepsilon(x, \nabla \tilde{u}_\varepsilon) \cdot \nabla p_\varepsilon.$$

Noting that  $p_\varepsilon = p_0 + \tilde{p}_\varepsilon$ , and defining

$$j_1(\varepsilon) := \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|\nabla u_0|)) \nabla u_0 \cdot (\nabla p_0 + \nabla \tilde{p}_\varepsilon), \quad (4.93)$$

$$j_2(\varepsilon) := \int_D S_{\nabla u_0}^\varepsilon(x, \nabla \tilde{u}_\varepsilon) \cdot (\nabla p_0 + \nabla \tilde{p}_\varepsilon), \quad (4.94)$$

together with (4.92), we get from (4.91) that

$$\mathcal{J}_\varepsilon(u_\varepsilon) - \mathcal{J}_0(u_0) = j_1(\varepsilon) + j_2(\varepsilon) + \delta_J \varepsilon^2 + o(\varepsilon^2). \quad (4.95)$$

Note that the operator  $S_{\nabla u_0}^\varepsilon$  represents the nonlinearity of the problem. Therefore, the term  $j_2$  vanishes in the linear case where the nonlinear function  $\hat{\nu}$  is replaced by a constant  $\nu_1$ .

In Sections 4.4.3.1 and 4.4.3.2 we will show that there exist numbers  $J_1, J_2$  such that

$$j_1(\varepsilon) = \varepsilon^2 J_1 + o(\varepsilon^2) \quad \text{and} \quad j_2(\varepsilon) = \varepsilon^2 J_2 + o(\varepsilon^2).$$

Comparing expansion (4.95) with (4.2) this will yield the final formula for the topological derivative,

$$G(x_0) = J_1 + J_2 + \delta_J$$

in Theorem 4.44.

#### 4.4.3.1 Expansion of Linear Term $j_1(\varepsilon)$

Following approximation steps 2 and 3 of Sections 4.4.2.3 and 4.4.2.4, respectively, we define

$$\tilde{j}_1(\varepsilon) := (\nu_0 - \hat{\nu}(|U_0|)) \int_{\omega_\varepsilon} U_0 \cdot (P_0 + \nabla k_\varepsilon), \quad (4.96)$$

$$J_1 := (\nu_0 - \hat{\nu}(|U_0|)) \int_\omega U_0 \cdot (P_0 + \nabla K). \quad (4.97)$$

**Lemma 4.37.** *Let Assumption 1 hold. Then it holds*

$$\tilde{j}_1(\varepsilon) - \varepsilon^2 J_1 = o(\varepsilon^2). \quad (4.98)$$

The proof of Lemma 4.37 is given in Section 4.4.4.3 on page 97.

**Lemma 4.38.** *Let Assumptions 1, 3, 5 and 6 hold. Then it holds*

$$j_1(\varepsilon) - \tilde{j}_1(\varepsilon) = o(\varepsilon^2). \quad (4.99)$$

The proof of Lemma 4.38 is given in Section 4.4.4.3 on page 98.

Considering (4.97), it follows from the linearity of equation (4.80) that the mapping

$$P_0 \mapsto (\nu_0 - \hat{\nu}(|U_0|)) \int_\omega (P_0 + \nabla K)$$

is linear from  $\mathbb{R}^2$  to  $\mathbb{R}^2$ . It only depends on the set  $\omega$ , and on the positive definite matrix  $DT(U_0)$ . Hence, there exists a matrix

$$\mathcal{M} = \mathcal{M}(\omega, DT(U_0)) \in \mathbb{R}^{2 \times 2}, \quad (4.100)$$

such that

$$(\nu_0 - \hat{\nu}(|U_0|)) \int_{\omega} (P_0 + \nabla K) = \mathcal{M} P_0.$$

This matrix  $\mathcal{M}$  is related to the well-studied concept of polarization matrices, see, e.g., [11], as will be explained in detail in Section 4.6. Eventually, it follows that

$$J_1 = U_0^\top \mathcal{M} P_0. \quad (4.101)$$

In Section 4.6.3.2, an explicit formula for the matrix  $\mathcal{M} = \mathcal{M}(\omega, DT(U_0))$  will be derived. Summing up estimates (4.98) and (4.99), as well as (4.101), we get the following result:

**Corollary 4.39.** *Let Assumptions 1, 3, 5 and 6 hold. Then, there exists a matrix  $\mathcal{M} = \mathcal{M}(\omega, DT(U_0)) \in \mathbb{R}^{2 \times 2}$  such that*

$$j_1(\varepsilon) = \varepsilon^2 U_0^\top \mathcal{M} P_0 + o(\varepsilon^2). \quad (4.102)$$

We will derive an explicit expression for the matrix  $\mathcal{M}$  in Section 4.6.3.

#### 4.4.3.2 Expansion of Nonlinear Term $j_2(\varepsilon)$

Recall the term  $j_2(\varepsilon)$  from (4.94),

$$j_2(\varepsilon) = \int_D S_{\nabla u_0}^\varepsilon(x, \nabla \tilde{u}_\varepsilon) \cdot (\nabla p_0 + \nabla \tilde{p}_\varepsilon).$$

According to the approximation steps taken for the variations of the direct and adjoint state in Sections 4.4.1.3, 4.4.1.4, as well as 4.4.2.3 and 4.4.2.4, we define

$$\tilde{j}_2(\varepsilon) := \int_D S_{U_0}^\varepsilon(x, \nabla h_\varepsilon) \cdot (P_0 + \nabla k_\varepsilon), \quad (4.103)$$

$$J_2 := \int_{\mathbb{R}^2} \tilde{S}_{U_0}(x, \nabla H) \cdot (P_0 + \nabla K). \quad (4.104)$$

Note that, under Assumption 3, both  $\tilde{j}_2(\varepsilon)$  and  $J_2$  are well-defined due to growth condition (4.41).

**Lemma 4.40.** *Let  $\omega = B(0, 1)$  and let Assumptions 1, 3, 4 and 5 hold. Then it holds*

$$\tilde{j}_2(\varepsilon) - \varepsilon^2 J_2 = o(\varepsilon^2). \quad (4.105)$$

The proof of Lemma 4.40 can be found in Section 4.4.4.3 on page 98.

**Lemma 4.41.** *Let  $\omega = B(0, 1)$  and let Assumptions 1, 4, 5 and 6 be satisfied. Then, it holds*

$$\int_D |\nabla p_0 - P_0| |\nabla h_\varepsilon|^2 = o(\varepsilon^2). \quad (4.106)$$

The proof of Lemma 4.41 can be found in Section 4.4.4.3 on page 100.

**Lemma 4.42.** *Let  $\omega = B(0, 1)$  and let Assumptions 1, 3, 4, 5 and 6 be satisfied. Then, it holds*

$$j_2(\varepsilon) - \tilde{j}_2(\varepsilon) = o(\varepsilon^2).$$

The proof of Lemma 4.42 can be found in Section 4.4.4.3 on page 100.

Eventually, combining Lemma 4.40 and Lemma 4.42, we get the following result:

**Corollary 4.43.** *Let  $\omega = B(0, 1)$  and let Assumptions 1, 3, 4, 5 and 6 be satisfied. Then we have*

$$j_2(\varepsilon) = \varepsilon^2 \left( \int_{\mathbb{R}^2} \tilde{S}_{U_0}(x, \nabla H) \cdot (P_0 + \nabla K) \right) + o(\varepsilon^2). \quad (4.107)$$

#### 4.4.3.3 Main Result

Finally, combining (4.95) with (4.102) and (4.107), we get the main result of this chapter, i.e., the topological derivative for the introduction of linear material (air) inside a region of nonlinear (ferromagnetic) material according to (4.2). We recall the notation needed for stating the result of Theorem 4.44:

- $x_0 \in \Omega^d$  denotes the point around which we perturb the material coefficient,
- $u_0 \in H_0^1(D)$  is the unperturbed direct state, i.e., the solution to (4.3), and  $U_0 = \nabla u_0(x_0)$ ,
- $p_0 \in H_0^1(D)$  is the unperturbed adjoint state, i.e., the solution to (4.75), and  $P_0 = \nabla p_0(x_0)$ ,
- $\mathcal{M} = \mathcal{M}(\omega, DT(U_0))$  denotes the matrix defined in (4.233) where  $\omega$  represents the shape of the inclusion and  $DT$  is the Jacobian of  $T$  defined in (2.24),
- $H \in \mathcal{H}(\mathbb{R}^2)$  denotes the variation of the direct state at scale 1, i.e., the solution to (4.49),
- $K \in \mathcal{H}(\mathbb{R}^2)$  denotes the variation of the adjoint state at scale 1, i.e., the solution to (4.80),
- $\tilde{S}$  is defined in (4.48),
- $\delta_J$  is according to (4.7).

**Theorem 4.44.** *Assume that*

- $\omega = B(0, 1)$  the unit disk in  $\mathbb{R}^2$
- the ferromagnetic material is such that Assumptions 1, 3 and 4 are satisfied,
- the functional  $\mathcal{J}_\varepsilon$  satisfies Assumption 2,
- the unperturbed direct state  $u_0$  satisfies Assumption 5, i.e.,  $u_0 \in C^{1,\beta}$  for some  $\beta > 0$ ,
- the unperturbed direct state  $p_0$  satisfies Assumption 6, i.e.,  $p_0 \in C^{1,\tilde{\beta}}$  for some  $\tilde{\beta} > 0$ .

Then the topological derivative for introducing air inside ferromagnetic material reads

$$\begin{aligned} G^{f \rightarrow \text{air}}(x_0) = & U_0^\top \mathcal{M} P_0 \\ & + \int_{\mathbb{R}^2} \tilde{S}_{U_0}(x, \nabla H) \cdot (P_0 + \nabla K) + \delta_J. \end{aligned} \quad (4.108)$$

**Remark 4.45.** The proof of Theorem 4.44 is valid only under the assumption that  $\omega = B(0, 1)$ . This is mainly due to the fact that the proof of Proposition 4.23 uses Lemma 4.20 which exploits the symmetry of  $\omega$  with respect to the line  $\{U_0 \cdot x = 0\}$ . Since we need to make sure that this condition is satisfied for any possible  $U_0$ , we have to assume that  $\omega$  is a disk in the sequel. Thus, the condition  $\omega = B(0, 1)$  could be relaxed to arbitrarily-shaped inclusions with  $C^2$  boundary if an asymptotic behavior of the form (4.63) can be guaranteed otherwise. Note that the proof of the asymptotic behavior of  $K$  in (4.85) is independent of the shape of  $\omega$ .

The second place where the shape of the inclusion influences the topological derivative is in the formula for the matrix  $\mathcal{M}(\omega, DT(U_0))$ . Here, an extension to ellipse-shaped inclusions is possible, cf. Remark 4.73 in Section 4.6.

#### 4.4.4 Proofs

##### 4.4.4.1 Proofs on the Variation of the Direct State

**Proof of Lemma 4.18:** This proof is following the lines of [20, 46].

*Proof.* 1. We begin by proving estimate (4.51). Testing the variational form (4.46) with test function  $\eta = \tilde{u}_\varepsilon \in H_0^1(D)$  yields

$$\int_D (T_\varepsilon(x, \nabla u_0 + \nabla \tilde{u}_\varepsilon) - T_\varepsilon(x, \nabla u_0)) \cdot \nabla \tilde{u}_\varepsilon = - \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|\nabla u_0|)) \nabla u_0 \cdot \nabla \tilde{u}_\varepsilon. \quad (4.109)$$

Property (4.33) together with Remark 4.9 yields that

$$c_2 |\nabla \tilde{u}_\varepsilon|^2 \leq (T_\varepsilon(x, \nabla u_0 + \nabla \tilde{u}_\varepsilon) - T_\varepsilon(x, \nabla u_0)) \cdot \nabla \tilde{u}_\varepsilon. \quad (4.110)$$

After (4.45), we have  $\nabla u_0|_{\Omega^d} \in L^\infty(\Omega^d)$ . Thus we define

$$M := \nu_0 \|\nabla u_0\|_{L^\infty(\Omega^d)}.$$

By the Cauchy-Schwarz inequality and (2.19a), the right hand side of (4.109) can be estimated as

$$\begin{aligned} \left| - \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|\nabla u_0|)) \nabla u_0 \cdot \nabla \tilde{u}_\varepsilon \right| & \leq \int_{\omega_\varepsilon} |\nu_0 - \hat{\nu}(|\nabla u_0|)| |\nabla u_0| |\nabla \tilde{u}_\varepsilon| \\ & \leq M \int_{\omega_\varepsilon} |\nabla \tilde{u}_\varepsilon| \leq M \left( \int_{\omega_\varepsilon} 1 \right)^{1/2} \left( \int_{\omega_\varepsilon} |\nabla \tilde{u}_\varepsilon|^2 \right)^{1/2} \\ & = M \varepsilon |\omega|^{1/2} \|\nabla \tilde{u}_\varepsilon\|_{L^2(D)} \end{aligned} \quad (4.111)$$

where we used that  $|\nu_0 - \hat{\nu}(|\nabla u_0|)| \leq \nu_0$  due to (2.19a). Combining (4.109), (4.110) and (4.111) and dividing by  $\|\nabla \tilde{u}_\varepsilon\|_{L^2(D)}$  and taking the square, we get

$$\|\nabla \tilde{u}_\varepsilon\|_{L^2(D)}^2 \leq C \varepsilon^2$$

with  $C = |\omega| M^2 / (c_2)^2$ .

2. The proof of estimate (4.52) is similar to the one of (4.51), starting from variational form (4.46).
3. Lastly, since  $\hat{H} \in \mathcal{H}^w(\mathbb{R}^2)$ , by definition it holds  $\nabla \hat{H} \in L^2(\mathbb{R}^2)$ . Thus, making a change of scale yields

$$\begin{aligned}\|\nabla H_\varepsilon\|_{L^2(D)}^2 &= \int_D |\nabla H_\varepsilon|^2 = \varepsilon^2 \int_{D/\varepsilon} |\nabla \hat{H}|^2 \\ &\leq \varepsilon^2 \int_{\mathbb{R}^2} |\nabla \hat{H}|^2 = \varepsilon^2 \|\nabla \hat{H}\|_{L^2(\mathbb{R}^2)}^2 = \mathcal{O}(\varepsilon^2),\end{aligned}$$

which completes the proof of estimate (4.53).  $\square$

### Proof of Lemma 4.20:

*Proof.* Let  $H$  be the unique solution to  $QH = 0$  with  $Q$  defined in (4.56) and  $\tilde{H}$  a representative of the class  $H$ . Consider the transformation

$$\begin{aligned}\phi(x_1, x_2) &:= (-x_1, x_2) \\ \text{with } J &= \nabla \phi(x_1, x_2) = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} = J^{-\top} \quad \text{and} \quad |\det J| = 1,\end{aligned}$$

and define the function  $\tilde{H}^s \in \mathcal{H}(\mathbb{R}^2)$  by

$$\tilde{H}^s(x_1, x_2) := -\tilde{H}(-x_1, x_2) = -(\tilde{H} \circ \phi)(x).$$

We show that also  $Q \tilde{H}^s = 0$ . Thanks to the symmetry of  $\omega$  with respect to the line  $\{x \in \mathbb{R}^2 : U_0 \cdot x = 0\}$ , we have  $\phi^{-1}(\mathbb{R}^2 \setminus \omega) = \mathbb{R}^2 \setminus \omega$  and  $\phi^{-1}(\omega) = \omega$ . Thus, we get

$$\begin{aligned}\langle Q \tilde{H}^s, \eta^s \rangle &= \int_{\omega} \left[ \nu_0 \left( U_0 + J^{-\top} \nabla_y \hat{\tilde{H}}^s(y) \right) - \hat{\nu}(|U_0|) U_0 \right] \cdot J^{-\top} \nabla_y \hat{\eta}^s(y) dy \\ &+ \int_{\mathbb{R}^2 \setminus \omega} \left[ \hat{\nu} \left( \left| U_0 + J^{-\top} \nabla_y \hat{\tilde{H}}^s(y) \right| \right) \left( U_0 + J^{-\top} \nabla_y \hat{\tilde{H}}^s(y) \right) - \hat{\nu}(|U_0|) U_0 \right] \cdot J^{-\top} \nabla_y \hat{\eta}^s(y) dy,\end{aligned}$$

where  $\hat{\tilde{H}}^s = \tilde{H}^s \circ \phi$  and  $\hat{\eta}^s = \eta^s \circ \phi$ . By definition of  $\tilde{H}^s$  and due to the fact that  $(\phi \circ \phi)(x) = x$ , we have that  $\hat{\tilde{H}}^s = -\tilde{H}$ . Since the basis was chosen such that  $e_1 = U_0/|U_0|$  and  $e_2 = e_1^\perp$ , it holds that

$$\begin{aligned}U_0 + J^{-\top} \nabla_y \hat{\tilde{H}}^s(y) &= \begin{pmatrix} |U_0| - \frac{\partial}{\partial y_1}(-\tilde{H}(y)) \\ 0 + \frac{\partial}{\partial y_2}(-\tilde{H}(y)) \end{pmatrix}, \\ \left| U_0 + J^{-\top} \nabla_y \hat{\tilde{H}}^s(y) \right| &= \left| U_0 + \nabla_y \tilde{H}(y) \right|, \\ \left( U_0 + J^{-\top} \nabla_y \hat{\tilde{H}}^s(y) \right) \cdot J^{-\top} \nabla_y \hat{\eta}^s(y) &= \left( U_0 + \nabla_y \tilde{H}(y) \right) \cdot \nabla_y (-\hat{\eta}^s(y)), \\ U_0 \cdot J^{-\top} \nabla_y \hat{\eta}^s(y) &= U_0 \cdot \nabla_y (-\hat{\eta}^s(y)).\end{aligned}$$

Using these relations, we conclude that

$$\begin{aligned}\langle Q \tilde{H}^s, \eta^s \rangle &= \int_{\mathbb{R}^2 \setminus \omega} \left[ \hat{\nu} \left( |U_0 + \nabla_y \tilde{H}(y)| \right) \left( U_0 + \nabla_y \tilde{H}(y) \right) - \hat{\nu}(|U_0|) U_0 \right] \cdot \nabla_y (-\hat{\eta}^s(y)) dy \\ &\quad + \int_{\omega} \left[ \nu_0 \left( U_0 + \nabla_y \tilde{H}(y) \right) - \hat{\nu}(|U_0|) U_0 \right] \cdot \nabla_y (-\hat{\eta}^s(y)) dy \\ &= \langle Q \tilde{H}, -\hat{\eta}^s \rangle.\end{aligned}$$

Since  $-\hat{\eta}^s \in \mathcal{H}(\mathbb{R}^2)$  for  $\eta^s \in \mathcal{H}(\mathbb{R}^2)$ , (4.57) yields  $\langle Q \tilde{H}, -\hat{\eta}^s \rangle = 0$  and therefore

$$\langle Q \tilde{H}^s, \eta^s \rangle = 0 \quad \text{for all } \eta^s \in \mathcal{H}(\mathbb{R}^2).$$

Thus, it follows from the uniqueness of a solution  $H \in \mathcal{H}(\mathbb{R}^2)$  to (4.49) established in Proposition 4.17 that also  $\tilde{H}^s$  is a representative of  $H$  and therefore  $\tilde{H}^s(x_1, x_2) = \tilde{H}(x_1, x_2) + C$  for all  $(x_1, x_2) \in \mathbb{R}^2$  where  $C$  is a constant. Restricted to the line  $\{x \in \mathbb{R}^2 : x_1 = 0\}$ , this yields that  $-2\tilde{H}(0, x_2) = C$  for all  $x_2 \in \mathbb{R}$ . Thus, choosing the representative  $\tilde{H}$  in such a way that  $\tilde{H}(0, 0) = 0$  yields that  $C = 0$  and thus  $\tilde{H}^s = \tilde{H}$ , which yields

$$\tilde{H}(-x_1, x_2) = -\tilde{H}(x_1, x_2)$$

for all  $(x_1, x_2) \in \mathbb{R}^2$ .  $\square$

**Proof of Proposition 4.21:** This proof follows the ideas of [20, 46], but requires some different calculations.

*Proof.* Let  $k \in (0, 1]$  be defined depending on the lower bound  $\delta_{\hat{\nu}} > -1/3$  from Assumption 4 as

$$k = \begin{cases} \frac{1}{2} \left( -\frac{1}{\delta_{\hat{\nu}}(1+\varepsilon)} - 3 \right) & -\frac{1}{3} < \delta_{\hat{\nu}} \leq -\frac{1}{5}, \\ 1 & \delta_{\hat{\nu}} > -\frac{1}{5}, \end{cases} \quad (4.112)$$

for some  $\varepsilon > 0$ . Furthermore, let  $\bar{\sigma}$  be defined as

$$\bar{\sigma} := \min\{\sigma_0, \sigma_1, \sigma_2\} > 1,$$

where  $\sigma_0$  and  $\sigma_1$  are given by

$$\begin{aligned}\sigma_0 &:= 2, \\ \sigma_1 &:= 1 + \frac{\underline{\nu}}{\nu_0 + \tilde{c}'|U_0|\sqrt{5}} \in (1, 2),\end{aligned} \quad (4.113)$$

with  $\underline{\nu}$  and  $\tilde{c}'$  given by (2.19) and (4.28), respectively, and  $\sigma_2$  is defined as the unique solution in  $(1, 2)$  to the equation  $\tilde{f}(t) = \delta_{\hat{\nu}}$  with  $\tilde{f}$  defined in (4.127) in the case where  $\delta_{\hat{\nu}} < 0$ , and  $\sigma_2 := 2$  else. We show that property (4.59) holds for  $k$  chosen according to (4.112) and any fixed  $\sigma$  satisfying

$$\sigma \in (1, \bar{\sigma}). \quad (4.114)$$

Let us first compute the first and second derivatives of the function  $R_1$  given in (4.58). We use the notation  $r = |x|$  and  $e_r = x/|x|$ . For  $x \in \mathbb{R}^2 \setminus \bar{\omega}$ , we have

$$\begin{aligned} R_1(x) &= k(U_0 \cdot x)r^{-\sigma}, \\ \nabla R_1(x) \cdot \varphi &= k r^{-\sigma}(U_0 - \sigma(U_0 \cdot e_r)e_r) \cdot \varphi, \\ D^2 R_1(x)(\varphi, \psi) &= k r^{-\sigma} \left( -\sigma \frac{1}{r^2} \left[ r(U_0 \cdot \psi) - (U_0 \cdot x) \frac{1}{r}(x \cdot \psi) \right] (e_r \cdot \varphi) \right. \\ &\quad \left. - \sigma(U_0 \cdot e_r) \frac{1}{r^2} \left[ r(\varphi \cdot \psi) - (\varphi \cdot x) \frac{1}{r}(x \cdot \psi) \right] \right) \\ &\quad - \sigma k r^{-\sigma-2} r(e_r \cdot \psi)(U_0 \cdot \varphi - \sigma(U_0 \cdot e_r)(e_r \cdot \varphi)), \end{aligned} \tag{4.115}$$

where we used that

$$\nabla(e_r \cdot \eta) = \nabla \left( \frac{x \cdot \eta}{|x|} \right) = \frac{1}{|x|^2} \left( |x|\eta - (x \cdot \eta) \frac{x}{|x|} \right)$$

for  $\eta \in \mathbb{R}^2$ . For  $\psi = \varphi$  we get

$$\begin{aligned} D^2 R_1(x)(\varphi, \varphi) &= \sigma k r^{-\sigma-2} \left( \begin{array}{ll} -r(U_0 \cdot \varphi)(e_r \cdot \varphi) & + (U_0 \cdot x)(e_r \cdot \varphi)^2 \\ -r(U_0 \cdot e_r)(\varphi \cdot \varphi) & + (U_0 \cdot x) \frac{1}{r} (\varphi \cdot x)^2 \frac{1}{r} \\ -r(e_r \cdot \varphi)(U_0 \cdot \varphi) & + r(e_r \cdot \varphi)^2 \sigma(U_0 \cdot e_r) \end{array} \right) \\ &= \sigma k r^{-\sigma-2} \left( (\sigma+2)(U_0 \cdot x)(e_r \cdot \varphi)^2 - 2(x \cdot \varphi)(U_0 \cdot \varphi) - (U_0 \cdot x)(\varphi \cdot \varphi) \right). \end{aligned} \tag{4.116}$$

In particular, we obtain

$$\begin{aligned} \Delta R_1(x) &= \sum_{i=1}^2 D^2 R_1(x)(e_i, e_i) \\ &= \sigma k r^{-\sigma-2} \left( (\sigma+2)(U_0 \cdot x) - 2(U_0 \cdot x) - 2(U_0 \cdot x) \right) \\ &= -\sigma k r^{-\sigma-2} (U_0 \cdot x)(2-\sigma) \end{aligned} \tag{4.117}$$

where  $e_i$  denotes the unit vector in Cartesian coordinates in direction  $x_i$ .

Integration by parts yields

$$\langle Q R_1, \eta \rangle = \langle Q_{int} R_1, \eta \rangle + \langle Q_{trans} R_1, \eta \rangle + \langle Q_{ext} R_1, \eta \rangle$$

with

$$\begin{aligned} \langle Q_{int} R_1, \eta \rangle &:= \int_{\omega} -\nu_0 \Delta R_1 \eta, \\ \langle Q_{trans} R_1, \eta \rangle &:= \int_{\partial\omega} [-\hat{\nu}(|U_0 + (\nabla R_1)_{ext}|)(U_0 + (\nabla R_1)_{ext}) + \nu_0(U_0 + (\nabla R_1)_{int})] \cdot n \eta, \\ \langle Q_{ext} R_1, \eta \rangle &:= \int_{\mathbb{R}^2 \setminus \omega} -\operatorname{div}(\hat{\nu}(|U_0 + \nabla R_1|)(U_0 + \nabla R_1)) \eta, \end{aligned}$$

where  $n$  denotes the unit normal vector pointing out of  $\omega$ .

Thus, we have that  $\langle Q R_1, \eta \rangle \geq 0$  for all  $\eta \in \mathcal{H}(\mathbb{R}^2)$  with  $\text{supp}(\eta) \subset \mathbb{R}_+^2$  such that  $\eta \geq 0$  almost everywhere, if (and only if) the following three conditions are satisfied:

$$-\nu_0 \Delta R_1 \geq 0 \quad \forall x \in \omega : U_0 \cdot x > 0, \quad (4.118)$$

$$\begin{aligned} [-\hat{\nu}(|U_0 + (\nabla R_1)_{ext}|)(U_0 + (\nabla R_1)_{ext}) \\ + \nu_0(U_0 + (\nabla R_1)_{int})] \cdot n \geq 0 \quad \forall x \in \partial\omega : U_0 \cdot x > 0, \end{aligned} \quad (4.119)$$

$$-\text{div}(\hat{\nu}(|U_0 + \nabla R_1|)(U_0 + \nabla R_1)) \geq 0 \quad \forall x \in \mathbb{R}^2 \setminus \bar{\omega} : U_0 \cdot x > 0. \quad (4.120)$$

1. The first condition (4.118) is satisfied by definition as  $R_1$  is linear inside  $\omega$  and thus  $\Delta R_1 = 0$  in  $\omega$ .
2. Next, we investigate the transmission condition (4.119). For  $x \in \partial\omega$ , we have

$$(\nabla R_1)_{int} = k U_0$$

and, by means of formula (4.115),

$$(\nabla R_1)_{ext} = k (U_0 - \sigma(U_0 \cdot x)x)$$

because  $r = |x| = 1$ . Using that

$$\begin{aligned} (U_0 + (\nabla R_1)_{int}) \cdot n &= (1+k)(U_0 \cdot x) > 0, \\ (-(\nabla R_1)_{ext} + (\nabla R_1)_{int}) \cdot n &= \sigma k (U_0 \cdot x) > 0, \end{aligned}$$

because  $n(x) = x$  for  $x \in \partial\omega$ , and exploiting the physical property (2.19a), we get

$$\begin{aligned} &(-\hat{\nu}(|U_0 + (\nabla R_1)_{ext}|)(U_0 + (\nabla R_1)_{ext}) + \nu_0(U_0 + (\nabla R_1)_{int})) \cdot n \\ &\geq (-\hat{\nu}(|U_0 + (\nabla R_1)_{ext}|)(U_0 + (\nabla R_1)_{ext}) + \hat{\nu}(|U_0 + (\nabla R_1)_{ext}|)(U_0 + (\nabla R_1)_{int})) \cdot n \\ &= \hat{\nu}(|U_0 + (\nabla R_1)_{ext}|)(-(\nabla R_1)_{ext} + (\nabla R_1)_{int}) \cdot n \\ &\geq \underline{\nu} \sigma k (U_0 \cdot x) > 0. \end{aligned}$$

for all  $x \in \partial\omega$  with  $(U_0 \cdot x) > 0$ .

3. Now, we consider the exterior condition (4.120). For better readability, we introduce the notation

$$\theta := \cos^{-1} \left( \frac{x}{|x|} \cdot \frac{U_0}{|U_0|} \right) = \cos^{-1} \left( e_r \cdot \frac{U_0}{|U_0|} \right),$$

as well as  $\tilde{\varphi} = \tilde{\varphi}(r, \theta) := U_0 + (\nabla R_1)_{ext}$ , and the auxiliary variables

$$\begin{aligned} q &= q(r) := \sigma k r^{-\sigma}, \\ d_0 &= d_0(r) := 1 + k r^{-\sigma}, \end{aligned} \quad (4.121)$$

$$d_1 = d_1(r) := 1 + k r^{-\sigma}(1 - \sigma) = d_0 - q, \quad (4.121)$$

$$d_2 = d_2(r, \theta) := 1 + k r^{-\sigma}(1 - \sigma \cos^2 \theta) = d_0 - q \cos^2 \theta, \quad (4.122)$$

$$s = s(r, \theta) := \sin^2 \theta d_0^2,$$

$$c = c(r, \theta) := \cos^2 \theta d_1^2,$$

$$d = d(r, \theta) := d_1 d_2.$$

Note that all of these symbols are actually functions of  $r$  and possibly  $\theta$ . For better presentation, we will drop these dependencies for the rest of the proof. It can be seen that

$$\begin{aligned} e_r \cdot \tilde{\varphi} &= |U_0| \cos \theta d_1, \\ x \cdot \tilde{\varphi} &= (U_0 \cdot x) d_1, \\ U_0 \cdot \tilde{\varphi} &= |U_0|^2 d_2, \\ |\tilde{\varphi}|^2 &= |U_0|^2 (s + c). \end{aligned} \quad (4.123)$$

Note that  $d_0$  and  $d_1$  are positive because of  $\sigma < \sigma_0 = 2$ ,  $k \in (0, 1]$  and  $r > 1$ . This implies that  $|\tilde{\varphi}| > 0$  since not both  $\sin \theta$  and  $\cos \theta$  can vanish at the same time. These relations, together with (4.116) and (4.117), yield that

$$\begin{aligned} -\operatorname{div}(\hat{\nu}(|U_0 + \nabla R_1|)(U_0 + \nabla R_1)) &= -\operatorname{div}(\hat{\nu}(|\tilde{\varphi}|)\tilde{\varphi}) \\ &= -\left(\hat{\nu}(|\tilde{\varphi}|)\Delta R_1 + \frac{\hat{\nu}'(|\tilde{\varphi}|)}{|\tilde{\varphi}|}\mathbf{D}^2 R_1(\tilde{\varphi}, \tilde{\varphi})\right) \\ &= -\sigma k r^{-\sigma-2}(U_0 \cdot x)\left(\hat{\nu}(|\tilde{\varphi}|)(\sigma-2) + \hat{\nu}'(|\tilde{\varphi}|)\frac{1}{|\tilde{\varphi}|}|U_0|^2 f(r, \theta)\right) \end{aligned}$$

with

$$f(r, \theta) = (\sigma+1)c - 2d - s. \quad (4.124)$$

Thus, condition (4.120) is satisfied if

$$\hat{\nu}(|\tilde{\varphi}|)(\sigma-2) + \hat{\nu}'(|\tilde{\varphi}|)\frac{1}{|\tilde{\varphi}|}|U_0|^2 f(r, \theta) \leq 0 \quad (4.125)$$

for all  $x \in \mathbb{R}^2 \setminus \bar{\omega}$  with  $U_0 \cdot x > 0$ , i.e., for all  $(r, \theta)$  with  $r > 1$ ,  $\cos \theta \in (0, 1)$ . We distinguish three different cases:

**Case 0:** The spatial coordinates  $(r, \theta)$  are such that  $\hat{\nu}'(|\tilde{\varphi}|) = 0$ :  
Condition (4.125) is satisfied since  $\sigma$  is smaller than  $\sigma_0 = 2$  due to (4.114), and since  $\hat{\nu}(|\tilde{\varphi}|) > 0$  by physical property (2.19a).

**Case 1:** The spatial coordinates  $(r, \theta)$  are such that  $\hat{\nu}'(|\tilde{\varphi}|) > 0$ :  
We insert (4.123) and (4.124) into the left hand side of (4.125) and get

$$\begin{aligned} \hat{\nu}(|\tilde{\varphi}|)(\sigma-2) + \hat{\nu}'(|\tilde{\varphi}|)\frac{1}{|\tilde{\varphi}|}|U_0|^2 f(r, \theta) &= \hat{\nu}(|\tilde{\varphi}|)(\sigma-2) + \hat{\nu}'(|\tilde{\varphi}|)|U_0|\frac{1}{\sqrt{s+c}}((\sigma+1)c - 2d - s) \\ &\leq \hat{\nu}(|\tilde{\varphi}|)(\sigma-2) + \hat{\nu}'(|\tilde{\varphi}|)|U_0|\frac{((\sigma-1)c-s)}{\sqrt{s+c}} \\ &= \hat{\nu}(|\tilde{\varphi}|)(\sigma-2) + \hat{\nu}'(|\tilde{\varphi}|)|U_0|\left(\frac{\sigma c}{\sqrt{s+c}} - \frac{c+s}{\sqrt{c+s}}\right) \\ &\leq \hat{\nu}(|\tilde{\varphi}|)(\sigma-2) + \hat{\nu}'(|\tilde{\varphi}|)|U_0|\left(\frac{\sigma(c+s)}{\sqrt{s+c}} - \sqrt{c+s}\right) \\ &= \hat{\nu}(|\tilde{\varphi}|)(\sigma-2) + \hat{\nu}'(|\tilde{\varphi}|)|U_0|(\sigma-1)\sqrt{c+s}, \end{aligned}$$

where we used that

$$-2d = -2d_1 d_2 \leq -2d_1^2 \leq -2d_1^2 \cos^2\theta = -2c.$$

So, (4.125) holds if  $\sigma$  satisfies

$$\begin{aligned} & \hat{\nu}(|\tilde{\varphi}|)(\sigma - 2) + \hat{\nu}'(|\tilde{\varphi}|)|U_0|(\sigma - 1)\sqrt{c+s} \leq 0 \\ \Leftrightarrow & \sigma(\hat{\nu}(|\tilde{\varphi}|) + \hat{\nu}'(|\tilde{\varphi}|)|U_0|\sqrt{c+s}) \leq 2\hat{\nu}(|\tilde{\varphi}|) + \hat{\nu}'(|\tilde{\varphi}|)|U_0|\sqrt{c+s} \\ \Leftrightarrow & \sigma \leq \frac{2\hat{\nu}(|\tilde{\varphi}|) + \hat{\nu}'(|\tilde{\varphi}|)|U_0|\sqrt{c+s}}{\hat{\nu}(|\tilde{\varphi}|) + \hat{\nu}'(|\tilde{\varphi}|)|U_0|\sqrt{c+s}} = 1 + \frac{\hat{\nu}(|\tilde{\varphi}|)}{\hat{\nu}(|\tilde{\varphi}|) + \hat{\nu}'(|\tilde{\varphi}|)|U_0|\sqrt{c+s}}. \end{aligned}$$

Since  $\sigma \in (1, \sigma_1)$  with  $\sigma_1$  defined in (4.113), the above inequality is satisfied because

$$\frac{\hat{\nu}(|\tilde{\varphi}|)}{\hat{\nu}(|\tilde{\varphi}|) + \hat{\nu}'(|\tilde{\varphi}|)|U_0|\sqrt{c+s}} \geq \frac{\nu}{\hat{\nu}(|\tilde{\varphi}|) + \hat{\nu}'(|\tilde{\varphi}|)|U_0|\sqrt{c+s}} \geq \frac{\nu}{\nu_0 + \tilde{c}'|U_0|\sqrt{5}},$$

where we used property (2.19a) as well as the facts that  $0 < \hat{\nu}'(|\tilde{\varphi}|)$  by assumption,  $\hat{\nu}'(|\tilde{\varphi}|) \leq \tilde{c}'$  by (4.28), and  $0 \leq c \leq 1$ ,  $0 \leq s \leq (1+k)^2 \leq 4$  for  $\sigma \in (1, 2)$  and  $k \in (0, 1)$  noting that  $r > 1$ . Thus, choosing  $\sigma$  according to (4.114), condition (4.120) is satisfied at points where  $\hat{\nu}'(|\tilde{\varphi}|) > 0$ .

**Case 2:** The spatial coordinates  $(r, \theta)$  are such that  $\hat{\nu}'(|\tilde{\varphi}|) < 0$ :

Case 2a:  $f(r, \theta) \geq 0$ : Condition (4.125) is satisfied since  $\sigma < \sigma_0 = 2$  due to (4.114) because both summands are non-positive.

Case 2b:  $f(r, \theta) < 0$ :

In this case, we must show that the positive contribution of the second summand on the left hand side of (4.125) is compensated by the negative first term. This is possible if Assumption 4 holds.

We introduce  $g(r, \theta) := (|U_0|^2/|\tilde{\varphi}|^2) f(r, \theta)$ , such that condition (4.125) can be rewritten as

$$\hat{\nu}(|\tilde{\varphi}|)(\sigma - 2) + \hat{\nu}'(|\tilde{\varphi}|)|\tilde{\varphi}| g(r, \theta) \leq 0,$$

and find a lower bound  $\underline{g}$  for  $g(r, \theta)$ ,

$$\underline{g} \leq g(r, \theta) < 0, \quad \forall r > 1, \theta \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right).$$

Then, since  $\hat{\nu}'(|\tilde{\varphi}|) < 0$ , it holds that

$$\hat{\nu}(|\tilde{\varphi}|)(\sigma - 2) + \hat{\nu}'(|\tilde{\varphi}|)|\tilde{\varphi}| g(r, \theta) \leq \hat{\nu}(|\tilde{\varphi}|)(\sigma - 2) + \hat{\nu}'(|\tilde{\varphi}|)|\tilde{\varphi}| \underline{g}$$

and (4.125) follows if the right hand side of the given estimate is non-positive. The condition that the right hand side of the expression above is non-positive is equivalent to the condition

$$\frac{\hat{\nu}'(|\tilde{\varphi}|)|\tilde{\varphi}|}{\hat{\nu}(|\tilde{\varphi}|)} \geq \frac{2 - \sigma}{\underline{g}}. \quad (4.126)$$

Let us now investigate the expression  $g(r, \theta)$  and find a bound  $\underline{g}$  from below. Using (4.123) and (4.124), we have

$$g(r, \theta) = \frac{(\sigma + 1) \cos^2 \theta d_1^2 - 2 d_1 d_2 - \sin^2 \theta d_0^2}{d_0^2 \sin^2 \theta + d_1^2 \cos^2 \theta} =: \frac{g_1}{g_2}.$$

Rewriting the nominator  $g_1$  in terms of  $d_0$  using (4.121) and (4.122), we get

$$\begin{aligned} g_1 &= (\sigma + 1) \cos^2 \theta (d_0^2 - 2qd_0 + q^2) - 2(d_0^2 - qd_0 - q \cos^2 \theta d_0 + q^2 \cos^2 \theta) - \sin^2 \theta d_0^2 \\ &= d_0^2 ((\sigma + 2) \cos^2 \theta - 3) + 2d_0 q(1 - \sigma \cos^2 \theta) + q^2 \cos^2 \theta (\sigma - 1). \end{aligned}$$

Similarly, we get for the denominator  $g_2$ ,

$$\begin{aligned} g_2 &= d_0^2 \sin^2 \theta + d_1^2 \cos^2 \theta \\ &= d_0^2 - 2d_0 q \cos^2 \theta + q^2 \cos^2 \theta. \end{aligned}$$

Note that  $g_2$  is positive and, therefore,  $g_1$  must be negative by the assumption of Case 2b. Together, we get

$$\begin{aligned} \frac{g_1}{g_2} &= \frac{d_0^2 ((\sigma + 2) \cos^2 \theta - 3) + 2d_0 q(1 - \sigma \cos^2 \theta) + q^2 \cos^2 \theta (\sigma - 1)}{d_0^2 - 2d_0 q \cos^2 \theta + q^2 \cos^2 \theta} \\ &= -\frac{g_2}{g_2} + \frac{1}{g_2} (d_0^2 ((\sigma + 2) \cos^2 \theta - 2) + 2d_0 q(1 - \cos^2 \theta (1 + \sigma)) + q^2 \cos^2 \theta \sigma) \\ &= -1 + \frac{1}{g_2} (\cos^2 \theta [d_0^2 (\sigma + 2) - 2(1 + \sigma) d_0 q + q^2 \sigma] - 2d_0 (d_0 - q)). \end{aligned}$$

Note that, for  $r > 1$  and  $\sigma > 1$ , we have

$$0 \leq r^{-\sigma} \leq 1 \quad \text{and} \quad 0 \leq \cos^2 \theta \leq 1,$$

and thus, for  $\sigma \in (1, \sigma_0)$  and  $k \in (0, 1]$ ,

$$1 \leq d_0 \leq 1 + k, \quad \text{and} \quad 0 \leq q \leq \sigma k, \quad \text{and} \quad 0 < 1 - k(\sigma - 1) \leq d_0 - q = d_1 \leq 1.$$

Hence, we can see that

$$d_0^2 (\sigma + 2) - (1 + \sigma) d_0 2q + q^2 \sigma = (d_0 - q)^2 + (\sqrt{\sigma} d_0 - \sqrt{\sigma} q)^2 + (d_0 + q)(d_0 - q) > 0,$$

and we can estimate

$$\frac{g_1}{g_2} \geq -1 - \frac{2d_0(d_0 - q)}{g_2} \geq -1 - \frac{2(1 + k)}{g_2}.$$

For the denominator  $g_2$ , it can be seen that

$$\begin{aligned} g_2 &= d_0^2 \sin^2 \theta + d_1^2 \cos^2 \theta \\ &\geq \sin^2 \theta + (1 - k(\sigma - 1))^2 \cos^2 \theta \\ &= 1 + \cos^2 \theta ((1 - k(\sigma - 1))^2 - 1) \\ &\geq 1 + ((1 - k(\sigma - 1))^2 - 1) = (1 - k(\sigma - 1))^2, \end{aligned}$$

because  $(1 - k(\sigma - 1))^2 - 1 < 0$  due to  $\sigma < \sigma_0 = 2$  and  $k \in (0, 1]$ , and thus

$$g(r, \theta) = \frac{g_1}{g_2} \geq -1 - \frac{2(1+k)}{(1-k(\sigma-1))^2} = -\frac{(1-k(\sigma-1))^2 + 2(1+k)}{(1-k(\sigma-1))^2} =: \underline{g}.$$

Note that  $\underline{g}$  depends on  $\sigma$  and  $k$ . For  $t \in (1, 2)$ , we define

$$\tilde{f}(t) := -\frac{(2-t)(1-k(t-1))^2}{(1-k(t-1))^2 + 2(1+k)}, \quad (4.127)$$

and note that  $\tilde{f}(\sigma) = (2-\sigma)/\underline{g}$ . If now  $\sigma$  satisfies that

$$\tilde{f}(\sigma) \leq \inf_{s>0} \frac{\hat{\nu}'(s)s}{\hat{\nu}(s)} = \delta_{\hat{\nu}},$$

then (4.126) is satisfied, which yields (4.125) and, therefore, (4.120) in Case 2b.

If  $\delta_{\hat{\nu}}$  is non-negative, this is satisfied because  $\tilde{f}(t) < 0$  for any  $t \in (1, 2)$  since  $k \in (0, 1]$ , and (4.126) and thus also (4.120) holds because  $\sigma < \sigma_2 = 2$  in this case.

In the case where  $\delta_{\hat{\nu}}$  is negative, recall that  $\delta_{\hat{\nu}} > -1/3$  by Assumption 4, so we have  $-1/3 < \delta_{\hat{\nu}} < 0$ . In (4.112), we defined  $k$  in such a way that  $\tilde{f}(1) = -1/(3+2k) = \delta_{\hat{\nu}}(1+\varepsilon) < \delta_{\hat{\nu}}$ . Since  $\tilde{f}(1) < \delta_{\hat{\nu}}$ ,  $\tilde{f}(2) = 0 > \delta_{\hat{\nu}}$  and since it can be seen that  $\tilde{f}$  is continuous and increasing in the interval  $(1, 2)$ , there exists a unique  $\sigma_2 \in (1, 2)$  such that  $\tilde{f}(\sigma_2) = \delta_{\hat{\nu}}$  and it holds that  $\tilde{f}(t) < \delta_{\hat{\nu}}$  for all  $t \in (1, \sigma_2)$ . Thus, if  $\sigma \in (1, \sigma_2)$ , inequality (4.126) is satisfied, which yields (4.125) and thus (4.120).

Hence, choosing  $\sigma$  and  $k$  according to (4.114) and (4.112), respectively, yields the statement of Proposition 4.21.  $\square$

**Proof of Proposition 4.22:** This proof follows the ideas of [20, 46], but requires some different calculations.

*Proof.* The proof is similar to the proof of Proposition 4.21. Again, we define  $\bar{\sigma}$  as

$$\bar{\sigma} := \min\{\hat{\sigma}_0, \hat{\sigma}_1, \hat{\sigma}_2\} > 1,$$

where  $\hat{\sigma}_0$  and  $\hat{\sigma}_1$  are given by

$$\begin{aligned} \hat{\sigma}_0 &:= 2, \\ \hat{\sigma}_1 &:= 1 + \frac{\nu}{\nu_0 + \tilde{c}'|U_0|\sqrt{5}} \in (1, 2). \end{aligned}$$

If the bound  $\delta_{\hat{\nu}}$  from Assumption 4 is non-negative, we define  $\hat{\sigma}_2 := 2$ . Otherwise, we define  $\hat{\sigma}_2 := \min\{\hat{\sigma}_2^{(i)}, \hat{\sigma}_2^{(ii)}\}$  where, for  $j \in \{i, ii\}$ ,  $\hat{\sigma}_2^{(j)}$  is the unique solution in  $(1, 2)$  to the equation  $\tilde{f}^{(j)}(t) = \delta_{\hat{\nu}}$  with  $\tilde{f}^{(j)}$  defined in (4.134) and (4.135), respectively.

We show that property (4.61) holds for any fixed  $\sigma$  satisfying

$$\sigma \in (1, \bar{\sigma}). \quad (4.128)$$

Similarly to the proof of Proposition 4.21, we have to show the three conditions

$$-\nu_0 \Delta R_2 \leq 0 \quad \forall x \in \omega : U_0 \cdot x > 0 \quad (4.129)$$

$$\begin{aligned} & [-\hat{\nu}(|U_0 + (\nabla R_2)_{ext}|)(U_0 + (\nabla R_2)_{ext}) \\ & \quad + \nu_0(U_0 + (\nabla R_2)_{int})] \cdot n \leq 0 \quad \forall x \in \partial\omega : U_0 \cdot x > 0 \end{aligned} \quad (4.130)$$

$$-\operatorname{div}(\hat{\nu}(|U_0 + \nabla R_2|)(U_0 + \nabla R_2)) \leq 0 \quad \forall x \in \mathbb{R}^2 \setminus \bar{\omega} : U_0 \cdot x > 0 \quad (4.131)$$

where  $n$  denotes the unit normal vector pointing out of  $\omega$ .

1. As in the proof of the Proposition 4.21, it is easily seen that the first condition (4.129) is trivially satisfied as  $\nabla R_2$  is linear inside  $\omega$ .
2. Next we consider the transmission condition (4.130). Exploiting that, for  $x \in \partial\omega$  with  $\omega = B(0, 1)$ , the outward unit vector  $n$  is equal to  $x$  and  $|x| = r = 1$ , and noting the formulas for the gradient of  $R_2$  inside and outside the inclusion  $\omega$ ,

$$\begin{aligned} (\nabla R_2)_{int} &= k U_0, \\ (\nabla R_2)_{ext} &= k r^{-\sigma} (U_0 - \sigma(U_0 \cdot e_r)e_r), \end{aligned}$$

we obtain

$$\begin{aligned} & (-\hat{\nu}(|U_0 + (\nabla R_2)_{ext}|)(U_0 + (\nabla R_2)_{ext}) + \nu_0(U_0 + (\nabla R_2)_{int})) \cdot n \\ & = (-\hat{\nu}(|U_0 + \nabla R_2|_{ext}|)(1 + k(1 - \sigma)) + \nu_0(1 + k))(U_0 \cdot x) \\ & \leq (-\underline{\nu}(1 + k(1 - \sigma)) + \nu_0(1 + k))(U_0 \cdot x) \\ & = \left( -\frac{\sigma \nu_0}{\nu_0 + \underline{\nu}(\sigma - 1)} \right) + \nu_0 \frac{\sigma \underline{\nu}}{\nu_0 + \underline{\nu}(\sigma - 1)} (U_0 \cdot x) = 0. \end{aligned}$$

In the estimation, we used that  $1 + k(1 - \sigma) > 0$  since  $k \in (-1, 0)$  and  $\sigma \in (1, \sigma_0)$ .

3. For the exterior condition (4.131), we need to verify that

$$-\operatorname{div}(\hat{\nu}(|U_0 + \nabla R_2|)(U_0 + \nabla R_2)) \leq 0 \quad \forall x \in \mathbb{R}^2 \setminus \bar{\omega} : U_0 \cdot x > 0.$$

Again, for better readability, we introduce the symbols

$$\theta := \cos^{-1} \left( \frac{x}{|x|} \cdot \frac{U_0}{|U_0|} \right) = \cos^{-1} \left( e_r \cdot \frac{U_0}{|U_0|} \right),$$

$$\begin{aligned} \hat{q} &= \hat{q}(r) := & k \sigma r^{-\sigma} \\ \hat{d}_0 &= \hat{d}_0(r) := & 1 + k r^{-\sigma}, \\ \hat{d}_1 &= \hat{d}_1(r) := & 1 + k r^{-\sigma}(1 - \sigma) & = \hat{d}_0 - \hat{q}, \\ \hat{d}_2 &= \hat{d}_2(r, \theta) := & 1 + k r^{-\sigma}(1 - \sigma \cos^2 \theta) & = \hat{d}_0 - \hat{q} \cos^2 \theta, \\ \hat{s} &= \hat{s}(r, \theta) := & \sin^2 \theta \hat{d}_0^2, \\ \hat{c} &= \hat{c}(r, \theta) := & \cos^2 \theta \hat{d}_1^2, \\ \hat{d} &= \hat{d}(r, \theta) := & \hat{d}_1 \hat{d}_2, \end{aligned}$$

and drop the dependencies on  $r$  and  $\theta$ . Note that, due to  $k < 0$ , the symbols introduced above are not the same as the corresponding symbols used in the proof of Proposition 4.21. Analogously to the proof of Proposition 4.21, we introduce the notation  $\hat{\varphi} = \hat{\varphi}(r, \theta) := U_0 + (\nabla R_2)_{ext}$  and get the relations

$$\begin{aligned} e_r \cdot \hat{\varphi} &= |U_0| \cos \theta \hat{d}_1, \\ x \cdot \hat{\varphi} &= (U_0 \cdot x) \hat{d}_1, \\ U_0 \cdot \hat{\varphi} &= |U_0|^2 \hat{d}_2, \\ |\hat{\varphi}|^2 &= |U_0|^2 (\hat{s} + \hat{c}). \end{aligned}$$

Again, we can deduce

$$\begin{aligned} -\operatorname{div}(\hat{\nu}(|U_0 + \nabla R_2|)(U_0 + \nabla R_2)) \\ &= -\left( \hat{\nu}(|\hat{\varphi}|)\Delta R_2 + \frac{\hat{\nu}'(|\hat{\varphi}|)}{|\hat{\varphi}|} D^2 R_2(\hat{\varphi}, \hat{\varphi}) \right) \\ &= \sigma k r^{-\sigma-2} (U_0 \cdot x) \left( \hat{\nu}(|\hat{\varphi}|)(\sigma-2) + \hat{\nu}'(|\hat{\varphi}|)|\hat{\varphi}| \frac{1}{|\hat{\varphi}|^2} |U_0|^2 \hat{f}(r, \theta) \right), \end{aligned}$$

with the function  $\hat{f}$  defined as

$$\hat{f}(r, \theta) = (\sigma+1)\hat{c} - 2\hat{d}_1\hat{d}_2 - \hat{s}.$$

Thus, since  $k < 0$ , it again suffices to show that

$$\hat{\nu}(|\hat{\varphi}|)(\sigma-2) + \hat{\nu}'(|\hat{\varphi}|) \frac{1}{|\hat{\varphi}|} |U_0|^2 \hat{f}(r, \theta) \leq 0 \quad (4.132)$$

for all  $x \in \mathbb{R}^2 \setminus \bar{\omega}$  with  $U_0 \cdot x > 0$ , i.e., for all  $(r, \theta)$  with  $r > 1$ ,  $\cos \theta \in (0, 1)$ . Again, we distinguish three different cases:

**Case 0:**  $\hat{\nu}'(|\hat{\varphi}|) = 0$ : Estimate (4.132) obviously holds for  $\sigma < \hat{\sigma}_0 = 2$  since  $\hat{\nu}(|\hat{\varphi}|) > 0$  by physical property (2.19a).

**Case 1:**  $\hat{\nu}'(|\hat{\varphi}|) > 0$ :

Also for  $k \in (-1, 0)$  and  $\sigma > 1$ , it holds that  $-2\hat{d}_1 \leq -2\hat{c}$  since  $d_1 > 0$  and

$$\begin{aligned} \hat{d}_2 &= 1 + k r^{-\sigma} - \sigma k r^{-\sigma} \cos^2 \theta \geq (1 + k r^{-\sigma}) \cos^2 \theta - \sigma k r^{-\sigma} \cos^2 \theta \\ &= \cos^2 \theta \hat{d}_1, \end{aligned}$$

because  $r > 1$ . Thus, we can perform the analogous estimations as in the proof of Proposition 4.21 on page 77. Using that, for  $\sigma \in (1, 2)$  and  $k \in (-1, 0)$ , we have

$$0 \leq \hat{c} \leq (1-k)^2 \leq 4, \quad 0 \leq \hat{s} \leq 1, \quad \text{and} \quad 0 \leq \sqrt{\hat{s} + \hat{c}} \leq \sqrt{5},$$

we can define

$$\hat{\sigma}_1 := 1 + \frac{\nu}{\nu_0 + \tilde{c}'|U_0|\sqrt{5}} \in (1, 2)$$

and again conclude that (4.132) and thus (4.131) hold since  $\sigma$  was chosen according to (4.128).

**Case 2:**  $\hat{\nu}'(|\hat{\varphi}|) < 0$ :

Case 2a:  $\hat{f}(r, \theta) \geq 0$ : Estimate (4.132) holds for any  $\sigma \in (1, 2)$  because both summands are non-positive.

Case 2b:  $\hat{f}(r, \theta) < 0$ :

Analogously to the proof of Proposition 4.21, we can introduce

$$g^{(2)}(r, \theta) := \frac{|U_0|^2}{|\hat{\varphi}|^2} \hat{f}(r, \theta) = \frac{1}{\hat{s} + \hat{c}} \hat{f}(r, \theta),$$

and rewrite condition (4.132) can as

$$\hat{\nu}(|\hat{\varphi}|)(\sigma - 2) + \hat{\nu}'(|\hat{\varphi}|)|\hat{\varphi}|g^{(2)}(r, \theta) \leq 0.$$

Again, we have to find a lower bound on the expression

$$g^{(2)}(r, \theta) = \frac{(\sigma + 1) \cos^2 \theta \hat{d}_1^2 - 2 \hat{d}_1 \hat{d}_2 - \sin^2 \theta \hat{d}_0^2}{\hat{d}_0^2 \sin^2 \theta + \hat{d}_1^2 \cos^2 \theta} =: \frac{g_1^{(2)}}{g_2^{(2)}}$$

which satisfies condition (4.126). The manipulations of the terms  $g_1^{(2)}$  and  $g_2^{(2)}$  are analogous to the proof of Proposition 4.21 and we arrive at the corresponding expression

$$\begin{aligned} \frac{g_1^{(2)}}{g_2^{(2)}} &= -1 + \frac{1}{g_2^{(2)}} \left[ \cos^2 \theta \left( \hat{d}_0^2 (\sigma + 2) - 2(1 + \sigma) \hat{d}_0 \hat{q} + \hat{q}^2 \sigma \right) - 2\hat{d}_0 (\hat{d}_0 - \hat{q}) \right] \\ &= -1 + \frac{1}{g_2^{(2)}} \left[ \cos^2 \theta \left( (\hat{d}_0 - \hat{q})^2 + (\sqrt{\sigma} \hat{d}_0 - \sqrt{\sigma} \hat{q})^2 + (\hat{d}_0 + \hat{q})(\hat{d}_0 - \hat{q}) \right) \right. \\ &\quad \left. - 2\hat{d}_0 (\hat{d}_0 - \hat{q}) \right]. \end{aligned}$$

Again, we will estimate this expression from below such that we can extract a condition on  $\sigma > 1$  that is sufficient for (4.132). For the estimation, we will use that, for  $r > 1$ ,  $\sigma \in (1, \hat{\sigma}_0)$  and  $k \in (-1, 0)$ , we have

$$\begin{aligned} \sigma k &\leq \hat{q} \leq 0, \\ 1 + k &\leq \hat{d}_0 \leq 1, \\ 1 &\leq \hat{d}_0 - \hat{q} = \hat{d}_1 \leq 1 + k(1 - \sigma), \\ 1 + k(1 + \sigma) &\leq \hat{d}_0 + \hat{q} \leq 1. \end{aligned}$$

Note that, for the denominator  $g_2^{(2)}$ , we have

$$\begin{aligned} g_2^{(2)} &= \hat{d}_0^2 \sin^2 \theta + \hat{d}_1^2 \cos^2 \theta \geq (1 + k)^2 \sin^2 \theta + \cos^2 \theta \\ &= \cos^2 \theta (1 - (1 + k)^2) + (1 + k)^2 \geq (1 + k)^2. \end{aligned} \tag{4.133}$$

For the estimation of  $g_1^{(2)}/g_2^{(2)}$ , we need to distinguish two more cases:

Case 2b (i): The spatial coordinates  $(r, \theta)$  are such that  $\hat{d}_0 + \hat{q} \geq 0$ :

In this case, recalling that  $g_2^{(2)} > 0$  since  $-1 < k < 0$ , we can estimate the above expression from below by dropping the positive cosine term and, taking into account (4.133), we get

$$\begin{aligned} \frac{g_1^{(2)}}{g_2^{(2)}} &\geq -1 - \frac{2\hat{d}_0(\hat{d}_0 - \hat{q})}{g_2^{(2)}} \geq -1 - \frac{2(1+k(1-\sigma))}{g_2^{(2)}} \\ &\geq -1 - \frac{2(1+k(1-\sigma))}{(1+k)^2} = \frac{-(1+k)^2 - 2(1+k(1-\sigma))}{(1+k)^2} =: \underline{g}^{(i)}. \end{aligned}$$

Case 2b (ii): The spatial coordinates  $(r, \theta)$  are such that  $\hat{d}_0 + \hat{q} < 0$ :

In this case, we get the estimate

$$\begin{aligned} \frac{g_1^{(2)}}{g_2^{(2)}} &\geq -1 + \frac{\cos^2\theta(\hat{d}_0 + \hat{q})(\hat{d}_0 - \hat{q}) - 2\hat{d}_0(\hat{d}_0 - \hat{q})}{g_2^{(2)}} \\ &\geq -1 + \frac{(\hat{d}_0 + \hat{q})(\hat{d}_0 - \hat{q}) - 2\hat{d}_0(\hat{d}_0 - \hat{q})}{g_2^{(2)}} = -1 + \frac{-(\hat{d}_0 - \hat{q})^2}{g_2^{(2)}} \\ &\geq -1 - \frac{(1+k(1-\sigma))^2}{g_2^{(2)}} \geq \frac{-(1+k)^2 - (1+k(1-\sigma))^2}{(1+k)^2} =: \underline{g}^{(ii)}. \end{aligned}$$

Recall that  $\sigma$  and  $k$  are fixed numbers only depending on the given data  $\underline{\nu}$ ,  $\nu_0$  and  $\delta_{\hat{\nu}}$ , and on  $|U_0|$ . In order to show that the condition  $\sigma < \hat{\sigma}_2$  yields (4.131) in Case 2, we introduce the functions

$$\hat{k}(t) := \frac{\underline{\nu} - \nu_0}{\nu_0 + \underline{\nu}(t-1)}, \quad (4.134)$$

$$\tilde{f}^{(i)}(t) := -\frac{(2-t)(1+\hat{k}(t))^2}{(1+\hat{k}(t))^2 + 2(1+\hat{k}(t)(1-t))}, \quad (4.134)$$

$$\tilde{f}^{(ii)}(t) := -\frac{(2-t)(1+\hat{k}(t))^2}{(1+\hat{k}(t))^2 + (1+\hat{k}(t)(1-t))^2}, \quad (4.135)$$

mapping from  $[1, 2]$  to  $\mathbb{R}$ . Note that  $\hat{k}(\sigma) = k$  and, for  $j \in \{i, ii\}$ , it holds  $\tilde{f}^{(j)}(\sigma) = (2-\sigma)/\underline{g}^{(j)}$ . It can be seen that

$$\inf_{t \in (1,2)} \hat{f}^{(i)}(t) = -\frac{(1+\hat{k}(1))^2}{(1+\hat{k}(1))^2 + 2} = \hat{f}^{(i)}(1) < 0 \quad \text{and}$$

$$\inf_{t \in (1,2)} \hat{f}^{(ii)}(t) = -\frac{(1+\hat{k}(1))^2}{(1+\hat{k}(1))^2 + 1} = \hat{f}^{(ii)}(1) < 0.$$

Note that, for  $j \in \{i, ii\}$ , we have that  $\hat{f}^{(j)}(1) < 0$  and  $\hat{f}^{(j)}(2) = 0$ , and it can be seen that  $\hat{f}^{(j)}$  is continuous and increasing in the interval  $(1, 2)$ . Thus, in the case where  $\delta_{\hat{\nu}} < 0$ , by the same arguments as in the proof of Proposition 4.21, under Assumption 4, we can define  $\hat{\sigma}_2^{(j)}$  such that  $\hat{f}^{(j)}(\hat{\sigma}) < \delta_{\hat{\nu}}$  for all  $\hat{\sigma} \in (1, \hat{\sigma}_2^{(j)})$ , which yields (4.132) and thus (4.131). If  $\delta_{\hat{\nu}}$  is non-negative, we set  $\sigma_2^{(i)} = \sigma_2^{(ii)} = 2$  like before. Setting  $\hat{\sigma}_2 = \min\{\hat{\sigma}_2^{(i)}, \hat{\sigma}_2^{(ii)}\}$  completes the proof of Case 2b.

