JOHANNES KEPLER
UNIVERSITÄT LINZ | JKU

TNF

Technisch-Naturwissenschaftliche
Fakultät

Topology Optimization in Electrical Engineering

MASTERARBEIT

zur Erlangung des akademischen Grades

Diplomingenieur

im Masterstudium

Industriemathematik

Eingereicht von:
Peter Gangl

Angefertigt am:
Institut für Numerische Mathematik

Beurteilung:
O.Univ.-Prof. Dipl.-Ing. Dr. Ulrich Langer

Linz, Februar, 2012

Abstract

Topology optimization is a well-established tool for finding optimal structures in mechanics. Similar techniques can also be useful for applications from electrical engineering. The optimization of electrical equipment such as electrical machines or electromagnets with respect to a given cost functional is usually done by means of shape optimization methods, which can only modify the shape of the boundary of a structure, but not its basic topology. Topology optimization methods distribute material in a given part of the computational domain in an optimal way and do not impose any a-priori conditions on the resulting structures. Hence, these methods allow for optimal designs which could not be realized using shape optimization methods.

Mathematically, topology optimization problems can be formulated as infinite-dimensional optimization problems constrained by a partial differential equation as well as additional inequality constraints. These optimization problems are typically *ill-posed*. This thesis deals with the analysis of general topology optimization problems, discusses various possible regularization approaches and presents an application to a concrete problem from electrical engineering.

In a topology optimization problem, the design of a structure is represented by a discontinuous function ρ, commonly refered to as the *density function*. Due to this discontinuous nature of the problem, the discretization technique investigated in this thesis is a *discontinuous Galerkin* (DG) method, which allows for easier handling of jumps of function values. This class of methods is introduced and analyzed in detail.

This thesis focuses on one regularization method, the *phase-field* method, which on the one hand regularizes the ill-posed optimization problem, and on the other hand enforces the density function to attain only values close to 0 and close to 1. Areas in the computational domain with $\rho(\mathbf{x}) = 0$ in the final solution have to be interpreted as void, whereas areas with $\rho(\mathbf{x}) = 1$ represent areas occupied by material.

The phase-field method is applied to a practical problem from electrical engineering where the aim is to find a geometry for an electromagnet such that the induced magnetic field is as homogeneous as possible in a given direction. A discretization of the resulting optimization problem by a DG method yields a system of nonlinear equations, which is solved by Newton's method. Finally, numerical results are presented and discussed.

Zusammenfassung

Topologieoptimierung ist eine gängige Methode zum Auffinden optimaler Strukturen in der Mechanik. Ähnliche Techniken können auch für Anwendungen aus der Elektrotechnik nützlich sein. Die Optimierung elektrischer Anlagen wie z.B. elektrische Maschinen oder Elektromagneten bezüglich eines gegebenen Kostenfunktionals wird üblicherweise mittels Verfahren aus der Formoptimierung realisiert. Diese Verfahren können nur die Gestalt des Randes einer Struktur verändern, nicht aber ihre grundlegende Topologie. Bei Topologieoptimierungsverfahren wird Material in einem gewissen Teil des Rechengebiets auf optimale Weise verteilt, wobei von vornherein keine Bedingungen an die resultierende Struktur gestellt werden. Daher ermöglichen diese Verfahren optimale Designs, die durch Formoptimierung nicht realisiert werden könnten.

Mathematisch können Topologieoptimierungsprobleme als unendlichdimensionale Optimierungsprobleme formuliert werden, die durch eine partielle Differentialgleichung und weitere Ungleichungen restringiert sind. Diese Probleme sind üblicherweise *schlecht gestellt*. Diese Arbeit behandelt die Analyse allgemeiner Topologieoptimierungsprobleme sowie verschiedene Möglichkeiten zur Regularisierung dieser und zeigt eine Anwendung auf ein konkretes Problem aus der Elektrotechnik.

In einem Topologieoptimierungsproblem wird die Gestalt einer Struktur durch eine unstetige Funktion ρ, häufig als *Dichtefunktion* bezeichnet, repräsentiert. Aufgrund dieser Unstetigkeit, wird in dieser Arbeit eine diskontinuierliche Galerkin (DG)-Methode zur Diskretisierung verwendet. Eine solche Diskretisierungsmethode ermöglicht eine einfachere Handhabung von Sprüngen der Funktionswerte und werden im Detail beschrieben und analysiert.

Das Hauptaugenmerk dieser Arbeit liegt auf einer Regularisierungsmethode, der *Phase-Field*-Methode, die einerseits das schlecht gestellte Optimierungsproblem regularisiert und andererseits die Dichtefunktion zwingt, nur Werte nahe 0 und nahe 1 anzunehmen. Teilgebiete des Rechengebiets Ω mit $\rho(\mathbf{x}) = 0$ in der Lösung müssen als Luft interpretiert werden, wohingegen Teilgebiete mit $\rho(\mathbf{x}) = 1$ Gebiete, die von Material besetzt sind, darstellen.

Dieses Verfahren wird auf ein praktisches Problem aus der Elektrotechnik angewendet, in dem die Geometrie für einen Elektromagneten gesucht ist, sodass das erzeugte Magnetfeld so homogen wie möglich in eine vorgegebene Richtung ist. Eine Diskretisierung des resultierenden Optimierungsproblems mittels DG führt auf ein nichtlineares Gleichungssystem, das mit dem Newtonverfahren gelöst wird. Schlussendlich werden noch numerische Resultate präsentiert und diskutiert.

Acknowledgments

I would like to express my deep thanks to my supervisor Prof. Ulrich Langer for giving me the opportunity to write this thesis, for numerous valuable discussions and for organizing financial support.

This work has been carried out at the Institute of Computational Mathematics, JKU Linz, in cooperation with *ACCM (Austrian Center of Competence in Mechatronics)*, a K2-Center of the COMET/K2 program of the Federal Ministry of Transport, Innovation, and Technology, and the Federal Ministry of Economics and Labour, Austria.

This work was supported by the Austrian Science Foundation - *Fonds zur Förderung der wissenschaftlichen Forschung (FWF)* - granted by the *Doctoral Program (Doktoratskolleg) "Computational Mathematics: Numerical Analysis and Symbolic Computation"*.

Special thanks goes to my parents and my family for supporting me throughout my studies. Finally, I would like to thank my colleagues Dipl.-Ing. Markus Eslitzbichler and Daniela Saxenhuber for various interesting discussions and for proofreading.

Peter Gangl
Linz, February 2012

Contents

Chapter 1

Introduction

Optimization is the branch of mathematics that is concerned with finding an element z out of a given set $Z = Z_{ad}$ of admissible elements that is in some sense "better" than all the others. Mathematically, the "quality" of these elements is described by a function $J : Z \to \mathbb{R}$, called the objective or cost function, and the aim is to minimize (or maximize) J.

In *structural optimization* these points z represent geometric structures and the goal is to find that geometry that minimizes a given functional. A classical example is to design a bridge in a way such that it is as stiff as possible under given loadings, i.e., to find an element z out of the set of admissible designs Z which maximizes the stiffness J. Maximizing the stiffness of a structure is equivalent to minimizing its compliance. Therefore, this problem is refered to as the minimal compliance problem. Structural optimization methods originate from applications in structural mechanics (cf. Bendsøe [2] and Bendsøe and Sigmund [3]), but have in recent years also been successfully applied to problems from electrical engineering and other disciplines. For previous results from electrical engineering we refer the reader to Lukáš [20] and the references therein.

The field of structural optimization comprises several different approaches. In *sizing optimization* the geometry of the structure is (in the simplest case) assumed to consist of straight bars and is fixed from the very beginning. The optimization process then consists in determining the optimal thickness for each bar. Figure 1.1 shows the initial and final designs of a bridge for the minimal compliance problem. See Section 4.2 for a more detailed description of the problem.

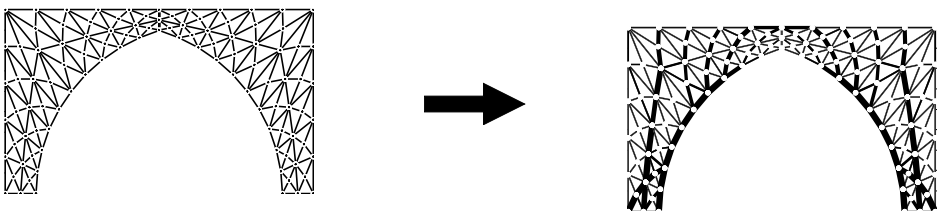

Figure 1.1: Initial and optimal design of a bridge using sizing optimization

In *shape optimization* the design parameter is some kind of parametrization of (a part of) the boundary of the structure. The goal is to determine the optimal shape of the boundary curve, whereas the basic topology of the design is prescribed. For a shape optimization

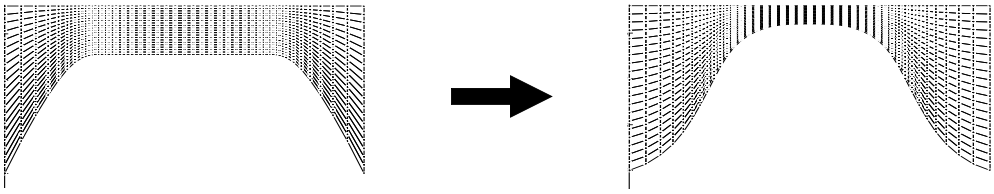

Figure 1.2: Initial and optimal design of a bridge using shape optimization

example see Figure 1.2.