Again, the overall statement of Proposition 4.22 follows because  $\sigma \in (1, \min\{\hat{\sigma}_0, \hat{\sigma}_1, \hat{\sigma}_2\})$ .  $\square$

**Proof of Proposition 4.23:** This proof follows the lines of [20, 46] and is adapted to our case here.

*Proof.* By assumption it holds  $\omega = B(0, 1)$ . Recall that, due to Lemma 4.20, there exists an element  $\tilde{H}$  of the class  $H$  such that

$$\tilde{H}(x) = 0 \quad \forall x \in \mathbb{R}^2 : U_0 \cdot x = 0.$$

We begin by showing the upper bound for  $\tilde{H}$  in (4.62). Let  $R_1$  the supersolution defined in Proposition 4.21. For all  $\eta \in \mathcal{H}(\mathbb{R}^2)$  with  $\text{supp}(\eta) \subset \mathbb{R}_+^2$  and  $\eta \geq 0$  almost everywhere, it holds that

$$\langle QR_1, \eta \rangle \geq 0.$$

As by definition of  $H$ , it holds  $QH = 0$ , it follows that

$$\langle QR_1 - QH, \eta \rangle \geq 0,$$

that is

$$\int_{\mathbb{R}^2} [\tilde{T}(x, U_0 + \nabla R_1) - \tilde{T}(x, U_0 + \nabla H)] \cdot \nabla \eta \geq 0. \quad (4.136)$$

Let  $\chi_{\mathbb{R}_+^2}$  denote the characteristic function of the half-space  $\mathbb{R}_+^2$ , i.e.,

$$\chi_{\mathbb{R}_+^2}(x) = \begin{cases} 1 & \text{if } x \in \mathbb{R}_+^2, \\ 0 & \text{else,} \end{cases}$$

for all  $x$  in  $\mathbb{R}^2$ . As  $R_1 = \tilde{H} = 0$  in the hyperplane  $(\mathbb{R}U_0)^\perp$ , the test function defined by

$$\eta(x) := \chi_{\mathbb{R}_+^2}(x) \max(0, \tilde{H}(x) - R_1(x)), \quad x \in \mathbb{R}^2$$

satisfies the conditions  $\eta \in \mathcal{H}(\mathbb{R}^2)$ ,  $\text{supp}(\eta) \subset \mathbb{R}_+^2$  and  $\eta \geq 0$  almost everywhere. Hence, it can be plugged into inequality (4.136), which yields

$$\int_{\{\tilde{H} > R_1\} \cap \mathbb{R}_+^2} [\tilde{T}(x, U_0 + \nabla R_1) - \tilde{T}(x, U_0 + \nabla H)] \cdot (\nabla H - \nabla R_1) \geq 0,$$

that is

$$\int_{\{\tilde{H} > R_1\} \cap \mathbb{R}_+^2} [\tilde{T}(x, U_0 + \nabla R_1) - \tilde{T}(x, U_0 + \nabla H)] \cdot (\nabla R_1 - \nabla H) \leq 0. \quad (4.137)$$

Moreover, after ellipticity condition (4.33), there exists  $c_2 > 0$  such that

$$\begin{aligned} 0 &\leq c_2 \int_{\{\tilde{H} > R_1\} \cap \mathbb{R}_+^2} |\nabla H - \nabla R_1|^2 \\ &\leq \int_{\{\tilde{H} > R_1\} \cap \mathbb{R}_+^2} [\tilde{T}(x, U_0 + \nabla R_1) - \tilde{T}(x, U_0 + \nabla H)] \cdot (\nabla R_1 - \nabla H). \end{aligned}$$

Combining this estimate with (4.137), we conclude

$$0 = \int_{\{\tilde{H} > R_1\} \cap \mathbb{R}_+^2} |\nabla H - \nabla R_1|^2 = \int_{\mathbb{R}^2} |\nabla \eta|^2.$$

Due to the Poincaré inequality stated in Lemma 4.12, it follows that  $\eta = 0$  in  $\mathcal{H}(\mathbb{R}^2)$ . Hence,  $R_1 \geq \tilde{H}$  almost everywhere in  $\mathbb{R}_+^2$ .

Analogously, one obtains from Proposition 4.22 that  $\tilde{H} \geq R_2$  almost everywhere on  $\mathbb{R}_+^2$ .  $\square$

#### Proof of Theorem 4.24:

*Proof.* From Proposition 4.23 we have that

$$R_2(x) \leq \tilde{H}(x) \leq R_1(x)$$

for almost every  $x \in \mathbb{R}_+^2$ . Since  $\tilde{H}$  as well as  $R_1$  and  $R_2$  are odd functions with respect to the first coordinate  $x_1$ , it follows immediately that

$$R_1(x) \leq \tilde{H}(x) \leq R_2(x)$$

for almost every  $x \in \mathbb{R}^2 \setminus \mathbb{R}_+^2$ . Let  $\sigma_{R_1}$  denote the value of the parameter  $\sigma$  in Proposition 4.21 and  $\sigma_{R_2}$  the value of  $\sigma$  in Proposition 4.22. Then it holds

$$\begin{aligned} R_1(y) &= \mathcal{O}(|y|^{-\tau}) \quad \text{as } |y| \rightarrow \infty \text{ and} \\ R_2(y) &= \mathcal{O}(|y|^{-\tau}) \quad \text{as } |y| \rightarrow \infty, \end{aligned}$$

with  $\tau := \min\{\sigma_{R_1}, \sigma_{R_2}\} - 1 > 0$ , and we conclude that

$$\tilde{H}(y) = \mathcal{O}(|y|^{-\tau}) \quad \text{as } |y| \rightarrow \infty.$$

$\square$

**Proof of Lemma 4.26:** This proof is following the lines of [20, 46] and is given here for sake of completeness.

*Proof.* Denote

$$C_\theta := \max(\|\theta(x)\|_{L^\infty(\mathbb{R}^2)}, \|\nabla \theta(x)\|_{L^\infty(\mathbb{R}^2)}) < \infty.$$

According to the Leibniz formula, it holds

$$\nabla \kappa_\varepsilon(x) = \nabla \theta(x) H_\varepsilon(x) + \theta(x) \nabla H_\varepsilon(x) \quad \text{for a.e. } x \in D.$$

Therefore, it follows that

$$|\nabla \kappa_\varepsilon(x)|^2 \leq 2C_\theta^2 (|H_\varepsilon(x)|^2 + |\nabla H_\varepsilon(x)|^2).$$

Since  $H_\varepsilon \in H^1(D)$ , it follows  $\kappa_\varepsilon \in H^1(D)$ .

Moreover, by definition of  $\theta$ , it holds that  $H_\varepsilon - \kappa_\varepsilon = 0$  on  $\partial D$ , and thus, according to the trace theorem in  $H_0^1(D)$ , it follows  $H_\varepsilon - \kappa_\varepsilon \in H_0^1(D)$ .

Let us now prove estimate (4.66).

1. In  $B(0, \rho)$ , it holds  $\theta \equiv 0$ . Thus

$$\int_{B(0, \rho)} |\nabla \kappa_\varepsilon|^2 = 0.$$

2. Integrating in  $B(0, R) \setminus B(0, \rho)$ , according to the asymptotic behavior of  $\tilde{H}$  given by (4.63) and since  $\nabla H \in L^2(\mathbb{R}^2)$  one obtains

$$\begin{aligned} & \frac{1}{2C_\theta^2} \int_{B(0, R) \setminus B(0, \rho)} |\nabla \kappa_\varepsilon|^2 \\ & \leq \int_{B(0, R) \setminus B(0, \rho)} |H_\varepsilon|^2 && + \int_{B(0, R) \setminus B(0, \rho)} |\nabla H_\varepsilon|^2 \\ & \leq \varepsilon^4 \int_{B(0, R/\varepsilon) \setminus B(0, \rho/\varepsilon)} |\tilde{H}|^2 && + \varepsilon^2 \int_{B(0, R/\varepsilon) \setminus B(0, \rho/\varepsilon)} |\nabla H|^2 \\ & \leq \varepsilon^4 \int_{B(0, R/\varepsilon) \setminus B(0, \rho/\varepsilon)} |x|^{-2\tau} && + \varepsilon^2 \int_{\mathbb{R}^2 \setminus B(0, \rho/\varepsilon)} |\nabla H|^2 \\ & \leq \varepsilon^4 \left(\frac{\varepsilon}{R}\right)^{2\tau} \int_{B(0, R/\varepsilon) \setminus B(0, \rho/\varepsilon)} 1 && + \varepsilon^2 \int_{\mathbb{R}^2 \setminus B(0, \rho/\varepsilon)} |\nabla H|^2 \\ & = \varepsilon^4 \mathcal{O}(\varepsilon^{2\tau}) \mathcal{O}(\varepsilon^{-2}) && + \varepsilon^2 \int_{\mathbb{R}^2 \setminus B(0, \rho/\varepsilon)} |\nabla H|^2 \\ & = o(\varepsilon^2), \end{aligned}$$

where  $\tau > 0$  is as in Theorem 4.24. In the last step we used the fact that the last integral approaches zero for  $\varepsilon \rightarrow 0$ .

3. Lastly it holds  $\kappa_\varepsilon = H_\varepsilon$  in  $D \setminus B(0, R)$ . Again,  $\nabla H \in L^2(\mathbb{R}^2)$  and thus

$$\int_{D \setminus B(0, R)} |\nabla \kappa_\varepsilon|^2 \leq \varepsilon^2 \int_{\mathbb{R}^2 \setminus B(0, R/\varepsilon)} |\nabla H|^2 = o(\varepsilon^2).$$

Gathering these three results, one obtains the claimed estimate (4.66).  $\square$

**Proof of Proposition 4.27:** This proof is following the lines of [20, 46] and is adapted to our case here.

*Proof.* 1. We begin with estimate (4.67). For all  $\eta \in H_0^1(D)$ , we define  $\eta_\varepsilon \in \mathcal{H}(\mathbb{R}^2)$  by  $\eta_\varepsilon := \varepsilon^{-1} \eta(\varepsilon y)$  for all  $y \in D/\varepsilon$  and  $\eta_\varepsilon(y) := 0$  for all  $y \in \mathbb{R}^2 \setminus (D/\varepsilon)$ . Applying variational formulation (4.49) to  $\eta_\varepsilon$  and making the change of scale backwards, one obtains

$$\int_D [T_\varepsilon(x, U_0 + \nabla H_\varepsilon) - T_\varepsilon(x, U_0)] \cdot \nabla \eta = - \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|U_0|)) U_0 \cdot \nabla \eta.$$

Calculating the difference with variational form (4.47) yields

$$\int_D [T_\varepsilon(x, U_0 + \nabla h_\varepsilon) - T_\varepsilon(x, U_0 + \nabla H_\varepsilon)] \cdot \nabla \eta = 0 \quad (4.138)$$

Recall function  $\kappa_\varepsilon$  studied in Lemma 4.26. Since  $H_\varepsilon - \kappa_\varepsilon \in H_0^1(D)$ , it holds that  $\eta = h_\varepsilon - (H_\varepsilon - \kappa_\varepsilon) \in H_0^1(D)$ . Plugging in this  $\eta$  in (4.138), one obtains

$$\begin{aligned} & \int_D [T_\varepsilon(x, U_0 + \nabla h_\varepsilon) - T_\varepsilon(x, U_0 + \nabla H_\varepsilon)] \cdot \nabla(h_\varepsilon - H_\varepsilon) \\ &= - \int_D [T_\varepsilon(x, U_0 + \nabla h_\varepsilon) - T_\varepsilon(x, U_0 + \nabla H_\varepsilon)] \cdot \nabla \kappa_\varepsilon. \end{aligned} \quad (4.139)$$

Looking at the term on the left hand side of (4.139) it follows from property (4.33) and Remark 4.9 that

$$\begin{aligned} & c_2 \|\nabla h_\varepsilon - \nabla H_\varepsilon\|_{L^2(D)}^2 \\ & \leq \int_D [T_\varepsilon(x, U_0 + \nabla h_\varepsilon) - T_\varepsilon(x, U_0 + \nabla H_\varepsilon)] \cdot \nabla(h_\varepsilon - H_\varepsilon). \end{aligned} \quad (4.140)$$

Looking at the term on the right hand side of (4.139) and applying property (4.34) in combination with Remark 4.9, one obtains

$$\begin{aligned} & \left| - \int_D [T_\varepsilon(x, U_0 + \nabla h_\varepsilon) - T_\varepsilon(x, U_0 + \nabla H_\varepsilon)] \cdot \nabla \kappa_\varepsilon \right| \\ & \leq c_3 \int_D |\nabla h_\varepsilon - \nabla H_\varepsilon| |\nabla \kappa_\varepsilon| \\ & \leq c_3 \|\nabla h_\varepsilon - \nabla H_\varepsilon\|_{L^2(D)} \|\nabla \kappa_\varepsilon\|_{L^2(D)} \end{aligned} \quad (4.141)$$

Gathering (4.139), (4.140) and (4.141) as well as estimates (4.55) and (4.66), it follows that

$$\begin{aligned} c_2 \|\nabla h_\varepsilon - \nabla H_\varepsilon\|_{L^2(D)}^2 & \leq c_3 \|\nabla h_\varepsilon - \nabla H_\varepsilon\|_{L^2(D)} \|\nabla \kappa_\varepsilon\|_{L^2(D)} \\ & = \mathcal{O}(\varepsilon) o(\varepsilon) = o(\varepsilon^2). \end{aligned}$$

2. Next, we show inequality (4.68). Let  $\alpha > 0$  and  $r \in (0, 1)$ . Since  $\nabla H \in L^2(\mathbb{R}^2)$  and  $r - 1 < 0$ , it holds

$$\int_{D \setminus B(0, \alpha \varepsilon^r)} |\nabla H_\varepsilon|^2 \leq \varepsilon^2 \int_{\mathbb{R}^2 \setminus B(0, \alpha \varepsilon^{r-1})} |\nabla H|^2 = o(\varepsilon^2).$$

This estimate together with (4.67) yields

$$\int_{D \setminus B(0, \alpha \varepsilon^r)} |\nabla h_\varepsilon|^2 = o(\varepsilon^2)$$

by application of the triangle inequality.

3. We now prove estimate (4.69). After (4.44),  $\nabla u_0$  is  $\beta$ -Hölder continuous at point  $x_0$  for some  $\beta > 0$ . Hence, since  $U_0 = \nabla u_0(x_0)$  with  $x_0 = (0, 0)^\top$ , there exist  $\delta > 0$  and  $L > 0$  such that

$$|\nabla u_0(x) - U_0| \leq L|x|^\beta, \quad \forall x \in B(0, \delta).$$

To apply estimate (4.68), we choose  $\alpha = 1$  and  $r = 1/2$ . For all  $\varepsilon \in (0, \delta^2)$ , according to estimates (4.52) and (4.68) it follows

$$\begin{aligned} & \int_D |\nabla u_0 - U_0| |\nabla h_\varepsilon|^2 \\ & \leq \int_{B(0, \alpha \varepsilon^r)} L|x|^\beta |\nabla h_\varepsilon|^2 + 2\|\nabla u_0\|_{L^\infty(D)} \int_{D \setminus B(0, \alpha \varepsilon^r)} |\nabla h_\varepsilon|^2 \\ & \leq L\alpha^\beta \varepsilon^{r\beta} O(\varepsilon^2) + o(\varepsilon^2) = o(\varepsilon^2), \end{aligned}$$

which completes the proof of (4.69).

4. Regarding estimate (4.70), Cauchy-Schwarz's inequality and estimates (4.67), (4.52) and (4.53) entail that

$$\begin{aligned} & \int_D |\nabla h_\varepsilon - \nabla H_\varepsilon| (|\nabla h_\varepsilon| + |\nabla H_\varepsilon|) \\ & \leq \|\nabla h_\varepsilon - \nabla H_\varepsilon\|_{L^2(D)} [\|\nabla h_\varepsilon\|_{L^2(D)} + \|\nabla H_\varepsilon\|_{L^2(D)}] \\ & = o(\varepsilon) \mathcal{O}(\varepsilon) = o(\varepsilon^2) \end{aligned}$$

which completes the proof of (4.70).  $\square$

**Proof of Proposition 4.28:** Also this proof is following the lines of [20, 46] and is adapted to our case here.

*Proof.* 1. We begin with the proof of (4.71), i.e.,

$$\|\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon\|_{L^2(D)}^2 = o(\varepsilon^2).$$

For any  $\eta \in H_0^1(D)$ , forming the difference between variational problems (4.46) and (4.47) gives

$$\begin{aligned} & \int_D [T_\varepsilon(x, \nabla u_0 + \nabla \tilde{u}_\varepsilon) - T_\varepsilon(x, U_0 + \nabla h_\varepsilon) + T_\varepsilon(x, U_0) - T_\varepsilon(x, \nabla u_0)] \cdot \nabla \eta \\ & + \int_{\omega_\varepsilon} [T(U_0) - T(\nabla u_0)] \cdot \nabla \eta + \int_{\omega_\varepsilon} \nu_0 (\nabla u_0 - U_0) \cdot \nabla \eta = 0, \end{aligned}$$

which can be rewritten as

$$\begin{aligned} & \int_D [T_\varepsilon(x, \nabla u_0 + \nabla \tilde{u}_\varepsilon) - T_\varepsilon(x, \nabla u_0 + \nabla h_\varepsilon)] \cdot \nabla \eta \\ & = - \int_{\omega_\varepsilon} [T(U_0) - T(\nabla u_0)] \cdot \nabla \eta - \int_{\omega_\varepsilon} \nu_0 (\nabla u_0 - U_0) \cdot \nabla \eta \\ & + \int_D [T_\varepsilon(x, \nabla u_0) - T_\varepsilon(x, U_0)] \cdot \nabla \eta \\ & + \int_D [T_\varepsilon(x, U_0 + \nabla h_\varepsilon) - T_\varepsilon(x, \nabla u_0 + \nabla h_\varepsilon)] \cdot \nabla \eta. \end{aligned}$$

For all  $\alpha > 0$  and for all  $r \in (0, 1)$  we can split the last two integrals on the right hand side into an integral over  $B(0, \alpha\varepsilon^r)$  and one over  $D \setminus B(0, \alpha\varepsilon^r)$ . Rearranging of the terms yields the equality

$$\begin{aligned} & \int_D [T_\varepsilon(x, \nabla u_0 + \nabla \tilde{u}_\varepsilon) - T_\varepsilon(x, \nabla u_0 + \nabla h_\varepsilon)] \cdot \nabla \eta \\ &= - \int_{\omega_\varepsilon} [T(U_0) - T(\nabla u_0)] \cdot \nabla \eta - \int_{\omega_\varepsilon} \nu_0 (\nabla u_0 - U_0) \cdot \nabla \eta \\ &+ \int_{B(0, \alpha\varepsilon^r)} [T_\varepsilon(x, \nabla u_0) - T_\varepsilon(x, U_0)] \cdot \nabla \eta \\ &+ \int_{B(0, \alpha\varepsilon^r)} [T_\varepsilon(x, U_0 + \nabla h_\varepsilon) - T_\varepsilon(x, \nabla u_0 + \nabla h_\varepsilon)] \cdot \nabla \eta \\ &+ \int_{D \setminus B(0, \alpha\varepsilon^r)} [T_\varepsilon(x, U_0 + \nabla h_\varepsilon) - T_\varepsilon(x, U_0)] \cdot \nabla \eta \\ &+ \int_{D \setminus B(0, \alpha\varepsilon^r)} [T_\varepsilon(x, \nabla u_0) - T_\varepsilon(x, \nabla u_0 + \nabla h_\varepsilon)] \cdot \nabla \eta. \end{aligned}$$

Testing this equation with  $\eta = \tilde{u}_\varepsilon - h_\varepsilon$  and exploiting property (4.33) together with Remark 4.9, it follows that

$$c_2 \|\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon\|_{L^2(D)}^2 \leq \sum_{i=1}^5 \mathcal{E}_i(\varepsilon),$$

with

$$\begin{aligned} \mathcal{E}_1(\varepsilon) &= - \int_{\omega_\varepsilon} [T(U_0) - T(\nabla u_0)] \cdot \nabla (\tilde{u}_\varepsilon - h_\varepsilon) - \int_{\omega_\varepsilon} \nu_0 (\nabla u_0 - U_0) \cdot \nabla (\tilde{u}_\varepsilon - h_\varepsilon) \\ \mathcal{E}_2(\varepsilon) &= \int_{B(0, \alpha\varepsilon^r)} [T_\varepsilon(x, \nabla u_0) - T_\varepsilon(x, U_0)] \cdot \nabla (\tilde{u}_\varepsilon - h_\varepsilon) \\ \mathcal{E}_3(\varepsilon) &= \int_{B(0, \alpha\varepsilon^r)} [T_\varepsilon(x, U_0 + \nabla h_\varepsilon) - T_\varepsilon(x, \nabla u_0 + \nabla h_\varepsilon)] \cdot \nabla (\tilde{u}_\varepsilon - h_\varepsilon) \\ \mathcal{E}_4(\varepsilon) &= \int_{D \setminus B(0, \alpha\varepsilon^r)} [T_\varepsilon(x, U_0 + \nabla h_\varepsilon) - T_\varepsilon(x, U_0)] \cdot \nabla (\tilde{u}_\varepsilon - h_\varepsilon) \\ \mathcal{E}_5(\varepsilon) &= \int_{D \setminus B(0, \alpha\varepsilon^r)} [T_\varepsilon(x, \nabla u_0) - T_\varepsilon(x, \nabla u_0 + \nabla h_\varepsilon)] \cdot \nabla (\tilde{u}_\varepsilon - h_\varepsilon). \end{aligned}$$

Thus, it suffices to show that there exist  $\alpha > 0$  and  $r \in (0, 1)$  such that

$$\mathcal{E}_i(\varepsilon) = o(\varepsilon^2), \quad \forall i \in \{1, 2, 3, 4, 5\}.$$

After (4.44) there exists  $\beta > 0$  such that  $\nabla u_0$  is  $\beta$ -Hölder continuous at point  $x_0 = (0, 0)^\top$ . After (4.4), it holds  $\omega \subset\subset B(0, \rho/\lambda)$  with  $\rho > 0$ ,  $\lambda \in (0, 1]$ . For applying estimate (4.68), we choose

$$r := \frac{1}{2} \left( \frac{1}{\beta + 1} + 1 \right) \in (0, 1). \quad (4.142)$$

and  $\alpha := \rho/\lambda$ . After property (4.34), the operator  $T$  is Lipschitz continuous. Hence, there exists  $\delta > 0$  and  $L > 0$  such that

$$\max(|\nabla u_0(x) - U_0|, |T(\nabla u_0(x)) - T(U_0)|) \leq L|x|^\beta \quad \forall x \in D, |x| \leq \delta. \quad (4.143)$$

In addition, for all  $\varepsilon \in (0, \min(1, (\frac{\delta}{\alpha})^{1/r}))$ , it holds

$$\omega_\varepsilon \subset B(0, \alpha\varepsilon) \subset B(0, \alpha\varepsilon^r) \subset B(0, \delta).$$

Then, one may estimate  $\mathcal{E}_i(\varepsilon)$  for  $i = 1, \dots, 5$  as follows:

(a) It follows from Cauchy-Schwarz's inequality, estimates (4.143) and (4.54) that

$$\begin{aligned} |\mathcal{E}_1(\varepsilon)| &\leq \int_{\omega_\varepsilon} |T(U_0) - T(\nabla u_0)| |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon| + \int_{\omega_\varepsilon} \nu_0 |\nabla u_0 - U_0| |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon| \\ &\leq L(1 + \nu_0) \int_{\omega_\varepsilon} |x|^\beta |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon| \\ &\leq L(1 + \nu_0) \int_{\omega_\varepsilon} \alpha^\beta \varepsilon^\beta |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon| \\ &\leq L(1 + \nu_0) \alpha^\beta \varepsilon^\beta \int_{\omega_\varepsilon} |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon| \\ &\leq L(1 + \nu_0) \alpha^\beta \varepsilon^\beta \left( \int_{\omega_\varepsilon} 1 \right)^{1/2} \|\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon\|_{L^2(D)} \\ &\leq L(1 + \nu_0) \alpha^\beta |\omega|^{1/2} \varepsilon^{\beta+1} O(\varepsilon) = o(\varepsilon^2). \end{aligned}$$

(b) Similarly to above, after (4.143), Cauchy-Schwarz's inequality, estimate (4.54), and the choice of  $r$  (4.142), it holds

$$\begin{aligned} |\mathcal{E}_2(\varepsilon)| &\leq \int_{B(0, \alpha\varepsilon^r)} |T_\varepsilon(x, \nabla u_0) - T_\varepsilon(x, U_0)| |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon| \\ &\leq L \int_{B(0, \alpha\varepsilon^r)} |x|^\beta |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon| \\ &\leq L \alpha^\beta \varepsilon^{r\beta} \int_{B(0, \alpha\varepsilon^r)} |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon| \\ &\leq L \alpha^\beta \varepsilon^{r\beta} \left( \int_{B(0, \alpha\varepsilon^r)} 1 \right)^{1/2} \|\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon\|_{L^2(D)} \\ &\leq L \alpha^{\beta+1} \sqrt{\pi} \varepsilon^{r(\beta+1)} O(\varepsilon) = O\left(\varepsilon^{\frac{r(2\beta+2)+2}{2}}\right) = o(\varepsilon^2). \end{aligned}$$

(c) After property (4.34) together with Remark 4.9, it holds

$$\begin{aligned} |\mathcal{E}_3(\varepsilon)| &\leq \int_{B(0, \alpha\varepsilon^r)} |T_\varepsilon(x, U_0 + \nabla h_\varepsilon) - T_\varepsilon(x, \nabla u_0 + \nabla h_\varepsilon)| |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon| \\ &\leq c_3 \int_{B(0, \alpha\varepsilon^r)} |U_0 - \nabla u_0| |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon| \\ &\leq c_3 L \int_{B(0, \alpha\varepsilon^r)} |x|^\beta |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon|. \end{aligned}$$

This expression can be estimated in the same way as in the case of  $\mathcal{E}_2$  using (4.143), Cauchy-Schwarz's inequality, (4.54), and the choice of  $r$  (4.142). Hence, we get  $\mathcal{E}_3(\varepsilon) = o(\varepsilon^2)$ .

(d) After property (4.34) and Cauchy-Schwarz's inequality, we have that

$$\begin{aligned} |\mathcal{E}_4(\varepsilon)| &\leq \int_{D \setminus B(0, \alpha\varepsilon^r)} |T_\varepsilon(x, U_0 + \nabla h_\varepsilon) - T_\varepsilon(x, U_0)| |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon| \\ &\leq c_3 \int_{D \setminus B(0, \alpha\varepsilon^r)} |\nabla h_\varepsilon| |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon| \\ &\leq c_3 \left( \int_{D \setminus B(0, \alpha\varepsilon^r)} |\nabla h_\varepsilon|^2 \right)^{1/2} \|\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon\|_{L^2(D)}. \end{aligned}$$

Estimates (4.68) and (4.54) now yield that

$$|\mathcal{E}_4(\varepsilon)| \leq c_3 o(\varepsilon) O(\varepsilon) = o(\varepsilon^2).$$

(e) Finally, we have

$$|\mathcal{E}_5(\varepsilon)| \leq \int_{D \setminus B(0, \alpha\varepsilon^r)} |T_\varepsilon(x, \nabla u_0) - T_\varepsilon(x, \nabla u_0 + \nabla h_\varepsilon)| |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon|,$$

and the same steps as for  $\mathcal{E}_4(\varepsilon)$  can be taken since estimate (4.34) holds for all vectors  $\varphi \in \mathbb{R}^2$ . Thus, we have that  $\mathcal{E}_5(\varepsilon) = o(\varepsilon^2)$ .

Hence, we have shown that  $\|\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon\|_{L^2(D)^2} = o(\varepsilon^2)$ , which concludes the proof of (4.71).

2. Regarding estimate (4.72), Cauchy-Schwarz's inequality and estimates (4.71), (4.51) and (4.52) entail that

$$\begin{aligned} \int_D |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon| (|\nabla \tilde{u}_\varepsilon| + |\nabla h_\varepsilon|) \\ \leq \|\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon\|_{L^2(D)} (\|\nabla \tilde{u}_\varepsilon\|_{L^2(D)} + \|\nabla h_\varepsilon\|_{L^2(D)}) \\ = o(\varepsilon) O(\varepsilon) = o(\varepsilon^2), \end{aligned}$$

which completes the proof of estimate (4.72).

3. Estimate (4.73) follows directly from (4.68) and (4.71) by application of the triangle inequality:

$$\begin{aligned} \int_{D \setminus B(0, \alpha\varepsilon^r)} |\nabla \tilde{u}_\varepsilon|^2 &\leq \|\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon\|_{L^2(D \setminus B(0, \alpha\varepsilon^r))}^2 + \|\nabla h_\varepsilon\|_{L^2(D \setminus B(0, \alpha\varepsilon^r))}^2 \\ &= o(\varepsilon^2) + o(\varepsilon^2) = o(\varepsilon^2). \end{aligned}$$

□

#### 4.4.4.2 Proofs on the Variation of the Adjoint State

**Proof of Lemma 4.32:** This proof is following the lines of [20, 46] and is adapted to our case here.

*Proof.* Due to properties (4.45) and (4.77), we have that  $\nabla u_0|_{\Omega^d} \in L^\infty(\Omega^d)$  and  $\nabla p_0|_{\Omega^d} \in L^\infty(\Omega^d)$ . Lemma 4.8(1) states that  $DT \in C^{0,1}(\mathbb{R}^2, \mathbb{R}^{2 \times 2})$ , thus we also have that  $DT(\nabla u_0) \in L^\infty(\Omega^d)$ . Testing (4.78) with test function  $\eta = \tilde{p}_\varepsilon$ , property (4.32) yields

$$\begin{aligned} \underline{c}_1 \int_D |\nabla \tilde{p}_\varepsilon|^2 &\leq \int_D DT_\varepsilon(x, \nabla u_0) \nabla \tilde{p}_\varepsilon \cdot \nabla \tilde{p}_\varepsilon \\ &= \int_{\omega_\varepsilon} (DT(\nabla u_0) - \nu_0 I) \nabla p_0 \cdot \nabla \tilde{p}_\varepsilon \\ &\leq \|DT(\nabla u_0) - \nu_0 I\|_{L^\infty(\Omega^d)} \|\nabla p_0\|_{L^\infty(\Omega^d)} |\omega|^{1/2} \varepsilon \|\nabla \tilde{p}_\varepsilon\|_{L^2(D)}. \end{aligned}$$

Let  $C := |\omega| (\|DT(\nabla u_0) - \nu_0 I\|_{L^\infty(\Omega^d)} \|\nabla p_0\|_{L^\infty(\Omega^d)} / \underline{c}_1)^2$ . Then

$$\|\nabla \tilde{p}_\varepsilon\|_{L^2(D)}^2 \leq C \varepsilon^2,$$

which finishes the proof of (4.82).

The upper bound (4.83) to  $\|\nabla k_\varepsilon\|_{L^2(D)}^2$  is obtained in the same way.

After a change of scale and since, by definition,  $\nabla K \in L^2(\mathbb{R}^2)$ , it holds

$$\|\nabla K_\varepsilon\|_{L^2(D)}^2 = \varepsilon^2 \int_{D/\varepsilon} |\nabla K|^2 \leq \varepsilon^2 \int_{\mathbb{R}^2} |\nabla K|^2 = O(\varepsilon^2),$$

which proves estimate (4.84).  $\square$

**Proof of Proposition 4.33:** We adapt the proof of [20, 46] which is a standard proof and based on, e.g, [10].

*Proof.* Note that (4.80) can be rewritten as

$$\int_{\mathbb{R}^2} DT(U_0) \nabla K \cdot \nabla \eta = - \int_{\omega} (\nu_0 I - DT(U_0)) (P_0 + \nabla K) \cdot \nabla \eta. \quad (4.144)$$

Performing the change of variables  $y = DT(U_0)^{1/2}x$  and exploiting the positive definiteness and symmetry of the matrix  $DT(U_0)$ , (4.144) becomes

$$\begin{aligned} &\int_{\mathbb{R}^2} \nabla_x \hat{K}(x) \cdot \nabla_x \hat{\eta}(x) dx \\ &= - \int_{\hat{\omega}} (\nu_0 DT(U_0)^{-1} - I) (\hat{P}_0 + \nabla_x \hat{K}) \cdot \nabla_x \hat{\eta}(x) dx \quad \forall \hat{\eta} \in \mathcal{H}(\mathbb{R}^2), \end{aligned} \quad (4.145)$$

where  $\hat{K}(x) = K(DT(U_0)^{1/2}x)$ ,  $\hat{\omega} = DT(U_0)^{-1/2}\omega$  and  $\hat{P}_0 = DT(U_0)^{1/2}P_0$ . Note that (4.145) amounts to a Laplace equation in  $\mathbb{R}^2$  with a source supported by  $\hat{\omega}$ . For sake of readability, we will drop the hats in the rest of this proof.

Let  $E$  be the elementary solution of the Laplace operator in  $\mathbb{R}^2$  given for all  $y \in \mathbb{R}^2, y \neq 0$ , by

$$E(y) := \frac{1}{2\pi} \log|y|.$$

In particular, we have for all  $y \in \mathbb{R}^2, y \neq 0$ , that

$$|\nabla E(y)| = \frac{1}{2\pi}|y|^{-1}. \quad (4.146)$$

Let  $\mathcal{T}$  be the distribution in  $\mathbb{R}^2$  defined for all  $\eta \in C_0^\infty(\mathbb{R}^2)$  by

$$\langle \mathcal{T}, \eta \rangle := \int_{\omega} (\nu_0 DT(U_0)^{-1} - I)(P_0 + \nabla K) \cdot \nabla \eta.$$

It follows from (4.145) that

$$\Delta K = \mathcal{T}.$$

Hence let an element  $\tilde{K}$  of the class  $K$  be given by

$$\tilde{K} = \mathcal{T} * E. \quad (4.147)$$

Let  $\rho > 0, \lambda \in (0, 1]$  such that  $\omega \subset B(0, \rho/\lambda)$ , see (4.4). To study the behavior of  $\tilde{K}$  at infinity, let  $y \in \mathbb{R}^2, |y| \geq 2(\rho/\lambda)$ . In particular  $|z|/|y| \leq 1/2$  for all  $z \in \omega$ .

The convolution (4.147) reads

$$\tilde{K}(y) = \int_{\omega} (\nu_0 DT(U_0)^{-1} - I)(P_0 + \nabla K(z)) \cdot \nabla E(y - z) dz.$$

Since  $P_0 + \nabla K \in L^2(\omega)$ , the Cauchy-Schwarz inequality yields

$$|\tilde{K}(y)| \leq C \left( \int_{\omega} |\nabla E(y - z)|^2 dz \right)^{1/2},$$

with  $C := \|\nu_0 DT(U_0)^{-1} - I\| \|P_0 + \nabla K\|_{L^2(\omega)}$ .

In addition, after (4.146) and using that  $|z| \leq |y|/2$ , it holds

$$\begin{aligned} \int_{\omega} |\nabla E(y - z)|^2 dz &= \frac{1}{(2\pi)^2} \int_{\omega} |y - z|^{-2} dz \\ &\leq \frac{1}{(2\pi)^2} \int_{\omega} (|y| - |z|)^{-2} dz \\ &\leq \frac{1}{(2\pi)^2} |\omega| \left(\frac{1}{2}\right)^{-2} |y|^{-2}. \end{aligned}$$

Hence,

$$|\tilde{K}(y)| \leq C'|y|^{-1}$$

with  $C' = C \frac{1}{\pi} |\omega|^{1/2}$ . This completes the proof of the claimed asymptotic behavior (4.85).  $\square$

**Proof of Lemma 4.35:** This proof is following the lines of [20, 46] and is adapted to our case here.

*Proof.* 1. We begin proving estimate (4.88). For all  $\eta \in H_0^1(D)$ , we define  $\eta_1 \in \mathcal{H}(\mathbb{R}^2)$  by

$$\eta_1(y) := \frac{1}{\varepsilon} \eta(\varepsilon y), \quad y \in D/\varepsilon \quad \text{and} \quad \eta_1(y) := 0, \quad y \in \mathbb{R}^2 \setminus (D/\varepsilon).$$

We test the variational equation (4.80) with  $\eta_1 \in \mathcal{H}(\mathbb{R}^2)$  and make the change of scale backwards. This yields the equality

$$\int_D DT_\varepsilon(x, U_0) \nabla K_\varepsilon \cdot \nabla \eta = - \int_{\omega_\varepsilon} (\nu_0 I - DT(U_0)) P_0 \cdot \nabla \eta,$$

which must hold true for all  $\eta \in H_0^1(D)$ . Then, calculating the difference with the variational equation (4.79) yields that

$$\int_D DT_\varepsilon(x, U_0) (\nabla k_\varepsilon - \nabla K_\varepsilon) \cdot \nabla \eta = 0, \quad \forall \eta \in H_0^1(D). \quad (4.148)$$

Recall the function  $\kappa_{a\varepsilon}$  defined in (4.86). It holds  $\kappa_{a\varepsilon} \in H^1(D)$  and  $K_\varepsilon - \kappa_{a\varepsilon} \in H_0^1(D)$ . Choosing  $\eta = k_\varepsilon - (K_\varepsilon - \kappa_{a\varepsilon}) \in H_0^1(D)$  and plugging into (4.148), it holds

$$\int_D DT_\varepsilon(x, U_0) (\nabla k_\varepsilon - \nabla K_\varepsilon) \cdot (\nabla k_\varepsilon - \nabla K_\varepsilon) = - \int_D DT_\varepsilon(x, U_0) (\nabla k_\varepsilon - \nabla K_\varepsilon) \cdot \nabla \kappa_{a\varepsilon}.$$

Then, applying condition (4.32) together with Remark 4.9, one obtains

$$\begin{aligned} \underline{c}_1 \int_D |\nabla k_\varepsilon - \nabla K_\varepsilon|^2 &\leq \int_D DT_\varepsilon(x, U_0) (\nabla k_\varepsilon - \nabla K_\varepsilon) \cdot (\nabla k_\varepsilon - \nabla K_\varepsilon) \\ &\leq \left| \int_D DT_\varepsilon(x, U_0) (\nabla k_\varepsilon - \nabla K_\varepsilon) \cdot \nabla \kappa_{a\varepsilon} \right| \\ &\leq \nu_0 \left( \int_D |\nabla k_\varepsilon - K_\varepsilon|^2 \right)^{1/2} \left( \int_D |\nabla \kappa_{a\varepsilon}|^2 \right)^{1/2}, \end{aligned}$$

where we used that  $\max\{\lambda_1(|U_0|), \lambda_2(|U_0|)\} \leq \nu_0$ , see (2.29) together with (2.19). This means that

$$\|\nabla k_\varepsilon - \nabla K_\varepsilon\|_{L^2(D)}^2 \leq \left( \frac{\nu_0}{\underline{c}_1} \right)^2 \|\nabla \kappa_{a\varepsilon}\|_{L^2(D)}^2,$$

which yields the statement (4.88) due to (4.87).

2. Let us now prove estimate (4.89). Let  $\alpha > 0$  and  $r \in (0, 1)$ . By the triangle inequality, we have

$$|\nabla k_\varepsilon|^2 \leq |\nabla k_\varepsilon - \nabla K_\varepsilon|^2 + |\nabla K_\varepsilon|^2.$$

After a change of scale, we obtain

$$\int_{D \setminus B(0, \alpha \varepsilon^r)} |\nabla k_\varepsilon|^2 \leq \int_{D \setminus B(0, \alpha \varepsilon^r)} |\nabla k_\varepsilon - \nabla K_\varepsilon|^2 + \varepsilon^2 \int_{\mathbb{R}^2 \setminus B(0, \alpha \varepsilon^{r-1})} |\nabla K|^2. \quad (4.149)$$

As in the proof of (4.68), we use that  $\nabla K \in L^2(\mathbb{R}^2)$  and  $r - 1 < 0$  entail that

$$\int_{\mathbb{R}^2 \setminus B(0, \alpha \varepsilon^{r-1})} |\nabla K|^2 = o(1).$$

Hence, due to (4.88), it follows from (4.149) that

$$\int_{D \setminus B(0, \alpha \varepsilon^r)} |\nabla k_\varepsilon|^2 = o(\varepsilon^2).$$

□

**Proof of Lemma 4.36:** This proof is following the lines of [20, 46] and is adapted to our case here.

*Proof.* Due to assumptions (4.44) and (4.76), we have that  $\nabla u_0$  and  $\nabla p_0$  are Hölder continuous at point  $x_0 = (0, 0)^\top$  with exponents  $\beta$  and  $\tilde{\beta}$ , respectively. Lemma 4.8(1) implies that  $DT$  is Lipschitz continuous at point  $U_0$ . Thus, we have that  $DT(\nabla u_0) \in L^\infty(D)$ . We denote the Hölder constants for  $\nabla u_0$ ,  $\nabla p_0$  and  $DT$  by  $L_1$ ,  $L_2$  and  $L_3$ , respectively. Let  $\tilde{\tau} := \min(\beta, \tilde{\beta}) > 0$ . Hence, there exist  $\delta \in (0, 1)$  and  $L > 0$  such that for all  $x \in B(0, \delta)$  it holds

$$\|DT(\nabla u_0(x)) - DT(U_0)\| + |DT(\nabla u_0(x))\nabla p_0(x) - DT(U_0)P_0| \leq L|x|^{\tilde{\tau}} \quad (4.150)$$

which can be seen as follows (recall that  $x_0 = 0$ ):

$$\begin{aligned} & \|DT(\nabla u_0(x)) - DT(U_0)\| + |DT(\nabla u_0(x))\nabla p_0(x) - DT(U_0)P_0| \\ & \leq L_3|\nabla u_0(x) - U_0| + |(DT(\nabla u_0(x)) - DT(U_0))\nabla p_0(x)| + |DT(U_0)(\nabla p_0(x) - P_0)| \\ & \leq L_3L_1|x - x_0|^\beta + L_3L_1|x - x_0|^\beta\|\nabla p_0\|_{L^\infty(D)} + \|DT(U_0)\|_{L^\infty(D)}L_2|x - x_0|^{\tilde{\beta}}. \end{aligned}$$

Let  $\rho > 0$  and  $\lambda \in (0, 1]$  as in (4.4) such that  $\omega \subset B(0, \rho/\lambda)$ . So as to apply estimate (4.89), we choose  $\alpha := \rho/\lambda$  and  $r := 1/2$ . Therefore, for all  $\varepsilon \in (0, \min(1, (\delta/\alpha)^2))$ , it holds

$$\omega_\varepsilon \subset B(0, \alpha \varepsilon) \subset B(0, \alpha \varepsilon^r) \subset B(0, \delta). \quad (4.151)$$

We can now start our estimations. According to condition (4.32) and Remark 4.9, it holds

$$\begin{aligned} \|\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon\|_{L^2(D)}^2 &= \int_D |\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon|^2 \\ &\leq \frac{1}{c_1} \int_D DT_\varepsilon(x, \nabla u_0)(\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon) \cdot (\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon). \end{aligned} \quad (4.152)$$

Calculating the difference between the variational forms (4.78) and (4.79) and choosing the test function  $\eta = \tilde{p}_\varepsilon - k_\varepsilon \in H_0^1(D)$ , one obtains

$$\begin{aligned} & \int_D DT_\varepsilon(x, \nabla u_0)(\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon) \cdot (\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon) \\ &= \int_D (DT_\varepsilon(x, U_0) - DT_\varepsilon(x, \nabla u_0))\nabla k_\varepsilon \cdot (\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon) \\ & \quad + \int_{\omega_\varepsilon} \nu_0(P_0 - \nabla p_0) \cdot (\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon) \\ & \quad + \int_{\omega_\varepsilon} (DT(\nabla u_0)\nabla p_0 - DT(U_0)P_0) \cdot (\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon). \end{aligned} \quad (4.153)$$

1. We begin by estimating the first term on the right hand side of (4.153).

$$\begin{aligned}
& \left| \int_D (\mathrm{DT}_\varepsilon(x, U_0) - \mathrm{DT}_\varepsilon(x, \nabla u_0)) \nabla k_\varepsilon \cdot (\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon) \right| \\
& \leq \left| \int_{B(0, \alpha \varepsilon^r) \setminus \omega_\varepsilon} (\mathrm{DT}(U_0) - \mathrm{DT}(\nabla u_0)) \nabla k_\varepsilon \cdot (\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon) \right| \\
& \quad + \left| \int_{D \setminus B(0, \alpha \varepsilon^r)} (\mathrm{DT}(U_0) - \mathrm{DT}(\nabla u_0)) \nabla k_\varepsilon \cdot (\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon) \right| \\
& \leq L \int_{B(0, \alpha \varepsilon^r) \setminus \omega_\varepsilon} |x|^{\tilde{\tau}} |\nabla k_\varepsilon| |\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon| \\
& \quad + \tilde{C} \left( \int_{D \setminus B(0, \alpha \varepsilon^r)} |\nabla k_\varepsilon|^2 \right)^{1/2} \|\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon\|_{L^2(D \setminus B(0, \alpha \varepsilon^r))} \\
& \leq L \alpha^{\tilde{\tau}} \varepsilon^{r\tilde{\tau}} \|\nabla k_\varepsilon\|_{L^2(D)} \|\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon\|_{L^2(D)} + o(\varepsilon) \|\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon\|_{L^2(D)}.
\end{aligned}$$

where  $\tilde{C} := 2 \|\mathrm{DT}(\nabla u_0)\|_{L^\infty(D)}$  and we used that  $\mathrm{DT}_\varepsilon(x, U_0) = \nu_0 I = \mathrm{DT}_\varepsilon(x, \nabla u_0)$  inside  $\omega_\varepsilon$  and (4.89). Using that  $\|\nabla k_\varepsilon\|_{L^2(D)} = \mathcal{O}(\varepsilon)$  due to (4.83), we get

$$\left| \int_D (\mathrm{DT}_\varepsilon(x, U_0) - \mathrm{DT}_\varepsilon(x, \nabla u_0)) \nabla k_\varepsilon \cdot (\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon) \right| = o(\varepsilon) \|\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon\|_{L^2(D)}. \quad (4.154)$$

2. Similarly, we get for the remaining two terms of (4.153) that

$$\begin{aligned}
& \left| \int_{\omega_\varepsilon} \nu_0 (P_0 - \nabla p_0) \cdot (\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon) \right| + \left| \int_{\omega_\varepsilon} (\mathrm{DT}(\nabla u_0) \nabla p_0 - \mathrm{DT}(U_0) P_0) \cdot (\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon) \right| \\
& \leq \nu_0 L_2 \int_{\omega_\varepsilon} |x|^{\tilde{\beta}} |\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon| + L \int_{\omega_\varepsilon} |x|^{\tilde{\tau}} |\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon| \\
& \leq \nu_0 L_2 \int_{B(0, \alpha \varepsilon)} |x|^{\tilde{\beta}} |\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon| + L \int_{B(0, \alpha \varepsilon)} |x|^{\tilde{\tau}} |\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon| \\
& \leq \left( \nu_0 L_2 \alpha^{\tilde{\beta}} \varepsilon^{\tilde{\beta}} + L \alpha^{\tilde{\tau}} \varepsilon^{\tilde{\tau}} \right) \left( \int_{B(0, \alpha \varepsilon)} 1 \right)^{1/2} \|\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon\|_{L^2(B(0, \alpha \varepsilon))} \\
& \leq \left( \nu_0 L_2 \alpha^{\tilde{\beta}} \varepsilon^{\tilde{\beta}} + L \alpha^{\tilde{\tau}} \varepsilon^{\tilde{\tau}} \right) \pi \alpha \varepsilon \|\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon\|_{L^2(D)} = o(\varepsilon) \|\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon\|_{L^2(D)}. \quad (4.155)
\end{aligned}$$

Collecting (4.152), (4.153), (4.154) and (4.155), and dividing by  $\|\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon\|_{L^2(D)}$ , we get the claimed estimate (4.90).  $\square$

#### 4.4.4.3 Proofs on the Expansion of the Cost Functional

**Proof of Lemma 4.37:** This proof is following the lines of [20, 46] and is adapted to our case here.

*Proof.* It follows from definitions (4.96) and (4.97) that

$$\begin{aligned}\tilde{j}_1(\varepsilon) - \varepsilon^2 J_1 &= (\nu_0 - \hat{\nu}(|U_0|)) \int_{\omega_\varepsilon} U_0 \cdot (P_0 + \nabla k_\varepsilon) - \varepsilon^2 (\nu_0 - \hat{\nu}(|U_0|)) \int_{\omega} U_0 \cdot (P_0 + \nabla K) \\ &= (\nu_0 - \hat{\nu}(|U_0|)) \int_{\omega_\varepsilon} U_0 \cdot (P_0 + \nabla k_\varepsilon) - \varepsilon^2 (\nu_0 - \hat{\nu}(|U_0|)) \int_{\omega_\varepsilon} U_0 \cdot (P_0 + \nabla K_\varepsilon) \varepsilon^{-2} \\ &= (\nu_0 - \hat{\nu}(|U_0|)) U_0 \cdot \int_{\omega_\varepsilon} (\nabla k_\varepsilon - \nabla K_\varepsilon).\end{aligned}$$

Hence, after property (2.19a), estimate (4.88) and Cauchy-Schwarz's inequality, it holds

$$\begin{aligned}|\tilde{j}_1(\varepsilon) - \varepsilon^2 J_1| &\leq \nu_0 |U_0| |\omega|^{1/2} \varepsilon \|\nabla k_\varepsilon - \nabla K_\varepsilon\|_{L^2(D)} \\ &\leq \mathcal{O}(\varepsilon) o(\varepsilon) = o(\varepsilon^2),\end{aligned}$$

which completes the proof of Lemma 4.37.  $\square$

**Proof of Lemma 4.38:** This proof is following the lines of [20, 46] and is adapted to our case here.

*Proof.* It follows from definitions (4.93) and (4.96) that

$$\begin{aligned}j_1(\varepsilon) - \tilde{j}_1(\varepsilon) &= \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|\nabla u_0|)) \nabla u_0 \cdot (\nabla p_0 + \nabla \tilde{p}_\varepsilon) - \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|U_0|)) U_0 \cdot (P_0 + \nabla k_\varepsilon) \\ &= \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|\nabla u_0|)) \nabla u_0 \cdot \nabla p_0 - (\nu_0 - \hat{\nu}(|U_0|)) U_0 \cdot P_0 \\ &\quad + \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|\nabla u_0|)) \nabla u_0 \cdot \nabla \tilde{p}_\varepsilon - (\nu_0 - \hat{\nu}(|U_0|)) U_0 \cdot \nabla k_\varepsilon.\end{aligned}$$

Since the mapping  $x \in D \mapsto (\nu_0 - \hat{\nu}(|\nabla u_0(x)|)) \nabla u_0(x) \cdot \nabla p_0(x)$  is continuous at point  $x_0 = (0, 0)^\top$  due to Assumptions 5 and 6 and 3(1), it holds

$$\int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|\nabla u_0|)) \nabla u_0 \cdot \nabla p_0 - (\nu_0 - \hat{\nu}(|U_0|)) U_0 \cdot P_0 = |\omega_\varepsilon| o(1) = o(\varepsilon^2).$$

Moreover, since  $x \in D \mapsto (\nu_0 - \hat{\nu}(|\nabla u_0(x)|)) \nabla u_0(x)$  is continuous at point  $x_0 = (0, 0)^\top$  due to Assumptions 5 and 3(1), after Cauchy-Schwarz's inequality and estimates (4.90) and (4.83), it holds

$$\begin{aligned}&\left| \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|\nabla u_0|)) \nabla u_0 \cdot \nabla \tilde{p}_\varepsilon - (\nu_0 - \hat{\nu}(|U_0|)) U_0 \cdot \nabla k_\varepsilon \right| \\ &\leq \int_{\omega_\varepsilon} |(\nu_0 - \hat{\nu}(|\nabla u_0|))| |\nabla u_0| |\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon| + \int_{\omega_\varepsilon} [(\nu_0 - \hat{\nu}(|\nabla u_0|)) \nabla u_0 - (\nu_0 - \hat{\nu}(|U_0|)) U_0] \cdot \nabla k_\varepsilon \\ &\leq |\omega|^{1/2} \varepsilon \left( \|(\nu_0 - \hat{\nu}(|\nabla u_0|)) \nabla u_0\|_{L^\infty(\Omega^d)} \|\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon\|_{L^2(D)} + o(1) \|\nabla k_\varepsilon\|_{L^2(D)} \right) \\ &= \mathcal{O}(\varepsilon) (o(\varepsilon) + o(1) \mathcal{O}(\varepsilon)) = o(\varepsilon^2).\end{aligned}$$

This completes the proof of Lemma 4.38.  $\square$

**Proof of Lemma 4.40:** This proof is following the lines of [20, 46] and is adapted to our case here.

*Proof.* With the help of (4.47), tested with  $\eta = k_\varepsilon$ , and (4.79), tested with  $\eta = h_\varepsilon$ , the term  $\tilde{j}_2(\varepsilon)$  defined in (4.103) can be rewritten as follows:

$$\begin{aligned}\tilde{j}_2(\varepsilon) &= \int_D S_{U_0}^\varepsilon(x, \nabla h_\varepsilon) \cdot (P_0 + \nabla k_\varepsilon) \\ &= \int_D S_{U_0}^\varepsilon(x, \nabla h_\varepsilon) \cdot P_0 + (T_\varepsilon(x, U_0 + \nabla h_\varepsilon) - T_\varepsilon(x, U_0) - DT_\varepsilon(x, U_0) \nabla h_\varepsilon) \cdot \nabla k_\varepsilon \\ &= \int_D S_{U_0}^\varepsilon(x, \nabla h_\varepsilon) \cdot P_0 - \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|U_0|)) U_0 \cdot \nabla k_\varepsilon + \int_{\omega_\varepsilon} (\nu_0 I - DT(U_0)) P_0 \cdot \nabla h_\varepsilon.\end{aligned}\tag{4.156}$$

Similarly, by (4.49), tested with  $\eta = K$ , and (4.80), tested with  $\eta = H$ , we get for  $J_2$  defined in (4.104) that

$$J_2 = \int_{\mathbb{R}^2 \setminus \omega} \tilde{S}_{U_0}(x, \nabla H) \cdot P_0 - \int_{\omega} (\nu_0 - \hat{\nu}(|U_0|)) U_0 \cdot \nabla K + \int_{\omega} (\nu_0 I - DT(U_0)) \nabla H_\varepsilon \cdot P_0.$$

Then, making the change of scale backwards, we get

$$\begin{aligned}\varepsilon^2 J_2 &= \varepsilon^2 \int_{\mathbb{R}^2 \setminus (D/\varepsilon)} S_{U_0}(\nabla H) \cdot P_0 + \int_{D \setminus \omega_\varepsilon} S_{U_0}(\nabla H_\varepsilon) \cdot P_0 \\ &\quad - \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|U_0|)) U_0 \cdot \nabla K_\varepsilon + \int_{\omega_\varepsilon} (\nu_0 I - DT(U_0)) \nabla H_\varepsilon \cdot P_0.\end{aligned}\tag{4.157}$$

The first integral on the right hand side is the remainder of a converging integral. Thus,

$$\int_{\mathbb{R}^2 \setminus (D/\varepsilon)} S_{U_0}(\nabla H) \cdot P_0 = o(1) \quad \text{as } \varepsilon \rightarrow 0.$$

Therefore, gathering (4.156) and (4.157) yields,

$$\begin{aligned}\tilde{j}_2(\varepsilon) - \varepsilon^2 J_2 &= \int_{D \setminus \omega_\varepsilon} (S_{U_0}(\nabla h_\varepsilon) - S_{U_0}(\nabla H_\varepsilon)) \cdot P_0 + o(\varepsilon^2) \\ &\quad - \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|U_0|)) U_0 \cdot (\nabla k_\varepsilon - \nabla K_\varepsilon)\end{aligned}\tag{4.158}$$

$$+ \int_{\omega_\varepsilon} (\nu_0 I - DT(U_0)) (\nabla h_\varepsilon - \nabla H_\varepsilon) \cdot P_0\tag{4.159}$$

Regarding the second term (4.158) on the right hand side, Cauchy-Schwarz's inequality and estimate (4.88) imply

$$\begin{aligned}\int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|U_0|)) U_0 \cdot (\nabla k_\varepsilon - \nabla K_\varepsilon) &\leq |(\nu_0 - \hat{\nu}(|U_0|)) U_0| |\omega|^{1/2} \varepsilon \|\nabla k_\varepsilon - \nabla K_\varepsilon\|_{L^2(\omega_\varepsilon)} \\ &\leq |(\nu_0 - \hat{\nu}(|U_0|)) U_0| |\omega|^{1/2} \varepsilon o(\varepsilon) = o(\varepsilon^2).\end{aligned}$$

Similarly, for the third term (4.159) on the right hand side, Cauchy-Schwarz's inequality and estimate (4.67) imply

$$\begin{aligned}\int_{\omega_\varepsilon} ((\nu_0 I - DT(U_0)) P_0) \cdot (\nabla h_\varepsilon - \nabla H_\varepsilon) &\leq |((\nu_0 I - DT(U_0)) P_0)| |\omega|^{1/2} \varepsilon \|\nabla h_\varepsilon - \nabla H_\varepsilon\|_{L^2(\omega_\varepsilon)} \\ &\leq |((\nu_0 I - DT(U_0)) P_0)| |\omega|^{1/2} \varepsilon o(\varepsilon) = o(\varepsilon^2).\end{aligned}$$

It follows

$$\tilde{j}_2(\varepsilon) - \varepsilon^2 J_2 = \int_{D \setminus \omega_\varepsilon} (S_{U_0}(\nabla h_\varepsilon) - S_{U_0}(\nabla H_\varepsilon)) \cdot P_0 + o(\varepsilon^2). \quad (4.160)$$

Condition (4.35) immediately yields

$$\int_{D \setminus \omega_\varepsilon} |S_{U_0}(\nabla h_\varepsilon) - S_{U_0}(\nabla H_\varepsilon)| \leq c_4 \int_{D \setminus \omega_\varepsilon} |\nabla h_\varepsilon - \nabla H_\varepsilon| (|\nabla H_\varepsilon| + |\nabla h_\varepsilon|).$$

Hence, it follows from (4.70) that

$$\int_{D \setminus \omega_\varepsilon} |S_{U_0}(\nabla h_\varepsilon) - S_{U_0}(\nabla H_\varepsilon)| = o(\varepsilon^2).$$

Therefore, (4.160) entails

$$\tilde{j}_2(\varepsilon) - \varepsilon^2 J_2 = o(\varepsilon^2),$$

which completes the proof of this lemma.  $\square$

**Proof of Lemma 4.41:** This proof is analogous to [20, 46] and is given here for completeness.

*Proof.* Since  $\nabla p_0$  is  $\tilde{\beta}$ -Hölder continuous at point  $x_0 = (0, 0)^\top$  for some  $\tilde{\beta} > 0$ , there exist  $\delta > 0$  and  $L > 0$  such that

$$|\nabla p_0 - P_0| \leq L|x|^{\tilde{\beta}}, \quad \forall x \in B(0, \delta).$$

To apply estimate (4.68), we choose  $\alpha := \delta$  and  $r := 1/2$ . Hence, for all  $\varepsilon \in (0, 1)$ , according to estimates (4.52) and (4.68) it follows

$$\begin{aligned} & \int_D |\nabla p_0 - P_0| |\nabla h_\varepsilon|^2 \\ & \leq \int_{B(0, \alpha \varepsilon^r)} L|x|^{\tilde{\beta}} |\nabla h_\varepsilon|^2 + 2\|\nabla p_0\|_{L^\infty(D)} \int_{D \setminus B(0, \alpha \varepsilon^r)} |\nabla h_\varepsilon|^2 \\ & \leq L\alpha^{\tilde{\beta}} \varepsilon^{r\tilde{\beta}} O(\varepsilon^2) + o(\varepsilon^2) = o(\varepsilon^2), \end{aligned}$$

which completes the proof of estimate (4.106).  $\square$

**Proof of Lemma 4.42:** This proof is following the lines of [20, 46] and is adapted to our case here.

*Proof.* As in the proof of Lemma 4.40, we can rewrite the terms  $j_2(\varepsilon)$  (4.94) and  $\tilde{j}_2(\varepsilon)$  (4.103) as

$$\begin{aligned} j_2(\varepsilon) &= \int_D S_{\nabla u_0}^\varepsilon(x, \nabla \tilde{u}_\varepsilon) \cdot \nabla p_0 \\ &\quad - \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|\nabla u_0|)) \nabla u_0 \cdot \nabla \tilde{p}_\varepsilon + \int_{\omega_\varepsilon} (\nu_0 I - DT(\nabla u_0)) \nabla \tilde{u}_\varepsilon \cdot \nabla p_0, \\ \tilde{j}_2(\varepsilon) &= \int_D S_{U_0}^\varepsilon(x, \nabla h_\varepsilon) \cdot P_0 - \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|U_0|)) U_0 \cdot \nabla k_\varepsilon + \int_{\omega_\varepsilon} (\nu_0 I - DT(U_0)) \nabla h_\varepsilon \cdot P_0. \end{aligned}$$

Calculating the difference, we get

$$\begin{aligned}
j_2(\varepsilon) - \tilde{j}_2(\varepsilon) &= \int_{D \setminus \omega_\varepsilon} S_{\nabla u_0}(x, \nabla \tilde{u}_\varepsilon) \cdot \nabla p_0 - S_{U_0}(x, \nabla h_\varepsilon) \cdot P_0 \\
&\quad - \int_{\omega_\varepsilon} [\nu_0(\nabla u_0 \cdot \nabla \tilde{p}_\varepsilon - U_0 \cdot \nabla k_\varepsilon) - (\hat{\nu}(|\nabla u_0|) \nabla u_0 \cdot \nabla \tilde{p}_\varepsilon - \hat{\nu}(|U_0|) U_0 \cdot \nabla k_\varepsilon)] \\
&\quad + \int_{\omega_\varepsilon} [\nu_0(\nabla \tilde{u}_\varepsilon \cdot \nabla p_0 - \nabla h_\varepsilon \cdot P_0) - (DT(\nabla u_0) \nabla \tilde{u}_\varepsilon \cdot \nabla p_0 - DT(U_0) \nabla h_\varepsilon \cdot P_0)].
\end{aligned} \tag{4.161}$$

Let  $\delta > 0$ . Due to the continuity of  $\nabla u_0$  and  $\nabla p_0$  at point  $x_0$  and to the continuity of  $DT$  and  $\hat{\nu}(|\cdot|)$ , for  $\varepsilon > 0$  small enough it holds

$$\begin{aligned}
&\max(|DT(\nabla u_0) \nabla p_0 - DT(U_0) P_0|, |\nabla p_0 - P_0|, \\
&|\hat{\nu}(|\nabla u_0|) \nabla u_0 - \hat{\nu}(|U_0|) U_0|), |\nabla u_0 - U_0|) \leq \delta \quad \text{in } \omega_\varepsilon.
\end{aligned}$$

Let us investigate the second integral on the right hand side of (4.161). After Cauchy-Schwarz's inequality, and estimates (4.90) and (4.83), we get

$$\begin{aligned}
&\left| \int_{\omega_\varepsilon} [\nu_0(\nabla u_0 \cdot \nabla \tilde{p}_\varepsilon - U_0 \cdot \nabla k_\varepsilon) - (\hat{\nu}(|\nabla u_0|) \nabla u_0 \cdot \nabla \tilde{p}_\varepsilon - \hat{\nu}(|U_0|) U_0 \cdot \nabla k_\varepsilon)] \right| \\
&\leq \int_{\omega_\varepsilon} \nu_0 |\nabla u_0| |\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon| + \nu_0 |\nabla u_0 - U_0| |\nabla k_\varepsilon| \\
&\quad + \int_{\omega_\varepsilon} \hat{\nu}(|\nabla u_0|) |\nabla u_0| |\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon| + |\hat{\nu}(|\nabla u_0|) \nabla u_0 - \hat{\nu}(|U_0|) U_0| |\nabla k_\varepsilon| \\
&\leq \int_{\omega_\varepsilon} \|(\nu_0 + \hat{\nu}(|\nabla u_0|)) \nabla u_0\|_{L^\infty(\Omega^d)} |\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon| \\
&\quad + \int_{\omega_\varepsilon} (\nu_0 |\nabla u_0 - U_0| - |\hat{\nu}(|\nabla u_0|) \nabla u_0 - \hat{\nu}(|U_0|) U_0|) |\nabla k_\varepsilon| \\
&\leq |\omega|^{1/2} \varepsilon \left[ \|(\nu_0 + \hat{\nu}(|\nabla u_0|)) \nabla u_0\|_{L^\infty(\Omega^d)} \|\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon\|_{L^2(\omega_\varepsilon)} + (\nu_0 + 1) \delta \|\nabla k_\varepsilon\|_{L^2(\omega_\varepsilon)} \right] \\
&\leq |\omega|^{1/2} \varepsilon \left[ \|(\nu_0 + \hat{\nu}(|\nabla u_0|)) \nabla u_0\|_{L^\infty(\Omega^d)} \|\nabla \tilde{p}_\varepsilon - \nabla k_\varepsilon\|_{L^2(D)} + (\nu_0 + 1) \delta \|\nabla k_\varepsilon\|_{L^2(D)} \right] \\
&\leq O(\varepsilon) o(\varepsilon) + O(\varepsilon) \delta O(\varepsilon) = o(\varepsilon^2).
\end{aligned}$$

For the third integral on the right hand side of (4.161), we get

$$\begin{aligned}
& \left| \int_{\omega_\varepsilon} [\nu_0(\nabla \tilde{u}_\varepsilon \cdot \nabla p_0 - \nabla h_\varepsilon \cdot P_0) - (DT(\nabla u_0)\nabla \tilde{u}_\varepsilon \cdot \nabla p_0 - DT(U_0)\nabla h_\varepsilon \cdot P_0)] \right| \\
& \leq \int_{\omega_\varepsilon} \nu_0 (|\nabla p_0| |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon| + |\nabla h_\varepsilon| |\nabla p_0 - P_0|) \\
& \quad + \int_{\omega_\varepsilon} |DT(\nabla u_0)\nabla p_0| |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon| + |DT(\nabla u_0)\nabla p_0 - DT(U_0)P_0| |\nabla h_\varepsilon| \\
& \leq \int_{\omega_\varepsilon} \|\nu_0|\nabla p_0| + |DT(\nabla u_0)\nabla p_0|\|_{L^\infty(\Omega^d)} |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon| \\
& \quad + \int_{\omega_\varepsilon} |\nabla h_\varepsilon| (\nu_0|\nabla p_0 - P_0| + |DT(\nabla u_0)\nabla p_0 - DT(U_0)P_0|) \\
& \leq \omega^{1/2}\varepsilon \left[ \|\nu_0|\nabla p_0| + |DT(\nabla u_0)\nabla p_0|\|_{L^\infty(\Omega^d)} \|\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon\|_{L^2(\omega_\varepsilon)} + (\nu_0 + 1)\delta \|\nabla h_\varepsilon\|_{L^2(\omega_\varepsilon)} \right] \\
& \leq \omega^{1/2}\varepsilon \left[ \|\nu_0|\nabla p_0| + |DT(\nabla u_0)\nabla p_0|\|_{L^\infty(\Omega^d)} \|\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon\|_{L^2(D)} + (\nu_0 + 1)\delta \|\nabla h_\varepsilon\|_{L^2(D)} \right] \\
& \leq O(\varepsilon)o(\varepsilon) + O(\varepsilon)\delta O(\varepsilon) = o(\varepsilon^2),
\end{aligned}$$

where we used Cauchy-Schwarz's inequality and estimated the terms  $\|\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon\|_{L^2(D)}$  and  $\|\nabla h_\varepsilon\|_{L^2(D)}$  by (4.71) and (4.52), respectively. Thus, (4.161) yields

$$j_2(\varepsilon) - \tilde{j}_2(\varepsilon) = \int_{D \setminus \omega_\varepsilon} S_{\nabla u_0}(\nabla \tilde{u}_\varepsilon) \cdot \nabla p_0 - S_{U_0}(\nabla h_\varepsilon) \cdot P_0 + o(\varepsilon^2). \quad (4.162)$$

The integral on the right hand side can be split into three terms,

$$\begin{aligned}
& \int_{D \setminus \omega_\varepsilon} S_{\nabla u_0}(\nabla \tilde{u}_\varepsilon) \cdot \nabla p_0 - S_{U_0}(\nabla h_\varepsilon) \cdot P_0 \\
& = \int_{D \setminus \omega_\varepsilon} (S_{\nabla u_0}(\nabla \tilde{u}_\varepsilon) - S_{\nabla u_0}(\nabla h_\varepsilon)) \cdot \nabla p_0 \\
& \quad + \int_{D \setminus \omega_\varepsilon} (S_{\nabla u_0}(\nabla h_\varepsilon) - S_{U_0}(\nabla h_\varepsilon)) \cdot \nabla p_0 + \int_{D \setminus \omega_\varepsilon} S_{U_0}(\nabla h_\varepsilon) \cdot (\nabla p_0 - P_0).
\end{aligned} \quad (4.163)$$

- Regarding the first term on the right hand side of (4.163), condition (4.35) and estimate (4.72) entail that

$$\begin{aligned}
\int_{D \setminus \omega_\varepsilon} |S_{\nabla u_0}(\nabla \tilde{u}_\varepsilon) - S_{\nabla u_0}(\nabla h_\varepsilon)| & \leq c_4 \int_{D \setminus \omega_\varepsilon} |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon| (|\nabla \tilde{u}_\varepsilon| + |\nabla h_\varepsilon|) \\
& \leq c_4 \int_D |\nabla \tilde{u}_\varepsilon - \nabla h_\varepsilon| (|\nabla \tilde{u}_\varepsilon| + |\nabla h_\varepsilon|) = o(\varepsilon^2).
\end{aligned}$$

As  $\nabla p_0 \in L^\infty(D)$ , it follows

$$\int_{D \setminus \omega_\varepsilon} (S_{\nabla u_0}(\nabla \tilde{u}_\varepsilon) - S_{\nabla u_0}(\nabla h_\varepsilon)) \cdot \nabla p_0 = o(\varepsilon^2).$$

2. Regarding the second term on the right-hand side of (4.163), condition (4.36) and estimate (4.69) entail that

$$\begin{aligned} \int_{D \setminus \omega_\varepsilon} |S_{\nabla u_0}(\nabla h_\varepsilon) - S_{U_0}(\nabla h_\varepsilon)| &\leq c_5 \int_{D \setminus \omega_\varepsilon} |\nabla u_0 - U_0| |\nabla h_\varepsilon|^2 \\ &\leq c_5 \int_D |\nabla u_0 - U_0| |\nabla h_\varepsilon|^2 = o(\varepsilon^2). \end{aligned}$$

Again, as  $\nabla p_0 \in L^\infty(D)$ , it follows

$$\int_{D \setminus \omega_\varepsilon} (S_{\nabla u_0}(\nabla h_\varepsilon) - S_{U_0}(\nabla h_\varepsilon)) \cdot \nabla p_0 = o(\varepsilon^2).$$

3. For the third term on the right-hand side of (4.163), according to (4.41) derived from (4.35), it holds

$$\int_{D \setminus \omega_\varepsilon} |S_{U_0}(\nabla h_\varepsilon)| |\nabla p_0 - P_0| \leq \int_{D \setminus \omega_\varepsilon} c_4 |\nabla p_0 - P_0| |\nabla h_\varepsilon|^2.$$

Hence it follows from estimate (4.106) that

$$\int_{D \setminus \omega_\varepsilon} |S_{U_0}(\nabla h_\varepsilon)| |\nabla p_0 - P_0| = o(\varepsilon^2).$$

Finally, we conclude from the estimates above of the three terms on the right-hand side of (4.163) that

$$\int_{D \setminus \omega_\varepsilon} S_{\nabla u_0}(\nabla \tilde{u}_\varepsilon) \cdot \nabla p_0 - S_{U_0}(\nabla h_\varepsilon) \cdot P_0 = o(\varepsilon^2),$$

and, therefore, equation (4.162) yields

$$j_2(\varepsilon) - \tilde{j}_2(\varepsilon) = o(\varepsilon^2)$$

which completes the proof of Lemma 4.42.  $\square$

## 4.5 Topological Asymptotic Expansion: Case II

In this section, we will derive the topological derivative for the reverse scenario to the case considered in Section 4.4, i.e., we want to know how the objective function is affected by the creation of a small region of ferromagnetic material inside a region of air; see Figure 4.3. Most parts of the derivation will be analogous to Section 4.4. What is different in this section is that the material behavior outside the inclusion  $\omega_\varepsilon$  is linear. This yields that in particular the result corresponding to Theorem 4.24 about the asymptotic behavior of the variation of the direct state at scale 1 simplifies significantly.

### 4.5.1 Simplified Model Problem

In analogy to Section 4.1.1, we introduce a simplified model problem with homogeneous background material. In this simplified setting, the unperturbed state equation reads

$$\text{Find } u_0^{(2)} \in H_0^1(D) \text{ such that } \int_D \nu_0 \nabla u_0^{(2)} \cdot \nabla \eta = \langle F, \eta \rangle \quad \forall \eta \in H_0^1(D). \quad (4.164)$$

Here, again  $F \in H^{-1}(D)$  represents the sources given by the permanent magnetization and electric currents, and  $\nu_0$  denotes the magnetic reluctivity of air. In order to avoid confusion with the corresponding quantities in Case I, we mark all quantities arising in Case II with a superscript (2).

### 4.5.2 Perturbed Nonlinear Equation

Recall the notation introduced in Section 4.1.2. In particular, note that  $x_0 \in \Omega^d \subset\subset D \setminus \text{supp}(F)$  is the point around which we perturb the material coefficient and, for  $\varepsilon > 0$ ,  $\omega_\varepsilon = x_0 + \varepsilon\omega$  is the inclusion around  $x_0$  of radius  $\varepsilon$  and shape  $\omega$ . We define the operator

$$T_\varepsilon^{(2)}(x, W) := \chi_{D \setminus \omega_\varepsilon}(x) \nu_0 W + \chi_{\omega_\varepsilon}(x) T(W), \quad (4.165)$$

for  $x \in D$  and  $W \in \mathbb{R}^2$  where  $T$  is defined in (2.24). The Jacobian of  $T_\varepsilon^{(2)}$  is given by

$$DT_\varepsilon^{(2)}(x, W) = \chi_{D \setminus \omega_\varepsilon}(x) \nu_0 I + \chi_{\omega_\varepsilon}(x) DT(W).$$

Furthermore, we introduce

$$S_W^{\varepsilon, (2)}(x, V) := \chi_{\omega_\varepsilon}(x) S_W(V), \quad (4.166)$$

for  $V, W \in \mathbb{R}^2$ , with the operator  $S$  defined in (4.24).

Thus, in the simplified setting introduced in Section 4.5.1, the perturbed state equation reads

$$\text{Find } u_\varepsilon^{(2)} \in H_0^1(D) \text{ such that } \int_D T_\varepsilon^{(2)}(x, \nabla u_\varepsilon^{(2)}) \cdot \nabla \eta = \langle F, \eta \rangle \quad \forall \eta \in H_0^1(D). \quad (4.167)$$

**Remark 4.46.** It is easy to see that the statements of Lemma 4.7 and Lemma 4.8 also hold for the  $x$ -dependent operators  $T_\varepsilon^{(2)}$ ,  $DT_\varepsilon^{(2)}$ ,  $S^{\varepsilon, (2)}$  for each point in their domain of definition with the same constants.

### 4.5.3 Expansion of Cost Functional

We make the same assumption (4.7) on the cost functional as in Section 4.1.3, where  $u_\varepsilon$  and  $u_0$  are replaced by  $u_\varepsilon^{(2)}$  and  $u_0^{(2)}$ , respectively.

### 4.5.4 Variation of Direct State

Again, we are interested in the difference between the perturbed and the unperturbed direct state,  $u_\varepsilon^{(2)} - u_0^{(2)} =: \tilde{u}_\varepsilon^{(2)}$ . We will refer to  $\tilde{u}_\varepsilon^{(2)}$  as the variation of the direct state at scale  $\varepsilon$  in Case II.

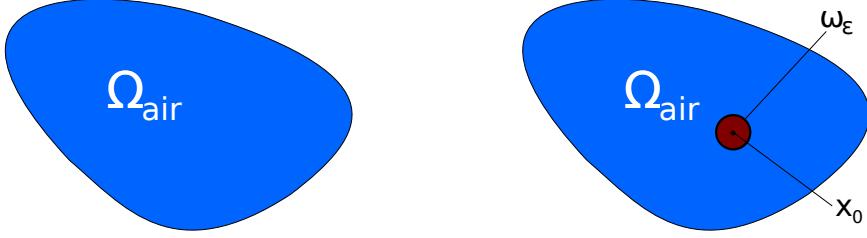


Figure 4.3: Left: Unperturbed configuration for Case II. Right: Perturbed configuration for Case II.

#### 4.5.4.1 Regularity Assumptions

We make the same regularity assumptions as in Section 4.4.1.1:

**Assumption 7.** There exists  $\beta > 0$  such that

$$u_0^{(2)}|_{\Omega^d} \in C^{1,\beta}(\Omega^d).$$

Again, we immediately get the following properties:

$$\begin{aligned} \nabla u_0^{(2)}|_{\Omega^d} &\in C^{0,\beta}(\Omega^d) \\ \nabla u_0^{(2)}|_{\Omega^d} &\in L^\infty(\Omega^d) \end{aligned} \tag{4.168}$$

#### 4.5.4.2 Step 1: Variation $u_\varepsilon^{(2)} - u_0^{(2)}$

Comparing the state equation (4.167) for the perturbed configuration ( $\varepsilon > 0$ ) with the state equation (4.164) of the unperturbed configuration in Case II (see Figure 4.3), we get

$$\begin{aligned} 0 &= \int_D T_\varepsilon^{(2)}(x, \nabla u_\varepsilon^{(2)}) \cdot \nabla \eta - \int_D T_0^{(2)}(\nabla u_0^{(2)}) \cdot \nabla \eta \\ &= \int_D (T_\varepsilon^{(2)}(x, \nabla u_\varepsilon^{(2)}) - T_\varepsilon^{(2)}(x, \nabla u_0^{(2)})) \cdot \nabla \eta + \int_{\omega_\varepsilon} (\hat{\nu}(|\nabla u_0^{(2)}|) - \nu_0) \nabla u_0^{(2)} \cdot \nabla \eta. \end{aligned}$$

This yields the boundary value problem for the variation of the direct state at scale  $\varepsilon$  in Case II,  $\tilde{u}_\varepsilon^{(2)}$ :

$$\begin{aligned} &\text{Find } \tilde{u}_\varepsilon^{(2)} \in H_0^1(D) \text{ such that} \\ &\int_D (T_\varepsilon^{(2)}(x, \nabla u_0^{(2)} + \nabla \tilde{u}_\varepsilon^{(2)}) - T_\varepsilon^{(2)}(x, \nabla u_0^{(2)})) \cdot \nabla \eta \\ &= \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|\nabla u_0^{(2)}|)) \nabla u_0^{(2)} \cdot \nabla \eta \quad \forall \eta \in H_0^1(D). \end{aligned} \tag{4.169}$$

Note the different sign on the right hand side compared to the corresponding boundary value problem (4.46) in Case I.

#### 4.5.4.3 Step 2: Approximation of Variation $u_\varepsilon^{(2)} - u_0^{(2)}$

We approximate boundary value problem (4.169) by the same boundary value problem where we replace the function  $\nabla u_0^{(2)}$  by the constant  $U_0^{(2)} := \nabla u_0^{(2)}(x_0)$ . We denote the solution to the arising boundary value problem by  $h_\varepsilon^{(2)}$ :

$$\begin{aligned} & \text{Find } h_\varepsilon^{(2)} \in H_0^1(D) \text{ such that} \\ & \int_D (T_\varepsilon^{(2)}(x, U_0^{(2)} + \nabla h_\varepsilon^{(2)}) - T_\varepsilon^{(2)}(x, U_0^{(2)})) \cdot \nabla \eta \\ & = \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|U_0^{(2)}|)) U_0^{(2)} \cdot \nabla \eta \quad \forall \eta \in H_0^1(D). \end{aligned} \quad (4.170)$$

#### 4.5.4.4 Step 3: Change of Scale

We proceed analogously to Section 4.4.1.4. We start out from boundary value problem (4.170) and, for fixed  $\varepsilon > 0$ , perform the change of scale  $x \mapsto x/\varepsilon$ . This means that, for small  $\varepsilon > 0$ , we rescale both  $D$  and the inclusion  $\omega_\varepsilon$ , resulting in a “much larger” domain  $D/\varepsilon$  and an inclusion of unit size. Then, by replacing the large bounded domain  $D/\varepsilon$  by the entire plane  $\mathbb{R}^2$ , we arrive at a transmission problem on  $\mathbb{R}^2$  which is an  $\varepsilon$ -independent approximation to problems (4.170) and (4.169) at scale 1. For that purpose, we introduce the  $\varepsilon$ -independent operators corresponding to (4.165) and (4.166), respectively:

$$\begin{aligned} \tilde{T}^{(2)}(x, W) &:= \chi_{\mathbb{R}^2 \setminus \omega}(x) \nu_0 W + \chi_\omega T(W), \\ \tilde{S}_W^{(2)}(x, V) &:= \chi_\omega(x) S_W(V), \end{aligned} \quad (4.171)$$

which are defined for all  $x \in \mathbb{R}^2$  and  $V, W \in \mathbb{R}^2$ . Note that the Jacobian of  $\tilde{T}^{(2)}$  is given by

$$\tilde{D}\tilde{T}^{(2)}(x, W) := \chi_{\mathbb{R}^2 \setminus \omega}(x) \nu_0 I + \chi_\omega D T(W).$$

The arising nonlinear transmission problem defining the variation of the direct state at scale 1 in Case II,  $H^{(2)}$ , reads

$$\begin{aligned} & \text{Find } H^{(2)} \in \mathcal{H}(\mathbb{R}^2) \text{ such that} \\ & \int_{\mathbb{R}^2} (\tilde{T}^{(2)}(x, U_0^{(2)} + \nabla H^{(2)}) - \tilde{T}^{(2)}(x, U_0^{(2)})) \cdot \nabla \eta \\ & = \int_\omega (\nu_0 - \hat{\nu}(|U_0^{(2)}|)) U_0^{(2)} \cdot \nabla \eta \quad \forall \eta \in \mathcal{H}(\mathbb{R}^2). \end{aligned} \quad (4.172)$$

**Remark 4.47.** It is easy to see that the statements of Lemma 4.7 and Lemma 4.8 also hold for the  $x$ -dependent operators  $\tilde{T}^{(2)}$ ,  $D\tilde{T}^{(2)}$ , and  $\tilde{S}^{(2)}$  for each point in their domain of definition with the same constants.

As in Section 4.4.1.4, we can show the existence of a unique solution to problem (4.172).

**Proposition 4.48.** Let Assumption 1 hold. Then there exists a unique solution  $H^{(2)} \in \mathcal{H}(\mathbb{R}^2)$  to (4.172).

*Proof.* The proof is analogous to the proof of Proposition 4.17 using the theorem of Zaran-tonello (Theorem 2.5). Only the right hand side  $L$  must be replaced by  $L^{(2)} := -L \in \mathcal{H}(\mathbb{R}^2)^*$ . The proof only uses properties (4.33) and (4.34) for  $\tilde{T}$ , which also hold for  $\tilde{T}^{(2)}$  by virtue of Remark 4.47.  $\square$

#### 4.5.4.5 Step 4: Asymptotic Behavior of Variations of Direct State

Similarly to Section 4.4.1.5 in Case I, we need to establish a sufficiently fast decay of the unique solution  $H^{(2)}$  as  $|x|$  tends to infinity. For that purpose, the procedure in Case II is significantly less tedious which is due to the fact that the material behavior in  $\mathbb{R}^2 \setminus \omega$  in (4.172) is linear.

Let  $H^{(2)} \in \mathcal{H}(\mathbb{R}^2)$  be the unique solution to problem (4.172). Again, given an element  $\hat{H}^{(2)} \in \mathcal{H}^w(\mathbb{R}^2)$  of the class  $H^{(2)} \in \mathcal{H}(\mathbb{R}^2)$  and  $\varepsilon > 0$ , we define  $H_\varepsilon^{(2)} : D \rightarrow \mathbb{R}$  by

$$H_\varepsilon^{(2)}(x) := \varepsilon \hat{H}^{(2)}(\varepsilon^{-1}x) \quad (4.173)$$

for  $x \in D$ . Again, noting that, for the weight function  $w$  defined in (4.42),

$$\underline{w} := \inf_{x \in D} w\left(\frac{x}{\varepsilon}\right) > 0,$$

it is easy to see that  $\hat{H}^{(2)} \in \mathcal{H}^w(\mathbb{R}^2)$  implies  $H_\varepsilon^{(2)} \in H^1(D)$ . We get estimates analogous to Lemma 4.18:

**Lemma 4.49.** *Let Assumption 1 as well as Assumption 7 hold. Then*

$$\|\nabla \tilde{u}_\varepsilon^{(2)}\|_{L^2(D)}^2 = \mathcal{O}(\varepsilon^2), \quad (4.174)$$

$$\|\nabla h_\varepsilon^{(2)}\|_{L^2(D)}^2 = \mathcal{O}(\varepsilon^2), \quad (4.175)$$

$$\|\nabla H_\varepsilon^{(2)}\|_{L^2(D)}^2 = \mathcal{O}(\varepsilon^2). \quad (4.176)$$

*Proof.* The proof is analogous to the proof of Lemma 4.18, which relies on property (4.33) for the operator  $T_\varepsilon$  and the regularity assumption on  $u_0$ . Property (4.33) also holds for  $T_\varepsilon^{(2)}$  by virtue of Remark 4.46 and we made the same regularity assumption on  $u_0^{(2)}$ , so the same proof steps can be conducted.  $\square$

**Remark 4.50.** *By application of the triangle inequality, it follows immediately from estimates (4.174), (4.175) and (4.176) that*

$$\|\nabla \tilde{u}_\varepsilon^{(2)} - \nabla h_\varepsilon^{(2)}\|_{L^2(D)}^2 = \mathcal{O}(\varepsilon^2), \quad (4.177)$$

$$\|\nabla h_\varepsilon^{(2)} - \nabla H_\varepsilon^{(2)}\|_{L^2(D)}^2 = \mathcal{O}(\varepsilon^2). \quad (4.178)$$

The asymptotic behavior of the solution  $H^{(2)}$  to (4.172) can be established similarly to Proposition 4.33, without the use of subsolution and supersolution as it was necessary for showing Theorem 4.24:

**Proposition 4.51.** *Let  $H^{(2)} \in \mathcal{H}(\mathbb{R}^2)$  be the unique solution to (4.172). Then, there exists an element  $\tilde{H}^{(2)}$  of the class  $H^{(2)} \in \mathcal{H}(\mathbb{R}^2)$  such that*

$$\tilde{H}^{(2)}(y) = \mathcal{O}(|y|^{-1}) \text{ as } |y| \rightarrow \infty. \quad (4.179)$$

*Proof.* Since in Case II, we have linear material outside the inclusion, this proof is similar to the proof of the asymptotic behavior of the variation of the adjoint state in both Cases I and II (Propositions 4.33 and 4.57, respectively).

The variation of the direct state at scale 1 in Case II,  $H^{(2)} \in \mathcal{H}(\mathbb{R}^2)$ , is given as the solution to (4.172), which can be rewritten as

Find  $H^{(2)} \in \mathcal{H}(\mathbb{R}^2)$  such that

$$\int_{\mathbb{R}^2} \nabla H^{(2)} \cdot \nabla \eta = \int_{\omega} \left( 1 - \frac{\hat{\nu}(|U_0^{(2)} + \nabla H^{(2)}|)}{\nu_0} \right) (U_0^{(2)} + \nabla H^{(2)}) \cdot \nabla \eta \quad \forall \eta \in \mathcal{H}(\mathbb{R}^2). \quad (4.180)$$

This amounts to a Laplace equation on  $\mathbb{R}^2$  supported on  $\omega$ . As in the proof of Proposition 4.33, let  $E$  the fundamental solution to the Laplace equation in  $\mathbb{R}^2$  which is given for all  $y \in \mathbb{R}^2$ ,  $y \neq 0$ , as

$$\begin{aligned} E(y) &= \frac{1}{2\pi} \log |y| \quad \text{with} \\ |\nabla E(y)| &= \frac{1}{2\pi} |y|^{-1}. \end{aligned}$$

Let  $\mathcal{F}$  denote the distribution defined by

$$\langle \mathcal{F}, \eta \rangle = \int_{\omega} \left( \frac{\hat{\nu}(|U_0^{(2)} + \nabla H^{(2)}|)}{\nu_0} - 1 \right) (U_0^{(2)} + \nabla H^{(2)}) \cdot \nabla \eta \quad \forall \eta \in C_0^\infty(\mathbb{R}^2).$$

Then, by (4.180) we have  $\Delta H^{(2)} = \mathcal{F}$ . Therefore, let an element  $\tilde{H}^{(2)}$  of the class  $H^{(2)}$  be given by  $\tilde{H}^{(2)} = \mathcal{F} * E$ , i.e.,

$$\tilde{H}^{(2)}(y) = \int_{\omega} \left( \frac{\hat{\nu}(|U_0^{(2)} + \nabla H^{(2)}(z)|)}{\nu_0} - 1 \right) (U_0^{(2)} + \nabla H^{(2)}(z)) \cdot \nabla E(y - z) dz.$$

Since  $U_0^{(2)} + \nabla H^{(2)} \in L^2(\omega)$ , the Cauchy-Schwarz inequality yields

$$|\tilde{H}^{(2)}(y)| \leq C \left( \int_{\omega} |\nabla E(y - z)|^2 dz \right)^{1/2},$$

with  $C := \left( 1 - \frac{\nu}{\nu_0} \right) \|U_0^{(2)} + \nabla H^{(2)}\|_{L^2(\omega)}$ .

The remainder of the proof is identical to the proof of Proposition 4.33.  $\square$

From now on, function  $H_\varepsilon^{(2)}$  is defined choosing  $\hat{H}^{(2)} = \tilde{H}^{(2)}$  in (4.173) where  $\tilde{H}^{(2)}$  is as in Proposition 4.51, i.e.,

$$H_\varepsilon^{(2)}(x) := \varepsilon \tilde{H}^{(2)}(\varepsilon^{-1}x), \quad x \in D.$$

#### 4.5.4.6 Estimates for the Variations of the Direct State

Recall the smooth function  $\theta : \mathbb{R}^2 \rightarrow \mathbb{R}$  defined in Section 4.4.1.6 as

$$\theta(x) = 0, \quad x \in B(0, \rho), \quad \text{and} \quad \theta(x) = 1, \quad x \in \mathbb{R}^2 \setminus B(0, R),$$

with  $0 < \rho < R$  defined in (4.4). Define the function  $\kappa_\varepsilon^{(2)} : D \rightarrow \mathbb{R}$  given by

$$\kappa_\varepsilon^{(2)}(x) = \theta(x) H_\varepsilon^{(2)}(x).$$

**Lemma 4.52.** *It holds that  $\kappa_\varepsilon^{(2)} \in H^1(D)$  and  $H_\varepsilon^{(2)} - \kappa_\varepsilon^{(2)} \in H_0^1(D)$ . Moreover,*

$$\|\nabla \kappa_\varepsilon^{(2)}\|_{L^2(D)}^2 = o(\varepsilon^2). \quad (4.181)$$

*Proof.* The proof is analogous to the proof of Lemma 4.26, exploiting the asymptotic behavior (4.179) of  $H^{(2)}$ . The only difference lies in the second step of the proof of (4.181) as the asymptotic behavior of  $H$  (4.63) is different from that of  $H^{(2)}$  (4.179):  $H = \mathcal{O}(|y|^{-\tau})$  with  $\tau > 0$  and  $H^{(2)} = \mathcal{O}(|y|^{-1})$ . Here, it suffices to replace the exponent  $\tau$  by 1 in the proof and the result (4.181) follows.  $\square$

**Proposition 4.53.** *Let Assumptions 1 and 7 hold. Then*

$$\|\nabla h_\varepsilon^{(2)} - \nabla H_\varepsilon^{(2)}\|_{L^2(D)}^2 = o(\varepsilon^2), \quad (4.182)$$

$$\forall \alpha > 0 \forall r \in (0, 1) : \int_{D \setminus B(0, \alpha \varepsilon^r)} |\nabla h_\varepsilon^{(2)}|^2 = o(\varepsilon^2), \quad (4.183)$$

$$\int_D |\nabla u_0^{(2)} - U_0^{(2)}| |\nabla h_\varepsilon^{(2)}|^2 = o(\varepsilon^2), \quad (4.184)$$

$$\int_D |\nabla h_\varepsilon^{(2)} - \nabla H_\varepsilon^{(2)}| (|\nabla h_\varepsilon^{(2)}| + |\nabla H_\varepsilon^{(2)}|) = o(\varepsilon^2). \quad (4.185)$$

*Proof.* The proof is analogous to the proof of Proposition 4.27 which, again, is based on properties (4.33) and (4.34), the regularity of  $u_0^{(2)}$  and the asymptotic behavior (4.181) of  $\nabla \kappa_\varepsilon^{(2)}$ . Properties (4.33) and (4.34) also hold for the operator  $T_\varepsilon^{(2)}$  by virtue of Remark 4.46. Note that, in Case II, the asymptotic behavior (4.181) of  $\kappa_\varepsilon^{(2)}$  holds without Assumption 4 and without the assumption that  $\omega = B(0, 1)$  because the asymptotic behavior (4.179) of  $H^{(2)}$  could be established without these assumptions.  $\square$

**Proposition 4.54.** *Let Assumptions 1 and 7 hold. Then*

$$\|\nabla \tilde{u}_\varepsilon^{(2)} - \nabla h_\varepsilon^{(2)}\|_{L^2(D)}^2 = o(\varepsilon^2), \quad (4.186)$$

$$\int_D |\nabla \tilde{u}_\varepsilon^{(2)} - \nabla h_\varepsilon^{(2)}| (|\nabla \tilde{u}_\varepsilon^{(2)}| + |\nabla h_\varepsilon^{(2)}|) = o(\varepsilon^2), \quad (4.187)$$

$$\forall \alpha > 0 \forall r \in (0, 1) : \int_{D \setminus B(0, \alpha \varepsilon^r)} |\nabla \tilde{u}_\varepsilon^{(2)}|^2 = o(\varepsilon^2). \quad (4.188)$$

*Proof.* Again, the proof is analogous to the proof of the corresponding proposition in Section 4.4.1.6, Proposition 4.28. For the proof of (4.186), first note that properties (4.33) and (4.34) also hold for the operator  $T_\varepsilon^{(2)}$  by virtue of Remark 4.46. In the same way as in Section 4.4.1.6, it can be seen that

$$\begin{aligned} c_2 \|\nabla \tilde{u}_\varepsilon^{(2)} - \nabla h_\varepsilon^{(2)}\|_{L^2(D)}^2 &\leq \int_D \left[ T_\varepsilon^{(2)}(x, \nabla u_0^{(2)} + \nabla \tilde{u}_\varepsilon^{(2)}) - T_\varepsilon^{(2)}(x, \nabla u_0^{(2)} + \nabla h_\varepsilon^{(2)}) \right] \cdot \nabla \eta \\ &= \sum_{i=1}^5 \mathcal{E}_i^{(2)}(\varepsilon), \end{aligned}$$

where the  $\mathcal{E}_i^{(2)}$  are defined analogously as in the proof of Proposition 4.28 with  $\mathcal{E}_1^{(2)}$  having a different sign. Note that, for  $\alpha = \rho/\lambda$  with  $\rho, \lambda$  defined in (4.4) and the same choices of  $\varepsilon$

and  $r$  as in the proof of Proposition 4.28,  $T_\varepsilon^{(2)}$  is linear outside  $B(0, \alpha\varepsilon^r)$ , which yields that  $\mathcal{E}_4^{(2)} + \mathcal{E}_5^{(2)} = 0$ . Due to Assumption 7 and the Lipschitz continuity of the operator  $T$  (4.34), we can make use of estimate (4.143) (with  $\nabla u_0$ ,  $U_0$  replaced by  $\nabla u_0^{(2)}$ ,  $U_0^{(2)}$ , respectively). The estimation of  $\mathcal{E}_1^{(2)}, \mathcal{E}_2^{(2)}, \mathcal{E}_3^{(2)}$  is then analogous to the estimation of  $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$  in the proof of Proposition 4.28, using estimate (4.177).

Estimates (4.187) and (4.188) can be shown in exactly the same way as in the proof of Proposition 4.28.  $\square$

#### 4.5.5 Variation of Adjoint State

The perturbed adjoint equation in the Case II reads

$$\begin{aligned} & \text{Find } p_\varepsilon^{(2)} \in H_0^1(D) \text{ such that} \\ & \int_D DT_\varepsilon^{(2)}(x, \nabla u_0^{(2)}) \nabla p_\varepsilon^{(2)} \cdot \nabla \eta = -\langle \tilde{G}, \eta \rangle \quad \forall \eta \in H_0^1(D). \end{aligned}$$

Here,  $\tilde{G}$  again denotes the bounded linear functional such that the functional  $\mathcal{J}_\varepsilon$  satisfies an expansion of the form (4.7). Note that, here, we exploited the symmetry of  $DT_\varepsilon$ . For  $\varepsilon = 0$ , we get the unperturbed adjoint equation,

$$\begin{aligned} & \text{Find } p_0^{(2)} \in H_0^1(D) \text{ such that} \\ & \int_D \nu_0 \nabla p_0^{(2)} \cdot \nabla \eta = -\langle \tilde{G}, \eta \rangle \quad \forall \eta \in H_0^1(D), \end{aligned} \tag{4.189}$$

where we used that  $DT_0^{(2)}(x, \nabla u_0^{(2)}) = \nu_0 I$  according to the definition of  $T_\varepsilon^{(2)}$  in (4.165).

##### 4.5.5.1 Regularity Assumptions

We again assume sufficient regularity of the solution  $p_0^{(2)}$  to the unperturbed adjoint equation.

**Assumption 8.** There exists  $\tilde{\beta} > 0$  such that

$$p_0^{(2)}|_{\Omega^d} \in C^{1,\tilde{\beta}}(\Omega^d).$$

Assumption 8 immediately yields the properties

$$\begin{aligned} \nabla p_0^{(2)}|_{\Omega^d} & \in C^{0,\tilde{\beta}}(\Omega^d), \\ \nabla p_0^{(2)}|_{\Omega^d} & \in L^\infty(\Omega^d). \end{aligned} \tag{4.190}$$

##### 4.5.5.2 Step 1: Variation $p_\varepsilon^{(2)} - p_0^{(2)}$

Subtracting (4.75) from (4.74) yields

$$\begin{aligned} 0 &= \int_D DT_\varepsilon^{(2)}(x, \nabla u_0^{(2)}) \nabla p_\varepsilon^{(2)} \cdot \nabla \eta - \int_D \nu_0 \nabla p_0^{(2)} \cdot \nabla \eta \\ &= \int_D DT_\varepsilon^{(2)}(x, \nabla u_0^{(2)}) \nabla \tilde{p}_\varepsilon^{(2)} \cdot \nabla \eta - \int_{\omega_\varepsilon} (\nu_0 I - DT_\varepsilon^{(2)}(x, \nabla u_0^{(2)})) \nabla p_0^{(2)} \cdot \nabla \eta, \end{aligned}$$

where  $\tilde{p}_\varepsilon^{(2)} = p_\varepsilon^{(2)} - p_0^{(2)}$  denotes the variation of the adjoint state at scale  $\varepsilon$  in Case II. Here we used that  $DT_\varepsilon^{(2)}(x, \nabla u_0^{(2)})$  is equal to  $\nu_0 I$  outside  $\omega_\varepsilon$ . Thus, noting that  $DT_\varepsilon^{(2)}(x, \nabla u_0^{(2)}) = DT(\nabla u_0^{(2)})$  inside  $\omega_\varepsilon$ , the boundary value problem for the variation of the adjoint state at scale  $\varepsilon$  in Case II reads

Find  $\tilde{p}_\varepsilon^{(2)} \in H_0^1(D)$  such that

$$\int_D DT_\varepsilon^{(2)}(x, \nabla u_0^{(2)}) \nabla \tilde{p}_\varepsilon^{(2)} \cdot \nabla \eta = \int_{\omega_\varepsilon} (\nu_0 I - DT(\nabla u_0^{(2)})) \nabla p_0^{(2)} \cdot \nabla \eta \quad \forall \eta \in H_0^1(D).$$

Also here, note the different sign of the right hand side in comparison with (4.78).

#### 4.5.5.3 Step 2: Approximation of Variation $p_\varepsilon^{(2)} - p_0^{(2)}$

As for the variation of the direct state, we approximate boundary value problem (4.191) by replacing the gradient of the adjoint and direct state by their values at the point  $x_0$ . Recall  $U_0^{(2)} := \nabla u_0^{(2)}(x_0)$  and let  $P_0^{(2)} := \nabla p_0^{(2)}(x_0)$ . Then the problem reads

Find  $k_\varepsilon^{(2)} \in H_0^1(D)$  such that

$$\int_D DT_\varepsilon^{(2)}(x, U_0^{(2)}) \nabla k_\varepsilon^{(2)} \cdot \nabla \eta = \int_{\omega_\varepsilon} (\nu_0 I - DT(U_0^{(2)})) P_0^{(2)} \cdot \nabla \eta \quad \forall \eta \in H_0^1(D). \quad (4.191)$$

#### 4.5.5.4 Step 3: Change of Scale

Here, we approximate problem (4.191) by performing a change of scale and sending the outer boundary of the rescaled domain  $D/\varepsilon$  to infinity in the same manner as for the variation of the direct state in Section 4.4.1.4. The arising linear transmission problem on the whole  $\mathbb{R}^2$  reads

Find  $K^{(2)} \in \mathcal{H}(\mathbb{R}^2)$  such that

$$\int_{\mathbb{R}^2} D\tilde{T}^{(2)}(x, U_0^{(2)}) \nabla K^{(2)} \cdot \nabla \eta = \int_{\omega} (\nu_0 I - DT(U_0^{(2)})) P_0^{(2)} \cdot \nabla \eta \quad \forall \eta \in \mathcal{H}(\mathbb{R}^2). \quad (4.192)$$

Again, we can show well-posedness of the above problem:

**Lemma 4.55.** *There exists a unique solution  $K^{(2)} \in \mathcal{H}(\mathbb{R}^2)$  of problem (4.192).*

*Proof.* The proof is completely analogous to the proof of Lemma 4.31 using the lemma of Lax-Milgram, exploiting property (4.32) in combination with Remark 4.47.  $\square$

#### 4.5.5.5 Step 4: Asymptotic Behavior of Variations of the Adjoint State

Let  $K^{(2)} \in \mathcal{H}(\mathbb{R}^2)$  be the unique solution to (4.192) and let  $\hat{K}^{(2)} \in \mathcal{H}^w(\mathbb{R}^2)$  denote a given element of the class  $K^{(2)}$ . For  $\varepsilon > 0$ , we introduce  $K_\varepsilon^{(2)} : D \rightarrow \mathbb{R}$  defined by

$$K_\varepsilon^{(2)}(x) := \varepsilon \hat{K}^{(2)}(\varepsilon^{-1}x).$$

As in Section 4.4.2.5, making the change of scale backward, as

$$\inf_{x \in D} w\left(\frac{x}{\varepsilon}\right) > 0,$$

it follows from  $\hat{K}^{(2)} \in \mathcal{H}^w(\mathbb{R}^2)$  that  $K_\varepsilon^{(2)} \in H^1(D)$ . The following estimates are analogous to Lemma 4.32:

**Lemma 4.56.** *Let Assumptions 1, 3, 7 and 8 hold. Then, it holds*

$$\|\nabla \tilde{p}_\varepsilon^{(2)}\|_{L^2(D)}^2 = O(\varepsilon^2), \quad (4.193)$$

$$\|\nabla k_\varepsilon^{(2)}\|_{L^2(D)}^2 = O(\varepsilon^2), \quad (4.194)$$

$$\|\nabla K_\varepsilon^{(2)}\|_{L^2(D)}^2 = O(\varepsilon^2). \quad (4.195)$$

*Proof.* Since we have the same regularity properties as in Section 4.4.2 and since property (4.32) also holds for the operator  $DT_\varepsilon^{(2)}$  by virtue of Remark 4.46, the proof can be conducted in the exactly same way as the proof of Lemma 4.32.  $\square$

Similarly to Proposition 4.33, we get the asymptotic behavior of an element of class  $K^{(2)}$ .

**Proposition 4.57.** *There exists an element  $\tilde{K}^{(2)}$  of the class  $K^{(2)} \in \mathcal{H}(\mathbb{R}^2)$  such that*

$$\tilde{K}^{(2)}(y) = \mathcal{O}(|y|^{-1}) \quad \text{as } |y| \rightarrow \infty. \quad (4.196)$$

*Proof.* By adding  $\int_\omega (\nu_0 I - DT(U_0^{(2)})) \nabla K^{(2)} \cdot \nabla \eta$  to both sides and dividing by  $\nu_0$ , transmission problem (4.192) can be reformulated as

$$\int_{\mathbb{R}^2} \nabla K^{(2)} \cdot \nabla \eta = \frac{1}{\nu_0} \int_\omega (\nu_0 I - DT(U_0^{(2)}))(P_0^{(2)} + \nabla K^{(2)}) \cdot \nabla \eta,$$

which can be rewritten as a Laplace equation in  $\mathbb{R}^2$  with a source supported by  $\omega$ . Then, the exactly same steps as in the proof of Proposition 4.33 can be conducted.  $\square$

Recall the smooth function  $\theta : \mathbb{R}^2 \rightarrow \mathbb{R}$  introduced in Section 4.4.1.6 satisfying

$$\theta(x) = 0, \quad x \in B(0, \rho), \quad \text{and} \quad \theta(x) = 1, \quad x \in \mathbb{R}^2 \setminus B(0, R),$$

and let the function  $\kappa_{a\varepsilon}^{(2)} : D \rightarrow \mathbb{R}$  be defined by

$$\kappa_{a\varepsilon}^{(2)}(x) := \theta(x) K_\varepsilon^{(2)}(x).$$

**Lemma 4.58.** *It holds  $\kappa_{a\varepsilon}^{(2)} \in H^1(D)$  and  $K_\varepsilon^{(2)} - \kappa_{a\varepsilon}^{(2)} \in H_0^1(D)$ . Moreover,*

$$\|\nabla \kappa_{a\varepsilon}^{(2)}\|_{L^2(D)}^2 = o(\varepsilon^2). \quad (4.197)$$

*Proof.* The proof is identical to the proof of Lemma 4.52, exploiting the asymptotic behavior (4.196) of  $K^{(2)}$ .  $\square$

**Lemma 4.59.** *Let Assumption 1 hold. Then, it holds*

$$\|\nabla k_\varepsilon^{(2)} - \nabla K_\varepsilon^{(2)}\|_{L^2(D)}^2 = o(\varepsilon^2). \quad (4.198)$$

*Proof.* For the proof of (4.198), we note that formally, the only difference between (4.80) and (4.192) is in the sign of the right hand side. The estimate that corresponds to (4.148) can easily be deduced. Since the function  $\kappa_{a\varepsilon}^{(2)}$  introduced above has the same asymptotic behavior (4.197) as the function  $\kappa_{a\varepsilon}$  (4.87), and since property (4.32) also holds for the operator  $DT_\varepsilon^{(2)}$  by virtue of Remark 4.46, the rest of the proof of (4.198) follows.  $\square$

**Lemma 4.60.** *Let Assumptions 1, 3, 7 and 8 be satisfied. Then, it holds*

$$\|\nabla \tilde{p}_\varepsilon^{(2)} - \nabla k_\varepsilon^{(2)}\|_{L^2(D)}^2 = o(\varepsilon^2). \quad (4.199)$$

*Proof.* The proof is analogous to the proof of Lemma 4.36 using regularity assumptions (4.168) and (4.190). We get the corresponding estimate to (4.152) whose right hand side can be split into three terms as in (4.153). The second and the third term can be estimated in the same way as in the proof of Lemma 4.36. For the term that corresponds to the first term on the right hand side of (4.153), note that the difference  $DT_\varepsilon^{(2)}(x, U_0) - DT_\varepsilon(x, \nabla u_0)$  vanishes outside  $\omega_\varepsilon$ . Therefore, it can be seen that, in Case II, we do not need an estimate of the type (4.89). The contribution from inside  $\omega_\varepsilon$  is treated as in Lemma 4.36 by exploiting that  $\omega_\varepsilon \subset B(0, \alpha\varepsilon^r)$  due to (4.151), and using (4.194) and the estimate corresponding to (4.150).  $\square$

#### 4.5.6 Topological Asymptotic Expansion

By Assumption 2 on the functional  $\mathcal{J}_\varepsilon : H_0^1(D) \rightarrow \mathbb{R}$ , we have

$$\mathcal{J}_\varepsilon(u_\varepsilon^{(2)}) = \mathcal{J}_0(u_0^{(2)}) + \langle \tilde{G}, u_\varepsilon^{(2)} - u_0^{(2)} \rangle + \delta_J \varepsilon^2 + R(\varepsilon),$$

where  $\tilde{G} \in H^{-1}(D)$ , and the remainder  $R(\varepsilon)$  is of the form (4.8). Again, it follows from estimate (4.188) that

$$R(\varepsilon) = o(\varepsilon^2).$$

Performing the analogous steps as in Section 4.4.3, we get

$$\mathcal{J}_\varepsilon(u_\varepsilon^{(2)}) - \mathcal{J}_0(u_0^{(2)}) = j_1^{(2)}(\varepsilon) + j_2^{(2)}(\varepsilon) + \delta_J \varepsilon^2 + o(\varepsilon^2), \quad (4.200)$$

where the terms  $j_1^{(2)}, j_2^{(2)}$  are defined as

$$j_1^{(2)}(\varepsilon) := - \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|\nabla u_0^{(2)}|)) \nabla u_0^{(2)} \cdot (\nabla p_0^{(2)} + \nabla \tilde{p}_\varepsilon^{(2)}), \quad (4.201)$$

$$j_2^{(2)}(\varepsilon) := S_{\nabla u_0^{(2)}}^{\varepsilon, (2)}(x, \nabla \tilde{u}_\varepsilon^{(2)}) \nabla (\nabla p_0^{(2)} + \nabla \tilde{p}_\varepsilon^{(2)}). \quad (4.202)$$

Here, again, the operator  $S_{\nabla u_0^{(2)}}^{\varepsilon, (2)}$  is as defined in (4.166) and represents the nonlinearity of the problem. Note that it vanishes in the linear case where the nonlinear function  $\hat{\nu}$  is replaced by a constant  $\nu_1$ .

##### 4.5.6.1 Expansion of Linear Term $j_1^{(2)}(\varepsilon)$

Following approximation steps 2 and 3 of Sections 4.5.5.3 and 4.5.5.4, respectively, we define

$$\tilde{j}_1^{(2)}(\varepsilon) := - \int_{\omega_\varepsilon} \left( \nu_0 - \hat{\nu}(|U_0^{(2)}|) \right) U_0^{(2)} \cdot \left( P_0^{(2)} + \nabla k_\varepsilon^{(2)} \right), \quad (4.203)$$

$$J_1^{(2)} := - \left( \nu_0 - \hat{\nu}(|U_0^{(2)}|) \right) \int_\omega U_0^{(2)} \cdot \left( P_0^{(2)} + \nabla K^{(2)} \right). \quad (4.204)$$

We get the same relations as in Section 4.4.3.1:

**Lemma 4.61.** *Let Assumption 1 hold. Then, for all  $\varepsilon \geq 0$  small enough, it holds*

$$\tilde{j}_1^{(2)}(\varepsilon) - \varepsilon^2 J_1^{(2)} = o(\varepsilon^2). \quad (4.205)$$

*Proof.* The proof can be conducted in the same way as the proof of Lemma 4.37 using estimate (4.198).  $\square$

**Lemma 4.62.** *Let Assumptions 1, 3, 7 and 8 hold. Then, for all  $\varepsilon \geq 0$  small enough, it holds*

$$j_1^{(2)}(\varepsilon) - \tilde{j}_1^{(2)}(\varepsilon) = o(\varepsilon^2). \quad (4.206)$$

*Proof.* The proof can be conducted in the same way as the proof of Lemma 4.38 using the continuity of  $\nabla u_0^{(2)}$  and  $\nabla p_0^{(2)}$  at  $x_0 = 0$  and estimates (4.199) and (4.194).  $\square$

As in Section 4.4.3.1, it follows from the linearity of equation (4.192) that the mapping

$$P_0^{(2)} \mapsto \left( \nu_0 - \hat{\nu}(|U_0^{(2)}|) \right) \cdot \left( \int_{\omega} P_0^{(2)} + \nabla K^{(2)} \right)$$

is a linear mapping from  $\mathbb{R}^2$  to  $\mathbb{R}^2$ , which only depends on the set  $\omega$ , and on the positive definite matrix  $DT(U_0^{(2)})$ . Therefore, there exists a matrix

$$\mathcal{M}^{(2)} = \mathcal{M}^{(2)}(\omega, DT(U_0^{(2)})), \quad (4.207)$$

such that

$$-\left( \nu_0 - \hat{\nu}(|U_0^{(2)}|) \right) \cdot \left( \int_{\omega} P_0^{(2)} + \nabla K^{(2)} \right) = \mathcal{M}^{(2)} P_0^{(2)}.$$

The relation between the matrix  $\mathcal{M}^{(2)}$  and the concept of polarization matrices, see, e.g., [11], will be explained in detail in Section 4.6. Eventually, it follows

$$J_1^{(2)} = (U_0^{(2)})^\top \mathcal{M}^{(2)} P_0^{(2)}. \quad (4.208)$$

An explicit formula for the matrix  $\mathcal{M}^{(2)} = \mathcal{M}^{(2)}(\omega, DT(U_0^{(2)}))$  is given in (4.234).

Summing up estimates (4.205) and (4.206), as well as (4.208), we get the following result:

**Proposition 4.63.** *Let Assumptions 1, 3, 7 and 8 hold. Then, there exists a matrix  $\mathcal{M}^{(2)} = \mathcal{M}^{(2)}(\omega, DT(U_0^{(2)})) \in \mathbb{R}^{2 \times 2}$  such that*

$$j_1^{(2)}(\varepsilon) = \varepsilon^2 (U_0^{(2)})^\top \mathcal{M}^{(2)} P_0^{(2)} + o(\varepsilon^2). \quad (4.209)$$

We will derive an explicit expression for the matrix  $\mathcal{M}^{(2)}$  in Section 4.6.4.

#### 4.5.6.2 Expansion of Nonlinear Term $j_2^{(2)}(\varepsilon)$

Recall the term  $j_2^{(2)}(\varepsilon)$  from (4.202),

$$j_2^{(2)}(\varepsilon) = \int_D S_{\nabla u_0^{(2)}}^{\varepsilon,(2)}(x, \nabla \tilde{u}_\varepsilon^{(2)}) \cdot \nabla (\nabla p_0^{(2)} + \nabla \tilde{p}_\varepsilon^{(2)}).$$

According to the approximation steps taken for the variations of the direct and adjoint state, we define

$$\tilde{j}_2^{(2)}(\varepsilon) := \int_D S_{U_0^{(2)}}^{\varepsilon,(2)}(x, \nabla h_\varepsilon^{(2)}) \cdot (P_0^{(2)} + \nabla k_\varepsilon^{(2)}), \quad (4.210)$$

$$J_2^{(2)} := \int_{\mathbb{R}^2} \tilde{S}_{U_0^{(2)}}^{\varepsilon,(2)}(x, \nabla H^{(2)}) \cdot (P_0^{(2)} + \nabla K^{(2)}). \quad (4.211)$$

Note that, under Assumption 3, both  $\tilde{j}_2^{(2)}(\varepsilon)$  and  $J_2^{(2)}$  are well-defined due to growth condition (4.41).

Also here, we get the analogous estimates to Section 4.4.3.2:

**Lemma 4.64.** *Let Assumptions 1, 3 and 7 hold. Then, for  $\varepsilon \geq 0$ , it holds*

$$\tilde{j}_2^{(2)}(\varepsilon) - \varepsilon^2 J_2^{(2)} = o(\varepsilon^2). \quad (4.212)$$

*Proof.* The proof of (4.212) works in the same way as the proof of (4.105) using (4.182) and (4.198) rather than (4.67) and (4.88). The only difference is, that in Case II the operators  $S_{U_0^{(2)}}^{\varepsilon,(2)}$  and  $\tilde{S}_{U_0^{(2)}}^{\varepsilon,(2)}$  vanish outside the inclusion  $\omega_\varepsilon$  and are non-zero inside  $\omega_\varepsilon$ . The estimation of the corresponding term is achieved by (4.35) and (4.185):

$$\begin{aligned} & \int_{\omega_\varepsilon} S_{U_0^{(2)}}(\nabla h_\varepsilon^{(2)}) - S_{U_0^{(2)}}(\nabla H_\varepsilon^{(2)}) \\ & \leq c_4 \int_{\omega_\varepsilon} |\nabla h_\varepsilon^{(2)} - \nabla H_\varepsilon^{(2)}| (|\nabla H_\varepsilon^{(2)}| + |\nabla h_\varepsilon^{(2)}|) \\ & \leq c_4 \int_D |\nabla h_\varepsilon^{(2)} - \nabla H_\varepsilon^{(2)}| (|\nabla H_\varepsilon^{(2)}| + |\nabla h_\varepsilon^{(2)}|) = o(\varepsilon^2). \end{aligned}$$

□

**Lemma 4.65.** *Let Assumptions 1 and 7 hold. Then, it holds*

$$\int_D |\nabla p_0^{(2)} - P_0^{(2)}|^2 |\nabla h_\varepsilon^{(2)}|^2 = o(\varepsilon^2) \quad (4.213)$$

*Proof.* The proof is analogous to the proof of (4.106), using estimates (4.175) and (4.183). □

**Lemma 4.66.** *Let Assumptions 1, 3, 7 and 8 be satisfied. Then, for  $\varepsilon \geq 0$ , it holds*

$$j_2^{(2)}(\varepsilon) - \tilde{j}_2^{(2)}(\varepsilon) = o(\varepsilon^2). \quad (4.214)$$

*Proof.* The proof is analogous to the proof of Lemma 4.42 using the counterparts (4.186), (4.175), (4.199), (4.194) of estimates (4.71), (4.52), (4.90), (4.83), respectively. The only difference lies in the fact that the operators  $S_{U_0^{(2)}}^{\varepsilon,(2)}$  and  $\tilde{S}_{U_0^{(2)}}^{(2)}$  representing the nonlinearity vanish outside the inclusion  $\omega_\varepsilon$  and are different from zero inside  $\omega_\varepsilon$ . The estimation of the term

$$\int_{\omega_\varepsilon} S_{\nabla u_0^{(2)}}(\nabla \tilde{u}_\varepsilon^{(2)}) \cdot \nabla p_0^{(2)} - S_{U_0^{(2)}}(\nabla h_\varepsilon^{(2)}) \cdot P_0^{(2)}$$

can be done in the same way as in the proof of Lemma 4.42 using properties (4.35), (4.41) and (4.36) together with the counterparts (4.187), (4.184) and (4.213) of estimates (4.72), (4.69) and (4.106).  $\square$

Eventually, summing estimates (4.212) and (4.214) yields the following result:

**Proposition 4.67.** *Let Assumptions 1, 3, 7 and 8 be satisfied. Then*

$$j_2^{(2)}(\varepsilon) = \varepsilon^2 \left( \int_{\mathbb{R}^2} \tilde{S}_{U_0^{(2)}}^{(2)}(x, \nabla H^{(2)}) \cdot (P_0^{(2)} + \nabla K^{(2)}) \right) + o(\varepsilon^2). \quad (4.215)$$

#### 4.5.6.3 Main Result in Case II

Finally, combining (4.200) with (4.209) and (4.215), we get the main result of this section, i.e., the topological derivative for the introduction of nonlinear (ferromagnetic) material inside a region of linear material (air) according to the definition (4.2). We recall the notation used in the statement of Theorem (4.68):

- $x_0 \in \Omega^d$  denotes the point around which we perturb the material coefficient,
- $u_0^{(2)} \in H_0^1(D)$  is the unperturbed direct state, i.e., the solution to (4.164), and  $U_0^{(2)} = \nabla u_0^{(2)}(x_0)$ ,
- $p_0^{(2)} \in H_0^1(D)$  is the unperturbed adjoint state, i.e., the solution to (4.189), and  $P_0^{(2)} = \nabla p_0^{(2)}(x_0)$ ,
- $\mathcal{M}^{(2)} = \mathcal{M}^{(2)}(\omega, DT(U_0^{(2)}))$  denotes the matrix defined in (4.234) where  $\omega$  represents the shape of the inclusion and  $DT$  is the Jacobian of  $T$  defined in (2.24),
- $H^{(2)} \in \mathcal{H}(\mathbb{R}^2)$  denotes the variation of the direct state at scale 1, i.e., the solution to (4.172),
- $K^{(2)} \in \mathcal{H}(\mathbb{R}^2)$  denotes the variation of the adjoint state at scale 1, i.e., the solution to (4.192),
- $\tilde{S}^{(2)}$  is defined in (4.171),
- $\delta_J$  is according to (4.7).