In contrast to these two approaches, the method of *topology optimization* does not make any a priori assumptions on the topology of the structure. This means, roughly speaking, that the number of holes in the final design is not known in the beginning. Clearly, this amounts to less restrictions during the optimization process and gives hope for better results compared to the shape and sizing optimization approaches. The basic idea in topology optimization is (in its most general, continuous setting) to find the optimal material distribution in the *design domain* Ω_d, which is usually a subset of the computational domain Ω, by deciding for every single point $\mathbf{x}$ in Ω_d whether it should contain material or not. This decision is guided by the objective functional, denoted by J, as well as by a constraining partial differential equation (PDE) and possibly some other constraints. Mathematically, we are facing a problem of PDE-constrained optimization which can be written as follows:

$$\min_{\rho,u} J(\rho, u) \tag{1.1a}$$

$$\text{subject to} \quad a(\rho; u, v) = \langle F, v \rangle \qquad \forall v \in V_0 \tag{1.1b}$$

$$\int_{\Omega_d} \rho \ \mathrm{d}\mathbf{x} \leq V_{max} \tag{1.1c}$$

$$\rho(\mathbf{x}) \in \{0, 1\} \tag{1.1d}$$

Here, the state equation (1.1b) is given in its variational form with the bilinear form $a(\rho; \cdot, \cdot)$ and the linear functional $\langle F, \cdot \rangle$. Of course, this infinite-dimensional optimization problem cannot be solved in its full generality, but has to be discretized at some point. In this thesis, we will follow the concept of "*first optimize, then discretize*" rather than the other way around, meaning that we will work in an infinite-dimensional setting as long as possible before performing a discretization.

Topology optimization problems are rather difficult to handle as they are very likely to be ill-posed, i.e., *existence* and *uniqueness* of a solution as well as its *continuous dependence* on the given data are not always guaranteed. Therefore, applying straightforward methods to problems like (1.1) often results in different kinds of numerical troubles. In order to obtain reasonable results, some regularization method has to be applied. We will discuss this issue in Chapter 4.

This thesis aims at applying one regularization approach, the *phase-field method* to a problem from Electrical Engineering, namely to the topology optimization of an electromagnet in two space dimensions. This method has been investigated in detail for problems from structural mechanics by Stainko in [30] and by Burger and Stainko in [5].

The idea behind most topology optimization approaches consists of two steps: First, a discrete optimization problem is avoided by allowing the function ρ in (1.1) to attain any

value between zero and one, i.e., condition (1.1d) is replaced by $0 \leq \rho(\mathbf{x}) \leq 1$. Second, intermediate values of ρ are penalized so that the final solution consists only of "black" and "white" areas, which have to be interpreted as material and void, respectively. The phase-field method also involves a regularization term that excludes different kinds of numerical anomalies by bounding the perimeter of the resulting structure. The idea of the phase-field method is to balance the regularization term and the penalization term in a proper way. In the beginning of the optimization process the focus is on avoiding numerical troubles, i.e., the regularization term dominates and intermediate values of ρ are not heavily penalized. Towards the end of the process the influence of the regularization term is decreased and the penalization term becomes dominant, driving $\rho(\mathbf{x})$ either to zero or to one for each point $\mathbf{x}$ in the design domain.

The benchmark problem we are going to investigate in Chapter 5 is the following: The aim is to design an electromagnet in such a way that the resulting magnetic field in a certain part of the computational domain is as homogeneous as possible in a given direction. These kinds of electromagnets are used for measurements of magneto-optic effects and have been developed at the Technical University of Ostrava, Czech Republic. For a more detailed problem description we refer to Section 5.1.

The remainder of this thesis is organized as follows: In Chapter 2 we will gather the mathematical machinery we are going to use. After a short introduction to general concepts of optimization in finite and infinite dimensions we will introduce Newton's method which is used for solving the resulting optimality system. After a brief overview on Sobolev spaces, we give a short introduction to mathematical modeling in electrical engineering and derive the equations of 2D magnetostatics, which will serve as state equations in our benchmark problem. Finally we will treat the question of existence and uniqueness of a solution to these equations.

In Chapter 3 we will present the discretization technique we are going to use, a *discontinuous Galerkin* (DG) method, discuss the existence and uniqueness issue, provide error estimates and present some numerical results.

Chapter 4 will give an overview on the general strategies for solving problems like (1.1), illustrate the different kinds of numerical anomalies arising in topology optimization as well as suggest possible cures such as the phase-field method.

In Chapter 5 we will describe in detail the benchmark problem of this thesis, which is to determine the optimal design for an electromagnet, and apply the phase-field method to it. We will set up the first-order necessary optimality conditions, discretize them using a DG method and finally solve the resulting system of nonlinear equations by Newton's method.

In Chapter 6 we will present the numerical results we obtained for the benchmark problem from Chapter 5.

Finally, in Chapter 7, we will summarize our findings and suggest possible future work.

Chapter 2

Preliminaries

2.1 Numerical Optimization

2.1.1 Constrained Optimization in Finite Dimensions

In this subsection we state the first-order necessary optimality conditions for an abstract finite-dimensional optimization problem with both equality and inequality constraints. This brief overview is based on STAINKO [30] and should only serve as a motivation for the infinite-dimensional case treated in the following subsection as, in this thesis, we will follow the concept of *"first optimize, then discretize"* and therefore will perform the optimization in the infinite-dimensional setting.

We consider the following optimization problem in $\mathbb{R}^n$:

$$\begin{aligned} & \min_{\mathbf{x}\in\mathbb{R}^n} J(\mathbf{x}) & \\ \text{subject to} \quad & c_i(\mathbf{x}) = 0 \qquad & i \in \mathcal{E} \\ & c_i(\mathbf{x}) \leq 0 \qquad & i \in \mathcal{I} \end{aligned} \tag{2.1}$$

Here, $\mathcal{E}$ and $\mathcal{I}$ are the disjoint sets of indices of *equality* and *inequality* constraints, respectively, and the objective function J as well as the constraints c_i, $i \in \mathcal{E} \cup \mathcal{I}$, are continuously differentiable mappings from $\mathbb{R}^n$ to $\mathbb{R}$. A point $\bar{\mathbf{x}}$ is said to be *feasible*, if it satisfies the constraints in (2.1). At any feasible point $\mathbf{x}$ we define the *active set*

$$\mathcal{A}(\mathbf{x}) = \{i \in \mathcal{E} \cup \mathcal{I} | c_i(\mathbf{x}) = 0\}.$$

Let m denote the number of constraints in (2.1), i.e., $m = |\mathcal{E} \cup \mathcal{I}|$. To formulate the first-order necessary optimality conditions, we have to introduce the Lagrange function or *Lagrangian* $\mathcal{L} : \mathbb{R}^{n+m} \to \mathbb{R}$ defined as

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = J(\mathbf{x}) + \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i c_i(\mathbf{x}), \tag{2.2}$$

with the *Lagrange multiplier vector* $\boldsymbol{\lambda}$. Furthermore, we have to make sure that the constraints do not show any degenerate behaviour. We do this by requiring the constraints to satisfy a *constraint qualification*:

Definition 2.1 (LICQ)**.** *The linear independence constraint qualification (LICQ) holds at the point* $\bar{\mathbf{x}}$ *if the set of gradients of the active constraints* $\{\nabla c_i(\bar{\mathbf{x}}) | i \in \mathcal{A}(\bar{\mathbf{x}})\}$ *is linearly independent.*

Now we can formulate the first-order necessary optimality conditions for a (local) solution of (2.1), commonly called the *Karush-Kuhn-Tucker (KKT)* conditions:

Theorem 2.2 (First-Order Necessary Conditions)**.** *Suppose that* $\bar{\mathbf{x}}$ *is a local solution of* (2.1) *with* J *and* $c_i, i \in \mathcal{E} \cup \mathcal{I}$*, continuously differentiable and that the LICQ holds at* $\bar{\mathbf{x}}$*. Then there exists a Lagrange multiplier* $\bar{\boldsymbol{\lambda}}$ *such that the following conditions are satisfied:*

$$\nabla_{\boldsymbol{x}} \mathcal{L}(\bar{\mathbf{x}}, \bar{\boldsymbol{\lambda}}) = 0 \tag{2.3a}$$

$$c_i(\bar{\mathbf{x}}) = 0 \qquad \forall i \in \mathcal{E} \tag{2.3b}$$

$$\bar{\lambda}_i \geq 0 \text{ and } c_i(\bar{\mathbf{x}}) \leq 0 \qquad \forall i \in \mathcal{I} \tag{2.3c}$$

$$\bar{\lambda}_i c_i(\bar{\mathbf{x}}) = 0 \qquad \forall i \in \mathcal{E} \cup \mathcal{I}. \tag{2.3d}$$

Proof. See, e.g., Nocedal and Wright [23]. □

Any point $\bar{\mathbf{x}}$ that satisfies (2.3) under the above assumptions is called a *first-order critical* or a *KKT* point for the problem (2.1).