**Theorem 4.68.** *Assume that*

- $\omega = B(0, 1)$  the unit disk in  $\mathbb{R}^2$

- the ferromagnetic material is such that Assumptions 1 and 3 are satisfied,
- the functional  $\mathcal{J}_\varepsilon$  satisfies Assumption 2,
- the unperturbed direct state  $u_0^{(2)}$  satisfies Assumption 7, i.e.,  $u_0^{(2)} \in C^{1,\beta}$  for some  $\beta > 0$ ,
- the unperturbed direct state  $p_0^{(2)}$  satisfies Assumption 8, i.e.,  $p_0^{(2)} \in C^{1,\tilde{\beta}}$  for some  $\tilde{\beta} > 0$ .

Then the topological derivative for introducing air inside ferromagnetic material reads

$$\begin{aligned} G^{air \rightarrow f}(x_0) = & (U_0^{(2)})^\top \mathcal{M}^{(2)} P_0^{(2)} \\ & + \int_{\mathbb{R}^2} \tilde{S}_{U_0^{(2)}}^{(2)}(x, \nabla H^{(2)}) \cdot (P_0^{(2)} + \nabla K^{(2)}) + \delta_J. \end{aligned} \quad (4.216)$$

## 4.6 Polarization Matrices

In this section we present a way to compute explicitly the matrices  $\mathcal{M}$  and  $\mathcal{M}^{(2)}$  introduced in (4.100) and (4.207), respectively. This approach is based on the notion of Pólya-Szegö polarization tensors which were introduced in the context of potential theory in connection with problems from hydrodynamics and electrostatics [185]. These matrices contain first order information about the perturbation of a field due to a small inclusion of a material with a conductivity that is different from the background conductivity. Later, the concept was extended to include also higher order information by introducing the so-called generalized polarization tensors (GPTs). Furthermore, also (generalized) anisotropic polarization tensors (APTs) were introduced in order to deal with the case where the conductivities of both the inclusion and the background medium are anisotropic. The GPTs of higher order can be used to more accurately reconstruct the shape and the conductivity of an inclusion in a background medium which is of great importance in many applications of electrical impedance imaging such as the detection of tumors in medical applications or of impurities in the material in applications from material science. For a thorough introduction to the notion of (generalized) polarization tensors, we refer the reader to [11] and the references therein.

The rest of this section is organized as follows: In Section 4.6.1 we give the definition and results for the anisotropic polarization tensors as presented in Chapter 4.12 of [11]. In order to apply the provided formulas we need to make a change of variables which is performed in Section 4.6.2. In Section 4.6.3, the term  $J_1$  introduced in (4.97) for Case I is rewritten by means of a polarization matrix. Finally, in Section 4.6.4 the same is done for the term  $J_1^{(2)}$  introduced in (4.204) for Case II.

### 4.6.1 Preliminaries

In this section, we assume that  $\omega$  is the unit ball,  $\omega = B(0, 1)$ , and we denote the conductivities in  $\omega$  and in  $\mathbb{R}^2 \setminus \bar{\omega}$  by  $\tilde{A}$  and  $A$ , respectively. Both  $\tilde{A}$  and  $A$  are constant  $2 \times 2$  positive definite, symmetric matrices and we assume that the matrix  $\tilde{A} - A$  is either positive definite or negative definite.

From the definition of APTs given in [11] it follows that for any pair of multi-indices  $p, q \in \mathbb{N}^2$ , the corresponding entry of the anisotropic polarization tensor is given by

$$\mathcal{P}_{pq} = \int_{\partial\omega} (n \cdot (\tilde{A} - A) \nabla x^q) \theta_p(A, \tilde{A}; \omega)|_- d\sigma,$$

where  $n$  denotes the outer unit normal vector to  $\partial\omega$  and, given a function  $v$  that is discontinuous across the interface  $\partial\omega$ ,  $v|_-$  denotes the limit coming from inside  $\omega$ . For a multi-index  $p \in \mathbb{N}^2$  we use the notation  $x^p := x_1^{p_1} x_2^{p_2}$ , and  $\theta_p(A, \tilde{A}; \omega)$  is the solution to the transmission problem

$$\begin{cases} \nabla \cdot (A \nabla \theta_p) = 0 & \text{in } \mathbb{R}^2 \setminus \bar{\omega}, \\ \nabla \cdot (\tilde{A} \nabla \theta_p) = 0 & \text{in } \omega, \\ \theta_p|_- - \theta_p|_+ = x^p & \text{on } \partial\omega, \\ n \cdot \tilde{A} \nabla \theta_p|_- - n \cdot A \nabla \theta_p|_+ = n \cdot A \nabla x^p & \text{on } \partial\omega, \\ \theta_p - \frac{1}{2\pi\sqrt{\det(A)}} \ln \|A^{-1/2}x\| \int_{\partial\omega} \theta_p(y) d\sigma(y) \rightarrow 0 & \text{as } |x| \rightarrow \infty. \end{cases} \quad (4.217)$$

We are only interested in the first order polarization tensors and therefore consider only multi-indices  $p, q$  with  $|p| = |q| = 1$ , i.e.,  $p = e_i$  and  $q = e_j$  for  $i, j \in \{1, 2\}$ , where  $(e_1, e_2)$  is the standard basis of  $\mathbb{R}^2$ . For  $p = e_i$  and  $q = e_j$  with  $i, j \in \{1, 2\}$ , we use the notation  $\theta_i$  instead of  $\theta_p$  and denote the corresponding entry by  $\mathcal{P}_{ij}$ . Integration by parts yields that

$$\mathcal{P}_{ij} = \int_{\omega} (\tilde{A} - A) e_j \cdot \nabla \theta_i(A, \tilde{A}; \omega), \quad (4.218)$$

and we define the first order APT

$$\mathcal{P}(A, \tilde{A}; \omega) := (\mathcal{P}_{ij})_{i,j=1,2}. \quad (4.219)$$

For the case where  $A = I$  and  $\omega$  is an ellipse which is aligned with the coordinate system, an explicit formula for the polarization matrix is available:

**Proposition 4.69** ([11], Proposition 4.31). *If  $\omega$  is an ellipse whose semi-axes are aligned with the  $x_1$ - and  $x_2$ -axes and of length  $a$  and  $b$ , respectively, then the first-order APT,  $\mathcal{P}(I, \tilde{A}; \omega)$ , takes the form*

$$\mathcal{P}(I, \tilde{A}; \omega) = |\omega| \left( I + (\tilde{A} - I) \left( \frac{1}{2} I - C \right) \right)^{-1} (\tilde{A} - I),$$

with the matrix

$$C = \frac{a-b}{2(a+b)} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (4.220)$$

In particular, if  $\omega$  is a disk, then

$$\mathcal{P}(I, \tilde{A}; \omega) = 2|\omega|(\tilde{A} + I)^{-1}(\tilde{A} - I). \quad (4.221)$$

Furthermore, we will use the following relation:

**Lemma 4.70** ([11], Lemma 4.30). *For any unitary transformation  $R$ , the following holds:*

$$\mathcal{P}(A, \tilde{A}; \omega) = R \mathcal{P}(R^\top A R, R^\top \tilde{A} R; R^{-1}\omega) R^\top.$$

### 4.6.2 Change of Variables

In order to apply Proposition 4.69 to the case of an anisotropic background conductivity  $A$ , we need to perform a change of variables such that the background conductivity  $A$  becomes the identity. We can show following relation:

**Lemma 4.71.** *Let  $\omega$  be bounded with smooth boundary,  $A, \tilde{A} \in \mathbb{R}^{2 \times 2}$  positive definite, symmetric and such that  $\tilde{A} - A$  is either positive definite or negative definite. Let the polarization matrix  $\mathcal{P}(A, \tilde{A}; \omega)$  be defined by (4.219), (4.218) and (4.217). Then it holds*

$$\mathcal{P}(A, \tilde{A}; \omega) = \det(A^{1/2}) (A^{1/2})^\top \mathcal{P}(I, A^{-1/2} \tilde{A} A^{-1/2}; A^{-1/2} \omega) A^{1/2}. \quad (4.222)$$

*Proof.* We perform a change of variables using weak formulations. To consider transmission problem (4.217) in weak form, we rewrite  $\theta_i(A, \tilde{A}; \omega)$  as

$$\theta_i(A, \tilde{A}; \omega) = K(A, \tilde{A}; \omega; e_i) + \chi_\omega(e_i \cdot x),$$

where  $K(A, \tilde{A}; \omega; e_i)$  is the solution to the transmission problem

Find  $K \in \mathcal{H}(\mathbb{R}^2)$  such that

$$\int_{\mathbb{R}^2 \setminus \omega} A \nabla K \cdot \nabla \eta + \int_\omega \tilde{A} \nabla K \cdot \nabla \eta = - \int_\omega (\tilde{A} - A) e_i \cdot \nabla \eta \quad \forall \eta \in \mathcal{H}(\mathbb{R}^2). \quad (4.223)$$

Note that  $K(A, \tilde{A}; \omega; e_i)$  is linear in the last argument  $e_i$ . Then, for  $P_0 = (p_1, p_2)^\top$ , let

$$\theta(A, \tilde{A}; \omega; P_0) := \sum_{i=1}^2 p_i \theta_i(A, \tilde{A}; \omega) = K(A, \tilde{A}; \omega; P_0) + \chi_\omega(P_0 \cdot x), \quad (4.224)$$

and note that  $\theta(A, \tilde{A}; \omega; P_0)$  is linear with respect to  $P_0$ . For  $U_0, P_0 \in \mathbb{R}^2$ , we then have

$$U_0^\top \mathcal{P}(A, \tilde{A}; \omega) P_0 = \int_\omega (\tilde{A} - A) U_0 \cdot \nabla_x \theta(A, \tilde{A}; \omega; P_0)(x).$$

We make the following coordinate transformation: Let

$$x = \varphi(y) = A^{1/2}y, \quad \text{with } \nabla \varphi = A^{1/2}, \quad \det(\nabla \varphi) = \det(A^{1/2}), \quad (4.225)$$

then for any function  $f$  it holds

$$\nabla_x f(\varphi(y)) = (\nabla \varphi)^{-\top} \nabla_y (f \circ \varphi)(y) = (A^{-1/2})^\top \nabla_y (f \circ \varphi)(y).$$

After a change of variables, transmission problem (4.223) becomes

Find  $K \in \mathcal{H}(\mathbb{R}^2)$  such that

$$\begin{aligned} & \int_{\mathbb{R}^2 \setminus \tilde{\omega}} \nabla K \cdot \nabla \eta + \int_{\tilde{\omega}} A^{-1/2} \tilde{A} A^{-1/2} \nabla K \cdot \nabla \eta \\ &= - \int_{\tilde{\omega}} (A^{-1/2} \tilde{A} A^{-1/2} - I) (A^{1/2} e_i) \cdot \nabla \eta \quad \forall \eta \in \mathcal{H}(\mathbb{R}^2), \end{aligned} \quad (4.226)$$

where  $\tilde{\omega} = A^{-1/2}\omega$ . Let  $\tilde{U}_0 = A^{1/2}U_0$  and  $\tilde{P}_0 = A^{1/2}P_0$ . For  $K(A, \tilde{A}; \omega; P_0)$  being the solution to (4.223) with  $e_i$  replaced by  $P_0$ , we see that

$$(K(A, \tilde{A}; \omega; P_0) \circ \varphi)(y) = K(I, A^{-1/2}\tilde{A}A^{-1/2}; \tilde{\omega}; \tilde{P}_0)(y),$$

and thus

$$\begin{aligned} (\theta(A, \tilde{A}; \omega; P_0) \circ \varphi)(y) &= K(A, \tilde{A}; \omega; P_0)(\varphi(y)) + \chi_\omega(\varphi(y))(P_0 \cdot \varphi(y)) \\ &= K(I, A^{-1/2}\tilde{A}A^{-1/2}; \tilde{\omega}; \tilde{P}_0)(y) + \chi_{\tilde{\omega}}(y)(\tilde{P}_0 \cdot y) \\ &= \theta(I, A^{-1/2}\tilde{A}A^{-1/2}; \tilde{\omega}; \tilde{P}_0)(y). \end{aligned} \quad (4.227)$$

Performing a change of coordinates, we obtain by (4.227),

$$\begin{aligned} U_0^\top \mathcal{P}(A, \tilde{A}; \omega) P_0 &= \int_{\omega} [(\tilde{A} - A) U_0] \cdot \nabla_x \theta(A, \tilde{A}; \omega; P_0)(x) dx \\ &= \int_{\tilde{\omega}} [(\tilde{A} - A) A^{-1/2} \tilde{U}_0] \cdot (A^{-1/2})^\top \nabla_y \theta(A, \tilde{A}; \omega; P_0)(\varphi(y)) \det(A^{1/2}) dy \\ &= \det(A^{1/2}) \int_{\tilde{\omega}} [(A^{-1/2}\tilde{A}A^{-1/2} - I) \tilde{U}_0] \cdot \nabla_y \theta(A, \tilde{A}; \omega; P_0)(\varphi(y)) dy \\ &= \det(A^{1/2}) \int_{\tilde{\omega}} [(A^{-1/2}\tilde{A}A^{-1/2} - I) \tilde{U}_0] \cdot \nabla_y \theta(I, A^{-1/2}\tilde{A}A^{-1/2}; \tilde{\omega}; \tilde{P}_0)(y) dy \\ &= \det(A^{1/2}) \sum_{i,j=1}^2 \tilde{u}_j \tilde{p}_i \int_{\tilde{\omega}} [(A^{-1/2}\tilde{A}A^{-1/2} - I) e_j] \cdot \nabla_y \theta(I, A^{-1/2}\tilde{A}A^{-1/2}; \tilde{\omega}; e_i)(y) dy \\ &= \det(A^{1/2}) \sum_{i,j=1}^2 \tilde{u}_j \tilde{p}_i \mathcal{P}_{ij}(I, A^{-1/2}\tilde{A}A^{-1/2}; \tilde{\omega}) \\ &= \det(A^{1/2}) \tilde{U}_0^\top \mathcal{P}(I, A^{-1/2}\tilde{A}A^{-1/2}; \tilde{\omega}) \tilde{P}_0 \\ &= \det(A^{1/2}) U_0^\top (A^{1/2})^\top \mathcal{P}(I, A^{-1/2}\tilde{A}A^{-1/2}; \tilde{\omega}) A^{1/2} P_0, \end{aligned}$$

where  $\tilde{U}_0 = (\tilde{u}_1, \tilde{u}_2)$  and  $\tilde{P}_0 = (\tilde{p}_1, \tilde{p}_2)$ . Thus, we have

$$\mathcal{P}(A, \tilde{A}; \omega) = \det(A^{1/2}) (A^{1/2})^\top \mathcal{P}(I, A^{-1/2}\tilde{A}A^{-1/2}; A^{-1/2}\omega) A^{1/2}.$$

□

### 4.6.3 Case I

Let us now compute the explicit formula of the matrix  $\mathcal{P}(A, \tilde{A}; \omega)$  for Case I studied in Section 4.4. In this case, the conductivities outside and inside the inclusion  $\omega$  read

$$DT(U_0) = R \begin{pmatrix} \lambda_2 & 0 \\ 0 & \lambda_1 \end{pmatrix} R^\top \quad \text{and} \quad \nu_0 I = \begin{pmatrix} \nu_0 & 0 \\ 0 & \nu_0 \end{pmatrix},$$

respectively, where  $\lambda_1 = \hat{\nu}(|U_0|)$ ,  $\lambda_2 = \hat{\nu}(|U_0|) + \hat{\nu}'(|U_0|)|U_0|$  and  $R$  denotes the rotation matrix around the angle between  $U_0$  and the x-axis such that

$$U_0 = R \begin{pmatrix} |U_0| \\ 0 \end{pmatrix},$$

see also (2.28)–(2.29).

### 4.6.3.1 Polarization Matrix

The explicit expression for the anisotropic first order polarization tensor  $\mathcal{P}(A, \tilde{A}; \omega)$  can now be calculated using Lemma 4.71 and Proposition 4.69.

**Proposition 4.72.** *Let  $\omega = B(0, 1)$  the unit disk in  $\mathbb{R}^2$ , and let Assumption 1 hold. Then, we get*

$$\mathcal{P}(\text{DT}(U_0), \nu_0 I; \omega) = |\omega| R \begin{pmatrix} \frac{(\lambda_2 + \sqrt{\lambda_1 \lambda_2})(\nu_0 - \lambda_2)}{\nu_0 + \sqrt{\lambda_1 \lambda_2}} & 0 \\ 0 & \frac{(\lambda_1 + \sqrt{\lambda_1 \lambda_2})(\nu_0 - \lambda_1)}{\nu_0 + \sqrt{\lambda_1 \lambda_2}} \end{pmatrix} R^\top.$$

*Proof.* We apply Lemma 4.71 and Proposition 4.69 with the conductivity matrices outside and inside the inclusion  $\omega$  given by  $\text{DT}(U_0)$  and  $\nu_0 I$ ,

$$A = \text{DT}(U_0) = RA_d R^\top \quad \text{and} \quad \tilde{A} = \nu_0 I = \tilde{A}_d,$$

respectively, where we introduced the diagonal matrices

$$A_d := \begin{pmatrix} \lambda_2 & 0 \\ 0 & \lambda_1 \end{pmatrix} \quad \text{and} \quad \tilde{A}_d := \begin{pmatrix} \nu_0 & 0 \\ 0 & \nu_0 \end{pmatrix}. \quad (4.228)$$

1. In a first step, we compute the matrix  $\mathcal{P}(A_d, \tilde{A}_d; \omega)$  with the matrices  $A_d, \tilde{A}_d$  defined in (4.228). Due to (2.19) which follows from Assumption 1, it holds  $\lambda_1 > 0, \lambda_2 > 0, \nu_0 - \lambda_1 > 0$  and  $\nu_0 - \lambda_2 > 0$ , and, therefore,  $A_d$  and  $\tilde{A}_d - A_d$  are positive definite. Thus, Lemma 4.71 yields that

$$\mathcal{P}(A_d, \tilde{A}_d; \omega) = \det(A_d^{1/2}) (A_d^{1/2})^\top \mathcal{P}(I, A_d^{-1/2} \tilde{A}_d A_d^{-1/2}; A_d^{-1/2} \omega) A_d^{1/2}. \quad (4.229)$$

The transformed domain  $A_d^{-1/2} \omega$  is an ellipse whose axes are aligned with the  $x_1$ - and  $x_2$ -axes. Thus, Proposition 4.69 yields

$$\begin{aligned} & \mathcal{P}(I, A_d^{-1/2} \tilde{A}_d A_d^{-1/2}; A_d^{-1/2} \omega) \\ &= |A_d^{-1/2} \omega| \left( I + (A_d^{-1/2} \tilde{A}_d A_d^{-1/2} - I) \left( \frac{1}{2} I - C \right) \right)^{-1} (A_d^{-1/2} \tilde{A}_d A_d^{-1/2} - I). \end{aligned}$$

The axes of the ellipse  $A_d^{-1/2} \omega$  have lengths  $1/\sqrt{\lambda_2}$  and  $1/\sqrt{\lambda_1}$ . Thus, we obtain for the matrix  $C$  in (4.220),

$$C = \frac{\frac{1}{\sqrt{\lambda_2}} - \frac{1}{\sqrt{\lambda_1}}}{2(\frac{1}{\sqrt{\lambda_2}} + \frac{1}{\sqrt{\lambda_1}})} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \frac{\sqrt{\lambda_1} - \sqrt{\lambda_2}}{2(\sqrt{\lambda_1} + \sqrt{\lambda_2})} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Noting that

$$A_d^{-1/2} \tilde{A}_d A_d^{-1/2} - I = \begin{pmatrix} \frac{\nu_0 - \lambda_2}{\lambda_2} & 0 \\ 0 & \frac{\nu_0 - \lambda_1}{\lambda_1} \end{pmatrix},$$

as well as

$$\left( I + (A_d^{-1/2} \tilde{A}_d A_d^{-1/2} - I) \left( \frac{1}{2} I - C \right) \right)^{-1} = \begin{pmatrix} \frac{\lambda_2 + \sqrt{\lambda_1 \lambda_2}}{\nu_0 + \sqrt{\lambda_1 \lambda_2}} & 0 \\ 0 & \frac{\lambda_1 + \sqrt{\lambda_1 \lambda_2}}{\nu_0 + \sqrt{\lambda_1 \lambda_2}} \end{pmatrix},$$

we get that

$$\mathcal{P}(I, A_d^{-1/2} \tilde{A}_d A_d^{-1/2}; A_d^{-1/2} \omega) = |A_d^{-1/2} \omega| \begin{pmatrix} \frac{\lambda_2 + \sqrt{\lambda_1 \lambda_2}}{\nu_0 + \sqrt{\lambda_1 \lambda_2}} \frac{\nu_0 - \lambda_2}{\lambda_2} & 0 \\ 0 & \frac{\lambda_1 + \sqrt{\lambda_1 \lambda_2}}{\nu_0 + \sqrt{\lambda_1 \lambda_2}} \frac{\nu_0 - \lambda_1}{\lambda_1} \end{pmatrix}. \quad (4.230)$$

Thus, it follows from (4.229) that

$$\mathcal{P}(A_d, \tilde{A}_d; \omega) = |\omega| \begin{pmatrix} \frac{(\lambda_2 + \sqrt{\lambda_1 \lambda_2})(\nu_0 - \lambda_2)}{\nu_0 + \sqrt{\lambda_1 \lambda_2}} & 0 \\ 0 & \frac{(\lambda_1 + \sqrt{\lambda_1 \lambda_2})(\nu_0 - \lambda_1)}{\nu_0 + \sqrt{\lambda_1 \lambda_2}} \end{pmatrix}. \quad (4.231)$$

2. Next, we consider the general case where  $U_0 = R(|U_0|, 0)^\top$  with a rotation matrix  $R$ . Then we have

$$A = R A_d R^\top \quad \text{and} \quad \tilde{A} = R \tilde{A}_d R^\top,$$

with the matrices  $A_d, \tilde{A}_d$  defined in (4.228). Note that

$$A^{-1/2} = R A_d^{-1/2} R^\top,$$

and

$$A^{-1/2} \tilde{A} A^{-1/2} = R A_d^{-1/2} \tilde{A}_d A_d^{-1/2} R^\top = R \begin{pmatrix} \frac{\nu_0}{\lambda_2} & 0 \\ 0 & \frac{\nu_0}{\lambda_1} \end{pmatrix} R^\top.$$

Thus, we have

$$\mathcal{P}(I, A^{-1/2} \tilde{A} A^{-1/2}; A^{-1/2} \omega) = \mathcal{P}(I, R \begin{pmatrix} \frac{\nu_0}{\lambda_2} & 0 \\ 0 & \frac{\nu_0}{\lambda_1} \end{pmatrix} R^\top; R A_d^{-1/2} R^\top \omega),$$

and Lemma 4.70 yields

$$\mathcal{P}(I, R \begin{pmatrix} \frac{\nu_0}{\lambda_2} & 0 \\ 0 & \frac{\nu_0}{\lambda_1} \end{pmatrix} R^\top; R A_d^{-1/2} R^\top \omega) = R \mathcal{P}(I, \begin{pmatrix} \frac{\nu_0}{\lambda_2} & 0 \\ 0 & \frac{\nu_0}{\lambda_1} \end{pmatrix}; A_d^{-1/2} R^\top \omega) R^\top,$$

where  $\mathcal{P}(I, \begin{pmatrix} \frac{\nu_0}{\lambda_2} & 0 \\ 0 & \frac{\nu_0}{\lambda_1} \end{pmatrix}; A_d^{-1/2} R^\top \omega)$  is as in (4.230). Since  $\omega$  is a ball, the rotation matrix  $R^\top$  leaves it unchanged, i.e.,  $R^\top \omega = \omega$ . Thus, we have by (4.222)

$$\begin{aligned} \mathcal{P}(A, \tilde{A}; \omega) &= \det(A^{1/2}) |A^{-1/2} \omega| A^{1/2} R \mathcal{P}(I, \begin{pmatrix} \frac{\nu_0}{\lambda_2} & 0 \\ 0 & \frac{\nu_0}{\lambda_1} \end{pmatrix}; A_d^{-1/2} R^\top \omega) R^\top A^{1/2} \\ &= \det(A^{1/2}) |A^{-1/2} \omega| R A_d^{1/2} \mathcal{P}(I, \begin{pmatrix} \frac{\nu_0}{\lambda_2} & 0 \\ 0 & \frac{\nu_0}{\lambda_1} \end{pmatrix}; A_d^{-1/2} \omega) A_d^{1/2} R^\top \\ &= R \mathcal{P}(A_d, \tilde{A}_d; \omega) R^\top, \end{aligned} \quad (4.232)$$

with  $\mathcal{P}(A_d, \tilde{A}_d; \omega)$  as given in (4.231), which completes the proof.

□

**Remark 4.73.** Similar results can be obtained if  $\omega$  is an ellipse. If the semi-axes of  $\omega$  are aligned with the coordinate system  $(U_0/|U_0|, U_0^\perp/|U_0^\perp|)$ , then  $A_d^{-1/2}R^\top\omega$  is an ellipse whose axes are aligned with the axes of the coordinate system  $(e_1, e_2)$  and we can apply Proposition 4.69 to obtain an explicit expression for  $\mathcal{P}(I, \begin{pmatrix} \frac{\nu_0}{\lambda_2} & 0 \\ 0 & \frac{\nu_0}{\lambda_1} \end{pmatrix}; A_d^{-1/2}R^\top\omega)$  in the last step of the calculation above. Otherwise it is possible to rotate the coordinate system by another application of Lemma 4.70 and then apply Proposition 4.69.

#### 4.6.3.2 Matrix $\mathcal{M}$ in the Topological Derivative

Now recall the term  $J_1 = J_1(U_0, P_0)$  arising from the topological asymptotic expansion of a functional  $\mathcal{J}$ , see Section 4.4.3.1. Let  $A = DT(U_0)$  and  $\tilde{A} = \nu_0 I$ . The definition of  $J_1$  (4.97) together with (4.224) and (4.218) yield

$$\begin{aligned} J_1(U_0, P_0) &= (\nu_0 - \lambda_1)U_0 \cdot \int_{\omega} P_0 + \nabla K \\ &= (\nu_0 - \lambda_1)(\tilde{A} - A)^{-1}U_0 \cdot \int_{\omega} (\tilde{A} - A)(\nabla\theta(A, \tilde{A}; \omega; P_0)) \\ &= (\nu_0 - \lambda_1)U_0^\top(\tilde{A} - A)^{-\top}\mathcal{P}(A, \tilde{A}; \omega)P_0, \end{aligned}$$

where  $\mathcal{P}(A, \tilde{A}; \omega)$  is given in (4.232). Thus, we finally get (4.101), i.e.,

$$J_1(U_0, P_0) = U_0^\top \mathcal{M} P_0,$$

with the matrix

$$\begin{aligned} \mathcal{M} &= (\nu_0 - \lambda_1)(\tilde{A} - A)^{-\top}\mathcal{P}(A, \tilde{A}; \omega) \\ &= (\nu_0 - \lambda_1)|\omega| R \begin{pmatrix} \frac{\lambda_2 + \sqrt{\lambda_1\lambda_2}}{\nu_0 + \sqrt{\lambda_1\lambda_2}} & 0 \\ 0 & \frac{\lambda_1 + \sqrt{\lambda_1\lambda_2}}{\nu_0 + \sqrt{\lambda_1\lambda_2}} \end{pmatrix} R^\top. \end{aligned} \quad (4.233)$$

Note that, in the linear case where  $\lambda_2 = \lambda_1 > 0$ , it holds  $(\nu_0 - \lambda_1)(\tilde{A} - A)^{-1} = I$ , and we obtain

$$\mathcal{M} = \mathcal{P}(A, \tilde{A}; \omega) = 2|\omega| \lambda_1 \frac{\nu_0 - \lambda_1}{\nu_0 + \lambda_1} I,$$

which coincides with the well-known formula derived in, e.g., [14]. Thus, here, unlike in the nonlinear case, the matrix appearing in the topological derivative is actually the polarization matrix according to [11].

**Remark 4.74.** Finally, we remark that the explicit form of the matrix  $\mathcal{M}$  satisfying relation (4.101) can also be obtained directly without exploiting Proposition 4.69 in the following way: Starting out from the transmission problem defining  $K$  after the coordinate transformation (4.225), i.e., (4.226), we can compute the solution  $K$  explicitly by a special ansatz similarly to [11, Proposition 4.6]. Noting that, by the coordinate transformation (4.225), the circular inclusion  $\omega$  becomes an ellipse  $\tilde{\omega}$ , we make the ansatz in elliptic coordinates. For that

purpose, let  $\tilde{R} \in \mathbb{R}$  and  $(r, \varphi) \in \mathbb{R}_0^+ \times [0, 2\pi]$  be such that  $x_1(r, \varphi) = \tilde{R} \cos \varphi \cosh r$  and  $x_2(r, \varphi) = \tilde{R} \sin \varphi \sinh r$ . For  $i = 1, 2$  we make the ansatz

$$K_{e_1}(r, \varphi) = \begin{cases} a_1 \tilde{R} \cos \varphi \cosh r & \text{in } \tilde{\omega}, \\ b_1 e^{-r} \cos \varphi & \text{in } \mathbb{R}^2 \setminus \bar{\omega}, \end{cases} \quad K_{e_2}(r, \varphi) = \begin{cases} a_2 \tilde{R} \sin \varphi \sinh r & \text{in } \tilde{\omega}, \\ b_2 e^{-r} \sin \varphi & \text{in } \mathbb{R}^2 \setminus \bar{\omega}, \end{cases}$$

for problem (4.226) involving the unit vector  $e_i$ , and choose the constants  $a_i, b_i$  such that  $K_{e_i}$  is continuous and satisfies the correct interface jump condition on  $\partial\tilde{\omega}$  which are incorporated in the variational formulation (4.226) for  $i = 1, 2$ . For a given  $P_0 \in \mathbb{R}^2$ , the solution  $K$  to (4.80) is then obtained as a linear combination of  $K_{e_1}$  and  $K_{e_2}$ . Plugging in this explicit solution  $K$  into (4.101), the matrix  $\mathcal{M}$  can be identified.

#### 4.6.4 Case II

We perform the same steps for Case II where the conductivities outside and inside the inclusion  $\omega$  read

$$\nu_0 I = \begin{pmatrix} \nu_0 & 0 \\ 0 & \nu_0 \end{pmatrix} \quad \text{and} \quad DT(U_0^{(2)}) = R \begin{pmatrix} \lambda_2 & 0 \\ 0 & \lambda_1 \end{pmatrix} R^\top,$$

respectively, where, again,  $\lambda_1 = \hat{\nu}(|U_0^{(2)}|)$ ,  $\lambda_2 = \hat{\nu}(|U_0^{(2)}|) + \hat{\nu}'(|U_0^{(2)}|)|U_0^{(2)}|$ , and  $R$  denotes the rotation matrix around the angle between  $U_0^{(2)}$  and the x-axis such that

$$U_0^{(2)} = R \begin{pmatrix} |U_0^{(2)}| \\ 0 \end{pmatrix},$$

see (2.28)–(2.29).

##### 4.6.4.1 Polarization Matrix

**Proposition 4.75.** *Let  $\omega = B(0, 1)$  the unit disk in  $\mathbb{R}^2$ , and let Assumption 1 hold. Then, we have*

$$\mathcal{P}(\nu_0 I, DT(U_0^{(2)}); \omega) = 2|\omega|\nu_0 R \begin{pmatrix} \frac{\lambda_2 - \nu_0}{\lambda_2 + \nu_0} & 0 \\ 0 & \frac{\lambda_1 - \nu_0}{\lambda_1 + \nu_0} \end{pmatrix} R^\top.$$

*Proof.* Let  $A = \nu_0 I$  and  $\tilde{A} = DT(U_0^{(2)})$ . Due to (2.19) which follows from Assumption 1, we get that  $\tilde{A}$  is positive definite and  $\tilde{A} - A$  is negative definite. Applying Lemma 4.71 yields

$$\mathcal{P}(A, \tilde{A}; \omega) = \det(A^{1/2})(A^{1/2})^\top \mathcal{P}(I, A^{-1/2}\tilde{A}A^{-1/2}; A^{-1/2}\omega) A^{1/2}.$$

Note that, since  $A$  is a scaled identity matrix,  $A^{-1/2}\omega = (1/\sqrt{\nu_0})\omega$  remains a disk and we get from (4.221)

$$\begin{aligned} \mathcal{P}(I, A^{-1/2}\tilde{A}A^{-1/2}; A^{-1/2}\omega) &= 2|A^{-1/2}\omega|(A^{-1/2}\tilde{A}A^{-1/2} + I)^{-1}(A^{-1/2}\tilde{A}A^{-1/2} - I) \\ &= 2|A^{-1/2}\omega|R \begin{pmatrix} \frac{\nu_0}{\lambda_2 + \nu_0} & 0 \\ 0 & \frac{\nu_0}{\lambda_1 + \nu_0} \end{pmatrix} \begin{pmatrix} \frac{\lambda_2 - \nu_0}{\nu_0} & 0 \\ 0 & \frac{\lambda_1 - \nu_0}{\nu_0} \end{pmatrix} R^\top \\ &= 2|A^{-1/2}\omega|R \begin{pmatrix} \frac{\lambda_2 - \nu_0}{\lambda_2 + \nu_0} & 0 \\ 0 & \frac{\lambda_1 - \nu_0}{\lambda_1 + \nu_0} \end{pmatrix} R^\top. \end{aligned}$$

Therefore, we get

$$\begin{aligned}\mathcal{P}(A, \tilde{A}; \omega) &= \det(A^{1/2}) 2|A^{-1/2}\omega| R(A^{1/2})^\top \begin{pmatrix} \frac{\lambda_2 - \nu_0}{\lambda_2 + \nu_0} & 0 \\ 0 & \frac{\lambda_1 - \nu_0}{\lambda_1 + \nu_0} \end{pmatrix} A^{1/2} R^\top \\ &= 2|\omega|\nu_0 R \begin{pmatrix} \frac{\lambda_2 - \nu_0}{\lambda_2 + \nu_0} & 0 \\ 0 & \frac{\lambda_1 - \nu_0}{\lambda_1 + \nu_0} \end{pmatrix} R^\top.\end{aligned}$$

□

#### 4.6.4.2 Matrix $\mathcal{M}^{(2)}$ in the Topological Derivative

Let  $A = \nu_0 I$  and  $\tilde{A} = DT(U_0^{(2)})$ . Just like in Case I in Section 4.6.3.2, we have to pre-multiply the derived polarization matrix  $\mathcal{P}(A, \tilde{A}; \omega)$  by the scalar contrast of the direct problem,  $\lambda_1 - \nu_0$ , and the transpose of the inverse of the matrix-valued contrast of the adjoint problem,  $(\tilde{A} - A)^{-\top}$ . This way, we get the explicit form of the matrix  $\mathcal{M}^{(2)}$  in (4.208), i.e., we get

$$J_1^{(2)} = (U_0^{(2)})^\top \mathcal{M}^{(2)} P_0^{(2)},$$

with the matrix

$$\mathcal{M}^{(2)} = (\lambda_1 - \nu_0)(\tilde{A} - A)^{-\top} \mathcal{P}(A, \tilde{A}; \omega) = 2|\omega|\nu_0 R \begin{pmatrix} \frac{\lambda_1 - \nu_0}{\lambda_2 + \nu_0} & 0 \\ 0 & \frac{\lambda_1 - \nu_0}{\lambda_1 + \nu_0} \end{pmatrix} R^\top. \quad (4.234)$$

Again, in the linear case, where  $\lambda_2 = \lambda_1 > 0$ , this gives the well-known formula

$$\mathcal{M}^{(2)} = 2\pi\nu_0 \frac{\lambda_1 - \nu_0}{\lambda_1 + \nu_0} I,$$

see e.g. [14].

**Remark 4.76.** *Similarly to Remark 4.74, also here we can compute the solution to transmission problem (4.172) explicitly by making a special ansatz. Unlike in Case I, here the conductivity matrix outside the inclusion is a scaled identity matrix,  $A = \nu_0 I$ , and the circular inclusion  $\omega$  does not become an ellipse. Therefore, the solution can be obtained by the following ansatz in polar coordinates:*

$$K_{e_1}^{(2)}(r, \varphi) = \begin{cases} a_1 r \cos \varphi & \text{in } \omega, \\ b_1 e^{-r} \cos \varphi & \text{in } \mathbb{R}^2 \setminus \overline{\omega}, \end{cases} \quad K_{e_2}^{(2)}(r, \varphi) = \begin{cases} a_2 r \sin \varphi & \text{in } \omega, \\ b_2 e^{-r} \sin \varphi & \text{in } \mathbb{R}^2 \setminus \overline{\omega}. \end{cases}$$

Again, the constants  $a_i, b_i$  must be chosen such that the interface conditions are satisfied, and the matrix  $\mathcal{M}^{(2)}$  can be identified from (4.208) after plugging in the explicit form of  $K$ .

## 4.7 Computational Aspects

In order to make use of formulas (4.108) and (4.216) in applications of shape and topology optimization, an efficient method to evaluate these formulas for every point  $x_0$  in the design region of the computational domain is of utter importance. For the rest of this section, we restrict our presentation to Case I, noting that analogous results hold true for Case II.

In particular, the evaluation of the second term  $J_2$  in (4.108) seems to be computationally very costly, as it involves the solutions  $H$  and  $K$  to the transmission problems (4.49) and (4.80), respectively. Both of these problems are defined on the unbounded domain  $\mathbb{R}^2$  and depend on  $U_0 = \nabla u_0(x_0)$ , i.e., the gradient of the unperturbed direct state  $u_0$  evaluated at the point of interest  $x_0$ . In addition, problem (4.80) also depends on  $P_0 = \nabla p_0(x_0)$ , i.e., the gradient of the unperturbed adjoint state  $p_0$  at point  $x_0$ . Recall the second term  $J_2$  defined in (4.104),

$$J_2 = J_2(x_0) = J_2(U_0, P_0) = \int_{\mathbb{R}^2} \tilde{S}_{U_0}(x, \nabla H(U_0)) \cdot (P_0 + \nabla K(U_0, P_0)), \quad (4.235)$$

where  $\tilde{S}$  is defined in (4.48),  $H = H(U_0)$  is the solution to problem (4.49),

Find  $H \in \mathcal{H}(\mathbb{R}^2)$  such that

$$\int_{\mathbb{R}^2} (\tilde{T}(x, U_0 + \nabla H) - \tilde{T}(x, U_0)) \cdot \nabla \eta = - \int_{\omega} (\nu_0 - \hat{\nu}(|U_0|)) U_0 \cdot \nabla \eta \quad \forall \eta \in \mathcal{H}(\mathbb{R}^2),$$

and  $K = K(U_0, P_0)$  the solution to (4.80),

Find  $K \in \mathcal{H}(\mathbb{R}^2)$  such that

$$\int_{\mathbb{R}^2} D\tilde{T}(x, U_0) \nabla K \cdot \nabla \eta = - \int_{\omega} (\nu_0 I - DT(U_0)) P_0 \cdot \nabla \eta \quad \forall \eta \in \mathcal{H}(\mathbb{R}^2).$$

At the first glance, this means that, for each point  $x_0 \in \Omega^d$  where one wants to evaluate the term  $J_2$ , one has to solve problems (4.49) and (4.80) in order to get the value for  $J_2$ . Topology optimization algorithms which are based on topological sensitivities usually require the values of these sensitivities at all points of the design domain  $\Omega^d$  simultaneously, see Algorithm 1 on page 36 and, later, Algorithm 2 on page 133, which would, of course, result in extremely inefficient optimization algorithms.

Luckily, the enormous computational effort of having to solve the nonlinear problem (4.49) for  $H$  and the linear problem (4.80) for a point  $x_0$  in order to evaluate  $J_2$  at  $x_0$  can be reduced with the help of the following observations.

**Lemma 4.77.** *Let  $U_0, P_0 \in \mathbb{R}^2$ ,  $R \in \mathbb{R}^{2 \times 2}$  an orthogonal matrix. Let  $H(U_0) \in \mathcal{H}(\mathbb{R}^2)$  be the solution to (4.49) and  $K(U_0, P_0)$  the solution to (4.80), and note that both these problems depend on  $U_0, P_0$ . Let further  $J_2 = J_2(U_0, P_0)$  be defined by (4.235). Then the following properties hold:*

1.  *$J_2$  is linear in the second argument, i.e., for all  $a, b \in \mathbb{R}$  and  $U_0, P_1, P_2 \in \mathbb{R}^2$ ,*

$$J_2(U_0, aP_1 + bP_2) = aJ_2(U_0, P_1) + bJ_2(U_0, P_2). \quad (4.236)$$

2. *Let  $y = Rx$ . For the solution  $H = H(U_0)$  to (4.49), we have*

$$R^\top \nabla_y H(U_0)(y) = \nabla H(R^\top U_0)(x). \quad (4.237)$$

3. *Let  $y = Rx$ . For the solution  $K = K(U_0, P_0)$  to (4.80), we have*

$$R^\top \nabla_y K(U_0, P_0)(y) = \nabla K(R^\top U_0, R^\top P_0)(x). \quad (4.238)$$

4. It holds

$$J_2(R^\top U_0, R^\top P_0) = J_2(U_0, P_0). \quad (4.239)$$

*Proof.* 1. It can easily be seen from (4.80) that  $K$  depends linearly on  $P_0$ .

2. For  $x \in \mathbb{R}^2$ , let  $H(U_0)(Rx)$  the solution to problem (4.49) after a coordinate transformation. Define  $\tilde{H}(U_0)(x) := H(U_0)(y(x))$ ,  $x \in \mathbb{R}^2$ , with  $y(x) = Rx$ . Then we have

$$\nabla_y H(U_0)(y) = R^{-\top} \nabla_x \tilde{H}(U_0)(x) = R \nabla_x \tilde{H}(U_0)(x),$$

since  $R$  is orthogonal. Similarly, for a test function  $\eta$  and  $x \in \mathbb{R}^2$ , we define  $\tilde{\eta}(x) := \eta(y(x))$  and get

$$\nabla_y \eta(y) = R \nabla_x \tilde{\eta}(x).$$

The left hand side of transmission problem (4.49) then becomes

$$\begin{aligned} & \int_{\mathbb{R}^2} \left( \tilde{T}(y, U_0 + \nabla_y H(y)) - \tilde{T}(y, U_0) \right) \cdot \nabla_y \eta(y) dy \\ &= \int_{\mathbb{R}^2 \setminus \omega} (\hat{\nu}(|U_0 + \nabla_y H(y)|)(U_0 + \nabla_y H(y)) - \hat{\nu}(|U_0|)U_0) \cdot \nabla_y \eta(y) dy \\ & \quad + \int_{\omega} \nu_0 \nabla_y H(y) \cdot \nabla_y \eta(y) dy \\ &= \int_{\mathbb{R}^2 \setminus \omega} \left( \hat{\nu}(|U_0 + R \nabla_x \tilde{H}(x)|)(U_0 + R \nabla_x \tilde{H}(x)) - \hat{\nu}(|U_0|)U_0 \right) \cdot (R \nabla_x \tilde{\eta}(x)) dx \\ & \quad + \int_{\omega} \nu_0 (R \nabla_x \tilde{H}(x)) \cdot (R \nabla_x \tilde{\eta}(x)) dx \\ &= \int_{\mathbb{R}^2 \setminus \omega} \left( \hat{\nu}(|R^\top U_0 + \nabla_x \tilde{H}(x)|)(R^\top U_0 + \nabla_x \tilde{H}(x)) - \hat{\nu}(|U_0|)R^\top U_0 \right) \cdot \nabla_x \tilde{\eta}(x) dx \\ & \quad + \int_{\omega} \nu_0 \nabla_x \tilde{H}(x) \cdot \nabla_x \tilde{\eta}(x) dx \end{aligned} \quad (4.240)$$

where we used that  $R\omega = \omega$ ,  $R^\top R = I$  with  $I$  the identity matrix in  $\mathbb{R}^2$ , that  $|\det R| = 1$  and that

$$|U_0 + R \nabla_x \tilde{H}(x)| = |R(R^\top U_0 + \nabla_x \tilde{H}(x))| = |R^\top U_0 + \nabla_x \tilde{H}(x)| \quad \forall x \in \mathbb{R}^2,$$

since  $R$  is orthogonal. Similarly, we get for the right hand side of (4.49),

$$\begin{aligned} - \int_{\omega} (\nu_0 - \hat{\nu}(|U_0|)) U_0 \cdot \nabla_y \eta(y) dy &= - \int_{\omega} (\nu_0 - \hat{\nu}(|U_0|)) U_0 \cdot (R \nabla_x \tilde{\eta}(x)) dx \\ &= - \int_{\omega} \left( \nu_0 - \hat{\nu}(|R^\top U_0|) \right) R^\top U_0 \cdot \nabla_x \tilde{\eta}(x) dx. \end{aligned} \quad (4.241)$$

On the other hand, considering the transmission problem obtained by replacing  $U_0$  in (4.49) by  $R^\top U_0$  and denote its solution by  $H(R^\top U_0)$ , we note that the left and right hand side are equal to (4.240) and (4.241), respectively. Thus, it follows from the uniqueness of a solution in  $\mathcal{H}(\mathbb{R}^2)$  to (4.49) (where  $U_0$  is replaced by  $R^\top U_0$ ) stated in

Proposition 4.17, that the solution of the original problem after a coordinate transformation  $y = Rx$  equals the solution to the problem in the original coordinates with the vector  $U_0$  rotated by application of  $R^\top$  in  $\mathcal{H}(\mathbb{R}^2)$ , i.e.,

$$H(U_0)(Rx) = H(R^\top U_0)(x) + c \quad (4.242)$$

for some constant  $c \in \mathbb{R}$ , for all  $x \in \mathbb{R}^2$ . From (4.242), it follows that

$$\nabla_x \left( H(R^\top U_0)(x) \right) = \nabla_x (H(U_0)(Rx)) = R^\top \nabla_y H(U_0)(y),$$

which finishes the proof of statement 2.

3. Statement 3 is shown in an analogous way by comparing the left and right hand side of transmission problem (4.80) first after the coordinate transformation  $y = Rx$  and second after replacing  $U_0$  and  $P_0$  by  $R^\top U_0$  and  $R^\top P_0$ , respectively. Note that

$$\begin{aligned} D\tilde{T}(x, R^\top U_0) &= R^\top D\tilde{T}(x, U_0)R \quad \text{and} \\ DT(R^\top U_0) &= R^\top DT(U_0)R. \end{aligned}$$

4. Note that, for  $V, W \in \mathbb{R}^2$ ,  $T$  defined in (2.24) and  $S$  defined in (4.24), we have

$$\begin{aligned} T(R^\top W) &= R^\top T(W), \\ DT(R^\top W) &= R^\top DT(W)R, \\ S_{R^\top V}(R^\top W) &= R^\top S_V(W). \end{aligned}$$

Then, it follows by (4.237) and (4.238) and the fact that  $|\det R| = 1$  that

$$\begin{aligned} J_2(R^\top U_0, R^\top P_0) &= \int_{\mathbb{R}^2 \setminus \omega} S_{R^\top U_0}(\nabla_x H(R^\top U_0)(x)) \cdot \left( R^\top P_0 + \nabla_x K(R^\top U_0, R^\top P_0)(x) \right) dx \\ &= \int_{\mathbb{R}^2 \setminus \omega} S_{R^\top U_0}(R^\top \nabla_y H(U_0)(y)) \cdot \left( R^\top P_0 + R^\top \nabla_y K(U_0, P_0)(y) \right) dy \\ &= \int_{\mathbb{R}^2 \setminus \omega} R^\top S_{U_0}(\nabla_y H(U_0)(y)) \cdot R^\top (P_0 + \nabla_y K(U_0, P_0)(y)) dy \\ &= \int_{\mathbb{R}^2 \setminus \omega} S_{U_0}(\nabla_y H(U_0)(y)) \cdot (P_0 + \nabla_y K(U_0, P_0)(y)) dy \\ &= J_2(U_0, P_0), \end{aligned}$$

which completes the proof of part 4 of Lemma 4.77. □

By means of properties 1 and 4 of Lemma 4.77, it is possible to efficiently evaluate  $J_2$  by first precomputing values in an offline stage and then looking them up and interpolating between them during the optimization procedure. Let  $t := |U_0|$ ,  $s := |P_0|$ ,  $e_i$  the unit vector in  $x_i$ -direction for  $i = 1, 2$ , and  $\theta$  and  $\varphi$  the angles between  $U_0$  and  $e_1$  and between  $P_0$  and  $e_1$ , respectively, i.e.,

$$U_0 = t R_\theta e_1 \quad \text{and} \quad P_0 = s R_\varphi e_1,$$

where  $R_\alpha$  denotes the counter-clockwise rotation matrix around an angle  $\alpha$ , i.e.,

$$R_\alpha = \begin{pmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{pmatrix}.$$

Then, by (4.239) and (4.236), we have

$$\begin{aligned} J_2(U_0, P_0) &= J_2(t R_\theta e_1, s R_\varphi e_1) \\ &= J_2(t e_1, s R_{\varphi-\theta} e_1) \\ &= J_2(t e_1, s \cos(\varphi - \theta) e_1 + s \sin(\varphi - \theta) e_2) \\ &= s \cos(\varphi - \theta) J_2(t e_1, e_1) + s \sin(\varphi - \theta) J_2(t e_1, e_2). \end{aligned} \quad (4.243)$$

Thus, by precomputing the values of  $J_2(t e_1, e_i)$  for  $i = 1, 2$  for a range of typical values of  $t = |U_0| = |\nabla u(x_0)| = |\mathbf{B}(x_0)|$  where  $\mathbf{B}$  denotes the magnetic flux density, the values of the term  $J_2$  can be efficiently approximated for any  $U_0$  and  $P_0$  by interpolation, without the need to solve a nonlinear problem for every evaluation.

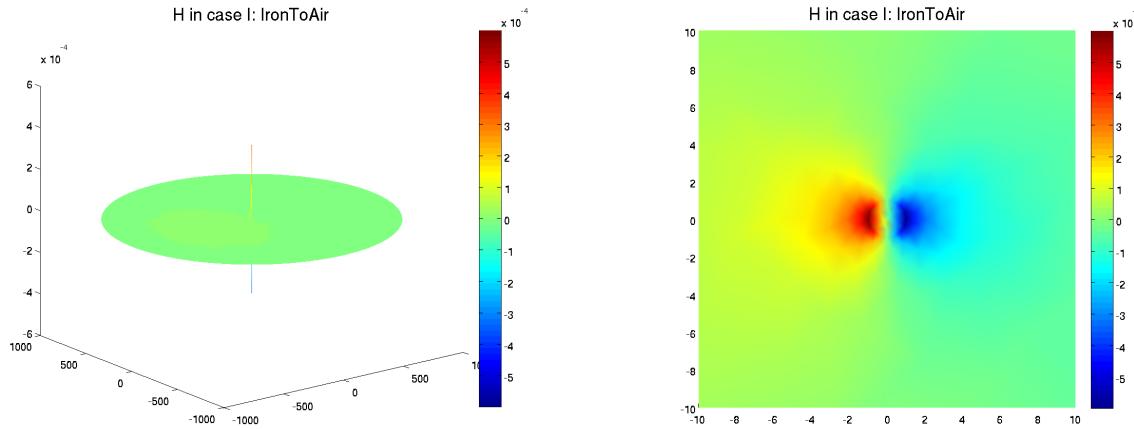
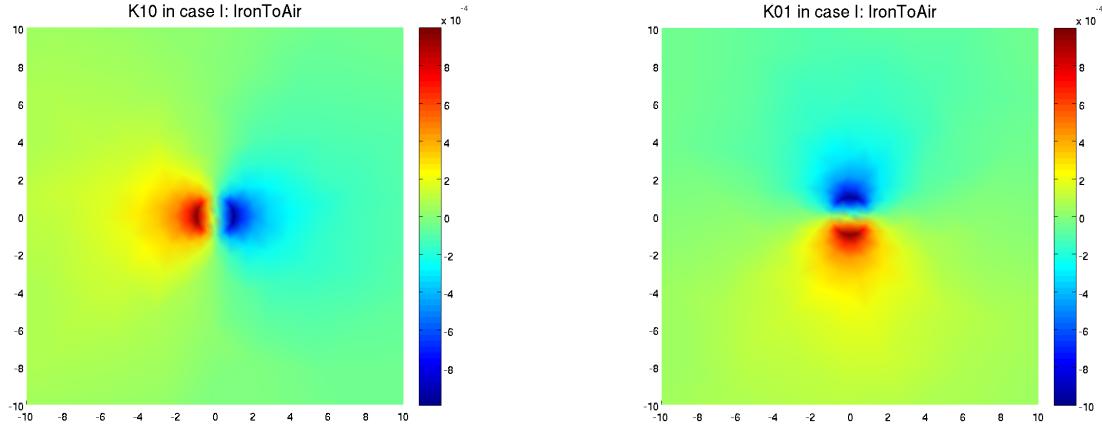
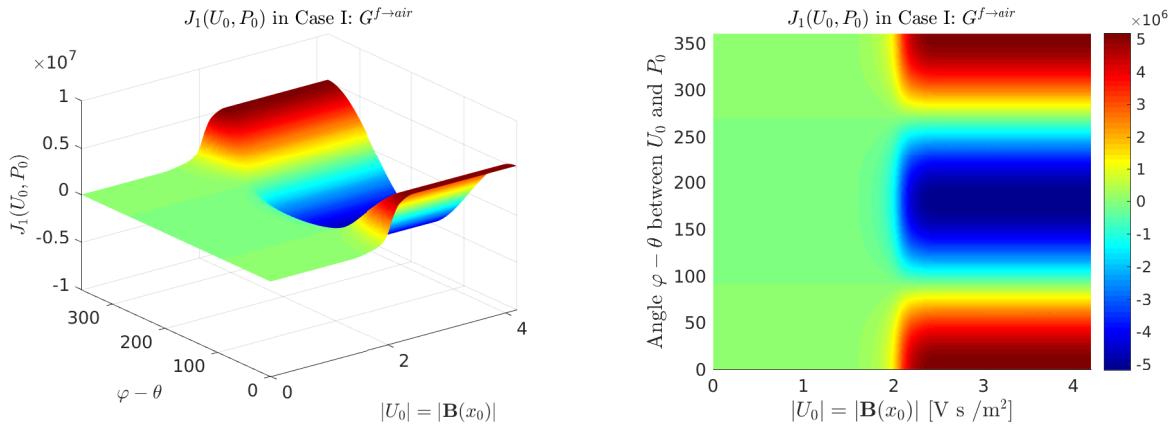
#### 4.7.1 Numerical Experiments

For given  $U_0, P_0 \in \mathbb{R}^2$ , we compute approximate solutions  $H_h$  and  $K_h$  to the elements  $\tilde{H}(U_0) \in H(U_0)$  and  $\tilde{K}(U_0, P_0) \in K(U_0, P_0)$  which satisfy the asymptotic behaviors (4.63) and (4.85), respectively, by the finite element method. We approximate problems (4.49) and (4.80), which are defined on the plane  $\mathbb{R}^2$ , by restricting the domains of integration to a circular domain of radius 1000 which is centered at the origin. The inclusion  $\omega$  is the unit disk,  $\omega = B(0, 1)$ . We use homogeneous Dirichlet boundary conditions for both problems. This approximation is justified by the asymptotic behavior of the solutions  $H$  and  $K$  derived in (4.63) and (4.85), respectively. We use piecewise linear finite elements on a triangular mesh. Figure 4.4 shows the obtained solutions for  $H_h \approx H(U_0)$ , and Figure 4.5 shows  $K_{h,10} \approx K(U_0, (1, 0)^\top)$  and  $K_{h,01} \approx K(U_0, (0, 1)^\top)$  with  $U_0 = (0.1, 0)^\top$ . Note that, for  $P_0 = (p_1, p_2)^\top$ , an approximation to  $K(U_0 P_0)$  is given by the linear combination  $p_1 K_{h,10} + p_2 K_{h,01}$ . We remark that there is no visual difference for  $U_0 = (|U_0|, 0)^\top$  with higher values for  $|U_0|$ .

Next, we compute and compare the terms  $J_1$  and  $J_2$  appearing in the topological derivative (4.108). The quantities  $J_1$  and  $J_2$  depend on  $U_0 = t R_\theta e_1$  and  $P_0 = s R_\varphi e_1$  and thus have, in two space dimensions, four degrees of freedom. Both  $J_1$  and  $J_2$  are linear in the second argument  $P_0$ , thus we can neglect  $s = |P_0|$ , as a scaling of  $P_0$  will result in the same scaling of  $J_1$  and  $J_2$ . Furthermore, in terms of the angles  $\theta, \varphi$ , both  $J_1$  and  $J_2$  only depend on the difference  $\varphi - \theta$ . For  $J_2$  this follows from (4.243) and for  $J_1$ , this can be seen from (4.101) and (4.233). Thus, we can visualize  $J_1$  and  $J_2$  in dependence of two degrees of freedom,  $|U_0|$  and  $\varphi - \theta$ . Figures 4.6–4.9 show  $J_1$  and  $J_2$  in Case I and Case II in dependence on these two degrees of freedom. Note that the difference in the order of magnitude of  $J_1$  compared to  $J_2$  suggests that, in the case of electrical machines, the second term  $J_2$  accounting for the nonlinearity of the function  $\hat{\nu}$  is of little importance compared the first term  $J_1$ .

## 4.8 Application to Model Problem

In this section, we employ the topological derivative derived in (4.108) and (4.216) for the model design optimization problem introduced in (2.17). We apply the level set algorithm

Figure 4.4:  $H_h$  for  $U_0 = (0.1, 0)^\top$  in Case I.Figure 4.5: Left: Zoom of  $K_{h,10}$  for  $U_0 = (0.1, 0)^\top$  in Case I. Right: Zoom of  $K_{h,01}$  for  $U_0 = (0.1, 0)^\top$  in Case I.Figure 4.6:  $J_1(U_0, P_0)$  in Case I.

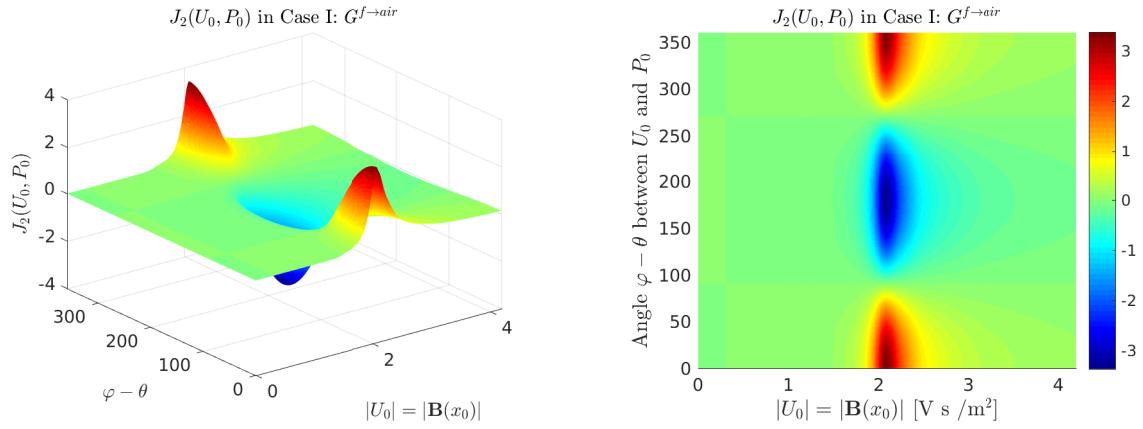


Figure 4.7:  $J_2(U_0, P_0)$  in Case I. Note the difference in the order of magnitude compared to  $J_1(U_0, P_0)$  in Figure 4.6.

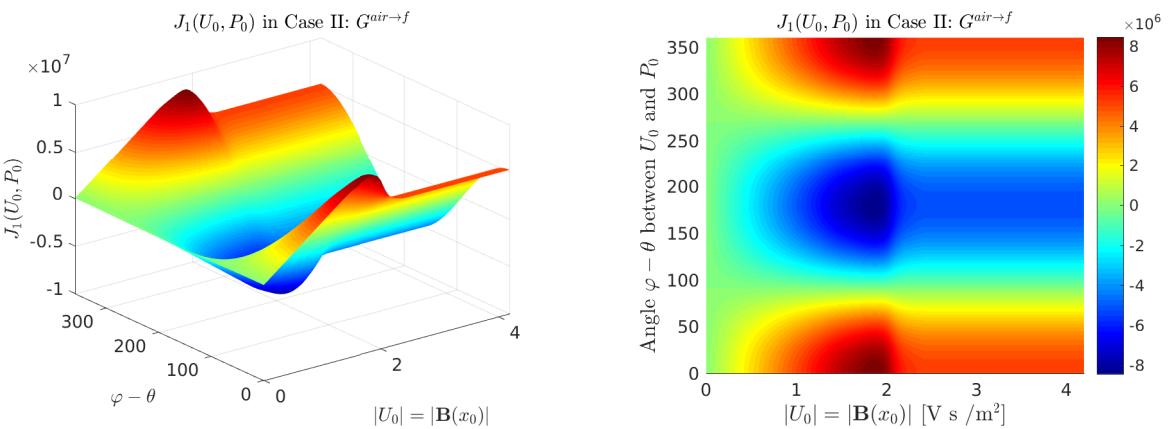


Figure 4.8:  $J_1^{(2)}(U_0, P_0)$  in Case II.

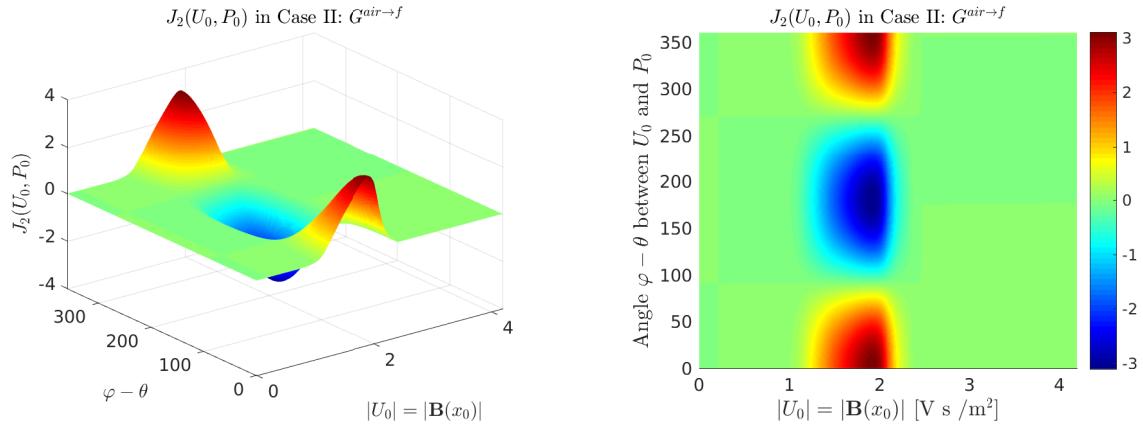


Figure 4.9:  $J_2^{(2)}(U_0, P_0)$  in Case II. Note the difference in the order of magnitude compared to  $J_1^{(2)}(U_0, P_0)$  in Figure 4.8.

introduced in [19], which is based on the topological derivative. This is in contrast to the level set method for shape optimization introduced in [165, 166] where the evolution of the interface is usually guided by shape sensitivity information and generally lacks a nucleation mechanism [60]. We give a short overview of the algorithm and refer to the references [16, 19] for a more detailed description as well as a convergence proof.

Recall the notation of Section 2.3. In particular, recall that the variable set  $\Omega_f$  was defined as that subset of  $D$  which is currently occupied with ferromagnetic material. The current design is represented by means of a level set function  $\psi : D \rightarrow \mathbb{R}$  which is positive in the ferromagnetic subdomain and negative in the air subdomain. The zero level set of  $\psi$  represents the interface between the two subdomains. Thus, we have

$$\psi(x) > 0 \Leftrightarrow x \in \Omega_f, \quad \psi(x) < 0 \Leftrightarrow x \in D \setminus \overline{\Omega}_f, \quad \psi(x) = 0 \Leftrightarrow x \in \partial\Omega_f. \quad (4.244)$$

The evolution of this level set function is guided by the generalized topological derivative, which, for a given design represented by  $\psi$ , is defined in the following way:

$$\tilde{G}_\psi(x) := \begin{cases} G^{f \rightarrow \text{air}}(x), & x \in \Omega_f, \\ -G^{\text{air} \rightarrow f}(x), & x \in D \setminus \overline{\Omega}_f. \end{cases} \quad (4.245)$$

Note that the topological derivative is only defined in the interior of  $\Omega_f$  and in the interior of  $D \setminus \overline{\Omega}_f$ , but not on the interface. The algorithm is based on the following observation:

**Lemma 4.78.** *Let  $\psi$  be a level set function representing the set  $\Omega_f$  via (4.244) and let  $\tilde{G}_\psi$  be the generalized topological derivative defined in (4.245). If, for all  $x \in \Omega_f \cup (D \setminus \overline{\Omega}_f)$ , it holds*

$$\psi(x) = c \tilde{G}_\psi(x) \quad (4.246)$$

*for a constant  $c > 0$ , then we are in the presence of a local minimum.*

*Proof.* Let  $\hat{x} \in \Omega_f$ . Then, we have by (4.244), (4.246) and (4.245) that

$$0 < \psi(\hat{x}) = \tilde{G}_\psi(\hat{x}) = G^{f \rightarrow \text{air}}(\hat{x}).$$

Thus, by definition (4.2), introducing air at point  $\hat{x}$  will yield an increase of the objective function. An analogous argument holds for  $\hat{x} \in D \setminus \overline{\Omega}_f$ .  $\square$

This observation motivates the following algorithm:

**Algorithm 2.** Initialization: Choose  $\psi_0$  with  $\|\psi_0\| = 1$ , compute  $\mathcal{J}(\psi_0)$  and  $\tilde{G}_{\psi_0}$  and set  $k = 0$ .

(i) Set  $\theta_k = \arccos \left( \psi_k, \frac{\tilde{G}_{\psi_k}}{\|\tilde{G}_{\psi_k}\|} \right)$  and

$$\psi_{k+1} = \frac{1}{\sin \theta_k} \left( \sin((1 - \kappa_k)\theta_k) \psi_k + \sin(\kappa_k \theta_k) \frac{\tilde{G}_{\psi_k}}{\|\tilde{G}_{\psi_k}\|} \right),$$

where  $\kappa_k = \max\{1, 1/2, 1/4, \dots\}$  such that  $\mathcal{J}(\psi_{k+1}) < \mathcal{J}(\psi_k)$

(ii) Compute  $\tilde{G}_{\psi_{k+1}}$  according to (4.245)

(iii) If  $\tilde{G}_{\psi_{k+1}} = \psi_{k+1}$  then stop, else set  $k \leftarrow k + 1$  and go to (ii)

Here, we identified the domain  $\Omega$  with the level set function  $\psi$  representing  $\Omega$  and wrote  $\mathcal{J}(\psi)$  instead of  $\mathcal{J}(\Omega)$ . Note that each evaluation of the objective function  $\mathcal{J}$  requires the solution of the state equation (2.17b) and each evaluation of the generalized topological derivative  $\tilde{G}_\psi$  additionally requires the adjoint state  $p$ , i.e., the solution to (2.31). Also note that the condition  $\tilde{G}_{\psi_{k+1}} = \psi_{k+1}$  in step (iii) could be replaced by the condition that  $\tilde{G}_{\psi_{k+1}}$  has the same sign as  $\psi_{k+1}$  almost everywhere in  $\Omega^d$ . Here, the norms and the scalar product are to be understood in the space  $L^2(D)$ . Algorithm 2 tries to minimize the angle between the level set function  $\psi$  and the generalized topological derivative  $\tilde{G}_\psi$  in an  $L^2$  sense. Note that, if a level set function  $\psi$  is multiplied by a positive constant, it still represents the same domain  $\Omega_f$ . Therefore, in order to get rid of this useless degree of freedom and to make the algorithm numerically more stable, the optimization is performed on the unit sphere of  $L^2$ , i.e., for all iterates  $\psi_k$  we have  $\|\psi_k\| = 1$ . Similarly to Algorithm 1, we use a line search in the parameter  $\kappa$  in order to ensure a decrease of the objective function in each iteration.

**Remark 4.79.** The generalized topological derivative  $\tilde{G}_\psi$  defined in (4.245) is not defined on the material interface  $\partial\Omega_f$ . However, we remark that the procedure of Algorithm 2 in a finite element setting is still well-defined. On the discrete level, we approximated the current design by deciding for each element whether it should belong to  $\Omega_f$  or  $\Omega_{air}$ . Proceeding like this yields a jagged interface, an issue which will be dealt with in Chapter 7. We use piecewise linear, globally continuous finite elements and evaluate the generalized topological derivative element-wise. In order to be able to add the level set function, which is given as nodal data, and the generalized topological derivative in step (ii) of the algorithm, we transferred this element data to the nodes of the mesh by a simple averaging. This can be seen as a smoothing of the topological derivative, which would otherwise be discontinuous across the interface. More details on the algorithm as well as its implementation can be found in [16, 19].

In our implementation, given a level set function  $\psi$ , we decided for each element of the mesh whether it should be part of (the approximations to)  $\Omega_f$  or  $\Omega_{air}$  by the position of its centroid, which gives a jagged interface. In Chapter 7, we will introduce a mesh modification strategy which yields a smoother resolution of the interface.

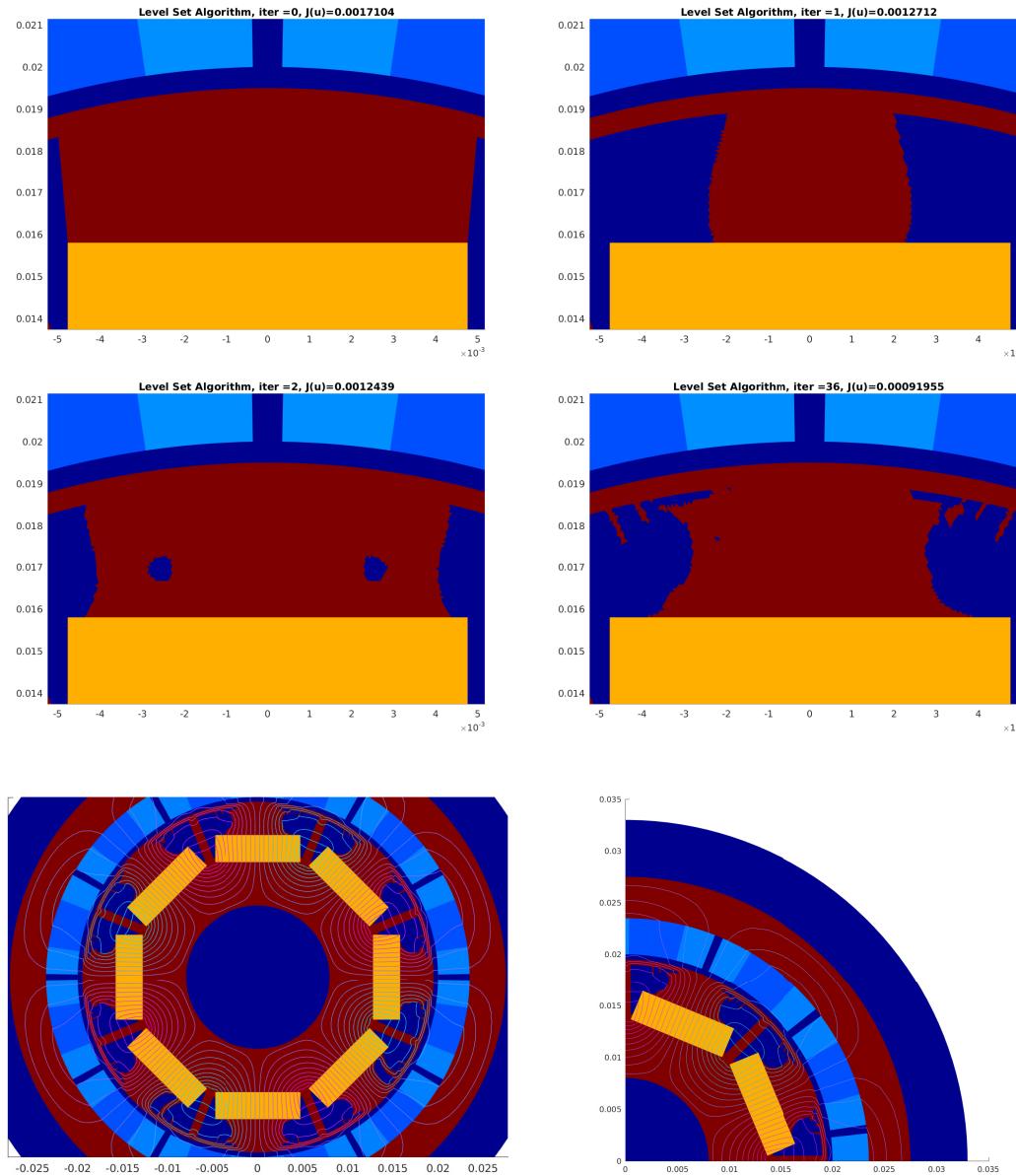


Figure 4.10: Evolution of the design in the course of the Algorithm 2. Top left: Initial design. Top right: Design after first iteration. Center left: Design after two iterations. Center right: Final design after 36 iterations. Bottom: Final design after 36 iterations of Algorithm 2 together with magnetic field.

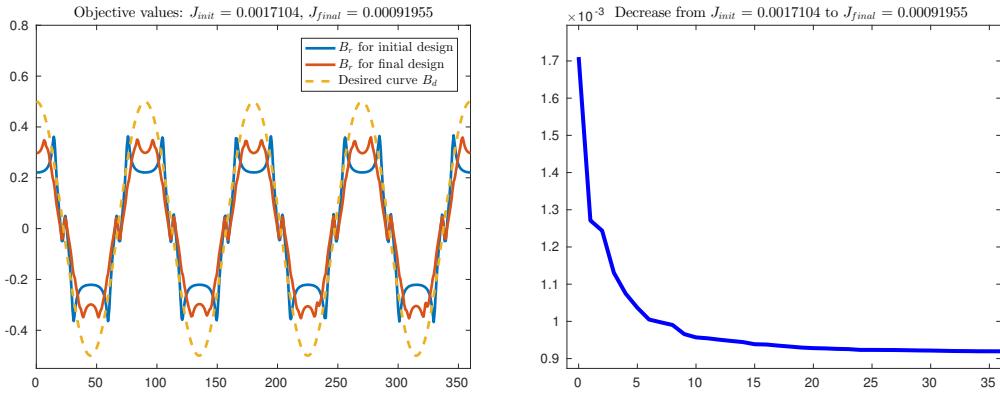


Figure 4.11: Left: Radial component of the magnetic flux density along the air gap for initial and final design compared with desired curve  $B_d$ . Right: Decrease of objective function (2.15) in the course of optimization procedure.

Figures 4.10–4.11 show the results obtained by applying Algorithm 2 to the design optimization problem (2.17). Here, we are only interested in altering the design in the eight parts of the design subdomain  $\Omega^d$ , see Figure 2.2, and, therefore, neglect both the level set function and the generalized topological derivative outside  $\Omega^d$ . Figure 4.10 shows the initial design, the design after one and two iterations as well as the final design obtained after a total of 36 iterations of Algorithm 2. Compared to the results obtained by the On/Off method in Section 3.3, it can be seen that a much smaller objective value was achieved. Figure 4.11 shows the radial component of the magnetic flux density for the initial and final design compared with the desired smooth sine curve  $B_d$ , as well as the decrease of the objective function  $\mathcal{J}$  defined in (2.15). Again, we remark that, due to the restriction on the design domain, the desired curve  $B_d$  cannot be reached. Also here we see that the final design is not perfectly periodic, i.e., we can observe different designs in some parts of the design subdomain  $\Omega^d$ . Again, this might be an effect coming from the mesh which is not perfectly symmetric. If one is only interested in symmetric final designs, we recommend not to consider the eight parts of the design subdomain individually, but rather to perform the optimization only in one part and use a rotated version of that level set function in the other parts.



# Chapter 5

## Comparison

In this chapter, we draw a comparison between the concepts of the material sensitivities used in the On/Off method introduced in Chapter 3 and the topological derivative derived in Chapter 4 in both the case of a linear and of a nonlinear state equation. This chapter is partly based on [87].

Recall that in Chapter 3, we first computed the On/Off sensitivities proposed in [163] on the discrete level in Section 3.1 and then generalized the idea to perturbations of the material coefficient  $\nu$  in arbitrary subdomains of  $\Omega^d$  by means of Fréchet derivatives in Section 3.2. Here, we will denote this arbitrary subdomain by  $\hat{\omega}$  in order to avoid confusion with the set  $\omega$  introduced in Section 4.1.2, which represents the shape of the inclusion for the topological derivative. We consider a fixed point  $x_0$  in  $\Omega^d$  and a smooth set  $\hat{\omega} \subset \Omega^d$  which contains  $x_0$ . We assume that the objective function  $\mathcal{J}$  does not depend explicitly on the reluctivity values inside the design domain  $\Omega^d$  such that  $\frac{\partial \mathcal{J}}{\partial \nu_k}$  and  $\delta_J$  in (3.2) and (4.7), respectively, vanish.

### 5.1 Linear Case

#### 5.1.1 Continuous Setting

Let  $u_0^{lin} \in H_0^1(D)$  and  $p_0^{lin} \in H_0^1(D)$  denote the solutions to the unperturbed state equation (2.17b) and the unperturbed adjoint equation (2.31), respectively, under the assumption of only linear material behavior, i.e.,  $\hat{\nu} = \nu_1 = const$ . Assume that  $u_0^{lin}$  and  $p_0^{lin}$  are sufficiently regular. By choosing  $\hat{\omega} = \omega_\varepsilon = x_0 + \varepsilon\omega$  with  $\varepsilon > 0$  as in the topological derivative, we get by the mean value theorem that there exists a  $\xi \in \omega_\varepsilon$  such that

$$\frac{1}{|\hat{\omega}|} \frac{d \mathcal{J}}{d \nu_{\hat{\omega}}} = \frac{1}{|\omega_\varepsilon|} \frac{d \mathcal{J}}{d \nu_{\omega_\varepsilon}} = \frac{1}{|\omega_\varepsilon|} \int_{\omega_\varepsilon} \nabla u_0^{lin} \cdot \nabla p_0^{lin} \, dx = \frac{1}{|\omega_\varepsilon|} |\omega_\varepsilon| \nabla u_0^{lin}(\xi) \cdot \nabla p_0^{lin}(\xi),$$

and, therefore, for  $\varepsilon \rightarrow 0$ ,

$$\frac{1}{|\omega_\varepsilon|} \frac{d \mathcal{J}}{d \nu_{\omega_\varepsilon}} \rightarrow \nabla u_0^{lin}(x_0) \cdot \nabla p_0^{lin}(x_0). \quad (5.1)$$

On the other hand, recall the formula for the topological derivative in the case of a circular inclusion  $\omega = B(0, 1)$  if  $x_0$  is in the ferromagnetic subdomain,

$$G^{f \rightarrow air}(x_0) = C^{f \rightarrow air} \nabla u_0^{lin}(x_0) \cdot \nabla p_0^{lin}(x_0), \quad (5.2)$$

with the constant  $C^{f \rightarrow air} = 2\nu_1 \frac{\nu_0/\nu_1 - 1}{\nu_0/\nu_1 + 1} \pi$ , see (4.22). Also recall the topological derivative for the case where  $x_0$  is in the air subdomain,

$$G^{air \rightarrow f}(x_0) = C^{air \rightarrow f} \nabla u_0^{lin}(x_0) \cdot \nabla p_0^{lin}(x_0),$$

with  $C^{air \rightarrow f} = 2\nu_0 \frac{\nu_1/\nu_0 - 1}{\nu_1/\nu_0 + 1} \pi$ , see (4.23), which is obtained in the exact same way as (4.22), but exchanging  $\nu_0$  and  $\nu_1$ .

We make the following observation: In the limiting case, the scaled version of the On/Off sensitivities (5.1) and the topological derivative (5.2) only differ by a constant factor. If one is only interested in removing material at the positions where these sensitivities are most negative, this constant factor does not play a role and therefore, the two sensitivities can be regarded as equivalent. However, if one wants to use a bi-directional optimization algorithm which can decide if it is better to remove material at one location in  $\Omega_f$  or to add material add another location in  $\Omega_{air}$ , it is important that the right scaling between the sensitivities in the ferromagnetic and in the air subdomain is used. From (4.22) and (4.23), it can be seen that

$$C^{air \rightarrow f} = -\frac{\nu_0}{\nu_1} C^{f \rightarrow air}.$$

In the case of electrical machines, the ratio  $\nu_0/\nu_1$  is usually of the order of  $10^3$ . In the absence of this factor, as it is the case for the On/Off sensitivities, a numerical optimization algorithm would always favor removing material even if an introduction of material at another position would yield a larger decrease of the cost function, see Figure 5.1 for the discussion of the same issue in the nonlinear case.

### 5.1.2 Discrete Setting

We compare the two different kinds of sensitivities for the setting which would be used in a numerical optimization algorithm where the state and adjoint equations are solved by a finite element method using globally continuous, piecewise linear finite elements on a triangular grid. Let  $u_h^{lin}$  and  $p_h^{lin}$ , denote the finite element approximations to the solutions of (2.17b) and (2.31) under the assumption of only linear material behavior,  $\hat{\nu} = \nu_1 = const$ .

For  $x_0 \in \Omega^d$  fixed, by choosing  $\hat{\omega}$  as the triangle  $T_k$  of the mesh which contains the point  $x_0$ , we get for the On/Off sensitivities as they were used in [163] and several other papers in the engineering community,

$$\frac{d\mathcal{J}}{d\nu_{\hat{\omega}}} = \frac{d\mathcal{J}}{d\nu_{T_k}} = \int_{T_k} \nabla u_h^{lin} \cdot \nabla p_h^{lin} dx = |T_k| \nabla u_h^{lin}(x_0) \cdot \nabla p_h^{lin}(x_0),$$

and for the topological derivative in the case of a circular inclusion  $\omega = B(0, 1)$ ,

$$\begin{aligned} G^{f \rightarrow air}(x_0) &= C^{f \rightarrow air} \nabla u_h^{lin}(x_0) \cdot \nabla p_h^{lin}(x_0), & x \in \Omega_f \\ G^{air \rightarrow f}(x_0) &= C^{air \rightarrow f} \nabla u_h^{lin}(x_0) \cdot \nabla p_h^{lin}(x_0), & x \in D \setminus \overline{\Omega}_f \end{aligned}$$

with  $C^{f \rightarrow air}$ ,  $C^{air \rightarrow f}$  as above.

Here, we make the following additional observation: In order to avoid effects coming from the mesh, it is advisable to scale the On/Off sensitivities by the area of the respective element,

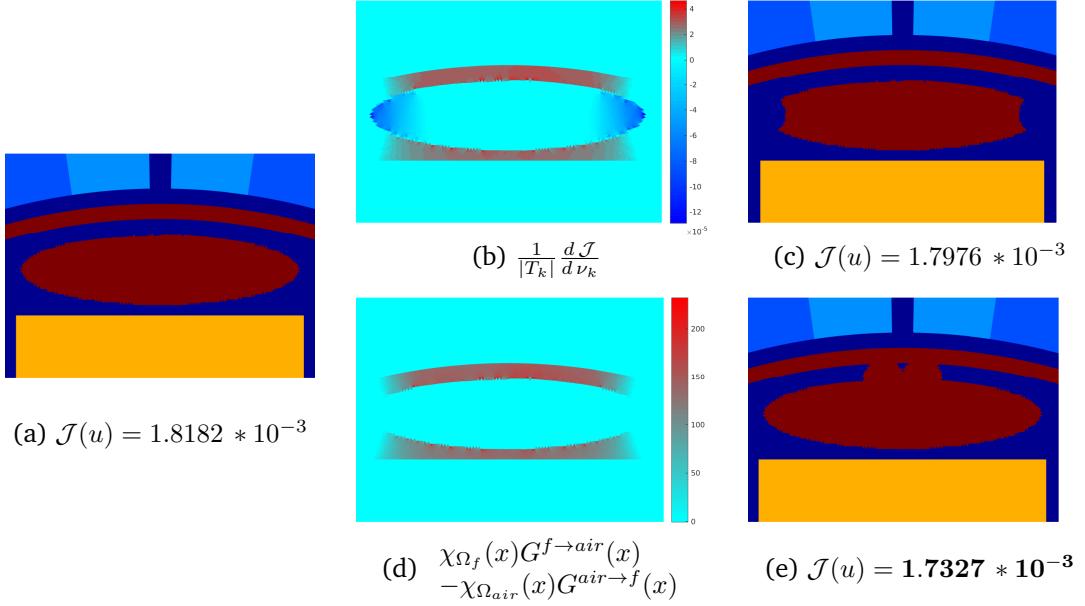


Figure 5.1: Illustration of difference between sensitivities in On/Off-Method (top row) and topological derivative (bottom row). (a) Ellipse-shaped initial design. (b) Scaled On/Off sensitivities. (c) Improved design according to On/Off sensitivities. (d) Topological derivative. (e) Improved design according to topological derivative giving a significantly larger decrease of the objective value.

i.e., we recommend to use the sensitivities

$$\frac{1}{|T_k|} \frac{d\mathcal{J}}{d\nu_{T_k}},$$

see also [49] for a comparison of the non-scaled and the scaled On/Off sensitivities.

## 5.2 Nonlinear Case

Again, let  $x_0 \in \Omega^d$  be fixed and  $\hat{\omega} = \omega_\varepsilon = x_0 + \varepsilon\omega$  with  $\omega = B(0, 1)$ . Let  $u_0^{nl} \in H_0^1(D)$  and  $p_0^{nl} \in H_0^1(D)$  the solutions to the unperturbed state and adjoint equation, respectively, where  $\hat{\nu}$  is a nonlinear function coming from a  $B$ - $H$ -curve, see Section 2.4. Assume that  $u_0^{nl}$  and  $p_0^{nl}$  are sufficiently regular. We noted in Chapter 3 that the On/Off sensitivities in the case of a nonlinear state equation are of the same form as in the linear case. Also here, we get in the limit that

$$\frac{1}{|\hat{\omega}|} \frac{d\mathcal{J}}{d\nu_{\hat{\omega}}} = \frac{1}{|\omega_\varepsilon|} \frac{d\mathcal{J}}{d\nu_{\omega_\varepsilon}} = \frac{1}{|\omega_\varepsilon|} \int_{\omega_\varepsilon} \nabla u_0^{nl} \cdot \nabla p_0^{nl} dx \rightarrow \nabla u_0^{nl}(x_0) \cdot \nabla p_0^{nl}(x_0). \quad (5.3)$$

On the other hand, the topological derivative for introducing air inside ferromagnetic material as derived in Chapter 4 reads

$$G^{f \rightarrow air}(x_0) = \nabla u_0^{nl}(x_0)^\top \mathcal{M}^{f \rightarrow air} \nabla p_0^{nl}(x_0) + \int_{\mathbb{R}^2} \tilde{S}_{\nabla u_0^{nl}(x_0)}(x, \nabla H) \cdot (\nabla p_0^{nl}(x_0) + \nabla K) dx,$$

	$\lim_{\varepsilon \rightarrow 0} \frac{1}{ \omega_\varepsilon } \frac{\partial \mathcal{J}}{\partial \nu_{\omega_\varepsilon}}$ (On/Off)	$G^{f \rightarrow air}(x_0)$ (Topological Derivative)
linear	$\nabla u_0^{lin}(x_0) \cdot \nabla p_0^{lin}(x_0)$	$C^{f \rightarrow air} \nabla u_0^{lin}(x_0) \cdot \nabla p_0^{lin}(x_0)$
nonlinear	$\nabla u_0^{nl}(x_0) \cdot \nabla p_0^{nl}(x_0)$	$\nabla u_0^{nl}(x_0)^\top \mathcal{M}^{f \rightarrow air} \nabla p_0^{nl}(x_0)$ $\int_{\mathbb{R}^2} \tilde{S}_{\nabla u_0^{nl}(x_0)}(x, \nabla H) \cdot (\nabla p_0^{nl}(x_0) + \nabla K) dx$

Table 5.1: Comparison of scaled On/Off sensitivities (5.1), (5.3) and topological derivative (4.22), (4.108) in case of a linear and a nonlinear state equation.

see (4.108), where  $\mathcal{M}^{f \rightarrow air}$  denotes the matrix  $\mathcal{M}$  for Case I given in (4.233). For the topological derivative for introducing ferromagnetic material inside air, i.e., in Case II, we obtained

$$\begin{aligned} G^{air \rightarrow f}(x_0) = & \nabla u_0^{nl}(x_0)^\top \mathcal{M}^{air \rightarrow f} \nabla p_0^{nl}(x_0) \\ & + \int_{\mathbb{R}^2} \tilde{S}_{\nabla u_0^{nl}(x_0)}^{(2)}(x, \nabla H^{(2)}) \cdot (\nabla p_0^{nl}(x_0) + \nabla K^{(2)}) dx, \end{aligned}$$

see (4.216), with  $\mathcal{M}^{air \rightarrow f}$  the matrix  $\mathcal{M}^{(2)}$  defined in (4.234).

In the nonlinear case, we observe that the topological derivative differs from the scaled version of the On/Off sensitivities (5.1) in two ways:

1. in the matrices  $\mathcal{M}^{f \rightarrow air}$  and  $\mathcal{M}^{air \rightarrow f}$ , which play the same role as the constants  $C^{f \rightarrow air}$  and  $C^{air \rightarrow f}$  in the linear case, but now are anisotropic  $2 \times 2$  matrices which depend on the gradient of the solution at the point  $x_0$ , and
2. in the presence of a new second term which accounts for the nonlinearity of the problem.

We saw in Section 4.7 that, in our case, the second term in (5.2) is negligible in comparison with the first term. Therefore, we neglect this term for the rest of this discussion. In Figure 5.1, we illustrate the importance of the polarization matrices  $\mathcal{M}^{f \rightarrow air}$  and  $\mathcal{M}^{air \rightarrow f}$ . For a hypothetical ellipse-shaped design, the scaled On/Off sensitivities suggest to remove material on the left and right ends of the ellipse, whereas the topological derivative suggests to add material at the top. Figure 5.1 gives numerical evidence that the latter yields a much larger decrease of the objective functional. This phenomenon might lead to final designs obtained by the On/Off method that are not optimal because the wrong sensitivities have been used. Note that we used the generalized topological derivative  $\tilde{G}(x) = \chi_{\Omega_f}(x)G^{f \rightarrow air}(x) - \chi_{\Omega_{air}}(x)G^{air \rightarrow f}(x)$  as introduced in Section 4.8 such that positive (or negative) values of  $\tilde{G}$  mean that adding (or removing) material would yield a decrease of the objective function.

In Table 5.1, we summarize the computed sensitivities.

### 5.3 General Remarks

The conceptual difference between the two kinds of sensitivities is the following:

- The scaled On/Off sensitivities (5.1) represent the sensitivities with respect to a small perturbation of the material property. This information is important, for instance, when one is interested in the sensitivity of a design with respect to manufacturing errors.
- On the other hand, the topological derivative is the sensitivity with respect to a change of material from  $\hat{\nu}$  directly to  $\nu_0$  in a small neighborhood of a point  $x_0$ .

From this comparison, it is clear that the topological derivative is the right sensitivity to be considered for topology optimization. As it was observed in [50], it might happen that the material sensitivity indicates a decrease of the objective functional if the material is modified only slightly, but that the functional actually increases when the material parameter is switched to the other material. This happens when the implicitly assumed monotonicity assumption mentioned in the beginning of Chapter 3 is violated.

A similar phenomenon was observed in [102, 103] where a density-based method is used for topology optimization of a coaxial-to-waveguide transition. In this particular application, the authors observed effects of self-penalization of intermediate density values. This means that intermediate density values yield a very bad performance of the structure and, therefore, when considering sensitivities with respect to the density variable, every “black-and-white” design seems to be locally optimal. The authors overcame this issue by the construction of a special density filter. However, this application is another example of a situation where the topological derivative should be preferred over material sensitivities.

We mention, however, that the (scaled) On/Off sensitivities are the right quantities in order to determine the sensitivity of the performance of a device with respect to small perturbations of the material coefficients, which often occur through manufacturing errors, as it was considered in [49].



## **Part II**

# **Shape Optimization**



# Chapter 6

## Shape Optimization

In this chapter, we treat model problem (2.17) introduced in Section 2.3 by means of shape optimization. The shape derivative represents the sensitivity of a cost function which depends on the shape of a (sub-)domain  $\Omega$  with respect to a smooth variation of the boundary. In contrast to optimization algorithms steered by the topological derivative, here, the optimal shape has a smooth boundary provided that the numerical algorithm is devised carefully. Computing the shape derivative for problems depending on linear partial differential equations is a well-understood topic, see for instance the monographs [75, 107, 196]. For nonlinear problems, the literature is scarcer and the computation of the shape derivative is often formal. In this chapter, we derive the shape derivative of the model problem (2.17), which is subject to a quasilinear PDE, in a rigorous way, based on a novel Lagrangian method for nonlinear problems, see [198]. Based on the shape derivative, we can obtain smooth deformation fields which we use as a descent direction in a gradient method. In the numerical algorithm, the interface is updated iteratively using this vector field until it reaches an equilibrium state.

The rest of this chapter is organized as follows: In Section 6.1, we introduce the concept of the shape derivative and discuss different ways how it can be computed as well as the two different possible representations. We prepare the setting of our nonlinear problem in Section 6.2. Existence of a solution to the shape optimization problem in a certain class of shapes is shown in Section 6.3. In Section 6.4, we give an abstract differentiability result which is used later to compute the shape derivative of the cost functional. Section 6.5 deals with the shape derivative of the cost functional of the model problem introduced in Section (2.3). Finally, in Section 6.6, a numerical algorithm is presented and numerical optimization results are presented.

This chapter is based on [90].

### 6.1 Introduction

In this section, we introduce some basic facts about the velocity method in shape optimization used to transform a reference shape, see [75, 196]. In the velocity method, also known as speed method, a domain  $\Omega \subset D \subset \mathbb{R}^2$  is deformed by the action of a velocity field  $V$  defined on  $D$ . Suppose that  $D$  is a Lipschitz domain and denote its boundary  $\Sigma := \partial D$ . The domain

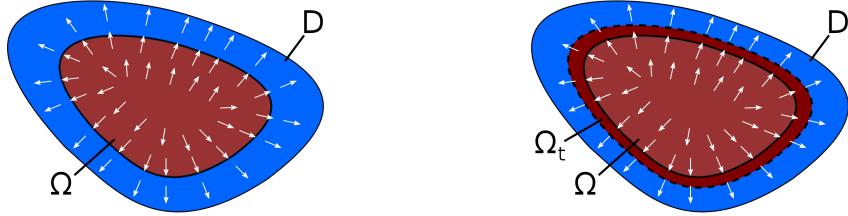


Figure 6.1: Domain  $\Omega$  is transformed to the domain  $\Omega_t$  for a given  $t > 0$  by means of the velocity field  $V$ .

evolution is described by the solution of the dynamical system

$$\frac{d}{dt}x(t) = V(x(t)), \quad t \in [0, \varepsilon], \quad x(0) = X \in \mathbb{R}^2, \quad (6.1)$$

for some real number  $\varepsilon > 0$ . Suppose that  $V$  is continuously differentiable and has compact support in  $D$ , i.e.,  $V \in \mathcal{D}^1(D, \mathbb{R}^2)$ . Then the ordinary differential equation (6.1) has a unique solution on  $[0, \varepsilon)$ . This enables us to define the diffeomorphism

$$T_t : \mathbb{R}^2 \rightarrow \mathbb{R}^2; X \mapsto T_t(X) := x(t), \quad (6.2)$$

which transforms the domain  $\Omega$  into  $\Omega_t := T_t(\Omega)$ , see Figure 6.1. Let  $\mathcal{P}(\mathbb{R}^2)$  denote the power set of  $\mathbb{R}^2$ , i.e., the set of all subsets of  $\mathbb{R}^2$ .

**Definition 6.1.** Suppose we are given a real valued shape function  $\mathcal{J}$  defined on a subset  $\Xi$  of  $\mathcal{P}(\mathbb{R}^2)$ . We say that  $\mathcal{J}$  is Eulerian semi-differentiable at  $\Omega \in \Xi$  in the direction  $V$  if the following limit exists in  $\mathbb{R}$

$$d\mathcal{J}(\Omega; V) := \lim_{t \searrow 0} \frac{\mathcal{J}(T_t(\Omega)) - \mathcal{J}(\Omega)}{t}.$$

If the map  $V \rightarrow d\mathcal{J}(\Omega; V)$  is linear and continuous with respect to the topology of  $\mathcal{D}(D, \mathbb{R}^2) := C_c^\infty(D, \mathbb{R}^2)$ , then  $\mathcal{J}$  is said to be shape differentiable at  $\Omega$  and  $d\mathcal{J}(\Omega; V)$  is called the shape derivative of  $\mathcal{J}$ .

**Remark 6.2.** In PDE-constrained shape optimization, the term shape derivative is also often used to refer to the sensitivity of the state with respect to a perturbation of the domain. Given a velocity field  $V$  and a family of transformations  $T_t$  as in (6.1)–(6.2), for  $t > 0$  small enough, let  $u_t$  be the solution to the constraining boundary value problem posed on the transformed domain  $\Omega_t = T_t(\Omega)$ . Then, the total derivative of the perturbed state  $u_t$  at  $t = 0$ ,

$$\dot{u}(x) := \frac{d}{dt}u_t(T_t(x))|_{t=0}, \quad x \in D, \quad (6.3)$$

is called the material derivative of  $u$ , and the partial derivative

$$u'(x) := \frac{d}{dt}u_t(x)|_{t=0}, \quad x \in D,$$

is called the shape derivative of  $u$ . By application of the chain rule, it is easily seen that these two concepts are related via

$$\dot{u} = u' + \nabla u \cdot V.$$

However, we remark that, in this thesis, the term shape derivative will always refer to the sensitivity of the objective function according to Definition 6.1.

### 6.1.1 Computation of Shape Derivative

We give a brief overview over the different approaches to the derivation of a shape derivative according to Definition 6.1 that are available in the literature. This overview is based on the paper [198].

One commonly used approach to shape differentiation in the presence of a PDE constraint involves the material derivative  $\dot{u}$  of the state variable, see (6.3), in intermediate steps of the derivation of the shape derivative of  $\mathcal{J}$ . Often, the proof of the differentiability of the mapping  $t \mapsto u_t \circ T_t$  is not straightforward and requires tools like the implicit function theorem. The material derivative  $\dot{u}$  is not present in the final formula for the shape derivative, but is only used in intermediate steps. This serves as a motivation for exploring other methods for shape differentiation.

In fact, there are several methods available to prove the shape differentiability of functions depending on the solution of a PDE without the need to compute the derivative of the state.

Several Lagrangian-type methods have been proposed in shape optimization. The approach of Céa [62] is frequently used to derive the formulas for the shape derivative, but it is formal and may lead to the wrong formulas if not used carefully, see [167]. The minimax approach of Correa and Seeger [69], which was first applied to shape optimization in [74], was the first Lagrangian method with a rigorous mathematical treatment. We mention that it is valid only for semi-convex cost functions and is based on the assumption of the existence of a saddle point, which is often difficult to verify. Lagrangian approaches also have the advantage of providing directly the adjoint equation.

Variational methods have also been proposed recently [118, 123, 124]. In these papers, a rearrangement technique is used to bypass the computation of the shape derivative of the state equation, which allows to reduce the regularity assumptions. The method proposed in this chapter of the thesis and the methods proposed in [118, 123, 124] are very general as they require little regularity assumptions on the domains and the solutions of the PDE involved. The approaches to compute the shape derivatives are quite different and it is not yet clear what is the exact relation between these methods.

The approach we use for deriving the shape derivative of the PDE-constrained shape optimization problem (2.17) in this chapter is based on the result in [198]. This result generalizes [74] and simplifies the assumptions as we can treat more general cost functionals. In contrast to the result in [167], this approach allows to deal with nonlinear PDEs. For a comprehensive comparison of the different approaches mentioned above, see [198].

### 6.1.2 Representation of Shape Derivative

As we already mentioned in the introduction of this thesis in Section 1.2.2, there are basically two ways how one can represent the shape derivative: either in the Hadamard form (1.2), i.e., as a distribution on the boundary which only depends on the normal component of the perturbation, or in the more general volume form (1.3), also called the distributed shape derivative. If the shape  $\Omega$  is regular enough, the Hadamard form can be rewritten as an integral over the boundary,

$$d\mathcal{J}(\Omega; V) = \int_{\partial\Omega} g_\Gamma V \cdot n \, ds, \quad (6.4)$$

with an integrable function  $g_\Gamma$ . The volume form can be written as

$$d\mathcal{J}(\Omega; V) = \int_{\Omega} g(V, DV) \, dx, \quad (6.5)$$

for some function  $g$ . One obvious advantage of the boundary-based form (6.4) is that a descent direction  $V = -g_\Gamma n$  is readily available whereas the extraction of a descent direction requires the solution of an additional PDE in the case of the volume-based form (6.5), see Section 1.2.2.2. For this reason, the distributed shape derivative has been widely neglected for a long time. In this section, we compare various aspects of these two representations and promote the use of the distributed shape derivative. For more details, see [134].

We start out by noting that the one advantage of the boundary-based form mentioned above does not always apply as the descent direction  $V = -g_\Gamma n$  might not be regular enough and has to be regularized. Furthermore, in many numerical procedures, it is not enough to have a descent direction that is only defined on the material interface. This is the case in the level set method, see Section 1.2.1.4, or when a material interface is aligned with an underlying finite element mesh and one wants to advect all mesh nodes in a descent direction. These difficulties are usually treated by solving an additional boundary value problem just like in the case of the distributed shape derivative, e.g., by solving the Laplace-Beltrami equation on the material interface in the former case.

On the other hand, the computational overhead of solving an additional boundary value problem in the case of the volume form may be reduced to a minimum by defining this auxiliary boundary value problem only in a small neighborhood of the material interface.

One benefit of the volume form is that, as mentioned above, it is more general, meaning that for shapes with lower regularity the distributed shape derivative may be well-defined whereas the Hadamard form is not.

A different aspect favoring the volume-based form (6.5) is concerned with numerical accuracy of the approximation of the shape derivative when the underlying state and adjoint equations are solved by the finite element method. In [111], the authors show that the finite element approximation to the volume-based form converges quadratically to the “true” shape derivative on the continuous level as the mesh size tends to zero, whereas the boundary-based form converges only linearly.

Another advantage pointed out in [44] is that the discretization process and the shape differentiation commute for the volume expression but not for the boundary expression, i.e., a

discretization of the boundary expression of the shape derivative does not generally lead to the same expression as the shape derivative computed when the problem is first discretized. For the above reasons, we will derive the shape derivative in its volume-based form (6.5).

## 6.2 Preliminaries

Recall the notation from Section 2.3, in particular recall the variable ferromagnetic set  $\Omega_f = (\Omega_f^{ref} \setminus \Omega^d) \cup \Omega$  where  $\Omega$  is the part to be optimized. For  $(x, \zeta) \in D \times \mathbb{R}$ , we define the nonlinear, piecewise smooth function  $\beta = \beta_\Omega$  as

$$\begin{aligned}\beta_\Omega(x, \zeta) &:= \beta_1(\zeta)\chi_{\Omega_f}(x) + \beta_2(\zeta)\chi_{D \setminus \Omega_f}(x) \\ &= \beta_1(\zeta)(\chi_\Omega(x) + \chi_{\Omega_f^{ref} \setminus \Omega^d}(x)) + \beta_2(\zeta)(\chi_{D \setminus \Omega_f^{ref}}(x) + \chi_{\Omega^d \setminus \Omega}(x)),\end{aligned}$$

where  $\chi$  is the indicator function of a given set. Again, note that the expression above is meaningful since  $\Omega \subset \Omega^d \subset \Omega_f^{ref}$ .

We will derive the shape derivative for a shape optimization problem constrained by the following boundary value problem:

$$\text{Find } u \in H_0^1(D) \text{ such that } \int_D \beta_\Omega(x, |\nabla u|^2) \nabla u \cdot \nabla \eta \, dx = \langle F, \eta \rangle \quad \forall \eta \in H_0^1(D), \quad (6.6)$$

where, in view of the model problem of Section 2.3, we assume that the right hand side  $F \in H^{-1}(D)$  is of the form

$$\langle F, \eta \rangle = \int_D f_0 \eta \, dx + \int_{\Omega_{mag}} M_1 \cdot \nabla \eta \, dx + \int_{D \setminus \Omega_{mag}} M_2 \cdot \nabla \eta \, dx, \quad (6.7)$$

where  $f_0 \in H^1(D)$ ,  $M_1$  is a constant vector on each of the eight magnets and  $M_2$  is a constant vector.

In the sequel, we make the following assumption for  $\beta_1$  and  $\beta_2$ :

**Assumption 9.** The functions  $\beta_1, \beta_2 : \mathbb{R} \rightarrow \mathbb{R}$  satisfy the following conditions:

1. For  $i = 1, 2$ , the function  $s \mapsto \beta_i(s^2)s$  is strongly monotone with monotonicity constant  $m$  and Lipschitz continuous with Lipschitz constant  $L$ :

$$\begin{aligned}(\beta_i(s^2)s - \beta_i(t^2)t)(s - t) &\geq m(s - t)^2 && \text{for all } s, t \geq 0, \\ |\beta_i(s^2)s - \beta_i(t^2)t| &\leq L|s - t| && \text{for all } s, t \geq 0.\end{aligned}$$

2. The functions  $\beta_1, \beta_2$  are in  $C^1(\mathbb{R})$ .

3. There exist constants  $\lambda, \Lambda > 0$  such that for  $i = 1, 2$ ,

$$\lambda|\varphi|^2 \leq \beta_i(|\rho|^2)|\varphi|^2 + 2\beta'_i(|\rho|^2)(\varphi \cdot \rho)^2 \leq \Lambda|\varphi|^2 \quad \text{for all } \varphi, \rho \in \mathbb{R}^2.$$

**Remark 6.3.** From Assumption 9.1, it can easily be seen by setting  $t = 0$  that, under Assumption 9.1, there exist constants  $c_1, c_2, c_3, c_4 > 0$ , such that

$$c_1 \leq \beta_1(\zeta) \leq c_2, \quad c_3 \leq \beta_2(\zeta) \leq c_4 \quad \text{for all } \zeta \in \mathbb{R}.$$

Note that, in the special case where  $\beta_1(\zeta) := \hat{\nu}(\sqrt{\zeta})$  and  $\beta_2(\zeta) := \nu_0$ , the function  $\beta_\Omega$  is related to  $\nu_\Omega$  defined in (2.11) as

$$\beta_\Omega(x, \zeta^2) = \nu_\Omega(x, \zeta),$$

and (6.6) coincides with the magnetostatic boundary value problem defined on the electric motor introduced in (2.14). Boundary value problem (6.6) models a nonlinear potential equation defined on a domain which consists of two different materials both of which may exhibit nonlinear material behavior. Note that the function  $\beta_\Omega$ , unlike  $\nu_\Omega$ , depends on the square of  $|\nabla u|$ , which is meant to simplify some of the forthcoming calculations.

We consider the PDE-constrained shape optimization problem consisting of finding a shape  $\Omega$  out of a set of admissible shapes  $\mathcal{O}$  which minimizes the functional  $\mathcal{J}$  defined in (2.15) subject to the PDE constraint (6.6). The problem reads

$$\begin{cases} \text{minimize } \mathcal{J}(u) = \int_{\Omega_g} |B_r(u) - B_d|^2 dx \\ \text{subject to } \Omega \in \mathcal{O} \text{ and } u \text{ solution of (6.6)} \end{cases} \quad (6.8)$$

where

$$\mathcal{O} = \left\{ \Omega \subset \Omega^d \subset \Omega_f^{ref}, \Omega \text{ open, Lipschitz with uniform Lipschitz constant } L_{\mathcal{O}} \right\}. \quad (6.9)$$

The sets  $\Omega^d$  and  $\Omega_f^{ref}$  are reference domains, see Section 2.3 and Figure 2.2. Here,  $B_r(u)$  denotes the radial component of the magnetic flux density  $\mathbf{B}(u)$  in the air gap, i.e.,  $B_r(u) = \mathbf{B}(u) \cdot n_g = \nabla u \cdot \tau_g$  where  $n_g, \tau_g$  are the vector fields defined in (2.16), and  $B_d \in C^1(\Omega_g)$  denotes the given desired radial component of the magnetic flux density along the air gap. In order to obtain the first-order optimality conditions for this minimization problem we compute the derivative of  $\mathcal{J}$  with respect to the shape  $\Omega$ . In the sequel, we emphasize the dependence of  $\mathcal{J}$  on the variable set  $\Omega \in \mathcal{O}$  by introducing  $\hat{\mathcal{J}}(\Omega) := \mathcal{J}(u(\Omega))$  where  $u(\Omega)$  denotes the solution to (6.6) for the given set  $\Omega$ .

Given Assumption 9, we can state existence and uniqueness of a solution  $u$  to the state equation (6.6).

**Theorem 6.4.** *Assume that Assumption 9.1 holds. Then problem (6.6) admits a unique solution  $u \in H_0^1(D)$  for any fixed right-hand side  $F \in H^{-1}(D)$  and we have the estimate*

$$\|u\|_{H_0^1(D)} \leq c \|F\|_{H^{-1}(D)}.$$

*Proof.* The proof is analogous to the proof of Theorem 2.7 using the Theorem of Zarantonello (Theorem 2.5). Note that Assumption 9.1 yields the strong monotonicity and Lipschitz continuity of the operator  $B_\Omega : H_0^1(D) \rightarrow H^{-1}(D)$  defined by

$$\langle B_\Omega(u), \eta \rangle := \int_D \beta_\Omega(|\nabla u|^2) \nabla u \cdot \nabla \eta \, dx,$$

by Lemma 2.6. Since the right hand side  $F$  is in  $H^{-1}(D)$ , Theorem 2.5 yields the existence of a unique solution  $u \in H_0^1(D)$  to the equation  $B_\Omega(u) = F$  in  $H^{-1}(D)$ , which is just (6.6), as well as the claimed stability estimate.  $\square$

### 6.3 Existence of Optimal Shapes

In this section, we show that problem (6.8) has a solution  $\Omega^*$  in the set of admissible solutions  $\mathcal{O}$  defined by (6.9). We make use of the following result [107, Theorem 2.4.10]:

**Theorem 6.5.** *Let  $\Omega_n$  be a sequence in  $\mathcal{O}$ . Then there exists  $\Omega^* \in \mathcal{O}$  and a subsequence  $\Omega_{n_k}$  which converges to  $\Omega^*$  in the sense of Hausdorff, and in the sense of characteristic functions. In addition,  $\bar{\Omega}_{n_k}$  and  $\partial\Omega_{n_k}$  converge in the sense of Hausdorff towards  $\bar{\Omega}^*$  and  $\partial\Omega^*$ , respectively.*

Let now  $\Omega_n \in \mathcal{O}$  be a minimizing sequence for problem (6.8). Note that such a sequence exists since  $\mathcal{J} \geq 0$ . According to Theorem 6.5, we can extract a subsequence, which we still denote by  $\Omega_n$ , which converges to some  $\Omega^* \in \mathcal{O}$ . Denote  $u_n$  and  $u^*$  the solutions of (6.6) with  $\Omega_n$  and  $\Omega^*$ , respectively. It can be shown that  $u_n \rightarrow u^*$  in  $H_0^1(D)$ :

**Proposition 6.6.** *Let  $\Omega_n \in \mathcal{O}$  be a minimizing sequence for problem (6.8) and  $\Omega^*$  be an accumulation point of this sequence as in Theorem 6.5. Assume there exists  $\varepsilon > 0$  such that the solution  $u$  of (6.6) satisfies*

$$\|u^*\|_{H^{1+\varepsilon}(D) \cap H_0^1(D)} \leq c$$

where  $c$  depends only on  $f$  and  $D$ . Then the sequence  $u_n \in H_0^1(D)$  corresponding to  $\Omega_n$  converges to  $u^*$  strongly in  $H_0^1(D)$ , where  $u^*$  is the solution of (6.6) in  $\Omega^*$ .

*Proof.* The proof can be found in [90]. □

Now, the existence of a solution  $\Omega^* \in \mathcal{O}$  can be seen as follows:

$$\inf_{\Omega \in \mathcal{O}} \hat{\mathcal{J}}(\Omega) = \lim_{n \rightarrow \infty} \hat{\mathcal{J}}(\Omega_n) = \lim_{n \rightarrow \infty} \mathcal{J}(u_n) = \mathcal{J}(u^*) = \hat{\mathcal{J}}(\Omega^*),$$

where we used the fact that  $(\Omega_n)$  is a minimizing sequence, the notation  $u_n = u(\Omega_n)$  for the solution to (6.6) with  $\Omega_n$ , and the continuity of  $\mathcal{J}$ .

From now on, for better readability, we will drop the hat and just write  $\mathcal{J}(\Omega)$  instead of  $\hat{\mathcal{J}}(\Omega)$ .

### 6.4 Derivation of the Shape Derivative

#### 6.4.1 Preliminaries

Recall that the continuously differentiable velocity field  $V$  has compact support,  $V \in \mathcal{D}^1(D, \mathbb{R}^2)$ . With this choice of  $V$ , the domain  $D$  is globally invariant under the transformation  $T_t$ , i.e.,  $T_t(D) = D$  and  $T_t(\partial D) = \partial D$ . For  $t \in [0, \varepsilon)$ ,  $T_t$  is invertible. Furthermore, for sufficiently small  $t > 0$ , the Jacobian determinant

$$\xi(t) := \det DT_t \tag{6.10}$$

of  $T_t$  is strictly positive. In the sequel, we use the notation  $DT_t^{-1}$  for the inverse of  $DT_t$  and  $DT_t^{-\top}$  for the transpose of the inverse. Then the following lemma holds [75]:

**Lemma 6.7.** *For  $\varphi \in W_{loc}^{1,1}(\mathbb{R}^2)$  and  $V \in \mathcal{D}^1(\mathbb{R}^2, \mathbb{R}^2)$  we have*

$$\begin{aligned} \nabla(\varphi \circ T_t) &= DT_t^\top (\nabla \varphi) \circ T_t, & \frac{d}{dt}(\varphi \circ T_t) &= (\nabla \varphi \cdot V) \circ T_t, \\ \frac{d\xi(t)}{dt} &= \xi(t) [\operatorname{div} V(t)] \circ T_t. \end{aligned}$$

### 6.4.2 An Abstract Differentiability Result

Let  $E$  and  $F$  be Banach spaces. Let  $G$  be a function

$$G : [0, \tau] \times E \times F \rightarrow \mathbb{R}, \quad (t, \varphi, \psi) \mapsto G(t, \varphi, \psi)$$

which is affine with respect to  $\psi$  and define

$$E(t) := \{u \in E \mid d_\psi G(t, u, 0; \hat{\psi}) = 0 \text{ for all } \hat{\psi} \in F\}.$$

Let us introduce the following hypotheses.

**Assumption 10** (H0). For every  $(t, \psi) \in [0, \tau] \times F$ , we assume that

- (i) the set  $E(t)$  contains only one element and we write  $E(t) = \{u^t\}$ ,
- (ii) the function  $[0, 1] \ni s \mapsto G(t, su^t + (1-s)u^0, \psi)$  is absolutely continuous,
- (iii) the function  $[0, 1] \ni s \mapsto d_\varphi G(t, su^t + (1-s)u^0, \psi; \eta)$  belongs to  $L^1(0, 1)$  for all  $\eta \in E$ .

For  $t \in [0, \tau]$  and  $u^t \in E(t)$ , let us introduce the set

$$Y(t, u^t, u^0) := \left\{ q \in F \mid \forall \eta \in E : \int_0^1 d_\varphi G(t, su^t + (1-s)u^0, q; \eta) \, ds = 0 \right\}, \quad (6.11)$$

which is called solution set of the *averaged adjoint equation* with respect to  $t$ ,  $u^t$  and  $u^0$ . Note that  $Y(0, u^0, u^0)$  coincides with the solution set of the usual adjoint state equation:

$$Y(0, u^0, u^0) = \{q \in F \mid d_\varphi G(0, u^0, q; \eta) = 0 \text{ for all } \eta \in E\}.$$

The following result, proved in [198], allows us to compute the Eulerian semi-derivative of Definition 6.1 without computing the material derivative  $\dot{u}$ . The key is the introduction of the set (6.11).

**Theorem 6.8.** *Let Assumption (H0) hold and the following conditions be satisfied:*

(H1) *For all  $t \in [0, \tau]$  and all  $\psi \in F$  the derivative  $\partial_t G(t, u^0, \psi)$  exists.*

(H2) *For all  $t \in [0, \tau]$ ,  $Y(t, u^t, u^0)$  is single-valued and we write  $Y(t, u^t, u^0) = \{p^t\}$ .*

(H3) *We have*

$$\lim_{t \searrow 0} \frac{G(t, u^0, p^t) - G(0, u^0, p^t)}{t} = \partial_t G(0, u^0, p^0).$$

*Then, for all  $\psi \in F$ , we obtain*

$$\frac{d}{dt}(G(t, u^t, \psi))|_{t=0} = \partial_t G(0, u^0, p^0).$$

### 6.4.3 Adjoint Equation

We introduce the Lagrangian associated to the minimization problem (6.8) for all  $\varphi, \psi \in H_0^1(D)$ ,

$$G(\Omega, \varphi, \psi) := \int_{\Omega_g} |B_r(\varphi) - B_d|^2 dx + \int_D \beta_\Omega(x, |\nabla u|^2) \nabla \varphi \cdot \nabla \psi dx - \langle F, \psi \rangle,$$

where  $F \in H^{-1}(D)$  is of the form (6.7). The adjoint state equation is obtained by differentiating  $G$  with respect to  $\varphi$  at  $\varphi = u$  and  $\psi = p$ ,

$$d_\varphi G(\Omega, u, p; \eta) = 0 \quad \text{for all } \eta \in H_0^1(D),$$

or, equivalently,

$$\begin{aligned} & 2 \int_D \partial_\zeta \beta_\Omega(x, |\nabla u|^2) (\nabla u \cdot \nabla \eta) (\nabla u \cdot \nabla p) dx + \int_D \beta_\Omega(x, |\nabla u|^2) \nabla p \cdot \nabla \eta dx \\ &= -2 \int_{\Omega_g} (B_r(u) - B_d) B_r(\eta) dx \quad \text{for all } \eta \in H_0^1(D). \end{aligned} \tag{6.12}$$

Furthermore, we introduce

$$\begin{aligned} \mathcal{A}(\nabla u) &:= \mathcal{A}_1(\nabla u) \chi_{\Omega_f} + \mathcal{A}_2(\nabla u) \chi_{D \setminus \Omega_f}, \\ \mathcal{A}_i(\nabla u) &:= \beta_i(|\nabla u|^2) I + 2\partial_\zeta \beta_i(|\nabla u|^2) \nabla u \otimes \nabla u \in \mathbb{R}^{2,2}, \quad i = 1, 2. \end{aligned}$$

Note that, with this notation, the variational form of the adjoint equation can be written as

$$\int_D \mathcal{A}(\nabla u) \nabla p \cdot \nabla \eta dx = -2 \int_{\Omega_g} (B_r(u) - B_d) B_r(\eta) dx \quad \text{for all } \eta \in H_0^1(D).$$

The existence of a unique solution  $p \in H_0^1(D)$  can be easily seen.

**Lemma 6.9.** *Let Assumption 9.3 hold. For given  $u \in H_0^1(D)$ , the problem*

$$\begin{aligned} &\text{Find } p \in H_0^1(D) \\ &\text{such that } \int_D \mathcal{A}(\nabla u) \nabla p \cdot \nabla \eta dx = -2 \int_{\Omega_g} (B_r(u) - B_d) B_r(\eta) dx \quad \forall \eta \in H_0^1(D), \end{aligned} \tag{6.13}$$

*has a unique solution  $p \in H_0^1(D)$ .*

*Proof.* The proof is analogous to the proof of Lemma 2.30 exploiting Assumption 9.3.  $\square$

## 6.5 Shape Derivative of the Cost Function

Now we give the result that the cost function  $\mathcal{J}$  given by (6.8) is shape differentiable in the sense of Definition 6.1 and provide a domain expression of the shape derivative using Theorem 6.8.

In this section we assume  $\Omega \subset \Omega^d$ ,  $V \in \mathcal{D}^1(\mathbb{R}^2, \mathbb{R}^2)$  and  $\text{supp}(V) \cap \Omega_g = \emptyset$ . Denote  $\Omega_k^d$ ,  $k = 1, \dots, 8$ , the connected components of  $\Omega^d$  (see Figure 2.2). Introduce  $\Gamma_k^d$  the boundary of  $\Omega_k^d$ . The four sides of  $\Gamma_k^d$  are denoted  $\Gamma_k^{d,N}, \Gamma_k^{d,W}, \Gamma_k^{d,E}, \Gamma_k^{d,S}$  where the exponents mean “north”, “south”, “east”, “west”, respectively. We assume  $V \cdot n = 0$  on  $\Gamma_k^{d,S}$  and  $V \cdot n \leq 0$  on  $\Gamma_k^{d,E} \cup \Gamma_k^{d,W} \cup \Gamma_k^{d,N}$ . These conditions guarantee that  $\Omega_t = T_t(\Omega) \subset \Omega^d$ . In addition, we assume that the vector field  $V$  is such that the transformation  $T_t$  satisfies  $T_t(\Omega_{mag}) = \Omega_{mag}$  for  $t$  small enough.

**Theorem 6.10.** *Let  $\beta_1$  and  $\beta_2$  satisfy Assumption 9. Then the functional  $\mathcal{J}$  is shape differentiable and its shape derivative in the direction  $V$  is given by*

$$\begin{aligned} d\mathcal{J}(\Omega; V) = & - \int_D (f_0 \operatorname{div}(V) + \nabla f_0 \cdot V) p \, dx \\ & - \int_{\Omega_{mag}} \mathbb{P}'(0) \nabla p \cdot M_1 \, dx - \int_{D \setminus \Omega_{mag}} \mathbb{P}'(0) \nabla p \cdot M_2 \, dx \\ & + \int_D \beta_\Omega(x, |\nabla u|^2) \mathbb{Q}'(0) \nabla u \cdot \nabla p \, dx \\ & - \int_D 2\partial_\zeta \beta_\Omega(x, |\nabla u|^2) (DV^\top \nabla u \cdot \nabla u) (\nabla u \cdot \nabla p) \, dx, \end{aligned} \quad (6.14)$$

where  $\mathbb{P}'(0) = (\operatorname{div} V)I - DV^\top$ ,  $\mathbb{Q}'(0) = (\operatorname{div} V)I - DV^\top - DV$ ,  $I \in \mathbb{R}^{2 \times 2}$  is the identity matrix, and  $u, p \in H_0^1(D)$  are the solutions of the problems (6.6) and (6.13), respectively.

*Proof.* Theorem 6.10 can be shown by applying Theorem 6.8. We only sketch the proof here and refer the reader to [90] for the details.