2.1.2 Constrained Optimization in Infinite Dimensions

Let us now turn to the, for us, more interesting issue of optimality conditions for optimization problems in an infinite-dimensional setting. This subsection is mainly based on Tröltzsch [31].

Now the optimization is not performed over $\mathbb{R}^n$, but over some Banach space Y. For sake of simplicity, we will restrict ourselves to the case of optimization problems with only equality constraints. A brief overview on possible ways to treat inequality constraints is given in Remark 2.7. We are facing a problem like

$$\begin{aligned} &\min_{y \in Y} J(y) \\ \text{subject to} \quad &G(y) = 0. \end{aligned} \tag{2.4}$$

Here, $J : Y \to \mathbb{R}$ is the objective functional and $G : Y \to Z$ represents the constraints, where Y and Z are real Banach spaces. Typical examples of problems of this type are minimization problems with a partial differential equation (PDE) in operator form as a constraint, e.g. optimal control problems (cf. Tröltzsch [31]) or structural optimization problems (cf. Bendsøe [2] and Bendsøe and Sigmund [3]). Again, the first-order optimality conditions are stated using the Lagrangian of the optimization problem (2.4):

Definition 2.3. *The function* $\mathcal{L} : Y \times Z^* \to \mathbb{R}$*,*

$$\mathcal{L}(y, z^*) = J(y) + \langle z^*, G(y) \rangle_{Z^* \times Z}$$

is called the Lagrange function of problem (2.4)*.*

The first-order necessary optimality conditions will involve a derivative of the functionals J and G, thus we have to introduce a notion of differentiation for mappings between Banach spaces. Let Y and Z be real Banach spaces with respective norms $\|\cdot\|_Y$ and $\|\cdot\|_Z$, let $\mathcal{Y} \subset Y$ be open and recall that $\mathcal{L}(Y, Z)$ denotes the set of all bounded linear operators from Y to Z.

Definition 2.4 (Fréchet derivative)**.** *A mapping $F : Y \supseteq \mathcal{Y} \to Z$ is said to be Fréchet differentiable at $y \in \mathcal{Y}$ if there exist an operator $A \in \mathcal{L}(Y, Z)$ and a mapping $r(u, \cdot) : Y \to Z$ with the following properties: For all $h \in Y$ such that $y + h \in \mathcal{Y}$ it holds*

$$F(y + h) = F(y) + Ah + r(u, h) \tag{2.5}$$

where

$$\frac{\|r(u, h)\|_Z}{\|h\|_Y} \to 0 \quad \text{as} \quad \|h\|_Y \to 0. \tag{2.6}$$

The operator A is said to be the Fréchet derivative of F at the point y, written as $A = F'(y)$.

In the following we will refer to the Fréchet derivative of an operator F simply by F'. Note that for $F : Y \to Z$ we have $F' : Y \to \mathcal{L}(Y, Z)$. Also note that any linear bounded operator B is Fréchet differentiable since $B(y + h) = By + Bh + 0$ with the Fréchet derivative being the operator itself. For more properties of Fréchet derivatives we refer to Tröltzsch [31] and Langer [18].

As in the finite-dimensional case, we again have to require some constraint qualification to ensure that the constraints do not show any degenerate behaviour.

Definition 2.5. *The constraint qualification of Zowe and Kurcyusz for* (2.4) *holds at $\bar{\mathbf{y}} \in Y$ if*

$$G'(\bar{\mathbf{y}})\, Y = Z, \tag{2.7}$$

i.e., if the operator $G'(\bar{\mathbf{y}})$ is surjective.

Note that Definition 2.5 applies to the case where we have only equality constraints in (2.4), but can be generalized to the case involving inequality constraints, again see Tröltzsch [31].

Now we can state the first-order necessary optimality conditions for a local solution to problem (2.4):

Theorem 2.6 (First-Order Necessary Optimality Conditions)**.** *Suppose that $\bar{\mathbf{y}} \in Y$ is a local solution of problem* (2.4) *with J and G continuously Fréchet-differentiable in an open neighborhood of $\bar{\mathbf{y}}$ and that the constraint qualification* (2.7) *holds. Then there exists a Lagrange multiplier $\mathbf{z}^* \in Z^*$ for $\bar{\mathbf{y}}$ satisfying the following conditions:*

$$\nabla_y \mathcal{L}(\bar{\mathbf{y}}, \mathbf{z}^*) = J'(\bar{\mathbf{y}}) + G'(\bar{\mathbf{y}})^* \mathbf{z}^* = 0 \quad \text{in } Y^* \tag{2.8a}$$

$$G(\bar{\mathbf{y}}) = 0 \quad \text{in } Z \tag{2.8b}$$

Proof. This theorem is a special case of Theorem 6.3 in Tröltzsch [31] where the more general case including inequality constraints is treated. □

Remark 2.7 (Box constraints)**.** *A frequently occurring class of inequality constraints are so-called box constraints. If we choose the Banach space Y in* (2.4) *to be $L^2(\Omega)$, the optimization problem involving box constraints looks as follows:*

$$\min_{y \in L^2(\Omega)} J(y) \tag{2.9a}$$

$$\text{subject to} \quad G(y) = 0 \tag{2.9b}$$

$$a(x) \leq y(x) \leq b(x) \quad \text{a.e. in } \Omega, \tag{2.9c}$$

where a and b are functions from $L^\infty(\Omega)$ and (2.9b) *represents a PDE constraint. It is then possible to eliminate these inequality constraints by introducing a projection operator and to reformulate the optimization problem as a non-smooth operator equation. Even though the resulting operator is not differentiable, a generalization of Newton's method can be applied (cf. Section 2.2 for the classical Newton's method). This generalization is called the Semismooth Newton Method and can be shown to be q-superlinearly convergent, cf.* Hinze et al. *[15] and* Gangl *[11]. We refer the reader also to* Hintermüller et al. *[14]. Box constraints typically arise in optimal control problems, cf.* Tröltzsch *[31]. In* Tröltzsch *[31] also more general kinds of inequality constraints are treated.*

2.2 Newton's Method

Here, we will briefly introduce Newton's method for solving (systems of) nonlinear equations which we will make use of in Chapter 5 as the optimality system of our topology optimization problem will turn out to be nonlinear. This section is mainly based on Jung and Langer [16]. We also refer the reader to Deuflhard [8].

Under certain conditions Newton's method can be applied to an equation of the type

$$G(\mathbf{x}) = 0, \tag{2.10}$$

with $G : \mathbb{R}^n \to \mathbb{R}^n$. The method is motivated via a first-order Taylor expansion as follows: For a given approximation $\mathbf{x}^{(k)}$ one would like to find a correction $\mathbf{w}^{(k)}$ such that $\mathbf{x}^{(k)} + \mathbf{w}^{(k)}$ is an exact solution of (2.10). For achieving that, one attempts to find such a Newton correction by performing a first-order Taylor expansion around $\mathbf{x}^{(k)}$:

$$G\left(\mathbf{x}^{(k)} + \mathbf{w}^{(k)}\right) \approx G\left(\mathbf{x}^{(k)}\right) + G'\left(\mathbf{x}^{(k)}\right)\mathbf{w}^{(k)} = 0. \tag{2.11}$$

The Newton correction is computed from (2.11) and added to the current iterate $\mathbf{x}^{(k)}$. The procedure looks as follows:

Algorithm 1. Newton's Method

0. Choose initial guess $\mathbf{x}^{(0)}$

For $k = 1$ until convergence do

1. Compute Newton correction $\mathbf{w}^{(k)}$ from

$$G'\left(\mathbf{x}^{(k)}\right)\mathbf{w}^{(k)} = -G\left(\mathbf{x}^{(k)}\right) \tag{2.12}$$

2. Update $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{w}^{(k)}$ and go to step 1

It seems natural that one must require G to be continuously Fréchet differentiable at the iterates $\mathbf{x}^{(k)}$ and that the Jacobian G' must be invertible at these points. The following theorem states the local q-quadratic convergence of Algorithm 1 under an additional smoothness assumption on G (cf. JUNG AND LANGER [16]). Note that a sequence $\{u^{(k)}\}$ is said to be q-quadratically convergent to a point u^* with respect to a norm $\|\cdot\|$ if $\|u^* - u^{(k+1)}\| \leq C\|u^* - u^{(k)}\|^2$ for some constant $C > 0$ independent of k.

Theorem 2.8. *Let $G : \mathbb{R}^n \to \mathbb{R}^n$ be twice continuously Fréchet differentiable and $\mathbf{x}^* \in \mathbb{R}^n$ be a solution of the system of nonlinear equations (2.10). Let the Jacobian $G'(\mathbf{x}^*)$ be regular in that solution point. If the initial guess $\mathbf{x}^{(0)}$ is sufficiently close to the solution $\mathbf{x}^*$ then Algorithm 1 converges q-quadratically to the solution $\mathbf{x}^*$ of (2.10).*

A proof for the case where $n = 1$ can be found in JUNG AND LANGER [16].