Given a vector field  $V \in \mathcal{D}^1(D, \mathbb{R}^2)$ , we consider the transformation  $T_t$  defined by (6.2) and introduce the transformed domain  $\Omega_t := T_t(\Omega)$ . The Lagrangian  $G(\Omega_t, \varphi, \psi)$  at the transformed domain  $\Omega_t$  for  $\varphi, \psi$  in  $H_0^1(D)$  is given by

$$G(\Omega_t, \varphi, \psi) = \int_{\Omega_g} |B_r(\varphi) - B_d|^2 \, dx + \int_D \beta_{\Omega_t}(x, |\nabla \varphi|^2) \nabla \varphi \cdot \nabla \psi \, dx - \langle F, \psi \rangle,$$

with  $F$  as in (6.7). We apply Theorem 6.8 to the function

$$\mathfrak{G}(t, \varphi, \psi) := G(\Omega_t, \varphi \circ T_t^{-1}, \psi \circ T_t^{-1}),$$

which, after the change of variables  $x = T_t(\hat{x})$ , reads

$$\begin{aligned} \mathfrak{G}(t, \varphi, \psi) = & \int_{\Omega_g} |B_r(\varphi) - B_d \circ T_t|^2 \xi(t) \, dx + \int_D \beta_\Omega(x, |\mathbb{M}(t) \nabla \varphi|^2) \mathbb{Q}(t) \nabla \varphi \cdot \nabla \psi \, dx \\ & - \int_D f_0 \circ T_t \psi \xi(t) \, dx - \int_{\Omega_{mag}} M_1 \cdot \mathbb{P}(t) \nabla \psi \, dx - \int_{D \setminus \Omega_{mag}} M_2 \cdot \mathbb{P}(t) \nabla \psi \, dx, \end{aligned}$$

where  $\mathbb{M}(t) = DT_t^{-\top}$ ,  $\mathbb{P}(t) = \xi(t)\mathbb{M}(t)$  and  $\mathbb{Q}(t) = \xi(t)\mathbb{M}(t)^\top \mathbb{M}(t)$ , and  $\xi(t)$  is defined in (6.10). Here, we used the assumption  $T_t(\Omega_{mag}) = \Omega_{mag}$ . It can be seen that  $\mathcal{J}(\Omega_t) = \mathfrak{G}(t, u^t, \psi)$  for all  $\psi \in H_0^1(D)$ , where  $u^t \in H_0^1(D)$  solves

$$\begin{aligned} & \int_D \beta_\Omega(x, |\mathbb{M}(t) \nabla u^t|^2) \mathbb{Q}(t) \nabla u^t \cdot \nabla \psi \, dx \\ & = \int_D f_0 \circ T_t \psi \xi(t) \, dx + \int_{\Omega_{mag}} M_1 \cdot \mathbb{P}(t) \nabla \psi \, dx + \int_{D \setminus \Omega_{mag}} M_2 \cdot \mathbb{P}(t) \nabla \psi \, dx. \end{aligned}$$

The verification of Hypotheses (H0)–(H3) of Theorem 6.8 with  $E = F = H_0^1(D)$  can be found in [90]. It follows from Theorem 6.8 that  $d\mathcal{J}(\Omega; V) = \partial_t \mathfrak{G}(0, u, p)$ , where  $u \in H_0^1(D)$  solves the state equation (6.6) and  $p \in H_0^1(D)$  is the solution of the adjoint equation (6.12). In order to compute  $\partial_t \mathfrak{G}(0, u, p)$ , we note that  $\mathbb{M}'(0) = -DV^\top$ ,  $\mathbb{P}'(0) = (\operatorname{div} V)I - DV^\top$ ,  $\mathbb{Q}'(0) = (\operatorname{div} V)I - DV^\top - DV$  and that the integrals on  $\Omega_g$  vanish since  $V = 0$  on  $\Omega_g$ .  $\square$

**Remark 6.11.** Note that the last integral in (6.14) is well-defined thanks to Assumption 9.3. To see this, note that  $V \in C^1(\overline{D}, \mathbb{R}^2)$  and that, for all  $\zeta \in \mathbb{R}^2$ , we have  $\beta'(|\zeta|^2)|\zeta|^2 \leq \Lambda$ . Hence

$$\begin{aligned} \left| \int_D 2\partial_\zeta \beta_\Omega(x, |\nabla u|^2) (DV^\top \nabla u \cdot \nabla u) (\nabla u \cdot \nabla p) dx \right| &\leq C \int_D \partial_\zeta \beta_\Omega(x, |\nabla u|^2) |\nabla u|^2 |\nabla u \cdot \nabla p| dx \\ &\leq C\Lambda \int_D |\nabla u \cdot \nabla p| dx < \infty. \end{aligned}$$

The other terms in (6.14) are obviously well-defined.

## 6.6 Application to Model Problem

In this section, we use the shape derivative derived in Theorem 6.10 for applying a gradient-based optimization algorithm to the model problem (2.17) introduced in Section 2.3. Recall that the problem consists in finding the shape  $\Omega \in \mathcal{O}$  of the ferromagnetic subdomain inside the design subdomain  $\Omega^d$  of the electric motor depicted in Figure 2.2 such that the cost functional (2.15), i.e.,

$$\mathcal{J}(\Omega) = \int_{\Omega_g} |B_r(u_\Omega) - B_d|^2 dx,$$

is minimized. In the notation of this chapter, we define

$$\begin{aligned} \beta_1(\zeta) &:= \hat{\nu}(\sqrt{\zeta}), & M_1 &:= \nu_{mag} M^\perp, & f_0 &:= J_3, \\ \beta_2(\zeta) &:= \nu_0, & M_2 &:= 0, \end{aligned} \tag{6.15}$$

where  $\zeta \in \mathbb{R}_0^+$ , such that boundary value problem (6.6) becomes the magnetostatic problem (2.14). In order to apply Theorem 6.10, we need to check Assumption 9 for  $\beta_1, \beta_2$  as chosen in (6.15). Obviously, all conditions are satisfied for  $\beta_2 = \nu_0$ . Concerning  $\beta_1$ , Assumption 9.1 follows from (2.20) and (2.21). The continuous differentiability of  $\beta_1$  as required in Assumption 9.2 follows from Lemma 2.3. Finally, we note that Assumption 9.3 for  $\beta_1$  is equivalent to

$$\exists \lambda, \Lambda > 0 : \lambda|\varphi|^2 \leq \varphi^\top DT(\rho)\varphi \leq \Lambda|\varphi|^2, \quad \forall \rho, \varphi \in \mathbb{R}^2.$$

It holds that  $\min\{\lambda_1(|\rho|), \lambda_2(|\rho|)\}|\varphi|^2 \leq \varphi^\top DT(\rho)\varphi \leq \max\{\lambda_1(|\rho|), \lambda_2(|\rho|)\}|\varphi|^2$  where  $\lambda_1(|\rho|)$ ,  $\lambda_2(|\rho|)$  denote the eigenvalues of  $DT(\rho)$ , see (2.29). Then, properties (2.19) yield Assumption 9.3 with  $\lambda = \underline{\nu}$  and  $\Lambda = \nu_0$ .

Thus we can apply Theorem 6.10 and the shape derivative reads

$$\begin{aligned} d\mathcal{J}(\Omega; V) = & - \int_D (J_3 \operatorname{div}(V) + \nabla J_3 \cdot V) p \, dx - \int_{\Omega_{mag}} \nu_{mag} \mathbb{P}'(0) \nabla p \cdot M^\perp \\ & + \int_D \nu_\Omega(x, |\nabla u|) \mathbb{Q}'(0) \nabla u \cdot \nabla p \, dx - \int_{\Omega_f} \frac{\hat{\nu}'(x, |\nabla u|)}{|\nabla u|} (DV^\top \nabla u \cdot \nabla u) (\nabla u \cdot \nabla p) \, dx \end{aligned}$$

where  $\mathbb{P}'(0) = (\operatorname{div} V)I - DV^\top$ ,  $\mathbb{Q}'(0) = (\operatorname{div} V)I - DV^\top - DV$ ,  $I \in \mathbb{R}^{2 \times 2}$  is the identity matrix, and  $u, p \in H_0^1(D)$  are the solutions of the problems (2.14) and (2.31), respectively. Note that for the model problem (2.17), there are no electric currents, thus we set  $J_3 = 0$ .

### 6.6.1 Numerical Method

Based on the shape derivative  $d\mathcal{J}(\Omega; V)$  derived in (6.14), we employ a gradient-based optimization algorithm. For that purpose, in each iteration we compute a vector field  $V$  which is a descent direction for  $\mathcal{J}$ , i.e., we compute a vector field  $V$  such that  $d\mathcal{J}(\Omega; V) < 0$ , and move the interface between the material and the air subdomain a certain distance into the direction of  $V$ .

#### 6.6.1.1 Setup of Interface

We represent the material interface by a polygon consisting of 191 points around each of the eight parts  $\Omega_k^d$  of the design subdomain  $\Omega^d$  (see Figure 6.2) and move the points of these polygons along the calculated velocity field  $V$  in the course of the optimization process. Each element of the design area whose center of gravity is inside this polygon is considered to contain ferromagnetic material, the others are considered to be air.

Note that we do not advect the mesh nodes as it is often done in shape optimization since, in our problem, it is important that some fixed parts of the motor such as the air gap or the magnets are not altered. For that reason, it is more practical to work with a representation of the interface which is decoupled from the mesh nodes.

#### 6.6.1.2 Descent Direction

In order to get a descent in the cost functional, we compute the velocity field as follows. We choose a symmetric and positive definite bilinear form

$$b : H_0^1(D_{\text{rot}}) \times H_0^1(D_{\text{rot}}) \rightarrow \mathbb{R}$$

defined on the subdomain  $D_{\text{rot}}$  of  $D$  representing the rotor and compute  $V$  as the solution of the following variational problem:

$$\text{Find } V \in P_h \text{ such that } b(V, W) = -d\mathcal{J}(\Omega; W) \text{ for all } W \in P_h. \quad (6.16)$$

Here,  $P_h \subset H_0^1(D_{\text{rot}})$  is a finite dimensional subspace. Outside  $D_{\text{rot}}$  we extend  $V$  by zero. Note that, by this choice, the condition  $V = 0$  on  $\Omega_g$ , which is assumed in Section 6.5, is satisfied. The obtained descent directions  $V \in P_h$  will also be in  $W^{1,\infty}(D, \mathbb{R}^2)$  and, consequently, they

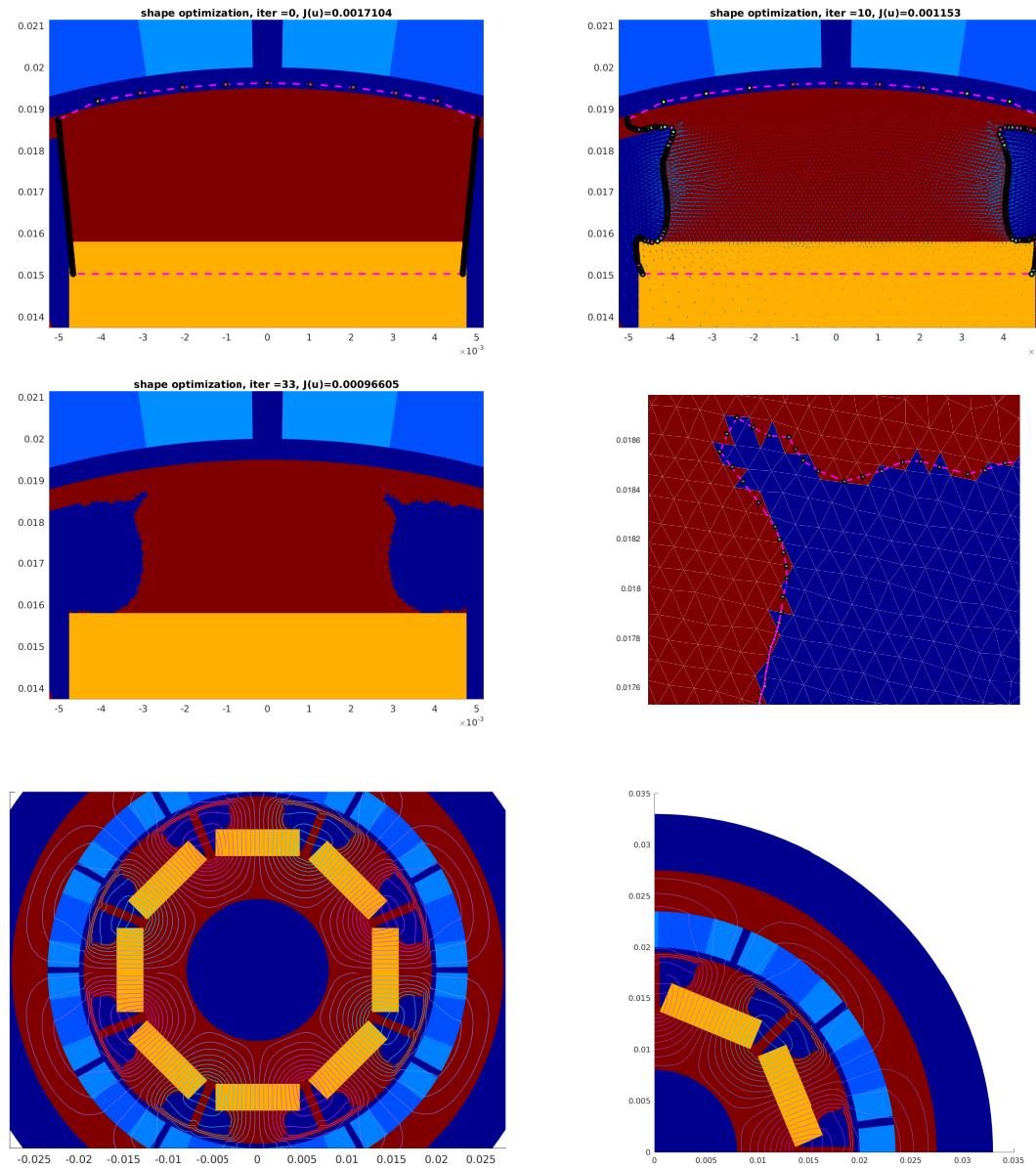


Figure 6.2: Top left: Initial design with polygon consisting of 191 points representing the material interface. Top right: Design after 10 iterations of Algorithm 3, together with vector field  $V$ . Center left: Final design after 33 iterations. Center right: Zoom of final design. Bottom: Final design after 33 iterations of Algorithm 3.

are admissible vector fields defining a flow  $T_t^V$ . The solution  $V$  computed this way is a descent direction for the cost functional since

$$d\mathcal{J}(\Omega; V) = -b(V, V) < 0.$$

For our numerical experiments, we choose the bilinear form

$$\begin{aligned} b : H_0^1(D_{\text{rot}}) \times H_0^1(D_{\text{rot}}) &\rightarrow \mathbb{R} \\ b(V, W) &= \int_{D_{\text{rot}}} (\alpha \nabla V : \nabla W + V \cdot W) \, dx. \end{aligned} \quad (6.17)$$

Here, the penalization function  $\alpha \in L^\infty(D_{\text{rot}})$  is chosen as

$$\alpha(x) = \begin{cases} 1 & x \in \Omega^d, \\ 10 & x \in \Omega^{d,\varepsilon} \setminus \Omega^d, \\ 10^2 & \text{else,} \end{cases}$$

where  $\Omega^{d,\varepsilon} = \{x \in D_{\text{rot}} : \text{dist}(x, \Omega^d) \leq \varepsilon\}$  for some small  $\varepsilon > 0$ . With this choice of  $\alpha$ , we ensure that the resulting velocity field  $V$  is small outside the design region  $\Omega^d$ .

We remark that the numerical overhead of using the volume form of the shape derivative compared to Hadamard's boundary-based form is the solution of the auxiliary boundary value problem (6.17). However, the additional numerical effort of solving this linear boundary value problem is negligible compared to the effort of solving the nonlinear state equation (2.17b). Furthermore, we solve this problem only on the reduced computational domain  $D_{\text{rot}} \subset D$ .

### 6.6.1.3 Updating the Interface

For updating the interface, we perform a backtracking line search algorithm: Once a descent direction  $V$  is computed, we move all interface points a step size  $\tau = \tau_{\text{init}}$  in the direction given by  $V$  and evaluate the cost function for the updated geometry. If the cost value has not decreased, the step size  $\tau$  is halved and the cost function is evaluated for the new, updated geometry. We repeat this step until a decrease of the cost function has been achieved. When the step size becomes too small such that no element switches its state, the algorithm is stopped.

## 6.6.2 Numerical Results

The procedure is summarized in Algorithm 3:

**Algorithm 3.** Initialization: Set  $k = 0$ , choose initial design  $\Omega_0$ , compute  $\mathcal{J}(\Omega_0)$ , set up interface, (6.16) for a descent direction  $V_k$ . Choose parameter  $\bar{d}$ .

- (i) Move interface a distance  $d_k$  in direction of  $V_k$  and set  $\Omega_{k+1}$  the updated domain, where  $d_k = \tilde{d}_k \bar{d} / \|V_k\|_{L^\infty(D_{\text{Rot}})}$  and  $\tilde{d}_k$  is chosen as  $\max\{1, 1/2, 1/4, \dots\}$  such that  $\mathcal{J}(\Omega_{k+1}) < \mathcal{J}(\Omega_k)$
- (ii) If no decrease could be achieved: Stop
- (iii) Compute descent direction  $V_{k+1}$  using (6.16) on  $\Omega_{k+1}$ , set  $k \leftarrow k + 1$  and go to (i)

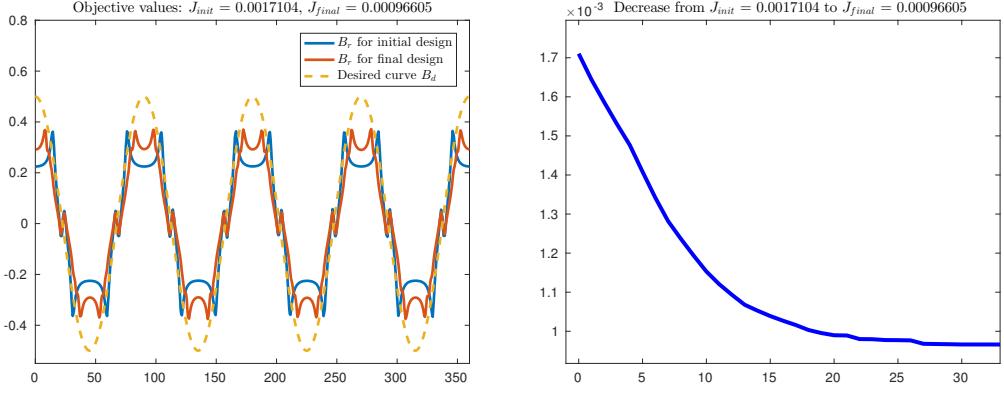


Figure 6.3: Left: Radial component of magnetic flux density along the air gap for initial and final design compared with desired curve. Right: Objective function during optimization process.

Note that for each evaluation of  $\mathcal{J}$ , the state equation (6.6) has to be solved, and for each computation of a descent direction  $V$  from (6.16), the state  $u$  and adjoint state  $p$ , which is the solution to (6.12), with the current shape  $\Omega$  are required. Here, the parameter  $\bar{d}$  must be set by the user and determines the size of the steps taken during the algorithm. In our experiments, we set  $\bar{d}$  to three times the minimal edge length of the mesh.

The final design after 33 iterations of Algorithm 3 is depicted in Figure 6.2. The cost function is reduced from  $1.7104 \times 10^{-3}$  to  $0.96605 \times 10^{-3}$ , see Figure 6.3 where also the radial component of the magnetic flux density for the initial and final design are shown. Figure 6.2 also shows the final design together with the magnetic field. Note that the result obtained by the shape optimization algorithm exhibits some symmetry even though we did not enforce it. From the right picture in the center of Figure 6.2, it can be seen that, due to the interface handling we chose, a jagged interface occurs. This issue serves as a motivation for introducing the interface finite element method of Chapter 7.



## **Part III**

# **Interface Handling**



## Chapter 7

# A Local Mesh Modification Strategy for Interface Problems

In the numerical experiments of Sections 3.3, 4.8 and 6.6, we were confronted with material interfaces that evolved in the course of the optimization procedure. In order to evaluate the corresponding sensitivities in the next iteration of the optimization algorithm, the state equation (2.14) and the adjoint equation (2.31) have to be solved on the updated geometry. This raises the question of how to resolve a marching material interface in each iteration of the given algorithm. In the experiments presented in Sections 3.3, 4.8 and 6.6, we neglected this aspect and decided for each element of the underlying triangulation whether it should contain ferromagnetic material or air by just looking at the position of its centroid. This approach does not only lead to an irregular, jagged interface as can be seen from Figure 6.2, but also yields a loss of accuracy as the mesh parameter  $h$  approaches zero. It is well-known that, when using standard finite element methods, the interface must be resolved by the mesh in order to obtain optimal convergence rates of the approximate solution  $u_h$  to the true solution  $u$  in the  $L^2(D)$  and  $H^1(D)$  norms, see also [29].

A straightforward approach to deal with this problem would be to create a new triangulation which resolves the updated geometry in each iteration of the algorithm. Of course, proceeding like this is computationally very expensive and should be avoided. Another approach which is very commonly used in shape optimization is to not only move the interface into the direction prescribed by the vector field  $V$  obtained from (6.16), but to advect all nodes of the finite element mesh in this direction, see e.g. [135, 187, 199, 200]. The vector field  $V$  is usually smooth enough such that self-intersection of the mesh does typically not occur. However, the approach can become problematic when more complex geometries with geometric constraints are involved, as it is the case for our model problem. Here, fixed parts of the electric motor like the circular air gap should not be altered under any circumstances. Although this could be dealt with by specifying problem (6.16) on a suitable subdomain with appropriate boundary conditions, it may be more convenient to have a description of the interface that is decoupled from the mesh. Another shortcoming of this kind of interface treatment is that no topological changes can be performed.

Besides the approaches mentioned above, there exist several other, more sophisticated interface methods that are well-studied. A brief overview over existing methods to deal with moving and non-matching interfaces, such as the extended finite element method (XFEM), the immersed FEM or the unfitted Nitsche method, was given in Section 1.3.

Another approach to deal with interfaces in the context of finite elements is presented in [82, 83]. This method is based on a patch hierarchy of a quadrilateral mesh and modifies the basis functions supported on patches which are cut by the interface. The bilinear basis functions on the four quadrilaterals of an intersected patch are replaced by piecewise linear functions on eight triangles in such a way that the interface is resolved more accurately. Therefore, the method can also be interpreted as a fitted finite element method on a mixed quadrilateral-triangular grid. However, the modification of the basis is done in an implicit parametric way without actually modifying the nodes of the mesh. One advantage of this kind of method over the ones mentioned above is that this method has a fixed number of unknowns independently of the position of the interface relative to the mesh. The given mesh is modified only locally near the material interface and the method is relatively easy to implement. Of course, this kind of interface finite element method is not restricted to applications from shape and topology optimization, but is applicable to a wide variety of physical problems involving material interfaces such as multiphase flows or even multiphysics problems like fluid-structure interaction.

The method we present in this chapter is an adaptation of the method introduced in [82, 83] to the case of triangular meshes. We start out from a triangulation with a hierarchical structure and modify some mesh nodes near the interface in such a way that, on the one hand, the interface is captured, and on the other hand, a maximum angle condition is satisfied. Similarly to [82], we show that this maximum angle condition yields the standard discretization error estimates in the  $L^2(D)$  and  $H^1(D)$  norms. Furthermore, we consider the condition number of the stiffness matrix and investigate requirements on the finite element basis which would guarantee optimal conditioning with respect to the discretization parameter  $h$ .

Finally, we mention that a very similar approach to ours is currently also being investigated independently in [112]. There, the modification of the finite element space is done in a parametric way without actually moving the mesh nodes. The authors also show a maximum angle condition and optimal order of convergence, but do not treat the issue of the condition number of the stiffness matrix with respect to small angles.

This chapter is an extension of the results presented in [89].

## 7.1 Preliminaries

We introduce the method for the linear potential equation in a bounded, polygonal computational domain  $D \subset \mathbb{R}^2$  consisting of two non-overlapping, open subdomains,  $\overline{D} = \overline{\Omega}_1 \cup \overline{\Omega}_2$ ,  $\Omega_1 \cap \Omega_2 = \emptyset$ , which represent two materials with different material coefficients  $\kappa_1, \kappa_2 > 0$ . On the material interface  $\Gamma := \overline{\Omega}_1 \cap \overline{\Omega}_2$ , we have to require that the solution as well as the normal flux are continuous. For simplicity, we assume homogeneous Dirichlet boundary conditions on  $\partial D$ . The problem reads as follows:

$$\begin{aligned} -\operatorname{div}(\kappa_i \nabla u) &= f && \text{in } \Omega_i, \quad i = 1, 2, \\ [[u]] &= 0 && \text{on } \Gamma, \\ \left[ \kappa \frac{\partial u}{\partial n} \right] &= 0 && \text{on } \Gamma, \\ u &= 0 && \text{on } \partial D, \end{aligned} \tag{7.1}$$

where, again,  $\llbracket \cdot \rrbracket$  denotes the jump of a function across the interface  $\Gamma$  and the material coefficient function  $\kappa$  given by

$$\kappa(x) = \begin{cases} \kappa_1, & x \in \Omega_1, \\ \kappa_2, & x \in \Omega_2. \end{cases}$$

The weak form of problem (7.1) reads

$$\text{Find } u \in H_0^1(D) \text{ such that } a(u, v) := \int_D \kappa \nabla u \cdot \nabla v \, dx = \int_D f v \, dx \quad \forall v \in H_0^1(D). \quad (7.2)$$

We assume that the boundaries of the two subdomains as well as the right hand side  $f$  are sufficiently regular such that

$$u \in H_0^1(D) \cap H^2(\Omega_1 \cup \Omega_2),$$

i.e., the restrictions of  $u \in H_0^1(D)$  to  $\Omega_1$  and  $\Omega_2$  belong to  $H^2(\Omega_1)$  and  $H^2(\Omega_2)$ , respectively, see e.g. [29].

Let  $\{\mathcal{T}_h\}$  be a regular family of subdivisions of  $D$  into triangular elements which is *shape-regular*, i.e., there exists a number  $\sigma_0 > 0$  such that, for all elements  $T$ , the ratio between the diameter  $h_T$  of  $T$  and the radius  $\rho_T$  of the largest inscribed circle in  $T$  is bounded from below,  $\rho_T/h_T \geq \sigma_0$ . This is equivalent to the condition that there exists  $\theta_0 > 0$  such that, for all elements  $T$ , all interior angles are bounded from below by  $\theta_0$ , see e.g. [47, 221]. Let now the mesh size  $h$  be fixed. We assume that  $\mathcal{T}_h$  has been obtained by one uniform refinement of a coarser mesh  $\mathcal{T}_{2h}$ . By this assumption, also  $\mathcal{T}_{2h}$  is shape-regular, and  $\mathcal{T}_h$  has a patch-hierarchy, i.e., always four elements  $T_1, T_2, T_3, T_4 \in \mathcal{T}_h$  can be combined to one larger triangle  $T \in \mathcal{T}_{2h}$ . We will refer to this larger element as the makro element or patch.

We need to make an assumption on the makro mesh  $\mathcal{T}_{2h}$ :

**Assumption 11.** We assume that the mesh of makro elements  $\mathcal{T}_{2h}$  is such that, for each makro element  $T \in \mathcal{T}_{2h}$ , the interface  $\Gamma$  either

1. does not intersect the interior of  $T$ , or
2. intersects  $\partial T$  in exactly two distinct edges, or
3. intersects  $\partial T$  in one vertex and in the opposite edge.

This assumption basically just excludes the case where one edge of the makro element is cut by the interface more than once. For a smooth interface  $\Gamma$ , this assumption can always be enforced by choosing a fine enough makro mesh  $\mathcal{T}_{2h}$ . We consider a makro element  $T \in \mathcal{T}_{2h}$  to be cut by the interface if the intersection of the interior of the makro element with the interface is not the empty set.

## 7.2 Description of the Method

The method presented in this chapter is a local mesh adaptation strategy, meaning that only makro elements close to the interface  $\Gamma$  are modified. Given the hierachic structure of the

mesh, on every makro element we have four elements of the mesh  $\mathcal{T}_h$  and six vertices, see Figure 7.1(a),(b). The idea of the method is the following: For each makro element that is cut by the interface, move the points which lie on the edges of the makro element along the corresponding edge. This is done in such a way that, on the one hand, the interface is resolved accurately, and, on the other hand, all interior angles in the four triangles are bounded away from  $180^\circ$ . For a makro element  $T$  that is cut by the interface, we distinguish four different configurations as follows:

In the case where the makro element is cut by the interface in two distinct edges, we denote the vertex of the makro element where these two edges meet by  $P_1$ , and the other two vertices in counter-clockwise order by  $P_2$  and  $P_3$ . The parameters  $s, t, r \in [0, 1]$  represent the positions of the points  $P_4, P_5, P_6$  along the corresponding edges by

$$P_4(s) = P_1 + s \frac{P_2 - P_1}{|P_2 - P_1|}, \quad P_5(t) = P_2 + t \frac{P_3 - P_2}{|P_3 - P_2|}, \quad P_6(r) = P_1 + r \frac{P_3 - P_1}{|P_3 - P_1|},$$

respectively. The parameters  $r$  and  $s$  will always be chosen in such a way that the intersection points of the interface and the edges  $P_1P_3$  and  $P_1P_2$  are the points  $P_6$  and  $P_4$ , respectively. Thus, we identify the position of the interface relative to the makro element  $T$  by the two parameters  $r, s$ . We choose the parameter  $t$  such that a maximum angle condition is satisfied as follows:

**Configuration A:**  $0 < r, s \leq 1/2$ . Set  $t = 1/2$ .

**Configuration B:**  $1/2 < r, s < 1$ . Set  $t = 1 - s$ .

**Configuration C:**  $0 < s \leq 1/2 < r < 1$  (*Configuration C1*) or  $0 < r \leq 1/2 < s < 1$  (*Configuration C2*). Set  $t = 1/2$ .

The case where the makro element is cut in one vertex and the opposite edge has to be considered separately. We denote the vertex of the makro element where it is cut by the interface by  $P_2$  and the other vertices, in counter-clockwise ordering, by  $P_3$  and  $P_1$ , see Figure 7.1(b). The location of the interface is given by the position of the point  $P_6$  on the edge between  $P_3$  and  $P_1$ . In this case, we also need to rearrange the triangles  $T_2$  and  $T_4$ .

**Configuration D:**

*Configuration D1:*  $0 < r \leq 1/2$ . Set  $s = r$  and  $t = 1/2$ .

*Configuration D2:*  $1/2 < r < 1$ . Set  $s = 1/2$  and  $t = r$ .

With this setting, it is possible to show the required maximum angle condition on the reference patch  $\hat{T}$  defined by the outer makro vertices  $\hat{P}_1 = (0, 0)^\top$ ,  $\hat{P}_2 = (1, 0)^\top$ ,  $\hat{P}_3 = (1/2, \sqrt{3}/2)^\top$ .

**Remark 7.1.** *The choice of  $t = 1 - s$  in Configuration B can without any problems be replaced by the choice  $t = r$ . This can be exploited if the position of  $P_5$  affects the neighboring patch in an unwanted way, see Figure 7.2. We refer to this choice as Configuration B'.*

### 7.3 Maximum Angle Condition

The convergence behavior of a finite element approximation  $u_h$  to the true solution  $u$  of problem (7.1) is usually shown by Céa's lemma and an interpolation error estimate. A condition that is sufficient and necessary for such an interpolation error estimate is that all interior angles of triangles in the family of meshes  $\{\mathcal{T}_h\}_{h>0}$  are bounded away from  $180^\circ$ , see [30].

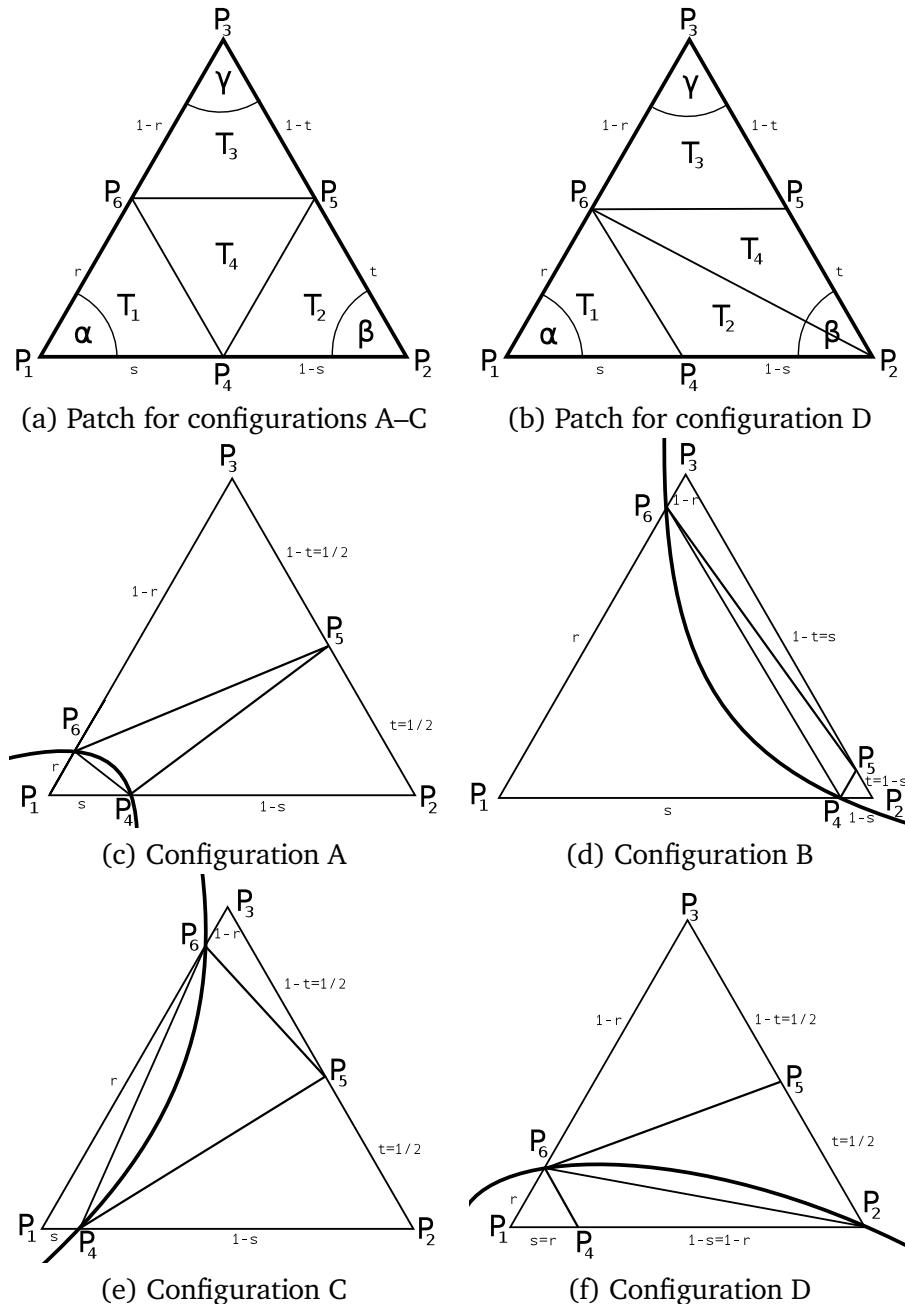


Figure 7.1: (a), (b): Patches for different configurations. (c)-(f): Different configurations of mesh points depending on position of the interface.

Note that this maximum angle condition is not necessary for optimal convergence rates. In fact, it may happen that the interpolation error is large, while the discretization error is very small, see [98] for an example. We show that the mesh modification strategy introduced in Section 7.2 does not violate the maximum angle condition.

We start with a maximum angle condition on the reference patch  $\hat{T}$ .

**Lemma 7.2.** *Let the reference patch  $\hat{T}$  be intersected in two distinct edges or in one vertex and the opposite edge and apply the mesh modification strategy described in Section 7.2. Then, all interior angles in triangles of the reference patch  $\hat{T}$  are bounded by  $150^\circ$ , independent of the parameters  $r, s \in [0, 1]$  representing the location of the interface.*

*Proof.* We have to ensure for each of the four subtriangles  $\hat{T}_1, \hat{T}_2, \hat{T}_3, \hat{T}_4$  that all of their three interior angles are not larger than  $150^\circ$ .

For three points  $P, Q, R$  in  $\mathbb{R}^2$ , define

$$\angle(P, Q, R) := \cos^{-1} \left( \frac{(P - Q, R - Q)}{|P - Q| |R - Q|} \right)$$

the interior angle of the triangle with vertices  $P, Q, R$  at point  $Q$ .

In Configuration A–C, the sub-triangles  $\hat{T}_1, \hat{T}_2$  and  $\hat{T}_3$  all have one angle of  $60^\circ$ . Obviously, the remaining two angles are bounded from above by  $120^\circ$ . It remains to check the angles in sub-triangle  $\hat{T}_4$ :

*Configuration A:* For  $r, s \in (0, 1/2]$ , we get for the angle in point  $P_4$  that

$$\angle(P_6, P_4, P_5) < \angle(P_1, P_4, P_5) = 180^\circ - \angle(P_5, P_4, P_2) \leq 180^\circ - \angle(P_5, P_1, P_2).$$

Since the reference patch  $\hat{T}$  is equilateral, it holds  $\angle(P_5, P_1, P_2) = \alpha/2$ . Analogously, we get for the angle in point  $P_6$  that  $\angle(P_5, P_6, P_4) < 180^\circ - \alpha/2$ . It is easy to see that the angle in point  $P_5$  increases with  $r, s$  and thus is maximized for  $r = s = 1/2$ , which yields that  $\angle(P_4, P_5, P_6) \leq \angle(P_4(1/2), P_5, P_6(1/2)) = 180^\circ - \beta - \gamma = \alpha$ . Here we used that, for  $r = s = t = 1/2$ , the four sub-triangles are congruent.

*Configuration B:* Note that, by the special choice of  $s, t$ , in this case we have that the line going through  $P_4$  and  $P_5$  is parallel to the edge connecting  $P_1$  and  $P_3$  for all values of  $s \in (1/2, 1)$ . Thus, we have

$$\begin{aligned} \angle(P_4, P_5, P_6) &\leq \angle(P_4, P_5, P_3) = 180^\circ - \gamma, \quad \text{and} \\ \angle(P_4, P_5, P_6) &= 180^\circ - \gamma - \angle(P_6, P_5, P_3) \\ &\geq 180^\circ - \gamma - \angle(P_6(1/2), P_5(1/2), P_3) = 180^\circ - \gamma - \beta = \alpha. \end{aligned}$$

The angles in  $P_4$  and in  $P_6$  must also be bounded from above by  $180^\circ - \alpha = 120^\circ$ .

*Configuration C:* We consider the case where  $r \in (1/2, 1)$  and  $s \in (0, 1/2]$ . The reverse case is treated analogously. For the angle in the fixed point  $P_5 = P_5(1/2) = (P_2 + P_3)/2$ , we get the estimates

$$\begin{aligned} \angle(P_4, P_5, P_6) &\leq \angle(P_4, P_5, P_3) \leq \angle((P_4(1/2), P_5, P_3) = 180^\circ - \gamma, \\ \angle(P_4, P_5, P_6) &\geq \angle(P_4, P_5, P_6(1/2)) \geq \angle(P_1, P_5, P_6(1/2)) = \angle(P_5, P_1, P_2) = \alpha/2. \end{aligned}$$

Thus, the angles  $\angle(P_6, P_4, P_5)$  and  $\angle(P_5, P_6, P_4)$  are also bounded from above by  $180^\circ - \beta/2$ .

*Configuration D:* We consider only Configuration D1, the corresponding result for Configuration D2 follows analogously. In Configuration D, the sub-triangles  $\hat{T}_1$  and  $\hat{T}_3$  both have one angle of  $60^\circ$  such that the remaining two angles of these triangles are bounded by  $120^\circ$ . We need to consider triangles  $\hat{T}_2$  and  $\hat{T}_4$ . Due to the choice of the parameter  $s$ , the line going through  $P_4$  and  $P_6$  is parallel to the edge connecting  $P_2$  and  $P_3$  for all values of  $r$ . In  $\hat{T}_2$ ,  $\angle(P_6, P_4, P_2) = 180^\circ - \beta$  and, therefore, the other two angles are bounded by  $\beta$ . In  $\hat{T}_4$ , we have for  $r \in (0, 1/2]$  that

$$\begin{aligned}\angle(P_6, P_2, P_5) &\leq \beta, \\ \angle(P_2, P_5, P_6) &\leq \angle(P_2, P_5, (P_3 + P_1)/2) = 180 - \beta, \\ \angle(P_5, P_6, P_2) &\leq \angle(P_3, P_6, P_4) = 180 - \gamma.\end{aligned}$$

Finally, noting that  $\alpha = \beta = \gamma = 60^\circ$  yields the statement of the lemma.  $\square$

Lemma 7.2 ensures that, for any smooth interface cutting the reference patch  $\hat{T}$ , there exists a subdivision of  $\hat{T}$  into  $\hat{T}_1, \hat{T}_2, \hat{T}_3, \hat{T}_4$  such that all interior angles are bounded away from  $180^\circ$ . However, it is important to note that in Configurations B and D, nodes which do not lie on the interface  $\Gamma$  are moved. This may cause the nodes of the patch adjacent to this non-intersected makro edge to move in an undesired way. In order to avoid troubles in this respect, we need to make another assumption on our makro mesh.

**Assumption 12.** If a makro element  $T \in \mathcal{T}_{2h}$  is cut by the interface then all patches adjacent to non-intersected edges of  $T$  are not cut by the interface, nor are they modified by any other neighboring patch other than  $T$ .

For an illustration of which cases are excluded by assumption 12, see Figure 7.2. Again, this assumption can be ensured by choosing a sufficiently fine makro mesh. Under this additional assumption we get the maximum angle condition on the whole modified mesh.

**Corollary 7.3.** *Let the makro mesh  $\mathcal{T}_{2h}$  be shape-regular and let Assumptions 11 and 12 hold. Then, after applying the mesh modification strategy described in Section 7.2, all interior angles in the modified mesh are bounded away from  $180^\circ$ .*

*Proof.* The maximum angle condition of Lemma 7.2 for the reference patch  $\hat{T}$  together with the shape-regularity of the makro mesh  $\mathcal{T}_{2h}$  yield that also the subtriangles of all physical patches  $T \in \mathcal{T}_{2h}$  which are cut by the interface satisfy a maximum angle condition. Due to Assumption 12, those patches which are not cut by the interface but whose nodes are modified due to the adjustment of one of the neighboring makro elements can be treated in the same way as in Configuration C in the proof of Lemma 7.2. Again, the maximum angle condition on these physical elements follows from the shape-regularity of the makro mesh. Finally, the maximum angle condition trivially holds for all patches that are not affected by the interface.  $\square$

**Remark 7.4.** *We mention that a violation of Assumption 12 does not automatically result in a failure of the method described in Section 7.2. In fact, e.g., the situation illustrated in the left picture of Figure 7.2 could be resolved by applying Configuration B' to the right patch, i.e., by putting the node on the common edge towards the bottom of the edge.*

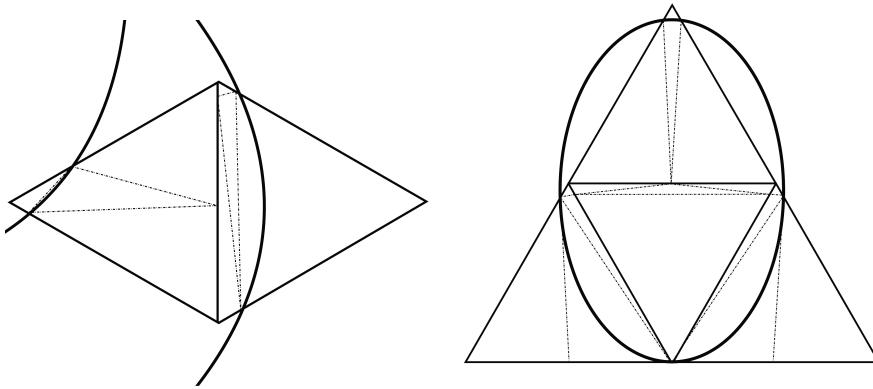


Figure 7.2: Examples of constellations that are excluded by Assumption 12. Left: Two adjacent patches are cut by the interface. Right: central patch is not cut by the interface, but affected, since lower left and lower right patch are in Configuration D.

From now on, let  $\mathcal{T}_h$  denote the locally modified mesh and  $V_h$  the corresponding space of globally continuous, piecewise linear basis functions on  $\mathcal{T}_h$ .

## 7.4 A Priori Error Estimates

This section follows the lines of [82] where the analogous results are shown for the case of a quadrilateral mesh. The proofs of Lemmas 7.6 and 7.8 as well as Theorem 7.9 are analogous to [82] and are given here for completeness of the presentation.

Now we are in the position to show an a priori error estimate for the finite element solution  $u_h$ . Since we have the maximum angle condition of Corollary 7.3, we get the interpolation error estimates for smooth functions  $v \in H^2(T) \cap C(\bar{T})$  on the triangle  $T \in \mathcal{T}_h$ ,

$$\|\nabla^k(v - I_h v)\|_{L^2(T)} \leq C h_{T,max}^{2-k} \|\nabla^2 v\|_T, \quad k = 0, 1, \quad (7.3)$$

where  $I_h : H^2(T) \rightarrow V_h|_{\bar{T}}$  denotes the Lagrangian interpolation operator,  $C$  is a positive generic constant, and  $h_{T,max}$  is the maximum edge length of  $T$ , see e.g. [25]. In the case where the interface  $\Gamma$  is not polygonal but smooth with  $C^2$  parametrization, and an element of the modified mesh  $\mathcal{T}_h$  is intersected by  $\Gamma$ , the solution  $u$  is not smooth across the interface and, hence, estimate (7.3) cannot be applied.

In Lemma 7.8 we will provide the corresponding interpolation error estimate for  $k = 1$  in the case of a  $C^2$ -smooth interface  $\Gamma$ , which will allow us to show the discretization error estimate of Theorem 7.9. By the procedure described in Section 7.2, we approximate the smooth interface  $\Gamma$  by a polygonal curve which we denote by  $\Gamma_h$ . This polygonal curve subdivides the computational domain  $D$  into the subdomains  $\Omega_h^1$  and  $\Omega_h^2$ . For  $i = 1, 2$ , let  $\mathcal{T}_h^i$  be the set of triangles in  $\Omega_h^i$ , i.e.,  $\mathcal{T}_h^i = \{T \in \mathcal{T}_h : T \subset \Omega_h^i\}$ . Then it holds

$$\overline{\Omega_h^i} = \bigcup_{T \in \mathcal{T}_h^i} \overline{T} \quad \text{and} \quad \overline{\Omega_h} = \bigcup_{T \in \mathcal{T}_h} \overline{T} = \overline{\Omega_h^1} \cup \overline{\Omega_h^2}.$$

In the case where the interface  $\Gamma$  is not resolved exactly by  $\Gamma_h$ , the difficulty arises in those triangles  $T$  whose interior is cut by the interface. Then the solution is not in  $H^2$  on these triangles. We will denote the union of such elements by  $S_h$ ,

$$\overline{S_h} = \bigcup_{T \in \mathcal{T}_h : \text{int}(T) \cap \Gamma \neq \emptyset} \overline{T},$$

and the set of such triangles on either side of the discrete interface by  $S_h^i$ ,

$$\overline{S_h^i} = \bigcup_{T \in \mathcal{T}_h^i : \text{int}(T) \cap \Gamma \neq \emptyset} \overline{T},$$

for  $i = 1, 2$ . Then it holds  $\overline{S_h} = \overline{S_h^1} \cup \overline{S_h^2}$ . For an illustration of this notation, see Figure 7.4. Recall the piecewise constant material coefficient function  $\kappa$ , and let  $\kappa_h$  be the piecewise constant discrete material coefficient function, i.e.,

$$\kappa(x) = \begin{cases} \kappa_1, & x \in \Omega_1, \\ \kappa_2, & x \in \Omega_2, \end{cases} \quad \kappa_h(x) = \begin{cases} \kappa_1, & x \in \Omega_h^1, \\ \kappa_2, & x \in \Omega_h^2. \end{cases}$$

The discrete approximation of the weak formulation (7.2) reads

$$\text{Find } u_h \in V_h \text{ such that } a_h(u_h, v_h) := \int_D \kappa_h \nabla u_h \cdot \nabla v_h \, dx = \int_D f v_h \, dx \quad \forall v_h \in V_h. \quad (7.4)$$

Noting that  $V_h \subset H_0^1(D)$  and that the right hand sides of problems (7.2) and (7.4) coincide for test functions  $v_h \in V_h$ , taking the difference of these two variational equations yields the Galerkin orthogonality relation

$$a(u, v_h) - a_h(u_h, v_h) = 0 \quad \forall v_h \in V_h, \quad (7.5)$$

for  $u, u_h$  the solutions to problems (7.2) and (7.4), respectively. Furthermore, we define  $S$  to be the area between the discrete and the continuous interface, i.e., the area where the discrete and the continuous bilinear forms differ, and  $S_T$  the intersection of  $S$  with a triangle  $T \in \mathcal{T}_h$ ,

$$S = \bigcup_{i=1}^2 (\Omega_h^i \setminus \overline{\Omega}_i), \quad \text{and} \quad S_T = S \cap T. \quad (7.6)$$

Moreover, for  $T \subset S_h$ , we introduce the parts of the discrete and continuous interface inside  $\overline{T}$  as

$$\Gamma_{h,T} = \Gamma_h \cap \overline{T} \quad \text{and} \quad \Gamma_T = \Gamma \cap \overline{T}.$$

As a first step for deriving the interpolation error estimate, we quantify the distance between the smooth interface  $\Gamma$  with a  $C^2$  parametrization and its discrete approximation  $\Gamma_h$ . For that purpose, we need the following lemma.

**Lemma 7.5.** *Let  $h > 0$  and  $f \in C^2(0, h)$  with  $f(0) = f(h) = 0$ . Then, for all  $x \in (0, h)$ , there is a constant  $C > 0$  such that*

$$f(x) \leq C h^2.$$

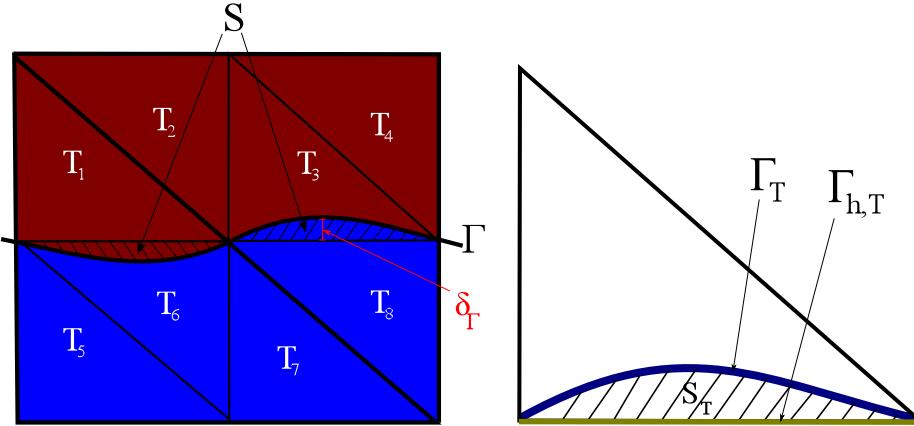


Figure 7.3: Illustration of notation: Domain is divided into  $\Omega_1$  (brown area) and  $\Omega_2$  (blue);  $\mathcal{T}_h^1 = \{T_1, T_2, T_3, T_4\}$ ,  $\mathcal{T}_h^2 = \{T_5, T_6, T_7, T_8\}$ ,  $\Omega_h^i$  is union of triangles in  $\mathcal{T}_h^i$ ,  $i = 1, 2$ ;  $S_h^1 = T_3$ ,  $S_h^2 = T_6$ ,  $S_h = T_3 \cup T_6$ ;  $\Gamma_T$  denotes the intersection of an element  $T$  with the interface  $\Gamma$  and  $\Gamma_{h,T}$  its polygonal approximation.

*Proof.* Let  $x \in (0, h)$ . A second order Taylor expansion around points  $x_0 = 0$  and  $x_1 = h$  yields that there exist number  $\xi_0, \xi_1 \in (0, h)$  such that

$$f(x) = f(0) + x f'(0) + \frac{1}{2} x^2 f''(\xi_0), \quad (7.7)$$

$$f(x) = f(h) + (x - h) f'(h) + \frac{1}{2} (x - h)^2 f''(\xi_1). \quad (7.8)$$

Since  $x \in (0, h)$  there exists  $\eta \in (0, 1)$  such that  $x = \eta h$  and thus  $h - x = (1 - \eta) h$ . Multiplying equation (7.7) by the factor  $\min\{\eta, 1 - \eta\}/\eta$  and equation (7.8) by  $\min\{\eta, 1 - \eta\}/(1 - \eta)$ , and summing up gives

$$\begin{aligned} \frac{1}{\max\{\eta, 1 - \eta\}} f(x) &= \frac{\min\{\eta, 1 - \eta\}}{\eta} \eta h f'(0) + \frac{\min\{\eta, 1 - \eta\}}{1 - \eta} ((1 - \eta) h) f'(h) \\ &\quad + \frac{\min\{\eta, 1 - \eta\}}{\eta} \frac{1}{2} \eta^2 h^2 f''(\xi_0) + \frac{\min\{\eta, 1 - \eta\}}{1 - \eta} \frac{1}{2} (1 - \eta)^2 h^2 f''(\xi_1) \\ &= \min\{\eta, 1 - \eta\} h^2 \frac{f'(0) - f'(h)}{h} \\ &\quad + \frac{\eta \min\{\eta, 1 - \eta\}}{2} h^2 f''(\xi_0) + \frac{(1 - \eta) \min\{\eta, 1 - \eta\}}{2} h^2 f''(\xi_1) \end{aligned}$$

where we used that  $f(0) = f(h) = 0$ . By the mean value theorem, there exists  $\xi_2 \in (0, h)$  such that  $(f'(0) - f'(h))/h = -f''(\xi_2)$ . Thus, we get

$$f(x) \leq \frac{3}{2} \max\{\eta, 1 - \eta\} \min\{\eta, 1 - \eta\} \max_{\xi \in (0, h)} |f''(\xi)| h^2,$$

which yields the statement of Lemma 7.5 with  $C = \frac{3}{8} \max_{\xi \in (0, h)} |f''(\xi)|$ .  $\square$

It follows directly from Lemma 7.5 that for the distance  $\delta_\Gamma$  between  $\Gamma$  and  $\Gamma_h$  it holds  $\delta_\Gamma = \mathcal{O}(h^2)$ . Using this observation, we get the following lemma on the  $L^2$  and  $H^1$  norms in  $S$ .

**Lemma 7.6** ([82]). *Let  $\Gamma$  be a smooth interface with  $C^2$ -parametrization, and let  $S$  be the region between the discrete and the continuous interface, see (7.6). Then, for a function  $\phi \in H^1(\Omega_1 \cup \Omega_2)$ , it holds that*

$$\|\phi\|_{L^2(S)} \leq C h \|\phi\|_{H^1(\Omega_1 \cup \Omega_2)}. \quad (7.9)$$

Moreover, for a discrete function  $\phi_h \in V_h$ , it holds

$$\|\nabla \phi_h\|_{L^2(S)} \leq C h^{1/2} \|\nabla \phi_h\|_{L^2(\Omega_1 \cup \Omega_2)}. \quad (7.10)$$

*Proof.* This proof is following the lines of [82].

Let  $T$  be an element of the mesh  $\mathcal{T}_h$  that is cut by the interface, i.e.,  $T \subset S_h$ . It holds that  $|T| = \mathcal{O}(h^2)$  and it follows from Lemma 7.5 that  $\delta_\Gamma = \mathcal{O}(h^2)$  and thus  $|S_T| = \mathcal{O}(h^3)$ . For a piecewise linear function  $\phi_h \in V_h$ , the gradient  $\nabla \phi_h$  is constant on  $T$ . Thus, a scaling argument yields that

$$\|\nabla \phi_h\|_{L^2(S_T)}^2 \leq C h \|\nabla \phi_h\|_{L^2(T)}^2,$$

which yields (7.10) after summation over all triangles  $T \subset S_h$ . For showing (7.9), we use that, due to  $\delta_\Gamma = \mathcal{O}(h^2)$ , we have the Poincaré-like estimate

$$\|\phi\|_{L^2(S_T)}^2 \leq C \left( h^2 \|\phi\|_{L^2(\Gamma_T)}^2 + h^4 \|\nabla \phi\|_{L^2(S_T)}^2 \right),$$

see e.g. [48], where the trace  $\Gamma_T$  is seen from  $S_T$ . Summing over all triangles in  $S_h$  gives

$$\|\phi\|_{L^2(S)}^2 \leq C \left( h^2 \|\phi\|_{L^2(\Gamma)}^2 + h^4 \|\nabla \phi\|_{L^2(S)}^2 \right),$$

which yields (7.9) by the trace inequality  $\|\phi\|_{L^2(\Gamma)} \leq C \|\phi\|_{H^1(\Omega_1 \cup \Omega_2)}$ .  $\square$

By means of Lemma 7.6, we can estimate the difference between the discrete and the continuous bilinear form, since it holds

$$\begin{aligned} |a(\phi, \psi) - a_h(\phi, \psi)| &= \left| \int_D (\kappa - \kappa_h) \nabla \phi \cdot \nabla \psi \, dx \right| = \left| \int_S (\kappa - \kappa_h) \nabla \phi \cdot \nabla \psi \, dx \right| \\ &\leq |\kappa_1 - \kappa_2| \|\nabla \phi\|_{L^2(S)} \|\nabla \psi\|_{L^2(S)} \end{aligned}$$

Lemma 7.6 yields the following result:

**Corollary 7.7** ([82]). *Let  $\Gamma$  be a smooth interface with  $C^2$  parametrization. Then the following estimates holds for the difference between the continuous and the discrete bilinear forms defined by (7.2) and (7.4), respectively: For all  $\phi_h, \psi_h \in V_h$ , it holds*

$$|a(\phi_h, \psi_h) - a_h(\phi_h, \psi_h)| \leq C h \|\nabla \phi_h\|_{L^2(D)} \|\nabla \psi_h\|_{L^2(D)},$$

and for all  $\phi \in H^{k+1}(\Omega_1 \cup \Omega_2), \psi \in H^{l+1}(\Omega_1 \cup \Omega_2)$ , we have that

$$|a(\phi, \psi) - a_h(\phi, \psi)| \leq C h^{k+l} \|\nabla \phi\|_{H^k(\Omega_1 \cup \Omega_2)} \|\nabla \psi\|_{H^l(\Omega_1 \cup \Omega_2)}.$$

Now we can show the interpolation error estimate for the case of a smooth interface that is not resolved by the triangulation. Recall that  $I_h$  denotes the nodal Lagrange interpolator which is piecewise linear on  $\Omega_h$  and satisfies  $I_h v(x_i) = v(x_i)$  for all mesh nodes  $x_i$  of the mesh  $\mathcal{T}_h$ .

**Lemma 7.8 ([82]).** *Let  $D \subset \mathbb{R}^2$  be a domain with convex, polygonal boundary and  $\Gamma$  a smooth interface with  $C^2$  parametrization. Let  $u \in H^2(\Omega_1 \cup \Omega_2) \cap H_0^1(D)$  and  $I_h$  the nodal Lagrange interpolation operator. Then it holds that*

$$\|\nabla(u - I_h u)\|_{L^2(D)} \leq C h \|u\|_{H^2(\Omega_1 \cup \Omega_2)}. \quad (7.11)$$

*Proof.* This proof follows the lines of [82], which is based on [33].

If the interface is resolved exactly by the modified finite element mesh, the result is standard, see (7.3). We divide the interpolation error into two parts,

$$\|\nabla(u - I_h u)\|_{L^2(D)}^2 = \|\nabla(u - I_h u)\|_{L^2(D \setminus S_h)}^2 + \|\nabla(u - I_h u)\|_{L^2(S_h)}^2. \quad (7.12)$$

For the first term, we can use estimate (7.3), since the solution  $u$  is smooth enough in  $D \setminus S_h$  to obtain

$$\|\nabla(u - I_h u)\|_{L^2(D \setminus S_h)}^2 \leq C h^2 \|\nabla^2 u\|_{L^2(D \setminus S_h)}^2 \leq C h^2 \|\nabla^2 u\|_{L^2(\Omega_1 \cup \Omega_2)}^2.$$

To deal with the second term, we introduce for  $u \in H^2(\Omega_i)$  ( $i = 1, 2$ ) a continuous extension  $\tilde{u}_i \in H^2(D)$  to the complete domain  $D$ . The existence of such an extension is guaranteed because  $\Gamma$  is smooth, see e.g. [211], and it holds

$$\|\tilde{u}_i - u\|_{H^2(\Omega_i)} = 0, \quad \text{and} \quad \|\tilde{u}_i\|_{H^2(D)} \leq \|\tilde{u}\|_{H^2(\Omega_i)}, \quad i = 1, 2. \quad (7.13)$$

We will derive an estimate for  $\|\nabla(u - I_h u)\|_{S_h^1}^2$ . The corresponding estimate on  $S_h^2$  follows analogously. The triangle inequality yields

$$\begin{aligned} \|\nabla(u - I_h u)\|_{L^2(S_h^1)} &\leq \|\nabla(u - \tilde{u}_1)\|_{L^2(S_h^1)} + \|\nabla(\tilde{u}_1 - I_h \tilde{u}_1)\|_{L^2(S_h^1)} \\ &\quad + \|\nabla(I_h \tilde{u}_1 - I_h u)\|_{L^2(S_h^1)}. \end{aligned} \quad (7.14)$$

The last term on the right hand side vanishes since  $I_h u = I_h \tilde{u}_1$  on  $S_h^1$  for the nodal Lagrange interpolator. The first term can be treated as follows: Noting that  $u - \tilde{u}_1$  vanishes everywhere on  $S_h^1$  except on  $S$ , we get by Lemma 7.6 and the continuity of the extension (7.13) that

$$\|\nabla(u - \tilde{u}_1)\|_{L^2(S_h^1)} \leq \|\nabla(u - \tilde{u}_1)\|_{L^2(S)} \leq \|\nabla u\|_{L^2(S)} + \|\nabla \tilde{u}_1\|_{L^2(S)} \leq C h \|u\|_{H^2(\Omega_1 \cup \Omega_2)}.$$

Finally, noting that  $\tilde{u}_1 \in H^2(D)$ , the interpolation error on the right hand side of (7.14) can be estimated in the standard way (7.3) for each triangle  $T \subset S_h^1$ . Summation over all  $T \subset S_h^1$  yields

$$\|\nabla(\tilde{u}_1 - I_h \tilde{u}_1)\|_{L^2(S_h^1)} \leq C h \|\nabla^2 \tilde{u}_1\|_{L^2(D)} \leq C h \|\nabla^2 u\|_{L^2(D)}. \quad (7.15)$$

In the last step, we again used the continuity of the extension (7.13). The statement follows by combining (7.12), (7.14) and (7.15) and the analogous estimate on  $S_h^2$ .  $\square$

**Theorem 7.9** ([82]). *Let  $D \subset \mathbb{R}^2$  be a domain with convex polygonal boundary, split into  $D = \Omega_1 \cup \Gamma \cup \Omega_2$ , where  $\Gamma$  is a smooth interface with  $C^2$ -parametrization. We assume that  $\Gamma$  divides  $D$  in such a way that the solution  $u$  belongs to  $H_0^1(D) \cap H^2(\Omega_1 \cup \Omega_2)$  and satisfies the stability estimate  $\|u\|_{H^2(\Omega_1 \cup \Omega_2)} \leq c_s \|f\|$ . Then, for the corresponding modified finite element solution  $u_h \in V_h$ , we have the estimates*

$$\|\nabla(u - u_h)\|_{L^2(D)} \leq C h \|f\| \quad \text{and} \quad \|u - u_h\|_{L^2(D)} \leq C h^2 \|f\|.$$

*Proof.* The proof is identical to [82], except that the patch size  $h_P$  is replaced by the global mesh size  $h$ . The proof is given for sake of completeness.

We start by estimating the  $H^1$ -seminorm error

$$\begin{aligned} c\|\nabla(u - u_h)\|_{L^2(D)}^2 &\leq a_h(u - u_h, u - u_h) \\ &= a_h(u - u_h, u - I_h u) + a_h(u - u_h, I_h u - u_h). \end{aligned}$$

For the first part, we use the Cauchy-Schwarz inequality and the interpolation estimate (7.11) as usual. The second part can be estimated with the Galerkin orthogonality (7.5) and Corollary 7.7:

$$\begin{aligned} a_h(u - u_h, I_h u - u_h) &= (a_h - a)(u, I_h u - u_h) \\ &\leq C h \|\nabla u\|_{H^1(\Omega_1 \cup \Omega_2)} \|\nabla(I_h u - u_h)\|_{L^2(D)} \\ &\leq C h \|f\|_{L^2(D)} (\|\nabla(I_h u - u)\|_{L^2(D)} + \|\nabla(u - u_h)\|_{L^2(D)}). \end{aligned}$$

Here, we have used the notation  $(a - a_h)(\cdot, \cdot) = a(\cdot, \cdot) - a_h(\cdot, \cdot)$  for better readability. By using Young's inequality and absorbing the last term into the left-hand side, we obtain

$$\|\nabla(u - u_h)\|_{L^2(D)}^2 \leq C \left( \|\nabla(u - I_h u)\|_{L^2(D)}^2 + h^2 \|f\|_{L^2(D)}^2 \right). \quad (7.16)$$

The  $H^1$ -seminorm estimate follows with Lemma 7.8.

To estimate the  $L^2$ -norm error, we make use of a dual problem. Let  $z \in H_0^1(D)$  be the solution of

$$a(\phi, z) = \|u - u_h\|_{L^2(D)}^{-1} (u - u_h, \phi)_{L^2(D)} \quad \forall \phi \in H_0^1(D). \quad (7.17)$$

The solution  $z$  lies in  $H_0^1(D) \cap H^2(\Omega_1 \cup \Omega_2)$  and  $\|z\|_{H^2(\Omega_1 \cup \Omega_2)} \leq c_s$ . By testing (7.17) with  $\phi = u - u_h$ , we have

$$\|u - u_h\|_{L^2(D)} = a(u - u_h, z) = a_h(u - u_h, z) + (a - a_h)(u - u_h, z). \quad (7.18)$$

For the second part, Corollary 7.7 gives

$$(a - a_h)(u - u_h, z) \leq C h \|\nabla(u - u_h)\|_{L^2(D)} \|z\|_{H^2(\Omega_1 \cup \Omega_2)}.$$

Next, we insert the interpolant  $I_h z$  into the first part of (7.18) and use the Galerkin orthogonality (7.5):

$$\begin{aligned} a_h(u - u_h, z) &= a_h(u - u_h, z - I_h z) + a_h(u - u_h, I_h z) \\ &= a_h(u - u_h, z - I_h z) + (a_h - a)(u, I_h z) \\ &= a_h(u - u_h, z - I_h z) + (a_h - a)(u, I_h z - z) + (a_h - a)(u, z). \end{aligned}$$

For the first term, we use the Cauchy-Schwarz inequality, the remaining terms can be handled with Corollary 7.7:

$$\begin{aligned} a_h(u - u_h, z - I_h z) &\leq C \|\nabla(u - u_h)\|_{L^2(D)} \|\nabla(z - I_h z)\|_{L^2(D)}, \\ (a_h - a)(u, I_h z - z) &\leq C h \|u\|_{H^2(\Omega_1 \cup \Omega_2)} \|\nabla(z - I_h z)\|_{L^2(D)}, \\ (a_h - a)(u, z) &\leq C h^2 \|u\|_{H^2(\Omega_1 \cup \Omega_2)} \|z\|_{H^2(\Omega_1 \cup \Omega_2)}. \end{aligned}$$

Finally, we use Lemma 7.8, the  $H^1$ -seminorm estimate (7.16) and stability estimates for the primal and dual solutions to obtain

$$\|u - u_h\|_{L^2(D)} \leq C h^2 \|u\|_{H^2(\Omega_1 \cup \Omega_2)} \|z\|_{H^2(\Omega_1 \cup \Omega_2)} \leq C h^2 \|f\|_{L^2(D)}.$$

□

#### 7.4.1 Numerical Experiments

We applied the method described in Section 7.2 to problem (7.1) where  $D = (-1, 1)^2$ ,  $\Omega_1 = B(x_m, 0.4)$  with  $x_m = (0.1, 0.2)^\top$ ,  $\Omega_2 = D \setminus \overline{\Omega}_1$ ,  $\kappa_1 = 1$ ,  $\kappa_2 = 10$ , and the right hand side as well as the Dirichlet data were chosen in such a way that the exact solution is given by

$$u(x) = \begin{cases} -4\kappa_1\kappa_2^2 R^2 \|x - x_m\|^2 + 2R^4\kappa_2(2\kappa_2\kappa_1 - 1) & x \in \Omega_1, \\ -2\kappa_2 \|x - x_m\|^4 & x \in \Omega_2. \end{cases}$$

The optimal order of convergence stated in Theorem 7.9 can be observed in Table 7.1. Note that all elements of the makro mesh have one right angle and two angles of  $45^\circ$ , which results in an upper bound for the maximum angle of  $180^\circ - 45^\circ/2 = 157.5^\circ$ .

nVerts	h	$\ u - u_h\ _{L^2(D)}$	rate $L_2$	$\ \nabla(u - u_h)\ _{L^2(D)}$	rate $H^1$	$\theta_{max}$
289	$h_0$	0.00724623	–	0.175665	–	140.334
1089	$h_0/2$	0.00180955	<b>2.0016</b>	0.087845	<b>0.9998</b>	138.116
4225	$h_0/4$	0.000453133	<b>1.9976</b>	0.0439104	<b>1.0004</b>	143.084
16641	$h_0/8$	0.000113451	<b>1.9979</b>	0.0219536	<b>1.0001</b>	152.223
66049	$h_0/16$	0.0000283643	<b>1.9999</b>	0.0109756	<b>1.0002</b>	149.110
263169	$h_0/32$	0.00000709548	<b>1.9991</b>	0.00548762	<b>1.0001</b>	155.643

Table 7.1: Convergence history of interface problem (7.1) using mesh adaptation strategy as mesh size approaches zero.

## 7.5 Condition Number

The procedure of Section 7.2 guarantees that no angle of the modified mesh becomes too large. However, it may happen that some angles in the triangulation are getting arbitrarily close to zero, which usually yields a bad condition of the finite element system matrix. Also this problem was addressed in [83] for the case of quadrilateral elements, and we adapt the procedure to the triangular case here.

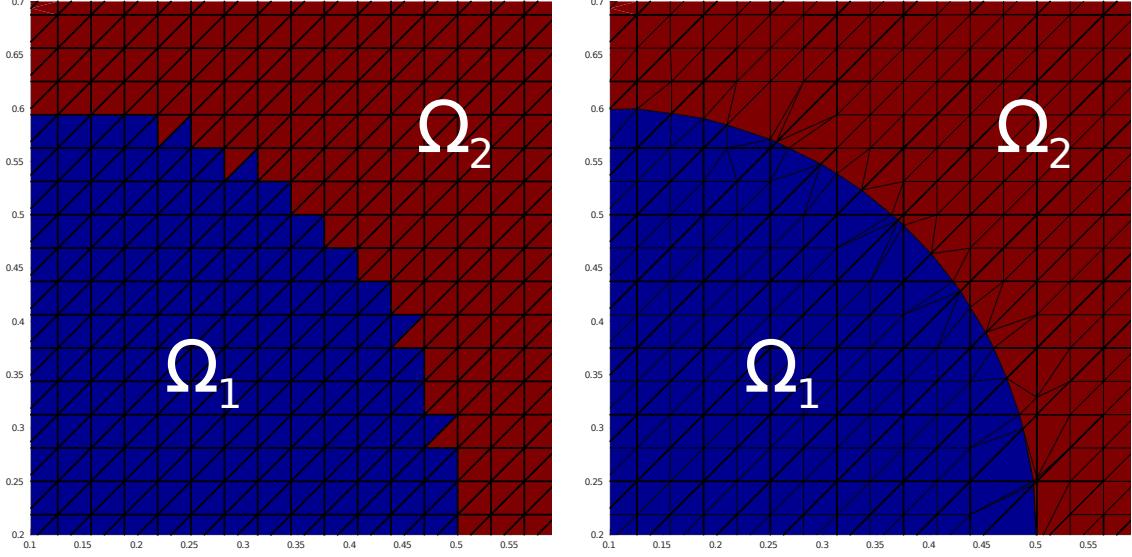


Figure 7.4: Left: Material coefficient  $\kappa$  without interface technique. Right:  $\kappa$  with interface technique

The idea consists in a hierarchical splitting of the finite element space  $V_h$  into the standard piecewise linear finite element space  $V_{2h}$  on the makro mesh  $\mathcal{T}_{2h}$  and the space of piecewise linear “bubble” functions  $V_b$  which vanish on the nodes of the makro elements,

$$V_h = V_{2h} + V_b. \quad (7.19)$$

Let  $\Phi_h := \{\phi_h^1, \dots, \phi_h^{N_h}\}$  be the nodal basis of the space  $V_h$ . Any function  $v_h \in V_h$  can be decomposed into the sum of a function  $v_{2h} \in V_{2h} = \text{span } \Phi_{2h}$  with  $\Phi_{2h} = \{\phi_{2h}^1, \dots, \phi_{2h}^{N_{2h}}\}$  and a function  $v_b \in V_b = \text{span } \Phi_b$  with  $\Phi_b = \{\phi_b^1, \dots, \phi_b^{N_b}\}$ ,

$$v_h = \sum_{i=1}^{N_h} \mathbf{v}_h^i \phi_h^i = \sum_{i=1}^{N_{2h}} \mathbf{v}_{2h}^i \phi_{2h}^i + \sum_{i=1}^{N_b} \mathbf{v}_b^i \phi_b^i = v_{2h} + v_b \in V_{2h} + V_b.$$

We denote the hierarchical basis of the space  $V_h = V_{2h} + V_b$  by  $\Phi_{hier} := \Phi_{2h} \cup \Phi_b = \{\phi_{2h}^1, \dots, \phi_{2h}^{N_{2h}}, \phi_b^1, \dots, \phi_b^{N_b}\}$ . In this setting, it is possible to show the usual bounds on the eigenvalues of the system matrix provided that the following two conditions are satisfied:

**Assumption 13.** We assume that

- there exists a constant  $C > 0$  independent of  $h, r, s$  such that

$$C^{-1} \leq \|\nabla \phi\| \leq C, \quad (7.20)$$

for all  $\phi \in \{\phi_{2h}^1, \dots, \phi_{2h}^{N_{2h}}, \phi_b^1, \dots, \phi_b^{N_b}\}$ , and

- there exists a constant  $C > 0$  independent of  $h, r, s$ , such that for all  $v_b \in V_b$

$$|\mathbf{v}_b^i| \leq C \|\nabla v_b\|_{\mathcal{N}_i}, \quad i = 1, \dots, N_b, \quad (7.21)$$

where  $\mathcal{N}_i$  denotes the patch around the vertex  $x_i$  of the mesh  $\mathcal{T}_h$ , i.e.,  $\mathcal{N}_i = \bigcup_{T \in \mathcal{T}_h: x_i \in \bar{T}} \bar{T}$ .

Given the basis  $\Phi_{hier}$ , let  $K_h$  denote the stiffness matrix of the model problem (7.1) and let  $\mathbf{v}_h$  be the coefficient vector of a function  $v_h \in V_h$  in this basis. Lemmas 7.10 and 7.11 give estimates on the maximal and minimal eigenvalues of the system matrix  $K_h$ , respectively. Both results are proven in [83] following the ideas of [32].

**Lemma 7.10** (large eigenvalues). *Let Assumption 13 hold. Then there exists a constant  $C > 0$  independent of the interface location such that it holds*

$$\mathbf{v}_h^\top K_h \mathbf{v}_h = a(v_h, v_h) \leq C \mathbf{v}_h^\top \mathbf{v}_h \quad \forall v_h \in V_h.$$

**Lemma 7.11** (small eigenvalues). *Let Assumption 13 hold. Then there exists a constant  $C > 0$  independent of the interface location such that it holds*

$$\mathbf{v}_h^\top K_h \mathbf{v}_h = a(v_h, v_h) \geq C h^2 \mathbf{v}_h^\top \mathbf{v}_h \quad \forall v_h \in V_h.$$

The combination of Lemma 7.10 and Lemma 7.11 gives the following result for the condition number of the system matrix  $K_h$ :

**Theorem 7.12.** *Let Assumption 13 hold. Then there exists a constant  $C > 0$  independent of the interface location such that*

$$\text{cond}_2(K_h) \leq C h^{-2}.$$

Here,  $\text{cond}_2(K_h)$  denotes the condition number of the stiffness matrix  $K_h$  associated with the Euclidean vector norm.

### 7.5.1 Scaling of the Basis Functions

It remains to check whether conditions (7.20) and (7.21) are satisfied for the finite element spaces arising in the method described in Section 7.2. Note that both conditions are satisfied for standard finite element spaces on regular grids, which can be shown by inverse inequalities. Thus, since the makro mesh  $\mathcal{T}_{2h}$  is not affected by the location of the interface, condition (7.20) is satisfied for all basis functions  $\phi \in \Phi_{2h}$ .

In this section, we will investigate whether, after a proper scaling, both conditions also hold for the basis functions  $\phi_b^i$ ,  $i = 1, \dots, N_b$ , independently of the location of the interface. This has to be checked for each of the four configurations A–D. Configuration A is analogous to what is called Configuration B1 in [83] and was treated in detail there.

#### 7.5.1.1 Configuration C

We will go through the details in the case of Configuration C1, i.e., in the case where  $r$  is close to 1,  $s$  is close to 0 and  $t = 1/2$ , see Figure 7.1(e). Note that the steps for Configuration C2 are analogous. We will show (7.20) and (7.21) on the equilateral reference patch  $\hat{T}$  with nodes

$$\begin{aligned} P_1 &= (0, 0)^\top, & P_2 &= (1, 0)^\top, & P_3 &= (1/2, \sqrt{3}/2)^\top, \\ P_4 &= (s, 0)^\top, & P_5 &= (3/4, \sqrt{3}/4)^\top, & P_6 &= r(1/2, \sqrt{3}/2)^\top. \end{aligned}$$

The corresponding condition on the physical patches follows from the shape-regularity of the patch mesh  $\mathcal{T}_{2h}$ . By  $\tilde{\phi}_4, \tilde{\phi}_5, \tilde{\phi}_6$ , we denote the nodal add-on basis functions which take the value 1 at the vertices  $P_4, P_5, P_6$ , respectively, and vanish on all other nodes. For the gradients of these basis functions, we have for  $r \rightarrow 1, s \rightarrow 0$  that

$$\begin{aligned} \|\nabla \tilde{\phi}_4\|_{T_1}^2 &= \mathcal{O}\left(\frac{1}{s}\right), & \|\nabla \tilde{\phi}_4\|_{T_2}^2 &= \mathcal{O}(1), & \|\nabla \tilde{\phi}_4\|_{T_4}^2 &= \mathcal{O}(1), \\ \|\nabla \tilde{\phi}_5\|_{T_2}^2 &= \mathcal{O}(1), & \|\nabla \tilde{\phi}_5\|_{T_3}^2 &= \mathcal{O}(1-r), & \|\nabla \tilde{\phi}_5\|_{T_4}^2 &= \mathcal{O}(1), \\ \|\nabla \tilde{\phi}_6\|_{T_1}^2 &= \mathcal{O}(s), & \|\nabla \tilde{\phi}_6\|_{T_3}^2 &= \mathcal{O}\left(\frac{1}{1-r}\right), & \|\nabla \tilde{\phi}_6\|_{T_4}^2 &= \mathcal{O}(1). \end{aligned}$$

Thus, introducing the scaling factors

$$\tau_4 = \sqrt{s}, \quad \tau_5 = 1, \quad \tau_6 = \sqrt{1-r},$$

and the scaled basis functions  $\phi_i := \tau_i \tilde{\phi}_i$ , for  $i = 4, 5, 6$ , we get condition (7.20).

Considering condition (7.21), we have to show for each  $i \in \{1, \dots, N_b\}$  that there exists a constant  $C$  independent of  $h, r$  and  $s$  such that

$$(v_b^i)^2 \leq C \|\nabla v_b\|_{L^2(\mathcal{N}_i)}^2 \quad \forall v_b \in V_b,$$

where  $\mathcal{N}_i$  denotes the union of all elements that have  $P_i$  as a node. Again, due to the shape-regularity of the makro mesh  $\mathcal{T}_{2h}$ , it is enough to show (7.21) on the reference patch  $\hat{T}$ . It is sufficient to show this condition for only one triangle  $T$  of the neighborhood  $\mathcal{N}_i$ , i.e.,

$$(v_b^i)^2 \leq C \|\nabla v_b\|_{L^2(T)}^2 \quad \forall v_b \in V_b.$$

We begin with the degree of freedom corresponding to point  $P_4$  and consider the subtriangle  $T_1$ , see Figure 7.1(a). A function  $v_b \in V_b$  restricted to  $T_1$  has contributions from the two degrees of freedom corresponding to  $P_4$  and  $P_6$ , thus the above condition is in this case equivalent to

$$1 \leq C \|\nabla \phi_4 + \mathbf{v} \nabla \phi_6\|_{L^2(T_1)}^2 \quad \forall \mathbf{v} \in \mathbb{R}. \quad (7.22)$$

We have

$$\nabla \phi_4|_{T_1} = \tau_4 \begin{pmatrix} \frac{1}{s} \\ -\frac{1}{\sqrt{3}s} \end{pmatrix}, \quad \nabla \phi_6|_{T_1} = \tau_6 \begin{pmatrix} 0 \\ \frac{2}{\sqrt{3}r} \end{pmatrix}, \quad |T_1| = \frac{1}{4} \sqrt{3} r s,$$

and

$$\|\nabla \phi_4\|_{L^2(T_1)}^2 = \tau_4^2 \frac{r}{\sqrt{3}s}, \quad \|\nabla \phi_6\|_{L^2(T_1)}^2 = \tau_6^2 \frac{s}{\sqrt{3}r}, \quad (\nabla \phi_4, \nabla \phi_6)_{T_1} = -\tau_4 \tau_6 \frac{1}{2\sqrt{3}},$$

and thus, for  $r$  close to 1 and  $s$  close to 0, we have that

$$\|\nabla \phi_4 + \mathbf{v} \nabla \phi_6\|_{L^2(T_1)}^2 = \|\nabla \phi_4\|_{L^2(T_1)}^2 + 2 \mathbf{v} (\nabla \phi_4, \nabla \phi_6)_{T_1} + \mathbf{v}^2 \|\nabla \phi_6\|_{L^2(T_1)}^2 \approx \frac{1}{\sqrt{3}}. \quad (7.23)$$

Thus, condition (7.22) is satisfied if  $C > \sqrt{3}$ .

Similarly, we consider the basis function corresponding to point  $P_5$  on  $T_2$  and show the existence of a constant  $C$  such that

$$1 \leq C \|\nabla \phi_5 + \mathbf{v} \nabla \phi_4\|_{L^2(T_2)}^2 \quad \forall \mathbf{v} \in \mathbb{R}. \quad (7.24)$$

Here, we have

$$\nabla\phi_5|_{T_2} = \begin{pmatrix} 0 \\ \frac{4}{\sqrt{3}} \end{pmatrix}, \quad \nabla\phi_4|_{T_2} = \tau_4 \begin{pmatrix} -\frac{1}{1-s} \\ -\frac{1}{\sqrt{3}(1-s)} \end{pmatrix}, \quad |T_2| = \frac{1}{8}\sqrt{3}(1-s),$$

and

$$\|\nabla\phi_5\|_{L^2(T_2)}^2 = \frac{2(1-s)}{\sqrt{3}}, \quad \|\nabla\phi_4\|_{L^2(T_2)}^2 = \tau_4^2 \frac{1}{2\sqrt{3}(1-s)}, \quad (\nabla\phi_5, \nabla\phi_4)_{T_2} = -\tau_4 \frac{1}{2\sqrt{3}},$$

and thus, again for  $r \approx 1$  and  $s \approx 0$ ,

$$\|\nabla\phi_5 + \mathbf{v}\nabla\phi_4\|_{L^2(T_2)}^2 \approx \frac{2}{\sqrt{3}}. \quad (7.25)$$

Condition (7.24) follows for  $C > \frac{\sqrt{3}}{2}$ .

In an analogous way, it can be shown that

$$\|\nabla\phi_6 + \mathbf{v}\nabla\phi_5\|_{L^2(T_3)}^2 \approx \frac{1}{2\sqrt{3}},$$

for  $r$  close to 1 and  $s$  close to 0. This, together with (7.23) and (7.25) yields that, in Configuration C1, condition (7.21) is satisfied for  $C > 2\sqrt{3}$  in the case of Configuration C.

### 7.5.1.2 Configuration B

Let us now have a look at Configuration B, i.e., the case where both  $r$  and  $s$  are close to 1. For the gradients of the basis functions, we get  $t = 1 - s, r, s \rightarrow 1$

$$\begin{aligned} \|\nabla\tilde{\phi}_4\|_{T_1}^2 &= \mathcal{O}(1), & \|\nabla\tilde{\phi}_4\|_{T_2}^2 &= \mathcal{O}(1), & \|\nabla\tilde{\phi}_4\|_{T_4}^2 &= \mathcal{O}(\frac{1}{1-s}), \\ \|\nabla\tilde{\phi}_5\|_{T_2}^2 &= \mathcal{O}(1), & \|\nabla\tilde{\phi}_5\|_{T_3}^2 &= \mathcal{O}(r-1), & \|\nabla\tilde{\phi}_5\|_{T_4}^2 &= \mathcal{O}(\frac{1}{1-s}), \\ \|\nabla\tilde{\phi}_6\|_{T_1}^2 &= \mathcal{O}(1), & \|\nabla\tilde{\phi}_6\|_{T_3}^2 &= \mathcal{O}(\frac{1}{1-r}), & \|\nabla\tilde{\phi}_6\|_{T_4}^2 &= \mathcal{O}(1-s), \end{aligned}$$

thus we get condition (7.20) by scaling the basis functions by

$$\tau_4 = \sqrt{1-s}, \quad \tau_5 = \sqrt{1-s}, \quad \tau_6 = \sqrt{1-r}.$$

Considering condition (7.21), we again treat the three basis functions corresponding to  $P_4, P_5, P_6$  individually:

- For the basis function corresponding to node  $P_6$ , analogous calculations as in Configuration C1 above yield that

$$\|\nabla\phi_6 + \mathbf{v}\nabla\phi_5\|_{L^2(T_3)}^2 \approx \frac{1}{\sqrt{3}}$$

for  $r, s \approx 1$ . Thus, condition (7.21) is satisfied.

- For point  $P_5$  we get

$$\|\nabla\phi_5 + \mathbf{v}\nabla\phi_4 + \mathbf{w}\nabla\phi_6\|_{L^2(T_2 \cup T_3 \cup T_4)}^2 \approx \frac{1}{\sqrt{3}} ((\mathbf{v}-1)^2 + \mathbf{w}^2)$$

for  $r, s \approx 1$ , which vanishes for  $\mathbf{v} = 1$  and  $\mathbf{w} = 0$ . In this case, condition (7.21) can be shown by looking at triangle  $T_5$  in the neighboring patch, see Figure 7.5(a). Due to

Assumption 12, point  $P_7$  stays the midpoint of the makro edge and triangle  $T_5$  has two edges with length of order 1 and one edge of length  $t = 1 - s$ , see Figure 7.6(a). Due to the scaling of the basis function  $\phi_5 = \tau_5 \tilde{\phi}_5$  with  $\tau_5 = \sqrt{1-s}$  and since the basis function  $\phi_7$  already has the correct scaling to fulfill (7.20), i.e.,  $\tau_7 = 1$  (cf. Configuration C), we get for  $r, s \approx 1$  that

$$\|\nabla \phi_5 + \mathbf{v} \nabla \phi_7\|_{L^2(T_5)}^2 \approx \frac{1}{2\sqrt{3}},$$

which yields condition (7.21).

- Similarly, for the point  $P_4$  we get for  $r, s \approx 1$  that

$$\|\nabla \phi_4 + \mathbf{v} \nabla \phi_5 + \mathbf{w} \nabla \phi_6\|_{L^2(T_1 \cup T_2 \cup T_4)}^2 \approx \frac{(\mathbf{v} - 1)^2}{\sqrt{3}},$$

which vanishes for  $\mathbf{v} = 1$ . Also here, we can have a look at the triangles coming from the adjacent patch, see Figure 7.5(a). This case is more involved since we have to distinguish several cases depending on where the interface leaves the adjacent patch, see Figures 7.6(b)–(f). Depending on the configuration of the adjacent patch (consisting of triangles  $T_6$ – $T_9$  in Figure 7.5), the basis functions corresponding to points  $P_8$  and  $P_{10}$  have different scaling factors. Therefore, a proof of condition (7.21) in this case becomes quite technical and remains to be shown in a rigorous way.

### 7.5.1.3 Configuration D

We consider Configuration D1, where the interface parameter  $r$  is close to 0 and remark that analogous findings hold for Configuration D2 due to symmetry. For the gradients of the basis functions, we obtain

$$\begin{aligned} \|\nabla \tilde{\phi}_4\|_{T_1}^2 &= \mathcal{O}(1), & \|\nabla \tilde{\phi}_4\|_{T_2}^2 &= O(\frac{1}{r}), & \|\nabla \tilde{\phi}_5\|_{T_3}^2 &= \mathcal{O}(1), & \|\nabla \tilde{\phi}_5\|_{T_4}^2 &= \mathcal{O}(1), \\ \|\nabla \tilde{\phi}_6\|_{T_1}^2 &= \mathcal{O}(1), & \|\nabla \tilde{\phi}_6\|_{T_2}^2 &= O(\frac{1}{r}), & \|\nabla \tilde{\phi}_6\|_{T_3}^2 &= \mathcal{O}(1), & \|\nabla \tilde{\phi}_6\|_{T_4}^2 &= \mathcal{O}(1). \end{aligned}$$

We scale the basis functions by the factors

$$\tau_4 = \sqrt{r}, \quad \tau_5 = 1, \quad \tau_6 = \sqrt{r}.$$

which yields condition (7.20). For the degree of freedom corresponding to the node  $P_5$ , we get that condition (7.21) is satisfied due to

$$\|\nabla \phi_5 + \mathbf{v} \nabla \phi_6\|_{L^2(T_3)}^2 = \frac{2}{\sqrt{3}}.$$

However, for the remaining two degrees of freedom, condition (7.21) cannot be shown by only considering the subtriangles  $T_1$ – $T_4$ . Similarly to Configuration B, the condition can be shown for the degree of freedom corresponding to  $P_4$  by looking at the triangle  $T_6$  (note that  $P_8$  and  $P_{10}$  must be the midpoints of the corresponding makro edges due to Assumption 12). For the degree of freedom corresponding to  $P_6$ , again we have to distinguish several cases depending on where the interface leaves the adjacent patch. Also this technical discussion remains open.

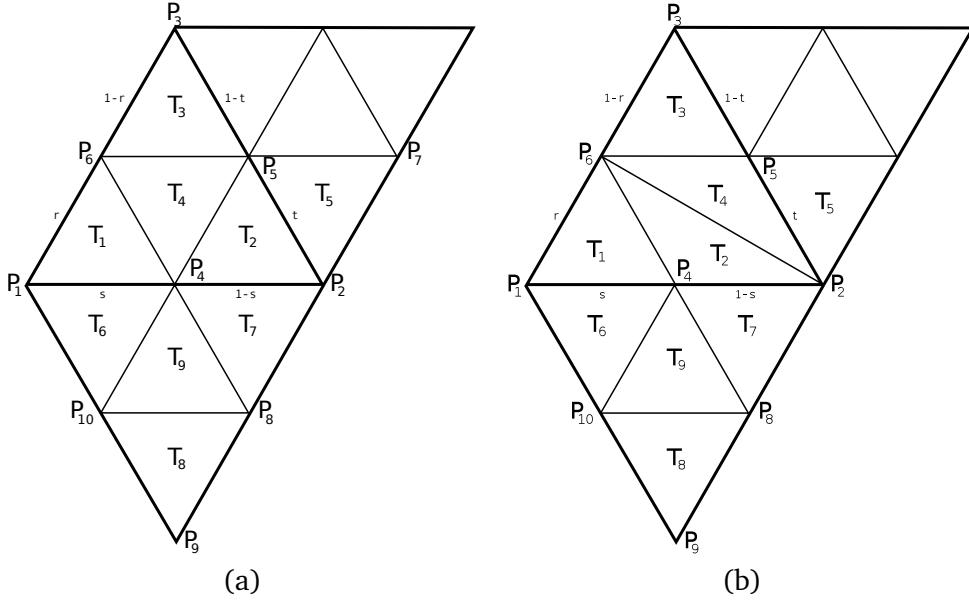


Figure 7.5: Reference patch together with two adjacent patches for Configurations A–C (left) and Configuration D (right).

Summarizing, we have seen that the suggested scaling of the basis functions yields condition (7.20) and also that condition (7.21) is satisfied for all bubble functions  $v_b \in V_b$  except in the two limiting cases in Configurations B and D mentioned above. We remark that (7.21) is only a sufficient condition, thus it may still be possible to show the bound on the condition number of Theorem 7.12 even if (7.21) does not hold.

In a practical implementation, an appropriate scaling of the basis functions can be achieved by a simple diagonal scaling of the stiffness matrix. Instead of solving the linear system  $K_h \mathbf{u}_h = \mathbf{f}_h$  we solve the system

$$\tilde{K}_h \tilde{\mathbf{u}}_h = \tilde{\mathbf{f}}_h \quad \text{with} \quad \tilde{K}_h = D^{-1/2} K_h D^{-1/2}, \quad \tilde{\mathbf{u}}_h = D^{1/2} \mathbf{u}_h, \quad \tilde{\mathbf{f}}_h = D^{-1/2} \mathbf{f}_h, \quad (7.26)$$

where  $D = \text{diag}(K_h)$ . By this scaling, we achieve that the diagonal entries of the scaled stiffness matrix  $\tilde{K}_h$  are equal to 1, which can be interpreted as a scaling of the basis functions such that  $\|\nabla \phi_i\|^2 = a(\phi_i, \phi_i) = (\tilde{K}_h)_{ii} = 1$ .

### 7.5.2 Numerical Experiments

In this section, we illustrate in two simple examples how the choice of the hierarchical basis (7.19) together with the diagonal scaling (7.26) of the stiffness matrix reduces the condition number of the system matrix in the presence of small angles. As a rough measure of the condition number, we consider the number of iterations of the conjugate gradient (CG) method which, given a right hand side vector, is needed to reduce the initial residual by a given factor (here by the factor  $10^{-10}$ ). Note that, for the CG method, the number of required iterations is of the order of the square root of the condition number, see e.g. [156, 222].

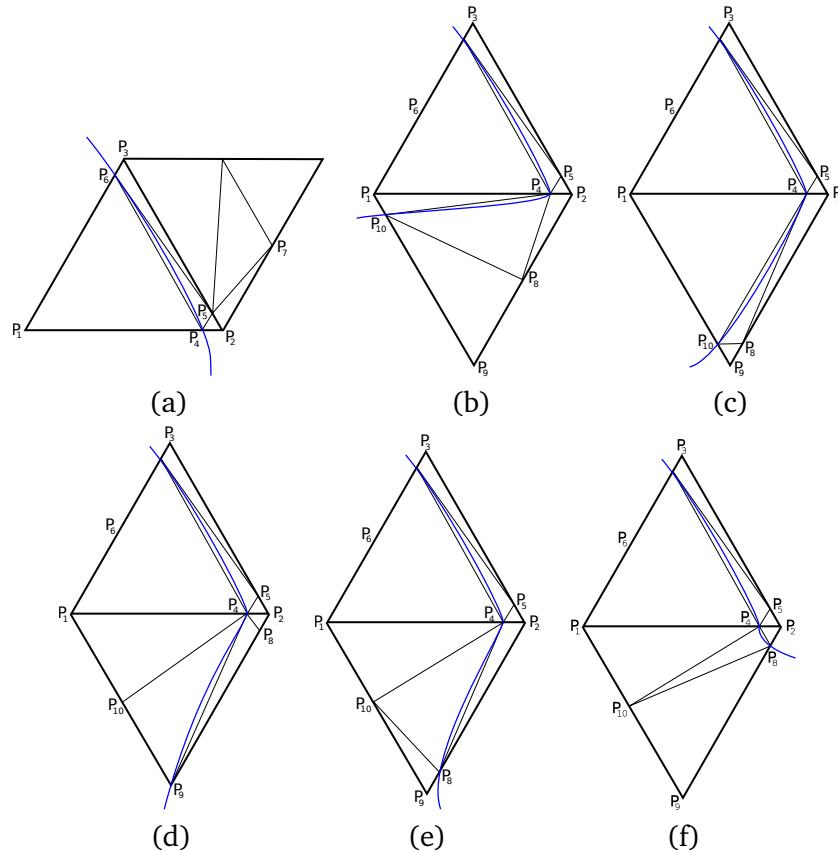


Figure 7.6: Possible different configurations of adjacent patches when patch consisting of points  $P_1-P_6$  is cut by the interface. (a) Patch adjacent to non-intersected makro edge in Configuration B. (b)–(f) Patch adjacent to intersected makro edge in Configuration B for different configurations of this patch. (g) Patch adjacent to non-intersected makro edge in Configuration D.

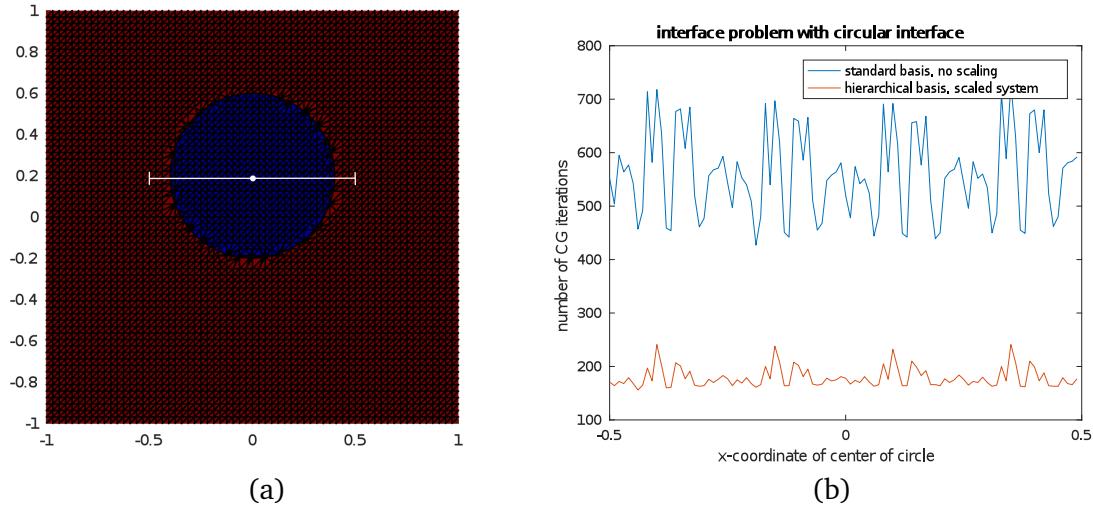


Figure 7.7: Interface problem with circular interface; x-coordinate of center of circle is incremented by steps of 0.01 between  $x = -0.5$  and  $x = 0.5$ . Right: Number of CG iterations needed to reduce the initial residual by a factor of  $10^{-10}$  depending on position of circle for standard FE approach (blue) and for hierarchical basis with scaling (7.26) (orange).

As a first example we consider a circular interface at different positions within the computational domain. The finite element system is solved for a range of values of the x-coordinate  $x_{M,1}$  of the midpoint of the circle between  $-0.5$  and  $0.5$ , see Figure 7.7(a). Figure 7.7(b) compares the number of CG iterations needed for solving the original system in the standard finite element basis and the scaled system together with the hierarchical basis. It can be seen that the latter approach yields a much better condition, though a dependence on the location of the interface is still observable.

In a second example, we investigate the condition of the system when some angles approach zero. For that purpose, we divide the computational domain  $D$  into the subdomains  $\Omega_1^{t_{IF}} := \{(x_1, x_2) \in D : x_2 > 0.5 + t_{IF}\}$  with coefficient  $\kappa = 1$ ,  $\Omega_2^{t_{IF}} := D \setminus \overline{\Omega_1^{t_{IF}}}$  with  $\kappa = 10$  as well as the interface  $\Gamma^{t_{IF}}$ , and let the parameter  $t_{IF}$  tend to zero. By this procedure, the horizontal material interface, which is parallel to the edges of the mesh at  $x_2 = 0.5$ , approaches these edges and produces very small angles (due to Configurations A and B), see the right picture of Figure 7.8. Table 7.2 compares the number of CG iterations for the different approaches as the interface parameter, and therefore the minimal angle, approaches zero.

From the numerical experiments conducted in this section, we can conclude that the condition number of the stiffness matrix  $K_h$  can be significantly improved by using a hierarchical basis and scaling the system of finite element equations according to (7.26). However, we also see that there is still a dependence of this condition number on the location of the interface relative to the makro mesh which is most likely due to the presence of very small angles.

We mention that, in this thesis, all linear systems arising in the discretization of the state and adjoint equations were solved using a direct solver, namely the PARDISO (parallel direct solver) project [131, 183, 184]. A bad condition of the system matrix also affects the accuracy of the solution obtained by direct methods. However, in the case of direct solvers, we have

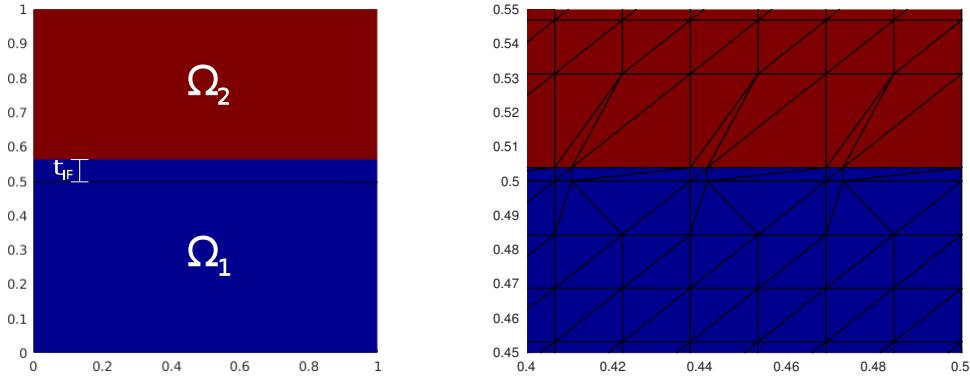


Figure 7.8: Horizontal interface approaching the line  $\{y = 0.5\}$  as  $t_{IF}$  goes to zero. As  $t_{IF}$  decreases, smallest angle in mesh gets smaller and condition of system matrix deteriorates (see Table 7.2).

control over the number of valid digits in the solution to a system of linear equations. As a rule of thumb, using floating point arithmetic with  $d$  digits (e.g.,  $d = 16$  when using double precision in C++), the first  $d - \log_{10}(\kappa(A)) - 1$  many digits of the solution are reliable where  $\kappa(A)$  denotes the condition number of the system matrix  $A$ , see e.g. [190].

It is subject of future investigation to design a robust preconditioner for iterative methods by replacing the node-wise diagonal scaling (7.26) by a patch-wise block diagonal scaling which can be understood as an additive Schwarz preconditioner.

$t_{IF}$	$\theta_{min}$	$\theta_{max}$	no scaling		diagonal scaling	
			$\Phi_h$	$\Phi_{2h} \cup \Phi_b$	$\Phi_h$	$\Phi_{2h} + \Phi_b$
1/8	45	90	537	357	210	134
1/16	45	90	507	339	208	139
1/32	45	90	492	323	210	141
1/64	45	90	488	323	212	142
1/128	15.255	108.43	493	320	228	143
1/256	6.5463	113.2	571	379	229	149
1/512	3.0556	115.02	705	474	239	160
1/1024	1.4786	115.82	789	566	253	177
1/2048	0.72756	116.2	854	655	284	197
1/4096	0.36092	116.38	1015	828	353	235
1/8192	0.17975	116.48	1072	1053	414	314
1/16384	0.0897	116.52	1176	1373	507	404
1/32768	0.044806	116.54	1234	1667	629	481
1/65536	0.022392	116.55	1284	1984	711	593

Table 7.2: Number of CG iterations needed for solving linear system of problem (7.1) on domain of Figure 7.8 for decreasing sequence of interface parameter  $t_{IF}$  without and with preconditioning (7.26), using the standard basis of the FE space  $V_h$  or the hierarchical splitting (7.19) into  $V_{2h} + V_b$ .



## **Part IV**

# **Combined Topology and Shape Optimization with Interface Handling**



## Chapter 8

# Numerical Optimization Results

In this chapter, we combine the techniques we developed so far in this thesis and apply them to two design optimization problems for electric motors which are, from a practical point of view, slightly more relevant than the model problem introduced in Section 2.3. The first practical problem is posed on the same motor geometry as in Section 2.3, but we optimize with respect to a different objective functional. In the second practical example, we deal with a synchronous reluctance motor, which does not have permanent magnets. There, the design subdomain consists of the whole rotor and the goal is to maximize the torque of the motor.

### 8.1 Implementation

We make a few remarks concerning the numerical simulation and optimization results obtained in this section as well as in Sections 3.3, 4.8 and 6.6:

- All arising PDEs were solved approximately by piecewise linear, globally continuous finite elements on triangular meshes.
- Given a geometry description of the electric motor, a triangular mesh was obtained by NETGEN, see [186]. In order to allow for different rotor-to-stator constellations, the rotor and the stator are meshed individually and are connected by a third mesh representing the air gap. This air gap mesh has particularly high resolution which yields a higher accuracy of the computed magnetic field in this very important region.
- The system of nonlinear equations arising from the finite element discretization of the nonlinear state equation was solved by a damped Newton method, see, e.g., [76, 156, 222].
- All systems of linear equations arising from the finite element discretization were solved using the parallel direct solver PARDISO [131, 183, 184].
- The model problem introduced in Section 2.3 was solved on a mesh with 44810 degrees of freedom and 89454 elements where we chose a particularly fine discretization in the design regions  $\Omega^d$  (53488 design elements). For the synchronous reluctance motor of Section 8.4, we used a mesh consisting of 24935 vertices and 49768 elements (17888 design elements).

- All numerical simulation and optimization results were obtained using our own implementation in C++.

## 8.2 Combined Topology and Shape Optimization with Interface Handling

One drawback of Algorithm 2 presented in Section 4.8, where the design is represented by a level set function whose evolution is steered by the topological derivative, is that the topological derivative is not defined on the interface and, therefore, the procedure is not well-defined there. In a numerical realization, the generalized topological derivative at the material interface is obtained by averaging over the neighboring elements of the mesh, see Remark 4.79. However, there is still no guarantee that the material interface arising from an optimization procedure that is guided by the topological derivative is smooth. In particular, when combining Algorithm 2 with the mesh adaptation strategy of Chapter 7, we observed that Assumption 11 was often violated.

For this reason, we recommend the following two-stage optimization procedure:

**Algorithm 4.** (Combined topology and shape optimization with interface handling)

**Stage I:** Apply Algorithm 2 to find an optimal topology.

**Stage II:**

- Approximate the final design of Stage II by polygonal interface.
- Apply Algorithm 3 with the following modification:  
For each solve of the state and adjoint equations for  $u$  and  $p$ , the local mesh adaptation strategy of Section 7.2 is applied for the updated polygonal interface.

**Remark 8.1.** A more natural version of Algorithm 4 consists in skipping step II(i) and performing shape optimization based on the level set method, see Section 1.2.1.4, using the level set representation of the final result of Stage I. However, due to the difficulties arising in the numerical treatment of the Hamilton-Jacobi equation (1.1) (in particular on non-structured grids) which we pointed out in Section 1.2.1.4, we used an explicit representation of the interface and the gradient-like method of Algorithm 3. The extension to a level-set based shape optimization method should cause no fundamental problems and is subject of future work.

## 8.3 Minimizing Total Harmonic Distortion

The goal of the model problem introduced in Section 2.3 was to achieve a smooth rotation of the rotor by having a smooth radial component of the magnetic flux density  $B_r$  in the air gap. With the choice of the objective functional (2.15), we tried to achieve this by minimizing the  $L^2$  distance between  $B_r$  and a given sine curve, see Figure 2.3. However, in practice it is not clear how big the amplitude of this given sine curve should be chosen.

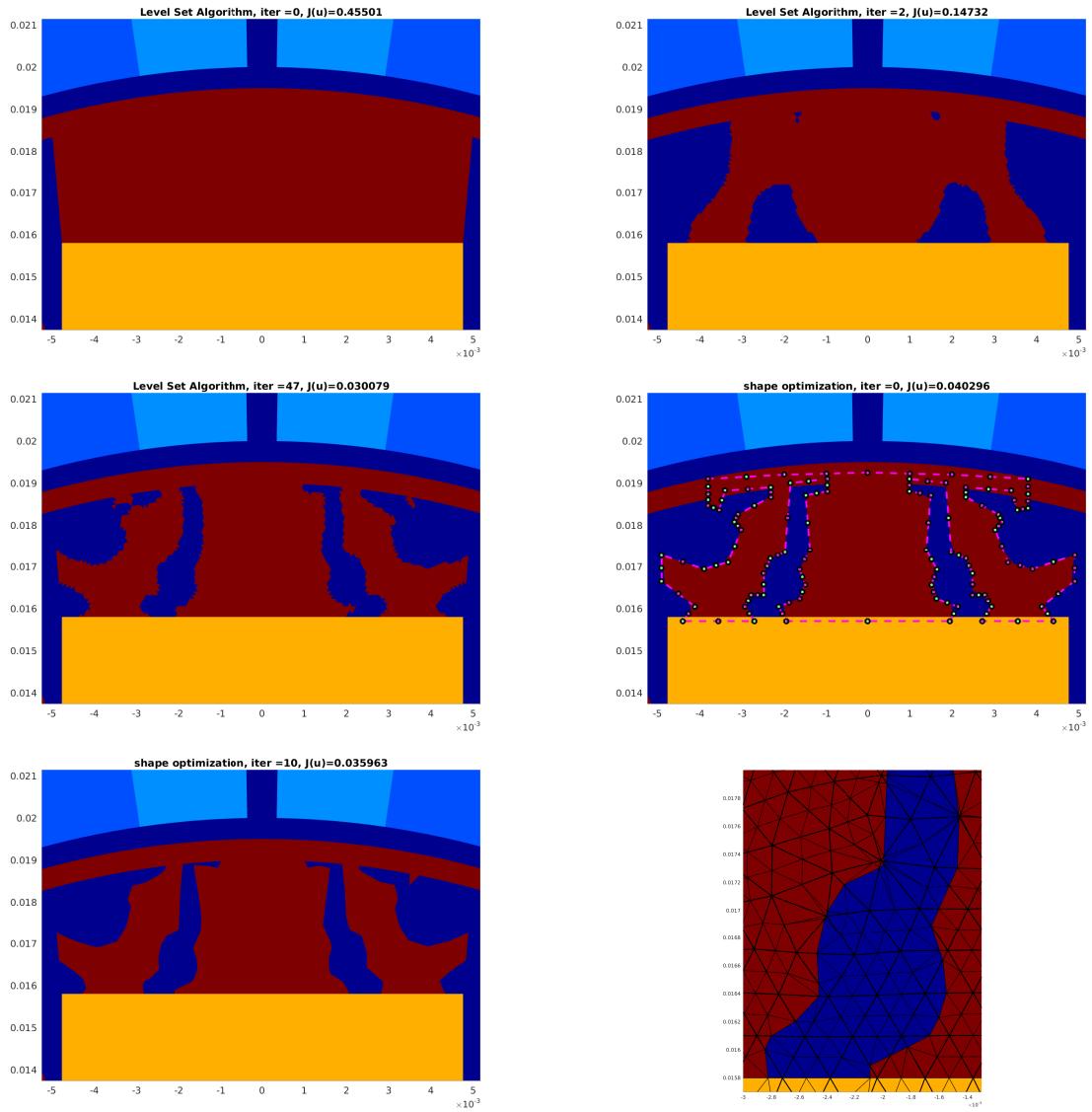


Figure 8.1: Top left: Initial Design. Top right: Design after two iterations of topology optimization by Algorithm 2. Center left: Final design of topology optimization after 47 iterations. Center right: Initial design for shape optimization by approximation of topology optimization result. Bottom left: Final design of shape optimization with mesh adaptation strategy after 10 iterations. Bottom right: Zoom on modified mesh.

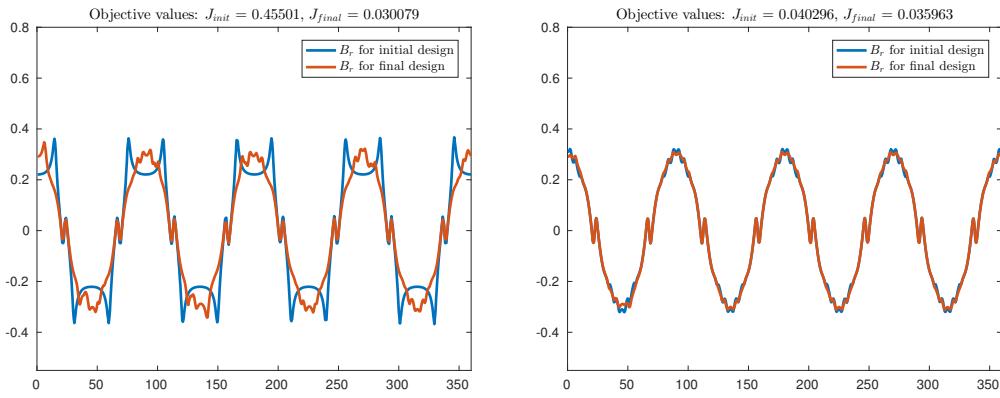


Figure 8.2: Radial component of magnetic flux density along the air gap for initial and final designs. Left: Stage I (Topology optimization). Right: Stage II (Shape Optimization)

A better way to achieve a smooth radial component of  $\mathbf{B}$  is the following: We consider  $B_r$  along a circular curve inside the air gap as a periodic signal and decompose it into its Fourier coefficients,

$$B_r(u)(\varphi) = \sum_{k=1}^{\infty} A_k \sin(\omega k \varphi) + B_k \cos(\omega k \varphi), \quad (8.1)$$

where  $A_k, B_k \in \mathbb{R}$ ,  $\varphi \in [0, 2\pi]$  and  $\omega$  denotes the number of pole pairs of the motor. In the motor introduced in Section 2.3, we have eight magnetic poles, thus  $\omega = 4$ . Due to the geometry of the motor, the coefficients  $A_k$  are approximately zero and will be neglected. The total harmonic distortion (THD) measures the contributions of higher harmonics (i.e.,  $k > 1$ ) to the total signal, see [45, 66, 87, 130]. For practical purposes, we only consider the first  $N = 20$  harmonics. Then, the total harmonic distortion of  $B_r$  reads

$$THD(B_r) = \sqrt{\frac{\sum_{k=2}^N B_k^2}{\sum_{k=1}^N B_k^2}},$$

where the coefficients  $B_k$  are according to (8.1). The minimization of the THD filters out all higher harmonics. In order to make sure that first harmonic does not become too small, we minimize the functional

$$\mathcal{J}(u) = \frac{THD(B_r(u))^2}{B_1(B_r(u))},$$

where  $B_1(B_r(u))$  denotes the coefficient  $B_1$  in (8.1). In our implementation, we computed the Fourier coefficients by a least square approach.

Figure 8.1 shows the evolution of the design by using Algorithm 4 starting from an initial design. The final design obtained after a total of 47 iterations is approximated by an explicit polygonal interface, which served as an initial guess for the shape optimization. The final design after the shape optimization procedure together with the local mesh modification

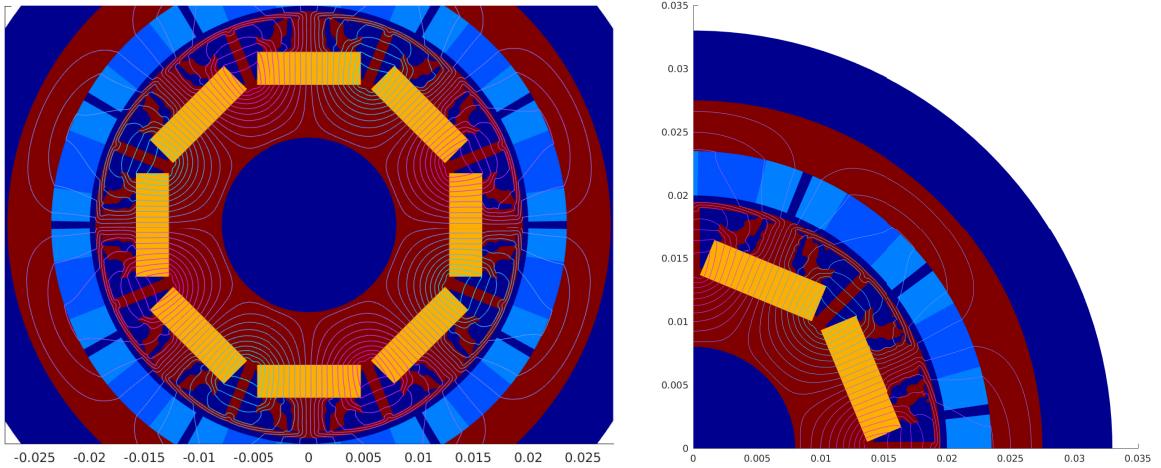


Figure 8.3: Final designs after Stage II together with magnetic field lines.

strategy introduced and analyzed in Chapter 7 can be seen in the bottom row of Figure 8.1. Figure 8.2 shows the curve  $B_r$  for the initial and the final design of both stages of the optimization procedure, and Figure 8.3 the final design together with the magnetic field.

We remark that in step II(i), we approximated the final design of Stage I by a polygonal interface by hand. Here, we enforced symmetry of the design among the eight parts of the design area  $\Omega^d$  and also within each of these eight parts. By proceeding like this, we introduce a rather large approximation error and the objective value increases. After shape optimization, we get a smoother and more symmetric design, however the objective value is still larger than for the final design of Stage I. As mentioned in Remark 8.1, Algorithm 4 could be improved by performing Stage II by a level set method. Moreover, in all of our experiments, we performed the optimization of the eight parts of the design region  $\Omega^d$  independently in order to allow for more general designs. If one is only interested in symmetric designs, we recommend to compute sensitivities only in one of the eight parts and to perform the same design updates in each part in every iteration. We remark that the source of non-symmetry in FE-based shape and topology optimization is often a non-symmetric mesh.

## 8.4 Maximizing Torque for Synchronous Reluctance Motor

In this final section of the thesis, we consider a different kind of electric motors, namely a synchronous reluctance motor which does not contain any permanent magnets. Compared to electrical machines with permanent magnets, this kind of motor performs generally worse, but is more robust and cheaper to manufacture. The key component of this kind of motors is that the magnetic flux is guided by so-called flux barriers, i.e., narrow air regions inside the ferromagnetic material in the rotor. The optimal number and shape of such flux barriers is subject of active research, see, e.g. [70, 126], but has, to the knowledge of the author, so far only been considered from a parametric point of view.

In contrast to the model problem used in the previous numerical experiments, here, we are

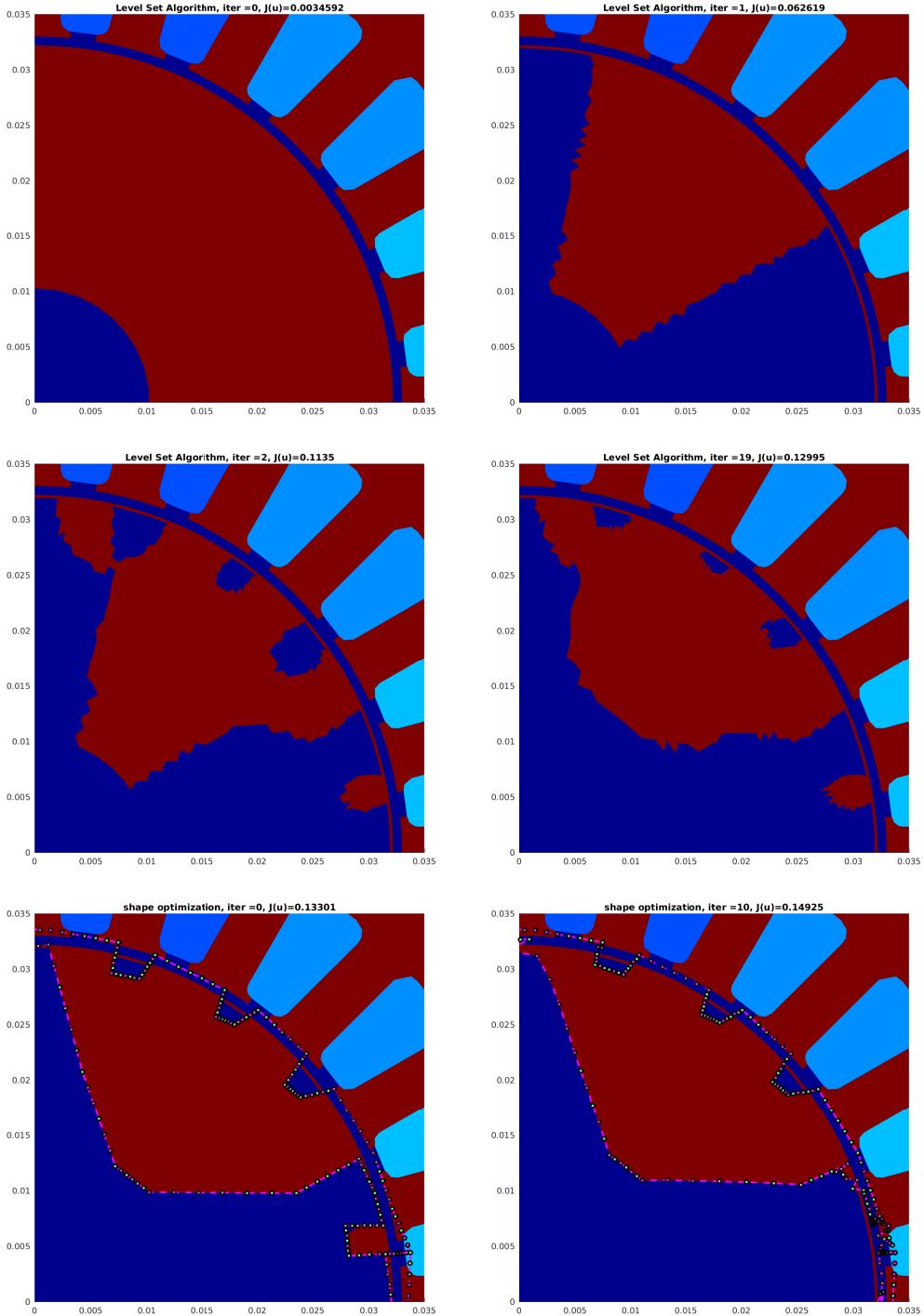


Figure 8.4: Top left: Initial Design. Top right: Design after one iteration of topology optimization by Algorithm 2. Center left: Design after two iterations of topology optimization by Algorithm 2. Center right: Final design of topology optimization after 19 iterations. Bottom left: Initial design for shape optimization by approximation topology optimization result. Bottom right: Final design of shape optimization with mesh adaptation strategy after 10 iterations.