Remark 2.9. *A sequence $\{u^{(k)}\}$ is said to be q-superlinearly convergent to a point u^* with respect to a norm $\|\cdot\|$ if $\|u^* - u^{(k+1)}\| = o(\|u^* - u^{(k)}\|)$ or, in other words, if*

$$\limsup_{k\to\infty} \frac{\|\mathbf{x}^* - \mathbf{x}^{(k+1)}\|}{\|\mathbf{x}^* - \mathbf{x}^{(k)}\|} = 0. \tag{2.13}$$

If G is not twice, but only once continuously Fréchet differentiable and regular at the solution $\mathbf{x}^*$, Newton's method still converges locally q-superlinearly (see, e.g., HINZE ET AL. [15]). Furtheron, if G is only once continuously Fréchet differentiable and the Jacobian G' is Lipschitz continuous then the method already converges q-quadratically. Hence, the requirement that G should be twice continuously Fréchet differentiable in Theoerem 2.8 is actually too strong. A proof of this statement can be found in LANGER [18], where Newton's method is treated for operators between general Banach spaces.

A non-trivial issue is the choice of a proper initial value $\mathbf{x}^{(0)}$ that is *sufficiently close* to the exact (in general unknown) solution $\mathbf{x}^*$. One way to deal with this inconvenience is to perform Algorithm 1 with a variable step size (*damped Newton's method*). In each iteration the Newton correction $\mathbf{w}^{(k)}$ is computed as in (2.12) and a step size $\tau = \tau^{(k)} \in (0, 1]$ is chosen such that

$$\|G\left(\mathbf{x}^{(k)} + \tau\mathbf{w}^{(k)}\right)\|^2 < \|G\left(\mathbf{x}^{(k)}\right)\|^2. \tag{2.14}$$

The next theorem guarantees the existence of a $\tau^{(k)} > 0$ such that (2.14) is satisfied (JUNG AND LANGER [16]).

Theorem 2.10. *Let $G : \mathbb{R}^n \to \mathbb{R}^n$ be differentiable. Furthermore, assume that the k-th iterate $\mathbf{x}^{(k)}$ is not a root of G and that the Jacobian G' is regular in $\mathbf{x}^{(k)}$, i.e., $G\left(\mathbf{x}^{(k)}\right) \neq 0$ and $G'\left(\mathbf{x}^{(k)}\right)^{-1}$ exists. Then there exists a positive real number τ^* such that inequality (2.14) holds for all τ in $(0, \tau^*)$.*

For a proof of this theorem we again refer the reader to JUNG AND LANGER [16]. A simple strategy would be to choose in every step $\tau \in \{1, 1/2, 1/4, 1/8, \dots\}$ maximal such that (2.14) is satisfied. In particular, choose $\tau = 1$ whenever it is admissible. For more sophisticated strategies for choosing $\tau^{(k)}$ like the *Wolfe* conditions or *Goldstein* conditions we refer to NOCEDAL AND WRIGHT [23].

2.3 Function Spaces

Next, we will introduce the function spaces we are going to use for formulating partial differential equations in *weak* or *variational* form. This section is based on Rivière [27] and Adams and Fournier [1].

Throughout this thesis, Ω denotes an open bounded Lipschitz domain in $\mathbb{R}^d$, where d denotes the space dimension. The space $L^2(\Omega)$ is the space of Lebesgue measurable, square-integrable functions over Ω,

$$L^2(\Omega) = \{v \text{ measurable} : \int_\Omega v(\mathbf{x})^2 \mathrm{d}\mathbf{x} < \infty\}.$$

Note that the elements of $L^2(\Omega)$ are actually equivalence classes of functions: Two functions v_1 and v_2 belong to the same equivalence class if they differ only on a set of measure zero. The space $L^2(\Omega)$, equipped with the norm

$$\|v\|_{L^2(\Omega)} = \left(\int_\Omega v(\mathbf{x})^2 \mathrm{d}\mathbf{x}\right)^{\frac{1}{2}},$$

and the inner product

$$(v, w)_{L^2(\Omega)} = \int_\Omega v(\mathbf{x})\, w(\mathbf{x})\ \mathrm{d}\mathbf{x},$$

is a Hilbert space. Given a multi-index $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$, a locally integrable function w is said to be the $\boldsymbol{\alpha}$-th *weak partial derivative* of a locally integrable function u if

$$\int_\Omega u(\mathbf{x}) D^{\boldsymbol{\alpha}} v(\mathbf{x})\ \ \mathrm{d}\mathbf{x} = (-1)^{|\boldsymbol{\alpha}|} \int_\Omega w(\mathbf{x}) v(\mathbf{x})\ \ \mathrm{d}\mathbf{x} \qquad \forall v \in C_0^\infty(\Omega), \tag{2.15}$$

where $C_0^\infty(\Omega)$ is the space of all $C^\infty(\Omega)$ functions with compact support and the partial derivative with respect to $\boldsymbol{\alpha}$ is given by

$$D^{\boldsymbol{\alpha}} v = \frac{\partial^{|\boldsymbol{\alpha}|} v}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}},$$

with $|\boldsymbol{\alpha}| = \sum_{i=1}^d \alpha_i$. For w, we again use the notation $w = D^{\boldsymbol{\alpha}} u$. Next, we will introduce the *Sobolev space* $H^s(\Omega)$, $s \in \mathbb{N}$, that is the space of $L^2(\Omega)$ functions with weak derivatives in $L^2(\Omega)$ up to order s:

$$H^s(\Omega) = \{v \in L^2(\Omega) | D^{\boldsymbol{\alpha}} v \in L^2(\Omega)\ \forall\ 0 \leq |\boldsymbol{\alpha}| \leq s\}.$$

The space $H^s(\Omega)$ is again a Hilbert space with the norm

$$\|v\|_{H^s(\Omega)} = \left(\sum_{0 \leq |\boldsymbol{\alpha}| \leq s} \|D^{\boldsymbol{\alpha}} v\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}},$$

and the inner product

$$(v, w)_{H^s(\Omega)} = \sum_{0 \leq |\boldsymbol{\alpha}| \leq s} (D^{\boldsymbol{\alpha}} v, D^{\boldsymbol{\alpha}} w)_{L^2(\Omega)}.$$

The $H^s(\Omega)$-seminorm is defined by

$$|v|_{H^s(\Omega)} = \left(\sum_{|\boldsymbol{\alpha}|=s} \|D^{\boldsymbol{\alpha}} v\|^2_{L^2(\Omega)} \right)^{\frac{1}{2}}.$$

The definition of Sobolev spaces can also be extended to $H^s(\Omega)$ with real s. Then we cannot interpret the space as the space with s weak derivatives any more, but it can be seen as an interpolated space, cf. Bergh and Löfström [4]. For instance, $H^{1/2}(\Omega)$ lies in some sense in between $H^1(\Omega)$ and $H^0(\Omega)(= L^2(\Omega))$. For a more detailed introduction to Sobolev spaces we refer to Adams and Fournier [1].

An important result is the embedding theorem that relates the Sobolev space $H^s(\Omega)$ to the space of r times continuously differentiable functions $C^r(\Omega)$, $r \in \mathbb{N}$:

Theorem 2.11 (Sobolev's Embedding Theoerem)**.** *For $\Omega \subset \mathbb{R}^d$, we have*

$$H^s(\Omega) \subset C^r(\Omega) \quad \text{if } s - r > \frac{d}{2}. \tag{2.16}$$

To be more precise, the theorem says that under certain conditions, if $v \in H^s(\Omega)$ then there is a representative in the equivalence class of v that is $C^r(\Omega)$. In particular ($r = 0$):

$$H^s(\Omega) \subset C^0(\Omega) \text{ if } \begin{cases} s > \frac{1}{2} \text{ for } d = 1, \\ s > 1 \text{ for } d = 2, \\ s > \frac{3}{2} \text{ for } d = 3. \end{cases}$$

For properly defining boundary conditions we have to introduce the notion of the trace of a Sobolev function:

Theorem 2.12 (Trace theorem, [27])**.** *Let Ω be a bounded domain with Lipschitz boundary Γ and outward normal vector $\mathbf{n}$. There exist trace operators $\gamma_0 : H^s(\Omega) \to H^{s-1/2}(\partial\Omega)$ for $s > \frac{1}{2}$ and $\gamma_1 : H^s(\Omega) \to H^{s-3/2}(\Gamma)$ for $s > 3/2$ that are extensions of the boundary values and boundary normal derivatives, respectively. The operators γ_0, γ_1 are surjective. Furthermore, if $v \in C^1(\bar{\Omega})$, then*

$$\gamma_0 v = v|_\Gamma, \quad \gamma_1 v = \nabla v \cdot \mathbf{n}|_\Gamma.$$

Remark 2.13. *For sake of convenience, when talking about function values and normal derivatives on boundaries we will always write v and $\nabla v \cdot \mathbf{n}$ instead of the traces $\gamma_0 v$ and $\gamma_1 v$, respectively, but always keep in mind the trace operators of Theorem 2.12.*

Note that the surjectivity of the operator γ_0 yields that any Dirichlet data $g_D \in H^{1/2}(\Gamma)$ can be extended to the interior of the domain by an $H^1(\Omega)$ function. Therefore, a variational problem with inhomogeneous Dirichlet boundary conditions can be converted to a problem with homogeneous ones if the Dirichlet data satisfies $g_D \in H^{1/2}(\Gamma)$ (*homogenization*).