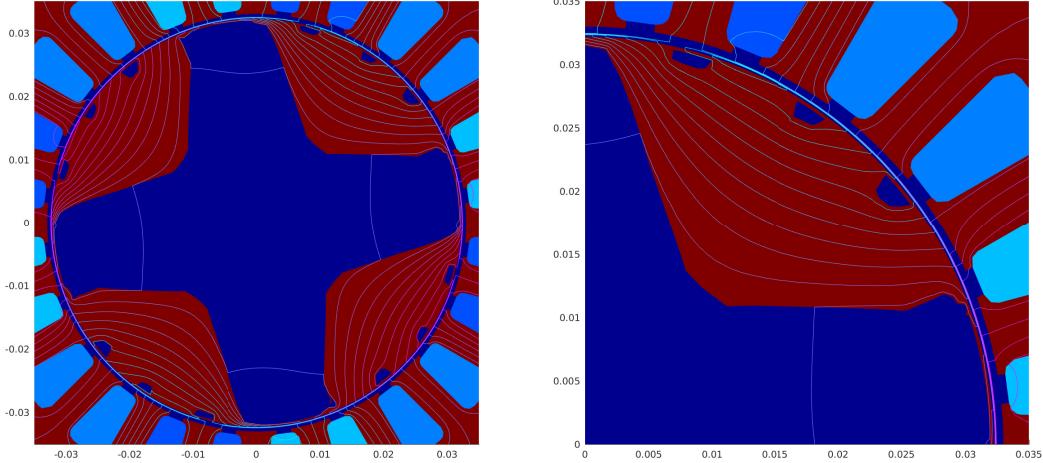


Figure 8.5: Final design after Stage II together with magnetic field lines.

interested in an objective which comprises all possible positions of the rotor (inner part of the motor) relative to the stator (outer part). We are interested in maximizing the average of the torque over all different rotor positions. Therefore, we consider the functional

$$\mathcal{J} = \frac{1}{N} \sum_{k=1}^N T(\varphi_k)$$

where  $T(\varphi)$  denotes the torque of the motor at rotor position  $\varphi \in [0, 2\pi]$  and  $\{\varphi_k\}_{k=1}^n$  is an equidistant subdivision of the interval  $[0, 2\pi]$ . For a fixed rotor position  $\varphi$ , the torque can be expressed as

$$T(\varphi) = \nu_0 \int_{\Gamma_0} \nabla u_\varphi^\top Q(x) \nabla u_\varphi \, ds,$$

with the symmetric matrix

$$Q(x) = \frac{1}{\sqrt{x_1^2 + x_2^2}} \begin{pmatrix} x_1 x_2 & \frac{x_2^2 - x_1^2}{2} \\ \frac{x_2^2 - x_1^2}{2} & -x_1 x_2 \end{pmatrix},$$

see [81], where  $\Gamma_0$  is a circular curve inside the air gap, and  $u_\varphi$  is the solution to the magnetostatic boundary value problem (2.17b) for the rotor-to-stator constellation given by the angle  $\varphi$ . Note that, compared to the previous examples, here the magnetization  $M^\perp$  in (2.13) vanishes and the current density  $J_3$  is piecewise constant in the coil areas. In each of the coil subdomains, the induced electric current  $J_3$  is a periodic sine-like function in terms of the rotation angle  $\varphi$ .

Since we are interested in a high mean torque, we minimize the functional  $-\mathcal{J}$ . Due to the periodicity of the motor and the form of the currents, it suffices to consider only a sixth of a full rotation. For the equidistant rotation angles we chose  $N = 15$  and  $\{\varphi_1, \dots, \varphi_N\} = \{0, 2, 4, \dots, 28\}$ .

Figure 8.4 shows the evolution of the design in the course of the two-stage optimization procedure summarized in Algorithm 4. Note that, here, the approximation of the final design of the topology optimization algorithm by an explicit, polygonal interface actually yields a slight improvement of the objective function. The final design after Stage II of Algorithm 4 is shown in Figure 8.5.

We remark that further studies have to be done including the choice of different initial configurations since, as it is usual in gradient-based optimization, we can only guarantee local optimality of the obtained designs. However, the presented result should serve as an illustration of the flexibility of the method.

# Conclusion and Outlook

## Conclusion

This thesis was motivated by a particular class of problems, namely by the task to find an optimal distribution of ferromagnetic material within a design area of an electric motor, where the word “optimal” is understood with respect to a given objective functional. Here, we considered the situation where the ferromagnetic material exhibits nonlinear material behavior, resulting in a PDE-constrained shape optimization problem with a quasilinear PDE constraint.

For this problem, we considered the On/Off method introduced in [163] for optimal design problems in electromagnetics. We generalized the method, which is based on a discretization of the state equation, to the continuous level and showed first numerical results for a model problem.

Motivated by this topology optimization method based on the sensitivity of the objective function with respect to a perturbation of the material coefficient, we investigated the connection between this kind of sensitivities and the mathematical concept of the topological derivative. We derived an explicit formula for the topological derivative of a shape functional which is constrained by the equation of two-dimensional nonlinear magnetostatics, which, in addition to a term that is well-known from the case of linear PDE constraints, also contains a second term which accounts for the nonlinearity of the problem. Numerical tests showed, however, that in the case of electrical machines this second term of the topological derivative is negligible compared to the first term. We derived the topological derivative for both the introduction of air inside ferromagnetic material and the other way around, which allowed us to use a level set algorithm based on the topological derivative.

We also considered the same problem from the perspective of shape optimization by means of smooth perturbations of the material interfaces and derived the shape derivative for the same PDE-constrained shape optimization problem. We obtained numerical optimization results by applying this sensitivity information in a gradient-like method.

Numerical approaches to shape and topology optimization usually yield material interfaces which evolve in the course of the optimization procedure. In order to accurately resolve these interfaces, which are not necessarily aligned with the underlying triangular finite element mesh, we developed a mesh adaptation strategy. The approach consists in modifying the finite element mesh only locally in a neighborhood of the interface while ensuring that none of the interior angles of the triangular mesh becomes too large. This allowed us to show optimal order of convergence as the mesh size tends to zero. Furthermore, we investigated the issue of the condition number of the finite element matrix in terms of small angles.

Finally, we combined all of the techniques developed in this thesis and applied them to two practically interesting design optimization problems. We used a two-stage optimization procedure consisting of topology optimization in the first stage, followed by shape optimization including the interface finite element method as a post-processing in the second stage. This optimization procedure allows to obtain interesting, smooth designs without any a priori knowledge of the optimal solution.

## Possible future work

The work presented in this thesis can be extended in the following directions:

- While the setting of two-dimensional magnetostatics is widely accepted for the simulation and optimization of electrical machines, there are applications where the three-dimensional model (2.6a) should preferred. This is the case for motors whose axial dimension is small compared to its diameter.

The generalization of the On/Off method using material sensitivities to the 3D case is straightforward and has been used in the literature [161].

To the knowledge of the author, the rigorous derivation of the topological derivative for three-dimensional nonlinear magnetostatics is an open problem.

The application of Theorem 6.8 to the derivation of the shape derivative for the curl-curl problem is possible under slight modifications, as it was noted in [198].

An extension of the interface finite element method of Chapter 7 to the case of globally continuous, piecewise linear finite elements on a tetrahedral mesh seems possible. Also here, the main ingredient for showing optimal convergence rates will be to prove a maximum angle condition. In the case of three-dimensional magnetostatics, however, where Nédélec elements are used, the situation is not clear.

- Isogeometric analysis (IgA) was introduced in [115] and has gained very much popularity in many fields of engineering since then. Isogeometric analysis uses spline functions for both representing the geometry and the solution to the PDE. The big advantage of IgA over the finite element method is that spline-based geometries can be represented exactly. This approach has particularly big potential for the simulation and optimization of electrical machines since the performance of electric machines depends on the magnetic field in the air gap, which is usually of circular shape and has to be approximated in the finite element method.

Several authors have considered shape optimization in the framework of isogeometric analysis in a parametric way [114, 154], where sensitivities are computed with respect to a certain (fixed) number of control points which represent the geometry. Nonparametric shape optimization in the framework of isogeometric analysis has also been considered in [71, 85].

- The magnetostatic model (2.6) we used throughout this thesis describes very well the setting of an electric motor which rotates at constant speed. However, in the starting

phase, when the motor is accelerated from rest to a rotation at constant speed, time-dependent effects have to be considered and the preferred model is the magnetoquasistatic model (2.5). A potential way to treat this parabolic problem is to use space-time methods such as [132] which has the great advantage that it can be parallelized in time, see [86]. We mention that, for the linear magnetoquasistatic model, the shape derivative has been derived in [110].

- Another active field of research in the context of optimization of electrical equipment is the field of uncertainty quantification and robust optimization. As we mentioned in Section 2.4, the  $B$ - $H$ -curve defining the reluctivity function  $\hat{\nu}$  is not known explicitly, but interpolated or approximated from measured data, which are subject to measurement errors. Therefore, the  $B$ - $H$ -curve should be modeled in a stochastic way. Uncertainty quantification for nonlinear magnetostatics has been considered in [177, 178]. In order to obtain robust designs, these uncertainties should also be accounted for in the optimization procedure as it was done in [72, 172].
- A further possible extension of the presented methods is multi-material optimization, i.e., finding an optimal configuration consisting of three or more different materials, as it has already been considered in [3, 148, 220]. This way, it would be possible to find optimal designs consisting of not only air and ferromagnetic material, but the right position and shape of the permanent magnets could be included into the optimization process, too. Although, in general, only magnets of rectangular shape are used in practice, topology optimization might give interesting insights into other possibilities. In particular, an extension of the level set algorithm based on the topological derivative [19] to the multi-material case would be interesting.
- A comparison of the results of this thesis with different, well-established topology and shape optimization methods such as density-based methods, the phase-field method or the original level-set method based on shape sensitivities would be interesting.
- Another issue that is very relevant for practical applications is the fact that it should be possible to actually manufacture the final designs of the optimization process. In particular, highly oscillating shapes should be avoided. This can be achieved by adding the perimeter of the structure, multiplied with a weighting factor, to the cost function, which serves as a regularization of the problem and which is common practice in shape optimization. It is, however, not possible to compute the topological derivative of this additional perimeter term. In [18], a promising way to incorporate the perimeter of a structure into a topology optimization process by the level set algorithm [19] is presented. Furthermore, we mention the topology optimization approaches dealing with manufacturing constraints and thickness control [7, 148] as well as the phase-field/level set approach [213] which allows to adjust the complexity of the arising designs by a parameter. Furthermore, we mention the recent work [4] where the level set method is employed under a constraint which ensures that the final design can be produced by additive manufacturing.
- While the numerical experiments conducted in Section 4.7 showed that the second term  $J_2$  in the topological derivative for the nonlinear case is negligible compared to the first

term  $J_1$  in the case of electrical machines, it would be interesting to investigate situations where this term can become more important. This might be the case in situations where the difference between the maximal and minimal value attained by the nonlinear function  $\hat{\nu}$  is smaller.

# List of Notation

$\Omega_{air}^{ref}$	Air reference domain (fixed)
$P_0$	Gradient of unperturbed adjoint state at point $x_0$
$D$	Hold-all domain representing the electric motor together with air regions
$\mathbf{B}$	Magnetic flux density / magnetic induction
$\mathbf{H}$	Magnetic field intensity / magnetizing force
$\mathcal{J}$	Objective function
$\mathcal{M}$	Matrix in first term of topological derivative in Case I; related to polarization matrix
$\mathcal{M}^{(2)}$	Matrix in first term of topological derivative in Case I; related to polarization matrix
$\nu_0$	Magnetic reluctivity in vacuum or air (constant)
$\nu_1$	Constant magnetic reluctivity in ferromagnetic domain in simplified linear case
$\nu_\Omega$	Global reluctivity function in $D$ given the ferromagnetic subset $\Omega$
$\hat{\nu}$	Magnetic reluctivity function in ferromagnetic subdomain
$\Omega_{mag}$	Magnet areas (fixed)
$\Omega$	Subdomain of $\Omega_f^{ref}$ that is currently occupied with ferromagnetic material (subject to optimization)
$\Omega^d$	Design subdomain (fixed)
$\Omega_c$	Coil areas
$\Omega_f$	Subset of motor that is currently occupied with ferromagnetic material (subject to optimization)
$\Omega_g$	Air gap region (fixed)
$\Omega_f^{ref}$	Ferromagnetic reference domain (fixed)
$\tilde{G}$	Linearization of $\mathcal{J}$
$\tilde{G}_\psi$	Generalized topological derivative for design represented by level set function $\psi$

- $A_\Omega$  Operator representing left hand side of magnetostatic boundary value problem for given design  $\omega$
- $B_d$  Desired radial component of magnetic flux density in air gap in model problem (sine curve)
- $d\mathcal{J}(\Omega; V)$  Shape derivative of  $\mathcal{J}$  in direction of vector field  $V$
- $F$  Right hand side of magnetostatic boundary value problem comprising magnetization and impressed currents
- $G(x_0)$  Topological derivative at spatial point  $x_0$
- $H$  Variation of direct state at scale 1
- $K$  Variation of adjoint state at scale 1
- $S_W(V)$  Operator representing nonlinearity of  $T$
- $T$  Operator representing flux in nonlinear material,  $T(W) = \hat{\nu}(|W|)W$  for  $W \in \mathbb{R}^2$
- $T_t$  Transformation in direction of vector field  $V$
- $U_0$  Gradient of unperturbed direct state at point  $x_0$
- $V$  Velocity field defining shape perturbation
- $u$  State variable in 2D magnetostatics (third component of magnetic vector potential)
- THD Total harmonic distortion

# List of Figures

1.1	Illustration of XFEM and Immersed FEM . . . . .	11
2.1	Real world motor . . . . .	14
2.2	Geometry of model problem . . . . .	17
2.3	Radial component of the magnetic flux density along air gap . . . . .	20
2.4	$B$ - $H$ -curve and magnetic reluctivity $\hat{\nu}$ . . . . .	22
3.1	Evolution of design in On/Off method . . . . .	35
3.2	Final design obtained by On/Off method . . . . .	36
3.3	Objective values in the course of On/Off method . . . . .	37
4.1	Magnetic reluctivity function in ferromagnetic subdomain . . . . .	50
4.2	Illustration of topological derivative in Case I . . . . .	58
4.3	Illustration of topological derivative in Case II . . . . .	105
4.4	Numerical approximation of variation of direct state at scale 1 . . . . .	130
4.5	Numerical approximations of variation of adjoint state at scale 1 . . . . .	130
4.6	First term in topological derivative in Case I . . . . .	130
4.7	Second term in topological derivative in Case I . . . . .	131
4.8	First term in topological derivative in Case II . . . . .	131
4.9	Second term in topological derivative in Case II . . . . .	132
4.10	Evolution of design by level set algorithm based on topological derivative . . . . .	134
4.11	Objective values in the course of level set algorithm based on topological derivative . . . . .	135
5.1	Illustration of difference between sensitivities in On/Off-Method and topological derivative . . . . .	139
6.1	Illustration of shape perturbation . . . . .	146
6.2	Evolution of design in shape optimization algorithm . . . . .	157
6.3	Evolution of objective function in shape optimization algorithm . . . . .	159
7.1	Different configurations of patches depending on position of interface . . . . .	167
7.2	Examples of constellations that are excluded . . . . .	170
7.3	Illustration of notation . . . . .	172
7.4	Material coefficient without and with interface technique . . . . .	177
7.5	Reference patch together with two adjacent patches . . . . .	182
7.6	Possible different configurations of adjacent patches . . . . .	183

7.7	Interface problem with circular interface . . . . .	184
7.8	Horizontal interface approaching a line . . . . .	185
8.1	Evolution of design for first practical problem by combined topology and shape optimization with interface treatment . . . . .	191
8.2	Radial component of magnetic flux density in the course of the combined algorithm . . . . .	192
8.3	Final designs for first practical problem by combined algorithm . . . . .	193
8.4	Evolution of design for second practical problem by combined topology and shape optimization with interface treatment . . . . .	194
8.5	Final design for second practical problem . . . . .	195

# Bibliography

- [1] M. Alb, P. Alotto, C. Magele, W. Renhart, K. Preis, and B. Trapp. Firefly algorithm for finding optimal shapes of electromagnetic devices. *IEEE Transactions on Magnetics*, 52(3):1–4, March 2016.
- [2] G. Allaire. *Shape optimization by the homogenization method*. Applied mathematical sciences. Springer, New York, 2002.
- [3] G. Allaire, C. Dapogny, G. Delgado, and G. Michailidis. Multi-phase structural optimization via a level set method. *ESAIM: COCV*, 20(2):576–611, 2014.
- [4] G. Allaire, C. Dapogny, A. Faure, and G. Michailidis. Shape optimization of a layer by layer mechanical constraint for additive manufacturing. preprint, Nov. 2016.
- [5] G. Allaire, F. de Gournay, F. Jouve, and A.-M. Toader. Structural optimization using topological and shape sensitivity via a level set method. *Control and Cybernetics*, 34(1):59–80, 2005.
- [6] G. Allaire and F. Jouve. Coupling the level set method and the topological gradient in structural optimization. In Springer, editor, *IUTAM Symposium on Topological Design Optimization of Structures, Machines and Materials*, pages 3–12, 2006.
- [7] G. Allaire, F. Jouve, and G. Michailidis. Thickness control in structural optimization via a level set method. *Structural and Multidisciplinary Optimization*, 53(6):1349–1382, 2016.
- [8] G. Allaire, F. Jouve, and A.-M. Toader. Structural optimization using sensitivity analysis and a level-set method. *Journal of Computational Physics*, 194(1):363 – 393, 2004.
- [9] H. W. Alt. *Lineare Funktionalanalysis. Eine anwendungsorientierte Einführung*. Berlin: Springer, 5th revised ed. edition, 2006.
- [10] H. Ammari. *An Introduction to Mathematics of Emerging Biomedical Imaging*. Springer, 2008.
- [11] H. Ammari and H. Kang. *Polarization and Moment Tensors*. Springer-Verlag New York, 2007.
- [12] C. Amrouche, V. Girault, and J. Giroire. Weighted Sobolev spaces for Laplace's equation in  $\mathbf{R}^n$ . *J. Math. Pures Appl.*, 73:579–606, 1994.

- [13] S. Amstutz. The topological asymptotic for the Navier-Stokes equations. *ESAIM: COCV*, 11(3):401–425, 2005.
- [14] S. Amstutz. Sensitivity analysis with respect to a local perturbation of the material property. *Asymptotic analysis*, 49(1), 2006.
- [15] S. Amstutz. Topological sensitivity analysis for some nonlinear PDE systems. *Journal de Mathématiques Pures et Appliquées*, 85(4):540 – 557, 2006.
- [16] S. Amstutz. Analysis of a level set method for topology optimization. *Optimization Methods and Software - Advances in Shape and Topology Optimization: Theory, Numerics and New Application Areas*, 26(4-5):555–573, 2011.
- [17] S. Amstutz. Connections between topological sensitivity analysis and material interpolation schemes in topology optimization. *Structural and Multidisciplinary Optimization*, 43(6):755–765, 2011.
- [18] S. Amstutz. Regularized perimeter for topology optimization. *SIAM Journal on Control and Optimization*, 51(3):2176–2199, 2013.
- [19] S. Amstutz and H. Andrä. A new algorithm for topology optimization using a level-set method. *Journal of Computational Physics*, 216(2):573–588, 2006.
- [20] S. Amstutz and A. Bonnafé. Topological derivatives for a class of quasilinear elliptic equations. *Journal de mathématiques pures et appliquées*.
- [21] S. Amstutz, C. Dapogny, and A. Ferrer. A consistent relaxation of optimal design problems for coupling shape and topological derivatives. Submitted, 2016.
- [22] S. Amstutz and A. A. Novotny. Topological optimization of structures subject to von Mises stress constraints. *Structural and Multidisciplinary Optimization*, 41(3):407–420, 2010.
- [23] S. Amstutz and A. A. Novotny. Topological asymptotic analysis of the Kirchhoff plate bending problem. *ESAIM: COCV*, 17(3):705–721, 2011.
- [24] S. Amstutz, A. A. Novotny, and N. V. Goethem. Topological sensitivity analysis for elliptic differential operators of order 2m. *Journal of Differential Equations*, 256(4):1735 – 1770, 2014.
- [25] T. Apel. *Anisotropic Finite Elements: Local Estimates and Applications*. Teubner, Stuttgart, 1999.
- [26] R. Arumugam, J. Lindsay, D. Lowther, and R. Krishnan. Magnetic field analysis of a switched reluctance motor using a two dimensional finite element model. *IEEE Trans. Magn.*, 21(5):1883–1885, 1985.
- [27] H. Azegami, M. Shimoda, E. Katamine, and Z. C. Wu. A domain optimization technique for elliptic boundary value problems. *WIT Transactions on The Built Environment*, 14, 1995.

- [28] H. Azegami and K. Takeuchi. A smoothing method for shape optimization: Traction method using the robin condition. *International Journal of Computational Methods*, 03(01):21–33, 2006.
- [29] I. Babuska. The finite element method for elliptic equations with discontinuous coefficients. *Computing*, 5:207–213, 1970.
- [30] I. Babuska and A. Aziz. On the angle condition in the finite element method. *SIAM J. Numer. Anal.*, 13:214–226, 1976.
- [31] I. Babuska and J. M. Melenk. The partition of unity method. *International Journal of Numerical Methods in Engineering*, 40:727–758, 1996.
- [32] R. E. Bank and L. R. Scott. On the conditioning of finite element equations with highly refined meshes. *SIAM Journal on Numerical Analysis*, 26(6):1383–1394, 1989.
- [33] S. Basting and R. Prignitz. An interface-fitted subspace projection method for finite element simulations of particulate flows. *Computer Methods in Applied Mechanics and Engineering*, 267:133–149, 2013.
- [34] S. Basting, A. Quaini, S. Canic, and R. Glowinski. An extended ale method for fluid-structure interaction problems with large structural displacements. Numerical Analysis and Scientific Computing Preprint Seria, Oct. 2014.
- [35] S. Basting and M. Weismann. A hybrid level set-front tracking finite element approach for fluid-structure interaction and two-phase flow applications. *J. Comput. Phys.*, 255(C):228–244, Dec. 2013.
- [36] S. Basting and M. Weismann. A hybrid level set/front tracking approach for finite element simulations of two-phase flows. *Journal of Computational and Applied Mathematics*, 270:471 – 483, 2014. Fourth International Conference on Finite Element Methods in Engineering and Sciences (FEMTEC 2013).
- [37] T. Belytschko, R. Gracie, and G. Ventura. A review of extended/generalized finite element methods for material modeling. *Model. Simul. Mater. Sci. Eng.*, 17(4), 2009.
- [38] M. P. Bendsøe. Optimal shape design as a material distribution problem. *Structural Optimization*, 1(4):193–202, 1989.
- [39] M. P. Bendsøe. *Optimization of structural topology, shape and material*. Springer, Berlin Heidelberg, 1995.
- [40] M. P. Bendsøe, A. Ben-Tal, and J. Zowe. Optimization methods for truss geometry and topology design. *Structural and Multidisciplinary Optimization*, 7(3):141–159, Apr. 1994.
- [41] M. P. Bendsoe and N. Kikuchi. Generating optimal topologies in structural design using a homogenization method. *Comput. Methods Appl. Mech. Eng.*, 71(2):197–224, Nov. 1988.
- [42] M. P. Bendsøe and O. Sigmund. Material interpolation schemes in topology optimization. *Archive of Applied Mechanics*, 69(9):635–654, 1999.

- [43] M. P. Bendsøe and O. Sigmund. *Topology Optimization: Theory, Methods and Applications*. Springer, Berlin, 2003.
- [44] M. Berggren. A unified discrete-continuous sensitivity analysis method for shape optimization. In *Applied and numerical partial differential equations*, volume 15 of *Comput. Methods Appl. Sci.*, pages 25–39. Springer, New York, 2010.
- [45] A. Binder. *Elektrische Maschinen und Antriebe: Grundlagen, Betriebsverhalten*. Springer-Lehrbuch. Springer, 2012.
- [46] A. Bonnafé. *Développements asymptotiques topologiques pour une classe d'équations elliptiques quasi-linéaires. Estimations et développements asymptotiques de p-capacités de condensateur. Le cas anisotrope du segment*. PhD thesis, Université de Toulouse, France, 2013.
- [47] D. Braess. *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*. Cambridge University Press, Cambridge, 3 edition, 004 2007.
- [48] J. H. Bramble and J. T. King. A robust finite element method for nonhomogeneous Dirichlet problems in domains with curved boundaries. *Math. Comp.*, 63(207):1–17, 1994.
- [49] G. Bramerdorfer, P. Gangl, A. Fohler, U. Langer, and W. Amrhein. Determination of the cogging torque sensitivity of brushless permanent magnet machines due to changes of the material characteristics of ferromagnetic components. In *7th IET International Conference on Power Electronics, Machines and Drives (PEMD 2014)*, Manchester, pages 1–6, 2014.
- [50] P. Browne. *Topology Optimization of Linear Elastic Structures*. PhD thesis, University of Bath, 2013.
- [51] D. Bucur and G. Buttazzo. *Variational Methods in Shape Optimization Problems*. Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser Boston, 2006.
- [52] C. Bui, C. Dapogny, and P. Frey. An accurate anisotropic adaptation method for solving the level set advection equation. *International Journal for Numerical Methods in Fluids*, 70(7):899–922, 2012.
- [53] M. Burger, B. Hackl, and W. Ring. Incorporating topological derivatives into level set methods. *Journal of Computational Physics*, 194(1):344–362, 2004.
- [54] M. Burger and S. J. Osher. A survey on level set methods for inverse problems and optimal design. *European Journal of Applied Mathematics*, 16(02):263–301, 2005.
- [55] M. Burger and R. Stainko. Phase-field relaxation of topology optimization with local stress constraints. *SIAM J. Control Optim.*, 45(4):1447–1466, 2006.
- [56] E. Burman. Ghost penalty. *Comptes Rendus Mathematique*, 348:1217–1220, 2010.

- [57] E. Burman, S. Claus, P. Hansbo, M. G. Larson, and A. Massing. Cutfem: Discretizing geometry and partial differential equations. *International Journal for Numerical Methods in Engineering*, 104(7):472–501, 2015.
- [58] E. Burman, D. Elfverson, P. Hansbo, M. G. Larson, and K. Larsson. Shape Optimization Using the Cut Finite Element Method. *ArXiv e-prints*, Nov. 2016.
- [59] J. W. Cahn and J. E. Hilliard. *Free Energy of a Nonuniform System. I. Interfacial Free Energy*, pages 29–38. John Wiley & Sons, Inc., 2013.
- [60] F. Campelo, J. Ramírez, and H. Igarashi. A survey of topology optimization in electromagnetics: considerations and current trends. 2010.
- [61] A. Canelas, A. A. Novotny, and J. R. Roche. Topology design of inductors in electromagnetic casting using level-sets and second order topological derivatives. *Structural and Multidisciplinary Optimization*, 50(6):1151–1163, 2014.
- [62] J. Céa. Conception optimale ou identification de formes, calcul rapide de la dérivée directionnelle de la fonction cout. *Math.Mod. and Num.Anal.*, 20:371–402, 1986.
- [63] J. Céa, S. Garreau, P. Guillaume, and M. Masmoudi. The shape and topological optimizations connection. *Computer Methods in Applied Mechanics and Engineering*, 188(4):713 – 726, 2000.
- [64] J. Cea, A. Gioan, and J. Michel. *Adaptation de la méthode du gradient à un problème d'identification de domaine*, pages 391–402. Springer Berlin Heidelberg, Berlin, Heidelberg, 1974.
- [65] S. Chaabane, M. Masmoudi, and H. Meftahi. Topological and shape gradient strategy for solving geometrical inverse problems. *Journal of Mathematical Analysis and Applications*, 400(2):724 – 742, 2013.
- [66] J. Choi, K. Izui, S. Nishiwaki, A. Kawamoto, and T. Nomura. Rotor pole design of IPM motors for a sinusoidal air-gap flux density distribution. *Structural and Multidisciplinary Optimization*, 46(3):445–455, 2012.
- [67] P. Christensen and A. Klarbring. *An Introduction to Structural Optimization*. Solid Mechanics and Its Applications. Springer Netherlands, 2008.
- [68] A. N. Christiansen, M. Nobel-Jørgensen, N. Aage, O. Sigmund, and J. A. Bærentzen. Topology optimization using an explicit interface representation. *Structural and Multidisciplinary Optimization*, 49(3):387–399, 2014.
- [69] R. Correa and A. Seeger. Directional derivative of a minimax function. *Nonlinear Anal.*, 9(1):13–22, 1985.
- [70] F. Cupertino, G. M. Pellegrino, E. Armando, and C. Gerada. A syr and ipm machine design methodology assisted by optimization algorithms. In *2012 IEEE Energy Conversion Congress and Exposition (ECCE)*, pages 3686–3691, Sept 2012.

- [71] A.-V. V. Daniela Fußeder, Bernd Simeon. Fundamental aspects of shape optimization in the context of isogeometric analysis. *Computer Methods in Applied Mechanics and Engineering*, 286:313–331, 2015.
- [72] F. de Gournay, G. Allaire, and F. Jouve. Shape and topology optimization of the robust compliance via the level set method. *ESAIM: COCV*, 14(1):43–70, 2008.
- [73] K. Deb. *Multi-objective optimization using evolutionary algorithms*. Wiley India Pvt. Limited, 2010.
- [74] M. C. Delfour and J.-P. Zolésio. Shape sensitivity analysis via min max differentiability. *SIAM J. Control Optim.*, 26(4):834–862, 1988.
- [75] M. C. Delfour and J.-P. Zolésio. *Shapes and geometries*, volume 22 of *Advances in Design and Control*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 2011. Metrics, analysis, differential calculus, and optimization.
- [76] P. Deuflhard. *Systems of Equations: Global Newton Methods*, pages 109–172. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.
- [77] M. B. Dühring, J. S. Jensen, and O. Sigmund. Acoustic design by topology optimization. *Journal of Sound and Vibration*, 317(3–5):557 – 575, 2008.
- [78] H. A. Eschenauer, V. V. Kobelev, and A. Schumacher. Bubble method for topology and shape optimization of structures. *Structural optimization*, 8(1):42–51, 1994.
- [79] A. Evgrafov, G. Pingen, and K. Maute. Topology optimization of fluid domains: kinetic theory approach. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift fr Angewandte Mathematik und Mechanik*, 88(2):129–141, 2008.
- [80] R. A. Feijóo, A. A. Novotny, C. Padra, and E. O. Taroco. The topological-shape sensitivity analysis and its applications in optimal design. In *Mecánica Computacional Vol XXI*, 2002.
- [81] E. Frank. Free-form optimization of electric machines based on shape derivatives. Master’s thesis, Johannes Kepler University Linz, 2010.
- [82] S. Frei. *Eulerian finite element methods for interface problems and fluid-structure-interactions*. PhD thesis, Universität Heidelberg, 2016.
- [83] S. Frei and T. Richter. A locally modified parametric finite element method for interface problems. *SIAM J. Numer. Anal.*, 52(5):2315–2334, 2014.
- [84] T.-P. Fries and T. Belytschko. The extended/generalized finite element method: An overview of the method and its applications. *Int. J. Numer. Meth. Eng.*, 84(3):253–304, 2010.
- [85] D. Fußeder and B. Simeon. Algorithmic aspects of isogeometric shape optimization. In B. Jüttler and B. Simeon, editors, *Isogeometric Analysis and Applications 2014*, volume 107 of *Lecture Notes in Computational Science and Engineering*, chapter Algorithmic Aspects of Isogeometric Shape Optimization, pages 183–207. Springer International Publishing, 2015.

- [86] M. J. Gander and M. Neumüller. Analysis of a New Space-Time Parallel Multigrid Algorithm for Parabolic Problems. *ArXiv e-prints*, Nov. 2014.
- [87] P. Gangl, S. Amstutz, and U. Langer. Topology optimization of electric motor using topological derivative for nonlinear magnetostatics. *IEEE Transactions on Magnetics*, 52(3):1–4, March 2016.
- [88] P. Gangl and U. Langer. Topology optimization of electric machines based on topological sensitivity analysis. *Computing and Visualization in Science*, 2014.
- [89] P. Gangl and U. Langer. A Local Mesh Modification Strategy for Interface Problems with Application to Shape and Topology Optimization. *ArXiv e-prints*, Sept. 2016.
- [90] P. Gangl, U. Langer, A. Laurain, H. Meftahi, and K. Sturm. Shape optimization of an electric motor subject to nonlinear magnetostatics. *SIAM Journal on Scientific Computing*, 37(6):B1002–B1025, 2015.
- [91] H. Garcke and C. Hecht. Applying a phase field approach for shape optimization of a stationary navier-stokes flow. *ESAIM: COCV*, 22(2):309–337, 2016.
- [92] H. Garcke, C. Hecht, M. Hinze, and C. Kahle. Numerical approximation of phase field based shape and topology optimization for fluids. *SIAM Journal on Scientific Computing*, 37(4):A1846–A1871, 2015.
- [93] D. Gilbarg and N. Trudinger. *Elliptic partial differential equations of second order*. Grundlehren der mathematischen Wissenschaften. Springer, 1998.
- [94] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition, 1989.
- [95] B. Hackl. *Geometry Variations, Level Sets and Phase-field Methods for Perimeter Regularized Geometric Inverse Problems*. PhD thesis, Johannes Kepler University Linz, 2006.
- [96] J. Hadamard. *Mémoire sur le problème d'analyse relativ à l'équilibre des plaques élastiques encastrées*. Mémoires présentés par divers savants à l'Académie des sciences de l'Institut de France: Extrait. Imprimerie nationale, 1907.
- [97] L. Hägg and E. Wadbro. Nonlinear filters in topology optimization: existence of solutions and efficient implementation for minimum compliance problems. *Structural and Multidisciplinary Optimization*, pages 1–12, 2016.
- [98] A. Hannukainen, S. Korotov, and M. Krížek. The maximum angle condition is not necessary for convergence of the finite element method. *Numerische Mathematik*, 120(1):79–88, 2012.
- [99] 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.
- [100] J. Haslinger and R. Mäkinen. *Introduction to Shape Optimization*. Society for Industrial and Applied Mathematics, 2003.

- [101] J. Haslinger and P. Neittaanmäki. *Finite Element Approximation for Optimal Shape, Material and Topology Design*. Wiley, 1996.
- [102] E. Hassan. *Topology Optimization of Antennas and Waveguide Transitions*. PhD thesis, Umeå University, 2015.
- [103] E. Hassan, D. Noreland, E. Wadbro, and M. Berggren. Topology optimization of coaxial-to-waveguide transitions. Submitted.
- [104] B. Hassani and E. Hinton. A review of homogenization and topology optimization III – topology optimization using optimality criteria. *Computers & Structures*, 69(6):739 – 756, 1998.
- [105] L. He, C.-Y. Kao, and S. Osher. Incorporating topological derivatives into shape derivatives based level set methods. *J. Comput. Phys.*, 225(1):891–909, July 2007.
- [106] B. Heise. Analysis of a fully discrete finite element method for a nonlinear magnetic field problem. *SIAM J. Numer. Anal.*, 31(3):745–759, 1994.
- [107] A. Henrot and M. Pierre. *Variation et optimisation de formes*, volume 48 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer, Berlin, 2005. Une analyse géométrique. [A geometric analysis].
- [108] M. Hintermüller. *13. A Combined Shape-Newton and Topology Optimization Technique in Real-Time Image Segmentation*. 2007.
- [109] M. Hintermüller, A. Laurain, and A. A. Novotny. Second-order topological expansion for electrical impedance tomography. *Advances in Computational Mathematics*, 36(2):235–265, 2012.
- [110] M. Hintermüller, A. Laurain, and I. Yousept. Shape sensitivities for an inverse problem in magnetic induction tomography based on the eddy current model. *Inverse Problems*, 31(6):065006, 2015.
- [111] R. Hiptmair, A. Paganini, and S. Sargheini. Comparison of approximate shape gradients. *BIT*, 55(2):459–485, 2015.
- [112] J. Hoffman, B. Holm, and T. Richter. The locally adapted patch finite element method for interface problems on triangular meshes. *ArXiv e-prints*, Sept. 2016.
- [113] L. Hörmander. *Linear Partial Differential Operators*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 1976.
- [114] B. Horn and S. Ulbrich. Shape optimization for contact problems based on isogeometric analysis. *Journal of Physics: Conference Series*, 734(3):032008, 2016.
- [115] T. Hughes, J. Cottrell, and Y. Bazilevs. Isogeometric analysis: Cad, finite elements, nurbs, exact geometry and mesh refinement. *Computer Methods in Applied Mechanics and Engineering*, 194(39–41):4135 – 4195, 2005.
- [116] N. Ida and J. Bastos. *Electromagnetics and Calculation of Fields*. Lecture Notes in Statistics; 120. Springer New York, 1997.

- [117] M. Iguernane, S. A. Nazarov, J.-R. Roche, J. Sokółowski, and K. Szulc. Topological derivatives for semilinear elliptic equations. *Int. J. Appl. Math. Comput. Sci.*, 19(2):191–205, 2009.
- [118] K. Ito, K. Kunisch, and G. H. Peichl. Variational approach to shape derivatives. *ESAIM Control Optim. Calc. Var.*, 14(3):517–539, 2008.
- [119] H. Ji, J. Chen, and Z. Li. A symmetric and consistent immersed finite element method for interface problems. *J. Sci. Comput.*, 61:533–557, 2014.
- [120] M. Jung and U. Langer. *Methode der finiten Elemente für Ingenieure: Eine Einführung in die numerischen Grundlagen und Computersimulation*. Springer-Vieweg-Verlag, Darmstadt, 2013. 2., überarb. u. erw. Aufl., 639 S.
- [121] B. Kaltenbacher, M. Kaltenbacher, and S. Reitzinger. Identification of nonlinear  $b-h$  curves based on magnetic field computations and multigrid methods for ill-posed problems. *European Journal of Applied Mathematics*, 14(1):15–38, 02 2003.
- [122] M. Kaltenbacher. *Numerical Simulation of Mechatronic Sensors and Actuators*. Springer, Berlin, 2nd edition edition, 2007.
- [123] H. Kasumba and Kunisch. On shape sensitivity analysis of the cost functional without shape sensitivity of the state variable. *Control and Cybernet.*, 14(4):989–1017, 2011.
- [124] H. Kasumba and K. Kunisch. On computation of the shape hessian of the cost functional without shape sensitivity of the state variable. *Journal of Optimization Theory and Applications*, pages 1–26, 2014.
- [125] A. Khoei. *Extended Finite Element Method: Theory and Applications*. Wiley Series in Computational Mechanics. Wiley, 2015.
- [126] J. Kolehmainen. Synchronous reluctance motor with form blocked rotor. *IEEE Transactions on Energy Conversion*, 25(2):450–456, June 2010.
- [127] J. Kolota and S. Steien. Analysis of 2D and 3D finite element approach of a switched reluctance motor. *Przeglad Elektrotechniczny (Electrical Review)*,, 87(12a):188–190, 2011.
- [128] A. Kost. *Numerische Methoden in der Berechnung elektromagnetischer Felder*. Springer-Verlag Berlin Heidelberg, 1994.
- [129] E. Kuci, F. Henrotte, P. Duysinx, P. Dular, and C. Geuzaine. Design sensitivity analysis for shape optimization of nonlinear magnetostatic systems. *IEEE Transactions on Magnetics*, 52(3):1–4, March 2016.
- [130] K. Kundert. *The Designer’s Guide to Spice and Spectre*. The Designer’s Guide Book Series. Springer US, 2006.
- [131] A. Kuzmin, M. Luisier, and O. Schenk. Fast methods for computing selected elements of the greens function in massively parallel nanoelectronic device simulations. In F. Wolf, B. Mohr, and D. Mey, editors, *Euro-Par 2013 Parallel Processing*, volume 8097 of *Lecture Notes in Computer Science*, pages 533–544. Springer Berlin Heidelberg, 2013.

- [132] U. Langer, S. E. Moore, and M. Neumüller. Space–time isogeometric analysis of parabolic evolution problems. *Computer Methods in Applied Mechanics and Engineering*, 306:342 – 363, 2016.
- [133] I. Larrabide, R. Feijóo, A. Novotny, and E. Taroco. Topological derivative: A tool for image processing. *Computers & Structures*, 86(13–14):1386 – 1403, 2008. Structural Optimization.
- [134] A. Laurain and K. Sturm. Distributed shape derivative via averaged adjoint method and applications. *ESAIM: M2AN*, 50(4):1241–1267, 2016.
- [135] A. Laurain and S. W. Walker. Droplet footprint control. *SIAM Journal on Control and Optimization*, 53(2):771–799, 2015.
- [136] C. Lehrenfeld. *On a Space-Time Extended Finite Element Method for the Solution of a Class of Two-Phase Mass Transport Problems*. Dr., RWTH Aachen University, Aachen, 2015. Aachen, Techn. Hochsch., Diss., 2015.
- [137] C. Lehrenfeld and A. Reusken. Optimal preconditioners for nitsche-xfem discretizations of interface problems. *Numerische Mathematik*, pages 1–20, 2016.
- [138] R. J. Leveque and Z. Li. The immersed interface method for elliptic equations with discontinuous coefficients and singular sources. *SIAM Journal on Numerical Analysis*, 31(4):1019–1044, 1994.
- [139] Z. Li. The immersed interface method using a finite element formulation. *Appl. Num. Math.*, 27:253–267, 1998.
- [140] Z. Li, T. Lin, and X. Wu. New cartesian grid methods for interface problems using the finite element formulation. *Numer. Math.*, 96:61–98, 2003.
- [141] J. Malý and W. Ziemer. *Fine Regularity of Solutions of Elliptic Partial Differential Equations*. Mathematical surveys and monographs. American Mathematical Society, 1997.
- [142] A. Marrocco and O. Pironneau. Optimum design with lagrangian finite elements: Design of an electromagnet. *Computer Methods in Applied Mechanics and Engineering*, 15(3):277 – 308, 1978.
- [143] M. Masmoudi. Topological optimization with dirichlet condition. In M. J. et J. Jaffré, editor, *Picof*, pages 121–127, 1998.
- [144] M. Masmoudi, J. Pommier, and B. Samet. The topological asymptotic expansion for the Maxwell equations and some applications. *Inverse Problems*, 31:545–564, 2005.
- [145] K. Maute and O. Sigmund. Topology optimization approaches: A comparative review. *Structural and Multidisciplinary Optimization*, 6, 2013.
- [146] J. C. Maxwell. *A Treatise on Electricity and Magnetism*:, volume 1. Cambridge University Press, Cambridge, 001 1873.
- [147] M. Melenk. *On generalized finite element methods*. PhD thesis, University of Maryland, 1995.

- [148] G. Michailidis. *Manufacturing Constraints and Multi-Phase Shape and Topology Optimization via a Level-Set Method*. Theses, Ecole Polytechnique X, Jan. 2014.
- [149] A. Michell. LVIII. the limits of economy of material in frame-structures. *Philosophical Magazine Series 6*, 8(47):589–597, 1904.
- [150] M. K. Misztal and J. A. Bærentzen. Topology-adaptive interface tracking using the deformable simplicial complex. *ACM Trans. Graph.*, 31(3):24:1–24:12, June 2012.
- [151] N. Moës, J. Dolbow, and T. Belytschko. A finite element method for crack growth without remeshing. *Int. J. Numer. Meth. Engng.*, 46(1):131–150, 1999.
- [152] F. Murat and J. Simon. *Quelques résultats sur le contrôle par un domaine géométrique*. VI Laboratoire d’Analyse Numérique, 1974.
- [153] A. Nägel, V. Schulz, M. Siebenborn, and G. Wittum. Scalable shape optimization methods for structured inverse modeling in 3d diffusive processes. *Computing and Visualization in Science*, 17(2):79–88, 2015.
- [154] D. Nguyen, A. Evgrafov, and J. Gravesen. *Isogeometric Analysis and Shape Optimization in Electromagnetism*. PhD thesis, Technical University of Denmark, 2012.
- [155] J. Nitsche. Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind. *Abh. Math. Sem. Univ. Hamburg*, 36:9–15, 1971. Collection of articles dedicated to Lothar Collatz on his sixtieth birthday.
- [156] J. Nocedal and S. Wright. *Numerical Optimization*. Springer-Verlag New York, 1999.
- [157] L. Noël, L. V. Miegroet, and P. Duysinx. Analytical sensitivity analysis using the extended finite element method in shape optimization of bimaterial structures. *International Journal for Numerical Methods in Engineering*, 107(8):669–695, 2016. nme.5181.
- [158] A. A. Novotny, R. A. Feijóo, E. Taroco, M. Masmoudi, and C. Padra. Topological sensitivity analysis for a nonlinear case: the p-Poisson problem. In *6th World Congress on Structural and Multidisciplinary Optimization*. 2005.
- [159] A. A. Novotny and J. Sokołowski. *Topological Derivatives in Shape Optimization*. Springer-Verlag Berlin Heidelberg, 2013.
- [160] A. Novruzi and J. R. Roche. Newton’s method in shape optimisation: A three-dimensional case. *BIT Numerical Mathematics*, 40(1):102–120, 2000.
- [161] Y. Okamoto, K. Akiyama, and N. Takahashi. 3-D topology optimization of single-pole-type head by using design sensitivity analysis. *IEEE Transactions on Magnetics*, 42(4):1087–1090, 2006.
- [162] Y. Okamoto, M. Otake, and N. Takahashi. Magnetic shield design of perpendicular magnetic recording head by using topology optimization technique. *IEEE Trans. Magn.*, 41(5):1788–1791, 2005.

- [163] Y. Okamoto and N. Takahashi. A novel topology optimization of nonlinear magnetic circuit using ON/OFF method. *IEEJ Transactions on Fundamentals and Materials*, 125(6):549–553, 2005.
- [164] S. Osher and R. Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*. Applied Mathematical Sciences. Springer New York, 2002.
- [165] S. Osher and J. A. Sethian. Fronts propagating with curvature dependent speed: Algorithms based on hamilton-jacobi formulations. *Journal of Computational Physics*, 79(1):12–49, 1988.
- [166] S. J. Osher and F. Santosa. Level set methods for optimization problems involving geometry and constraints - i. frequencies of a two-density inhomogeneous drum. *J. Comput. Phys*, 171:272–288, 2001.
- [167] O. Pantz. Sensibilité de l'équation de la chaleur aux sauts de conductivité. *C. R. Math. Acad. Sci. Paris*, 341(5):333–337, 2005.
- [168] C. Pechstein. Multigrid-Newton-methods for nonlinear magnetostatic problems. Master's thesis, Johannes Kepler University Linz, 2004.
- [169] C. Pechstein and B. Jüttler. Monotonicity-preserving interproximation of B-H-curves. *J. Comp. App. Math.*, 196:45–57, 2006.
- [170] O. Pironneau. On optimum design in fluid mechanics. *Journal of Fluid Mechanics*, 64(1):97–110, 1974.
- [171] W. Prager and G. Rozvany. Optimization of the structural geometry. In B. A.R. and C. L., editors, *Dynamical systems*, pages 265–293. Academic Press New York, 1977.
- [172] P. Putek, K. Gausling, A. Bartel, K. M. Gawrylczyk, E. J. W. t. Maten, R. Pulch, and M. Günther. *Robust Topology Optimization of a Permanent Magnet Synchronous Machine Using Multi-Level Set and Stochastic Collocation Methods*, pages 233–242. Springer International Publishing, Cham, 2016.
- [173] P. Putek, P. Paplicki, and R. Pałka. Topology optimization of rotor poles in a permanent-magnet machine using level set method and continuum design sensitivity analysis. *COMPEL - The international journal for computation and mathematics in electrical and electronic engineering*, 33(3):711–728, 2014.
- [174] P. Putek, R. Pulch, A. Bartel, E. J. W. ter Maten, M. Günther, and K. M. Gawrylczyk. Shape and topology optimization of a permanent-magnet machine under uncertainties. *Journal of Mathematics in Industry*, 6(1):11, 2016.
- [175] O. Querin, G. Steven, and Y. Xie. Evolutionary structural optimisation (ESO) using a bidirectional algorithm. *Engineering Computations*, 15(8):1031–1048, 1998.
- [176] B. Rivière. *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations*. Society for Industrial and Applied Mathematics, 2008.
- [177] U. Römer. *Numerical Approximation of the Magnetoquasistatic Model with Uncertainties: Applications in Magnet Design*. Springer International Publishing, 2016.

- [178] U. Römer, S. Schöps, and T. Weiland. Stochastic modeling and regularity of the non-linear elliptic curl–curl equation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):952–979, 2016.
- [179] M. Rossow and J. Taylor. A finite element method for the optimal design of variable thickness sheets. *AIAA Journal*, 11(11):1566–1569, 1973.
- [180] G. Rozvany. Grillages of maximum strength and maximum stiffness. *International Journal of Mechanical Sciences*, 14(10):651 – 666, 1972.
- [181] G. Rozvany. Optimal load transmission by flexure. *Computer Methods in Applied Mechanics and Engineering*, 1(3):253 – 263, 1972.
- [182] G. I. N. Rozvany. A critical review of established methods of structural topology optimization. *Structural and Multidisciplinary Optimization*, 37(3):217–237, 2009.
- [183] O. Schenk, M. Bollhöfer, and R. A. Römer. On large-scale diagonalization techniques for the anderson model of localization. *SIAM Rev.*, 50(1):91–112, Feb. 2008.
- [184] O. Schenk, A. Wächter, and M. Hagemann. Matching-based preprocessing algorithms to the solution of saddle-point problems in large-scale nonconvex interior-point optimization. *Computational Optimization and Applications*, 36(2-3):321–341, 2007.
- [185] M. Schiffer and G. Szegö. Virtual mass and polarization. *Trans. Amer. Math. Soc.*, pages 130–205, 1949.
- [186] J. Schöberl. Netgen an advancing front 2D/3D-mesh generator based on abstract rules. *Computing and Visualization in Science*, 1(1):41–52, 1997.
- [187] V. Schulz and M. Siebenborn. Computational comparison of surface metrics for pde constrained shape optimization. *Comp. Meth. Appl. Math.*, 16(3), 2016.
- [188] V. H. Schulz, M. Siebenborn, and K. Welker. *Towards a Lagrange–Newton Approach for PDE Constrained Shape Optimization*, pages 229–249. Springer International Publishing, Cham, 2015.
- [189] A. Schumacher. *Topologieoptimierung von Bauteilstrukturen unter Verwendung von Lochpositionierungskriterien*. PhD thesis, Univ. Siegen, 1995.
- [190] H. Schwarz and N. Köckler. *Numerische Mathematik*. Vieweg+Teubner Verlag, 2011.
- [191] J. Sethian. *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 1999.
- [192] O. Sigmund and K. Maute. Topology optimization approaches: A comparative review. *Structural and Multidisciplinary Optimization*, 48(6):1031–1055, 2013.
- [193] O. Sigmund and J. Petersson. Numerical instabilities in topology optimization: A survey on procedures dealing with checkerboards, mesh-dependencies and local minima. *Structural Optimization*, 16(1):68–75, 1998.

- [194] J. Simon. Differentiation with respect to the domain in boundary value problems. *Numerical Functional Analysis and Optimization*, 2(7-8):649–687, 1980.
- [195] J. Sokołowski and A. Zochowski. On the topological derivative in shape optimization. *SIAM Journal on Control and Optimization*, 37(4):1251–1272, 1999.
- [196] J. Sokołowski and J.-P. Zolésio. *Introduction to shape optimization*, volume 16 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1992. Shape sensitivity analysis.
- [197] R. Stainko. *Advanced Multilevel Techniques to Topology Optimization*. PhD thesis, Johannes Kepler University Linz, 2006.
- [198] K. Sturm. Minimax Lagrangian approach to the differentiability of nonlinear PDE constrained shape functions without saddle point assumption. *SIAM Journal on Control and Optimization*, 53(4):2017–2039, 2015.
- [199] K. Sturm. Convergence of Newton’s method in shape optimisation via approximate normal functions. <https://arxiv.org/abs/1608.02699>, August 2016.
- [200] K. Sturm, M. Hintermüller, and D. Hömberg. Distortion compensation as a shape optimisation problem for a sharp interface model. *Computational Optimization and Applications*, 64(2):557–588, 2016.
- [201] K. Svanberg. The method of moving asymptotes – a new method for structural optimization. *International Journal for Numerical Methods in Engineering*, 24(2):359–373, 1987.
- [202] N. Takahashi, S. Nakazaki, and D. Miyagi. Examination of optimal design method of electromagnetic shield using ON/OFF method. *IEEE Transactions on Magnetics*, 45(3):1546–1549, March 2009.
- [203] N. Takahashi, S. Nakazaki, and D. Miyagi. Optimization of electromagnetic and magnetic shielding using ON/OFF method. *IEEE Transactions on Magnetics*, 46(8):3153–3156, 2010.
- [204] N. Takahashi, T. Yamada, and D. Miyagi. Examination of optimal design on IPM motor using ON/OFF method. *IEEE Trans. Magn.*, 46(8):3149–3152, 2010.
- [205] N. Takahashi, T. Yamada, S. Shimose, and D. Miyagi. Optimization of rotor of actual IPM motor using ON/OFF method. *IEEE Trans. Magn.*, 47(5):1262–1265, 2011.
- [206] H. Torkaman and E. Afjei. Comprehensive study of 2-D and 3-D finite element analysis of a switched reluctance motor. *J. Appl. Sciences*, 8(15):2758–2763, 2008.
- [207] R. Touzani and J. Rappaz. *Mathematical Models for Eddy Currents and Magnetostatics: With Selected Applications*. Springer Netherlands, 2014.
- [208] N. P. van Dijk, K. Maute, M. Langelaar, and F. van Keulen. Level-set methods for structural topology optimization: a review. *Structural and Multidisciplinary Optimization*, 48(3):437–472, 2013.

- [209] M. Y. Wang, X. Wang, and D. Guo. A level set method for structural topology optimization. *Computer Methods in Applied Mechanics and Engineering*, 192(1–2):227 – 246, 2003.
- [210] W. Wang, Y. Lu, J. S. Fu, and Y. Z. Xiong. Particle swarm optimization and finite-element based approach for microwave filter design. *IEEE Transactions on Magnetics*, 41(5):1800–1803, May 2005.
- [211] J. Wloka. *Partial Differential Equations*. Cambridge University Press, 1987.
- [212] Y. Xie and G. Steven. A simple evolutionary procedure for structural optimization. *Computers & Structures*, 49(5):885 – 896, 1993.
- [213] T. Yamada, K. Izui, S. Nishiwaki, and A. Takezawa. A topology optimization method based on the level set method incorporating a fictitious interface energy. *Computer Methods in Applied Mechanics and Engineering*, 199(45):2876–2891, 2010.
- [214] X.-S. Yang. *Nature-Inspired Optimization Algorithms*. Elsevier Science Publishers B. V., Amsterdam, The Netherlands, 1st edition, 2014.
- [215] J. Yoo and N. Kikuchi. Topology optimization in magnetic fields using the homogenization design method. *International Journal for Numerical Methods in Engineering*, 48(10):1463–1479, 2000.
- [216] J. Yoo and N. Kikuchi. Topology optimization for reduction of vibration caused by magnetic harmonic excitation. *IEEE Transactions on Magnetics*, 38(6):3643–3649, Nov 2002.
- [217] S. Zaglmayr. *High Order Finite Elements for Electromagnetic Field Computation*. PhD thesis, Johannes Kepler University Linz, 2006.
- [218] A. C. Zavioianu, E. Lughofer, G. Bramerdorfer, W. Amrhein, and S. Saminger-Platz. A surrogate-based strategy for multi-objective tolerance analysis in electrical machine design. In *2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 195–203, Sept 2015.
- [219] E. Zeidler. *Nonlinear Functional Analysis and its Applications II/B: Nonlinear Monotone Operators*. Springer-Verlag New York, 1990.
- [220] S. Zhou and M. Y. Wang. Multimaterial structural topology optimization with a generalized Cahn–Hilliard model of multiphase transition. *Structural and Multidisciplinary Optimization*, 33(2):89, 2006.
- [221] M. Zlámal. On the finite element method. *Numer. Math.*, 12:394–409, 1968.
- [222] W. Zulehner. *Numerische Mathematik: Eine Einführung anhand von Differentialgleichungsproblemen; Band 1: Stationäre Probleme*. Birkhäuser Basel, 2008.



# **Eidesstattliche Erklärung**

Ich erkläre an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.  
Die vorliegende Dissertation ist mit dem elektronisch übermittelten Textdokument identisch.

Linz, Dezember 2016

---

Peter Gangl



# Curriculum Vitae

**Name:** Peter Gangl

**Nationality:** Austria

**Date of Birth:** 22 February 1988

**Place of Birth:** Schärding, Austria

## **Education:**

Sep. 1998 – Jun. 2006	<b>High School</b> , A-level with distinction Bundesgymnasium Schärding
Oct. 2007 – Oct. 2010	<b>Bachelor in Technical Mathematics</b> , JKU Linz, with distinction, Bachelor thesis: “Exact and Inexact Semismooth Newton Methods for Elliptic Optimal Control Problems” Supervisor: Prof. Ulrich Langer
Oct. 2010 – Mar. 2012	<b>Master in Industrial Mathematics</b> , JKU Linz, with distinction, Master Thesis: “Topology Optimization in Electrical Engineering” Supervisor: Prof. Ulrich Langer
Jan.–Jun. 2011 Apr. 2012 – Feb. 2017	<b>ERASMUS Exchange Semester Lund University, Sweden</b> <b>PhD in Computational Mathematics</b> , JKU Linz PhD Thesis: “Sensitivity-Based Topology and Shape Optimization with Application to Electrical Machines” within Doctoral Program “Computational Mathematics” by FWF Supervisor: Prof. Ulrich Langer

## **Research Stays:**

Apr. 2013–Jul. 2013 and May 2014–Jun. 2014	<b>Technical University Berlin</b> , visiting Prof. Fredi Tröltzscher
Jun. 2014–Jul. 2014 and May 2015	<b>Université d'Avignon</b> , visiting Prof. Samuel Amstutz

## **Selected Presentations:**

- Jun. 2016 **Invited Talk** at IMA Special Workshop “Frontiers in PDE-Constrained Optimization”, Title: “Sensitivity-Based Topology and Shape Optimization for Electrical Machines subject to Nonlinear Magnetostatics”, Minneapolis, USA

**Teaching:**

Fall 2012	<b>Proseminar</b> “Mathematical Modeling in Engineering”
Feb. 2013	<b>Project instruction</b> at Project Week “Applied Mathematics” for High School Students, Project: Shape optimization of a bridge
Fall 2013	<b>Proseminar</b> “Mathematical Modeling in Engineering”
Jan. 2014	<b>Supervision of Bachelor Thesis</b> “Some Benchmark Problems in Electromagnetics” by B. Oberndorfer
Feb. 2014	<b>Project instruction</b> at Project Week “Applied Mathematics” for High School Students, Project: Optimal Control of a Heat Source
Spring 2014	<b>Tutorials</b> “Numerical Methods for Elliptic PDEs”
Fall 2014	<b>Proseminar</b> “Mathematical Modeling in Engineering”
Feb. 2015	<b>Project instruction</b> at Project Week “Applied Mathematics” for High School Students, Project: Mathematics and Music
Spring 2015	<b>Tutorials</b> “Numerical Methods for Elliptic PDEs”
Fall 2015	<b>Proseminar</b> “Mathematical Modeling in Engineering”
Mar. 2016	<b>Supervision of Bachelor Thesis</b> “Introduction to Magnetohydrodynamics” by A. Schafelner
Apr. 2016	<b>Supervision of Bachelor Thesis</b> “An Overview of Different Models described by Diffusion-Convection-Reaction Equations” by R. Schneckenleitner
Jul. 2016	<b>Project instruction</b> at 29th ECMI Modeling Week, Sofia, Bulgaria Project: Design Optimization of an Electric Motor

**Awards:**

June 2015	<b>Rita Trowbridge Award</b> , COMPUMAG Conference, Montréal Commendation as one of four finalists for best student presentation
Nov. 2012	<b>Würdigungspreis</b> , Austrian Ministry for Science and Research Awarded to the best 50 graduates of Austrian universities per year
Nov. 2011	<b>JKU Study Abroad Excellence Award</b> , JKU Linz, Award for excellent studies during exchange semester
Dec. 2008 – Dec. 2011	<b>Merit-based Scholarship</b> , JKU Linz, every year between 2008 and 2011

**Work Experience:**

Aug. 2009 – Jan. 2011	<b>Siemens Transformers Austria</b> , Linz (part time)
Dec. 2015 – Oct. 2016	<b>Linz Center of Mechatronics GmbH</b> , Linz (part time)

**Other:**

Social Activities	<b>Participation in Student Union (ÖH)</b> , JKU Linz Student representative for PhD students of technical and natural sciences Participation in <b>Science Slam</b> (Science Communication)
Sports	Football, Running, Volleyball, Snowboard, Hiking