2.4 Mathematical Modeling in Electrical Engineering

In this section, we will give a short introduction to the physical background of the topology optimization problem we are going to investigate. We state the full Maxwell equations and derive from these the equations of 2D magnetostatics which will be the governing equations in the optimization problem. This introduction is mainly based on Langer [19], Pechstein [24], Zaglmayr [32] and Schöberl [28]. We refer the reader also to the monographs Kaltenbacher [17] and Monk [22].

Electromagnetic fields are described by Maxwell's equations established by James C. Maxwell in 1862. The full Maxwell's field equations in classical (differential) form read

$$\text{curl } \boldsymbol{H} = J + \frac{\partial D}{\partial t} \tag{2.17a}$$

$$\text{div } \boldsymbol{B} = 0 \tag{2.17b}$$

$$\text{curl } \boldsymbol{E} = -\frac{\partial B}{\partial t} \tag{2.17c}$$

$$\text{div } \boldsymbol{D} = \rho \tag{2.17d}$$

with the electromagnetic quantities

$\mathbf{H} = (H_1, H_2, H_3)^T$	magnetic field intensity	$[A/m]$
$\mathbf{B} = (B_1, B_2, B_3)^T$	magnetic flux density	$[Vs/m^2]$
$\mathbf{E} = (E_1, E_2, E_3)^T$	electric field intensity	$[V/m]$
$\mathbf{D} = (D_1, D_2, D_3)^T$	electric flux density	$[As/m^2]$
$\mathbf{J} = (J_1, J_2, J_3)^T$	electric current density	$[As/m^2]$
ρ	electric charge density	$[As/m^3]$

All these quantities are functions of the spatial variable $\mathbf{x} = (x_1, x_2, x_3)^T$ and the time variable t. The units given in brackets are the SI units. Since the system of equations (2.17) is underdetermined (8 equations for 12 unknowns), we have to add material laws.

The electromagnetic fields are related via the following constitutive laws:

$$\mathbf{B} = \mu\mathbf{H} + \mu_0\mathbf{M} \tag{2.18a}$$

$$\mathbf{D} = \varepsilon\mathbf{E} + \mathbf{P} \tag{2.18b}$$

$$\mathbf{J} = \sigma\mathbf{E} + \mathbf{J_i} \tag{2.18c}$$

with

$\mathbf{M}$	permanent magnetization	$[Vs/m^2]$
$\mathbf{P}$	electric polarization	$[As/m^2]$
μ	magnetic permeability	$[Vs/Am]$
μ_0	permeability of vacuum $(= 4\pi 10^{-7})$	$[Vs/Am]$
ε	electric permittivity	$[As/Vm]$
σ	electric conductivity	$[A/Vm]$
$\mathbf{J_i}$	impressed current density	$[A/m^2]$.

For sake of simplicity we will neglect the effects of permanent magnetization and electric polarization and set $\mathbf{M} = \mathbf{P} = 0$. In the general case, the parameters μ, ε and σ are tensors depending on space and time as well as on the electromagnetic fields. However, we will

neglect the effects of hysteresis and restrict ourselves to the case of isotropic materials where these tensors reduce to scalar functions.

In many practical applications it is not necessary to treat the full Maxwell equations (2.17) as under some physical assumptions they can be reduced to special electromagnetic regimes. In non-conducting regions ($\sigma = 0$) the equations for the magnetic and electric fields decouple into two independent systems. If we further assume all fields to be time-independent (i.e. $\frac{\partial}{\partial t} = 0$), we arrive at the equations of magnetostatics and electrostatics. The last assumption is satisfied if the electromagnetic field is generated only by static or uniformly moving charges (cf. Zaglmayr [32]). For the rest of this thesis we will deal only with the magnetostatic regime:

$$\text{curl } \boldsymbol{H} = \mathbf{J_i} \tag{2.19a}$$

$$\text{div } \boldsymbol{B} = 0 \tag{2.19b}$$

$$\boldsymbol{B} = \mu \boldsymbol{H} \tag{2.19c}$$

Applying the divergence operator to (2.19a) and noting that div curl $\mathbf{u} = 0$ for any twice continuously differentiable vector field $\mathbf{u}$, we immediately obtain the necessary solvability condition

$$\text{div } \mathbf{J_i} = 0. \tag{2.20}$$

Assuming that the computational domain Ω is simply connected, equation (2.19b) implies the existence of a vector potential $\mathbf{A} = (A_1(\mathbf{x}), A_2(\mathbf{x}), A_3(\mathbf{x}))^T$ such that $\mathbf{B} = \text{curl}\mathbf{A}$. Introducing the magnetic reluctivity

$$\nu(\mathbf{x}, |\boldsymbol{B}|) := \frac{1}{\mu(\mathbf{x}, |\boldsymbol{B}|)},$$

system (2.19) can be written in the vector potential formulation

$$\text{curl}\left(\nu(\mathbf{x}, |\text{curl}\mathbf{A}|)\text{curl}\mathbf{A}(\mathbf{x})\right) = \mathbf{J_i}, \tag{2.21a}$$

together with the boundary conditions

$$\mathbf{A}(\mathbf{x}) \times \mathbf{n} = 0 \qquad \text{on } \Gamma_B \text{ and} \tag{2.21b}$$

$$\nu(\mathbf{x}, |\text{curl}\mathbf{A}|)\text{curl}\mathbf{A}(\mathbf{x}) \times \mathbf{n} = 0 \qquad \text{on } \Gamma_H, \tag{2.21c}$$

for two disjoint sets Γ_B, Γ_H such that $\overline{\Gamma}_B \cup \overline{\Gamma}_H = \partial\Omega$. Note that (2.21b) implies that $\mathbf{B} \cdot \mathbf{n} = 0$ (see, e.g., Pechstein [24]) which is the so-called *induction* boundary condition. The condition (2.21c) is equivalent to the condition $\mathbf{H} \times \mathbf{n} = 0$ which is called the *perfect magnetic conductors* (PMC) condition. Note that equation (2.21a) does not admit a unique solution since for any solution $\mathbf{A}$, a gradient field $\nabla\Phi$ can be added such that $\mathbf{A} + \nabla\Phi$ is a solution, too. A solution $\mathbf{A}$ can be made unique by additionally requiring, e.g., that div $A = 0$. This condition is called *Coulomb's gauge*. In the following, we will, for simplicity, assume the magnetic reluctivity ν (and also the magnetic permeability μ) to be homogeneous, i.e.,

$$\nu(\mathbf{x}, |\text{curl}\mathbf{A}|) = \nu(|\text{curl}\mathbf{A}|). \tag{2.22}$$

We remark that all results of this chapter are still valid if the computational domain Ω consists of an arbitrary, but finite number of materials that have uniform behaviour, i.e.,

$$\overline{\Omega} = \bigcup_{i \in I} \overline{\Omega}_i \text{ with } \Omega_i \text{ open and pairwise disjoint,} \tag{2.23}$$

$$\nu(\mathbf{x}, |\mathrm{curl}\mathbf{A}|) = \nu^{(i)}(|\mathrm{curl}\mathbf{A}|) \text{ for } \mathbf{x} \in \Omega_i, \tag{2.24}$$

for some functions $\nu^{(i)} : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ describing the behaviour of the material with index $i \in I$ (cf. Pechstein [24]).

2.4.1 Reduction to 2D

In this thesis we will deal with the topology optimization of an electromagnet in a two-dimensional setting. The underlying state equations will be the equations of 2D magnetostatics in vector potential formulation which can be derived from system (2.21) under certain assumptions:

1. $\Omega = \Omega_{2D} \times (-l, l)$ with $l \gg \mathrm{diam}(\Omega_{2D})$ or $\Omega = \Omega_{2D} \times (-\epsilon, \epsilon)$ with $\epsilon \ll \mathrm{diam}(\Omega_{2D})$,
2. $\mathbf{J_i} = \begin{pmatrix} 0 \\ 0 \\ J_3(x_1, x_2) \end{pmatrix}, \quad (x_1, x_2)^T \in \Omega_{2D},$
3. $\mathbf{H} = \begin{pmatrix} H_1(x_1, x_2) \\ H_2(x_1, x_2) \\ 0 \end{pmatrix}, \quad (x_1, x_2)^T \in \Omega_{2D}.$

Note that, because of the second assumption, the necessary solvability condition (2.20) is automatically satisfied. Let, from now on, $\mathbf{x} = (x_1, x_2)^T$ be a vector in $\mathbb{R}^2$, let $\Omega = \Omega_{2D}$ as well as $\Gamma_B = \Gamma_B \cap \overline{\Omega}_{2D}$ and $\Gamma_H = \Gamma_H \cap \overline{\Omega}_{2D}$. Under the above assumptions, using (2.19c), we further conclude

$$\mathbf{B} = \begin{pmatrix} B_1(\mathbf{x}) \\ B_2(\mathbf{x}) \\ 0 \end{pmatrix}, \quad \mathbf{x} \in \Omega_{2D},$$

and therefore

$$0 = B_3(\mathbf{x}) = (\mathrm{curl}\mathbf{A}(\mathbf{x}))_3 = \frac{\partial A_1}{\partial x_2}(\mathbf{x}) - \frac{\partial A_2}{\partial x_1}(\mathbf{x}).$$

This is fulfilled with the ansatz

$$\mathbf{A} = \mathbf{A}(\mathbf{x}) = \begin{pmatrix} 0 \\ 0 \\ u(\mathbf{x}) \end{pmatrix}, \tag{2.25}$$

which yields

$$|\mathbf{B}(\mathbf{x})| = |\mathrm{curl}\mathbf{A}(\mathbf{x})| = \left| \begin{pmatrix} \partial_2 u(\mathbf{x}) \\ -\partial_1 u(\mathbf{x}) \\ 0 \end{pmatrix} \right| = |\nabla u(\mathbf{x})|,$$

where $\nabla = (\partial_1, \partial_2)^T$ denotes the gradient operator in two space dimensions. The left hand side of equation (2.21a) now becomes

$$\begin{aligned}\mathrm{curl}(\nu(|\mathrm{curl}\mathbf{A}|)\mathrm{curl}\mathbf{A}(\mathbf{x})) &= \mathrm{curl}\begin{pmatrix}\nu(|\nabla u|)\partial_2 u(\mathbf{x})\\ -\nu(|\nabla u|)\partial_1 u(\mathbf{x})\\ 0\end{pmatrix}\\ &= \begin{pmatrix}0\\0\\ -\mathrm{div}\left(\nu(|\nabla u|)\,\nabla u(\mathbf{x})\right)\end{pmatrix}.\end{aligned}$$

Note that with ansatz (2.25) the Coulomb gauge as introduced above is automatically satisfied. A closer look at the boundary conditions (2.21b) and (2.21c) yields

$$0 = \mathbf{A}(\mathbf{x}) \times \mathbf{n} = \begin{pmatrix}-u(\mathbf{x})n_2\\ u(\mathbf{x})n_1\\ 0\end{pmatrix} \Leftrightarrow u(\mathbf{x}) = 0 \quad \text{on } \Gamma_B, \tag{2.26}$$

and

$$\begin{aligned}0 &= \nu(|\mathrm{curl}\mathbf{A}|)\mathrm{curl}\mathbf{A}(\mathbf{x}) \times \mathbf{n}\\ &= \nu(|\nabla u|)\begin{pmatrix}0\\0\\ \partial_2 u(\mathbf{x})n_2(\mathbf{x}) + \partial_1 u(\mathbf{x})n_1(\mathbf{x})\end{pmatrix} \quad \text{on } \Gamma_H,\end{aligned} \tag{2.27}$$

which leads to the equations of (nonlinear) 2D magnetostatics in classical formulation:

Find $u \in C^2(\Omega) \cap C^1(\Omega \cup \Gamma_H) \cap C(\Omega \cup \Gamma_B)$ such that

$$-\mathrm{div}\left(\nu(|\nabla u|)\nabla u(\mathbf{x})\right) = J_3(\mathbf{x}) \qquad \mathbf{x} \in \Omega, \tag{2.28a}$$

$$u(\mathbf{x}) = 0 \qquad \mathbf{x} \in \Gamma_D = \Gamma_B, \tag{2.28b}$$

$$\nu(|\nabla u|)\nabla u(\mathbf{x}) \cdot \mathbf{n} = 0 \qquad \mathbf{x} \in \Gamma_N = \Gamma_H. \tag{2.28c}$$

Assuming that the magnetic reluctivity does not depend on the magnetic field itself, problem (2.28) becomes a linear problem:

Find $u \in C^2(\Omega) \cap C^1(\Omega \cup \Gamma_H) \cap C(\Omega \cup \Gamma_B)$ such that

$$-\mathrm{div}\left(\nu\nabla u(\mathbf{x})\right) = J_3(\mathbf{x}) \qquad \mathbf{x} \in \Omega, \tag{2.29a}$$

$$u(\mathbf{x}) = 0 \qquad \mathbf{x} \in \Gamma_D = \Gamma_B, \tag{2.29b}$$

$$\nu\nabla u(\mathbf{x}) \cdot \mathbf{n} = 0 \qquad \mathbf{x} \in \Gamma_N = \Gamma_H. \tag{2.29c}$$

This assumption is justified for so-called *linear* materials. In this thesis, we will deal with ferromagnetic materials as iron, which are nonlinear materials. Nevertheless, for sake of simplicity, we will use the model (2.29) rather than (2.28), which will naturally result in a modeling error. Throughout this thesis, we will assume that Γ_D has positive measure because we want to avoid the case of the pure Neumann problem, a solution to which can be only unique up to an additive constant.

2.4.2 Existence and Uniqueness of Equations of Nonlinear 2D Magnetostatics

It is easily checked that the variational formulation of (2.28) reads as follows:

$$\text{Find } u \in V_g : \quad a(u; u, v) = \langle F, v \rangle \quad \forall v \in V_0, \tag{2.30}$$

with

$$V_g = V_0 = \{v \in H^1(\Omega) : v|_{\Gamma_D} = 0\}, \tag{2.31}$$

$$a(w; u, v) = \int_\Omega \nu(|\nabla w|)\nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x}) \, \mathrm{d}\mathbf{x} \text{ and} \tag{2.32}$$

$$\langle F, v \rangle = \int_\Omega f(\mathbf{x}) v(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \tag{2.33}$$

Remark 2.14. *For a function u in $H^1(\Omega)$, the second part of Theorem 2.12 is actually not applicable. This means that normal derivatives of $H^1(\Omega)$ functions cannot be represented by a function from $L^2(\Omega)$. However, for the case where $s = 1$, Neumann boundary data can be interpreted as functionals from the Sobolev space $H^{-1/2}(\partial\Omega)$ (cf.* Adams and Fournier *[1]). In our particular case where we only treat homogeneous Neumann boundary conditions we need not worry about this issue.*

The variational equation (2.30) is equivalent to the operator equation in the dual space,

$$A(u) = F \quad \text{ in } V_0^*, \tag{2.34}$$

with the nonlinear operator $A : V_0 \to V_0^*$ defined by

$$\langle A(u), v \rangle = a(u; u, v). \tag{2.35}$$

The well-posedness of a nonlinear variational problem of the form (2.30) is guaranteed by Theorem 2.15 which is refered to as the *nonlinear Lax-Milgram* theorem or theorem of *Zarantonello*.

Theorem 2.15 (Zarantonello)**.** *Let $(V, (\cdot,\cdot)_V, \|\cdot\|_V)$ be a Hilbert space, $F \in V^*$ and let $A : V \to V^*$ be a nonlinear operator fulfilling the following conditions:*

1. *A is strongly monotone:*

$$\exists c_1^A = const > 0 : \langle A(u) - A(v), u - v \rangle \geq c_1^A \|u - v\|_V^2 \quad \forall u, v \in V \tag{2.36}$$

2. *A is Lipschitz continuous*

$$\exists c_2^A = const > 0 : \|A(u) - A(v)\|_{V^*} \leq c_2^A \|u - v\|_V \quad \forall u, v \in V. \tag{2.37}$$

Then the operator equation

$$A(u) = F \tag{2.38}$$

has a uniquely determined solution $u^* \in V$.

Proof. A proof can be found in, e.g., PECHSTEIN [24]. □

Using this theorem, the well-posedness of the nonlinear variational problem (2.30) can be shown:

Theorem 2.16. *Let the mapping defined by $s \to \nu(s)s$ from $\mathbb{R}_0^+$ to $\mathbb{R}_0^+$ be strongly monotone with monotonicity constant m, i.e.,*

$$(\nu(s)s - \nu(t)t)(s-t) \geq m(s-t)^2 \quad \forall s,t \geq 0 \tag{2.39}$$

and Lipschitz continuous with Lipschitz constant $L > 0$,

$$|\nu(s)s - \nu(t)t| \leq L|s-t| \quad \forall s,t \geq 0. \tag{2.40}$$

Then the nonlinear variational problem (2.30) *has a unique solution $u^* \in V_0$ for any given $F \in V_0^*$.*

Proof. The conditions (2.39) and (2.40) imposed on ν yield the strong monotonicity and Lipschitz continuity of the operator A in (2.38). Thus, Theorem 2.15 is applicable. For more details, we again refer to PECHSTEIN [24]. □

Remark 2.17. *The conditions* (2.39) *and* (2.40) *are naturally satisfied as a consequence of certain physical properties (cf.* PECHSTEIN *[24]).*

2.4.3 Existence and Uniqueness of Equations of Linear 2D Magnetostatics

The variational formulation of (2.29) reads as follows:

$$\text{Find } u \in V_g: \quad a(u,v) = \langle F, v\rangle \quad \forall v \in V_0 \tag{2.41}$$

with

$$V_g = V_0 = \{v \in H^1(\Omega) : v|_{\Gamma_D} = 0\}, \tag{2.42}$$

$$a(u,v) = \int_\Omega \nu \nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x}) \, \mathrm{d}\mathbf{x} \text{ and} \tag{2.43}$$

$$\langle F, v\rangle = \int_\Omega f(\mathbf{x}) v(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \tag{2.44}$$

The lemma of *Lax-Milgram* states the conditions for well-posedness of a linear variational problem like (2.41).

Lemma 2.18 (Lax-Milgram)**.** *Let V be a normed linear space with norm $\|\cdot\|_V$, let the bilinear form $a : V \times V \to \mathbb{R}$ be coercive (elliptic), i.e.,*

$$a(v,v) \geq \mu_1 \|v\|_V^2 \qquad \forall v \in V.$$

for some constant $\mu_1 > 0$, and bounded (continuous), i.e.,

$$a(v,w) \leq \mu_2 \|v\|_V \|w\|_V \qquad \forall v,w \in V$$

for some constant $\mu_2 > 0$. Furthermore, let F be an element of V^*. Then there exists a unique solution to a variational problem of the form (2.41) *and*

$$\frac{1}{\mu_2}\|F\|_{V^*} \leq \|u\|_V \leq \frac{1}{\mu_1}\|F\|_{V^*}. \tag{2.45}$$

Proof. A proof can be found in most books on the finite element method, e.g., ZULEHNER [33]. □

To establish the well-posedness of (2.29) we have to check the coercivity and boundedness of the bilinear form (2.43), as well as the boundedness of the linear functional (2.44). Requiring the reluctivity to be bounded by two positive constants ν_0, ν_1 from below and above,

$$\nu_0 \leq \nu(\mathbf{x}) \leq \nu_1, \tag{2.46}$$

and using Cauchy's inequality on $L^2(\Omega)$, we immediately obtain the boundedness of the bilinear form:

$$a(u,v) = \int_\Omega \nu \nabla u \cdot \nabla v \, \mathrm{d}\mathbf{x} \leq \nu_1 \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}\mathbf{x} \leq \nu_1 \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} \tag{2.47}$$

$$\leq \nu_1 \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}. \tag{2.48}$$

As we require $meas_1(\Gamma_D) > 0$ we can use Friedrichs' inequality,

$$\exists C_F > 0 : \ \|v\|_{L^2(\Omega)} \leq C_F |v|_{H^1(\Omega)}, \tag{2.49}$$

which is valid for all v in V_0 (see, e.g., SCHÖBERL [29]), to show the coercivity of the bilinear form (2.43). First we note that adding $|v|_{H^1(\Omega)}$ on both sides of (2.49) gives

$$|v|_{H^1(\Omega)} \geq \frac{1}{C_F + 1} \|v\|_{H^1(\Omega)}. \tag{2.50}$$

Then we get

$$a(v,v) = \int_\Omega \nu |\nabla v|^2 \, \mathrm{d}\mathbf{x} \geq \nu_0 \int_\Omega |\nabla v|^2 \, \mathrm{d}\mathbf{x} = \nu_0 |v|^2_{H^1(\Omega)} \tag{2.51}$$

$$\geq \frac{\nu_0}{C_F + 1} \|v\|^2_{H^1(\Omega)}. \tag{2.52}$$

The boundedness of the linear form F (2.44) is easily verified using Cauchy's inequality:

$$\int_\Omega f v \, \mathrm{d}\mathbf{x} \leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)} \|v\|_{H^1(\Omega)} \tag{2.53}$$

Thus, Lemma 2.18 is applicable and yields existence and uniqueness of a solution to (2.41) as well as its continuous dependence on the right hand side F.

Chapter 3

Discontinuous Galerkin Methods

Discontinuous Galerkin (DG) methods are non-conforming finite element methods, which means that the approximation space V_h is not a subspace of the solution space V. More precisely, although we are searching for a solution in a space V of functions with continuous traces across the element interfaces, we make an ansatz with functions that are smooth only in the element interiors and allow for discontinuities across element interfaces. We try to regain continuity of the final solution by adding a penalization term to the bilinear form. The reason why we chose a DG method for the discretization of our topology optimization problem lies in the discontinuous nature of the design variable ρ. In the final (optimal) design, ρ should jump from 0 (void) to 1 (material). Using a conforming (continuous) version of the finite element method (FEM) would result in functions with very steep gradients at the interface between material and void, which we want to avoid.

In this chapter, we will first introduce the function spaces involved and then derive the DG variational formulation for a model problem. There exist several different versions of discontinuous Galerkin methods, but we will restrict ourselves to the class of *interior penalty* (IP) methods. We will discretize the obtained variational formulation of the model problem and treat the question of existence and uniqueness of a solution in the discrete setting. We will provide a priori error estimates and finally some numerical results for a model problem on the one hand and for a real world problem coming from computational electromagnetics on the other hand. The majority of this chapter is based on Rivière [27].

3.1 Preliminaries

An important tool in the analysis of DG methods are trace inequalities. For the rest of this thesis, we will restrict ourselves to the case of two space dimensions, i.e., $d = 2$. Let T be a bounded polygonal domain (e.g. a triangle in a FE mesh) with area $|T|$ and diameter $h_T = sup_{\mathbf{x},\mathbf{y}\in T}\|\mathbf{x}-\mathbf{y}\|$, where $\|\cdot\|$ denotes the Eucledian norm in $\mathbb{R}^d$. Furthermore, let $|e|$ denote the length of the edge e. Then there exists a constant C independent of h_T and v such that for any $v \in H^s(T)$ it holds

$$\|v\|_{L^2(e)} \leq C|e|^{1/2}|T|^{-1/2}(\|v\|_{L^2(T)} + h_T\|\nabla v\|_{L^2(T)}) \qquad \forall e \subset \partial T \quad \text{if } s \geq 1, \tag{3.1}$$

$$\|\nabla v \cdot \mathbf{n}\|_{L^2(e)} \leq C|e|^{1/2}|T|^{-1/2}(\|\nabla v\|_{L^2(T)} + h_T\|\nabla^2 v\|_{L^2(T)}) \quad \forall e \subset \partial T \quad \text{if } s \geq 2. \tag{3.2}$$

The constant C in (3.1) and (3.2) is a generic constant that is in general unknown. In the case where v is a polynomial we can exploit the equivalence of norms in finite-dimensional

spaces to obtain the trace inequalities

$$\|v\|_{L^2(e)} \leq C_t h_T^{-1/2} \|v\|_{L^2(T)} \; \forall e \subset \partial T \qquad \forall v \in P_k(T) \quad \text{and} \tag{3.3}$$

$$\|\nabla v \cdot \mathbf{n}\|_{L^2(e)} \leq C_t h_T^{-1/2} \|\nabla v\|_{L^2(T)} \; \forall e \subset \partial T \qquad \forall v \in P_k(T), \tag{3.4}$$

where $P_k(T)$ denotes the space of polynomials of total degree less than or equal to k on T,

$$P_k(T) = \text{span}\{x_1^I x_2^J : I + J \leq k, \; (x_1, x_2) \in T\}.$$

The constant C_t in (3.3) and (3.4) is independent of h_T and v, but depends on the polynomial degree k. For a polynomial $v \in P_k(T)$ on a planar triangle T, Hesthaven and Warburton [13] derived the trace inequality

$$\|v\|_{L^2(e)} \leq \sqrt{\frac{(k+1)(k+2)}{2} \frac{|e|}{|T|}} \|v\|_{L^2(T)}, \tag{3.5}$$

which does not include any unknown constants and which will help us in finding a proper coercivity constant (see Remark 3.8).

As already pointed out in the introduction of this chapter, in a DG method, given a subdivision $\mathcal{T}_h$ of the computational domain Ω, we are searching for solutions that are smooth in the interior of the elements of $\mathcal{T}_h$ and are possibly discontinuous across element interfaces. For a proper treatment of such functions we define the *broken Sobolev space* $H^s(\mathcal{T}_h)$, $s \in \mathbb{N}$, as

$$H^s(\mathcal{T}_h) = \{v \in L^2(\Omega) : v|_T \in H^s(T) \; \forall \; T \in \mathcal{T}_h\}, \tag{3.6}$$

equipped with the broken Sobolev norm

$$|||v|||_{H^s(\mathcal{T}_h)} = \left(\sum_{T \in \mathcal{T}_h} \|v\|^2_{H^s(T)} \right)^{\frac{1}{2}}. \tag{3.7}$$

Clearly, we have

$$H^s(\Omega) \subset H^s(\mathcal{T}_h) \quad \text{and} \quad H^{s+1}(\mathcal{T}_h) \subset H^s(\mathcal{T}_h).$$

Again, the definition can be extended to general indices $s \in \mathbb{R}$.

Now we are ready to define the interior penalty Galerkin (IPG) methods. We will illustrate the main ideas by deriving the variational formulation of the equations of 2D magnetostatics, which will be the governing equations in our optimization problem.

3.2 Variational Formulation of Interior Penalty Galerkin Methods

As a model problem, we will consider the two-dimensional problem (2.29) in a slightly more general setting, including inhomogeneous Dirichlet and Neuman boundary conditions:

Let Ω be a polygonal domain in $\mathbb{R}^2$ and $\mathcal{T}_h$ a subdivision of Ω into triangles. Let $\partial\Omega$ be partitioned into two disjoint sets, $\partial\Omega = \Gamma_D \cup \Gamma_N$, $\Gamma_D \cap \Gamma_N = \emptyset$, and let $\mathbf{n}$ be the outer unit

normal vector on $\partial\Omega$. In the case where $\Gamma_D = \emptyset$ a solution can be unique only up to an additive constant. Therefore, we assume $meas_{d-1}(\Gamma_D) > 0$. For the rest of this chapter we will restrict ourselves to the case of two space dimensions, i.e., $d = 2$. For given $f \in L^2(\Omega)$, $g_D \in H^{1/2}(\Gamma_D)$ and $g_N \in L^2(\Gamma_N)$, we formally consider the following elliptic boundary value problem:

$$-\mathrm{div}\,(\nu\nabla u) = f \quad \text{in } \Omega, \tag{3.8a}$$

$$u = g_D \quad \text{on } \Gamma_D, \tag{3.8b}$$

$$\nu\nabla u \cdot \mathbf{n} = g_N \quad \text{on } \Gamma_N. \tag{3.8c}$$

In our case, the real-valued coefficient function ν represents the magnetic reluctivity and has to be bounded from above and below by two positive constants ν_0, ν_1:

$$\nu_0 \leq \nu(\mathbf{x}) \leq \nu_1 \quad \text{a.e. in } \Omega. \tag{3.9}$$

We do not make any smoothness assumptions on ν as the magnetic reluctivity in problem (2.29) has jumps at the transition between different materials.

Remark 3.1. *The following procedures can be performed in an analogous way if we replace the real-valued function ν by a matrix valued function $\boldsymbol{\nu} = \boldsymbol{\nu}(\mathbf{x}) = (\nu_{ij}(\mathbf{x}))_{1\leq i,j\leq 2}$ that is symmetric ($\nu_{ij} = \nu_{ji}$), positive definite and uniformly bounded from above and below by two positive constants ν_0, ν_1:*

$$\nu_0\boldsymbol{\xi}\cdot\boldsymbol{\xi} \leq \boldsymbol{\nu}(\mathbf{x})\,\boldsymbol{\xi}\cdot\boldsymbol{\xi} \leq \nu_1\boldsymbol{\xi}\cdot\boldsymbol{\xi} \quad \forall\boldsymbol{\xi}\in\mathbb{R}^2 \quad \textit{a.e. in } \Omega.$$

Remark 3.2. *A subdivision $\mathcal{T}_h$ is called admissible if for any two elements T_i, $T_j \in \mathcal{T}_h$ the intersection of their closures, $\overline{T}_i \cap \overline{T}_j$, is either empty, a common vertex, a common edge or (in 3D) a common face. In particular, this property excludes the occurrence of so-called hanging nodes, which are vertices of elements that lie on an edge (or in 3D: face) of another element, see Figure 3.1. In a finite element method with continuous ansatz functions, hanging nodes must be avoided as they would destroy continuity across edges, whereas this admissibility requirement is not necessary when dealing with discontinuous ansatz functions. This is a major advantage of DG methods over conforming FE methods which allows for local refinement without accounting for neighboring elements.*

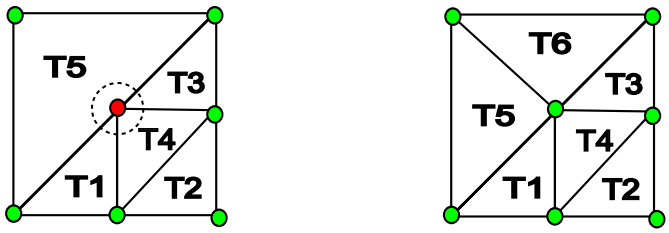

Figure 3.1: The red vertex in the left picture is a hanging node as it lies on an edge of T_5. In the right picture the neighboring element T_5 has been refined yielding an admissible subdivision again.

3.2.1 Derivation of the interior penalty variational formulation

Applying a DG method to a PDE means searching for a solution $u_h \in H^s(\mathcal{T}_h)$, i.e., a function that is smooth on element interiors but in general discontinuous across element interfaces. Thus, we cannot directly follow the procedure of deriving variational formulations in the conforming case where integration by parts is performed over the whole computational domain Ω. As we also treat the case of Neumann boundary conditions we need differentiability on the boundary. Due to Theorem 2.12, we get well-posedness of the boundary conditions for $s > 3/2$. For convenience we will assume $s = 2$.

Let us fix the following notation:

$\mathcal{T}_h$	set of triangles
Γ_h^0	set of interior edges
Γ_h^D	set of Dirichlet boundary edges
Γ_h^N	set of Neumann boundary edges

The derivation of the IP formulation of problem (3.8) consists of 5 steps. Note that ν, u, v and f are functions of $\mathbf{x}$ even though we will skip the arguments for better readability.

1. **Choice of approximation space $\mathbf{V_g}$ and test space $\mathbf{V_0}$:**

$$V_g = V_0 = V := H^s(\mathcal{T}_h)$$

2. **Multiply PDE with test function $\mathbf{v} \in \mathbf{V_0}$ integrate over each element and sum up:**

$$-\sum_{T\in\mathcal{T}_h} \int_T \operatorname{div}(\nu\,\nabla u)\, v \,\mathrm{d}\mathbf{x} = \sum_{T\in\mathcal{T}_h} \int_T f\, v \,\mathrm{d}\mathbf{x}$$

3. **Perform integration by parts (Ibp) in the principal part:**
Since the functions u and v are not continuously differentiable on the whole of Ω, we cannot apply integration by parts on Ω, but rather have to do it on each element separately:

$$-\sum_{T\in\mathcal{T}_h} \int_T \operatorname{div}(\nu\,\nabla u)\, v \,\mathrm{d}\mathbf{x} \overset{Ibp}{=} \sum_{T\in\mathcal{T}_h} \left[\int_T \nu\,\nabla u\cdot\nabla v \,\mathrm{d}\mathbf{x} - \int_{\partial T} \nu\nabla u\cdot\mathbf{n}\, v \,\mathrm{d}s \right].$$

Taking a closer look at the sum over the boundary integrals, we note that it can be rewritten as a sum over all edges. For an edge e, let T_r and T_s be the two adjacent elements, $\mathbf{n}_r$ and $\mathbf{n}_s$ be the corresponding outward unit normal vectors, and w_r, w_s denote the function values of a function w on T_r and T_s, respectively. Using this notation, we obtain

$$\begin{aligned}
-\sum_{T\in\mathcal{T}_h} \int_{\partial T} \nu\,\nabla u\cdot\mathbf{n}\, v \,\mathrm{d}s = &-\sum_{e\in\Gamma_h^0} \int_e \left((\nu\,\nabla u)|_{T_r}\cdot\mathbf{n}_r\, v_r + (\nu\,\nabla u)|_{T_s}\cdot\mathbf{n}_s\, v_s\right) \mathrm{d}s \\
&-\sum_{e\in\Gamma_h^D} \int_e (\nu\,\nabla u)|_{T_r}\cdot\mathbf{n}_r\, v_r \,\mathrm{d}s \\
&-\sum_{e\in\Gamma_h^N} \int_e (\nu\,\nabla u)|_{T_r}\cdot\mathbf{n}_r\, v_r \,\mathrm{d}s.
\end{aligned}$$

Now, on interior and Dirichlet boundary edges both $(\nu \nabla u)|_{T_r}$ and $(\nu \nabla u)|_{T_s}$ are replaced by an average value over the edge e, defined by

$$\{w\}_e := \begin{cases} \frac{1}{2}(w_r + w_s) & e \in \Gamma_h^0 \\ w_r & e \in \Gamma_h^D, \end{cases} \tag{3.10}$$

which gives

$$\begin{aligned} -\sum_{T\in\mathcal{T}_h} \int_{\partial T} \nu \nabla u \cdot \mathbf{n}\, v \, \mathrm{d}s = &- \sum_{e\in\Gamma_h^0} \int_e (\{\nu \nabla u\}_e \cdot \mathbf{n}_r\, v_r + \{\nu \nabla u\}_e \cdot \mathbf{n}_s\, v_s) \, \mathrm{d}s \\ &- \sum_{e\in\Gamma_h^D} \int_e \{\nu \nabla u\}_e \cdot \mathbf{n}_r\, v_r \, \mathrm{d}s - \sum_{e\in\Gamma_h^N} \int_e (\nu \nabla u)|_{T_r} \cdot \mathbf{n}_r\, v_r \, \mathrm{d}s. \end{aligned}$$

Proceeding like this is possible because the exact solution has continuous traces across the edge interfaces. In particular, for $u \in H^2(\Omega)$ it holds $(\nu\nabla u)|_{T_r} = (\nu\nabla u)|_{T_s}$. Furthermore, noting that $\mathbf{n}_s$ equals $-\mathbf{n}_r$ and introducing the jump of a function v on an interior or Dirichlet edge e as

$$[\![v]\!]_e := \begin{cases} v_r - v_s & e \in \Gamma_h^0 \\ v_r & e \in \Gamma_h^D, \end{cases} \tag{3.11}$$

the above sum can be written as

$$- \sum_{e\in\Gamma_h^0\cup\Gamma_h^D} \int_e \{\nu \nabla u \cdot \mathbf{n}_r\}_e\, [\![v]\!]_e \mathrm{d}s - \sum_{e\in\Gamma_h^N} \int_e (\nu \nabla u)|_{T_r} \cdot \mathbf{n}_r\, v_r \mathrm{d}s.$$

Since a boundary edge e can have only one adjacent element, we skip the subscripts for the Neumann edges. Combining our findings, we arrive at the following identity:
Find $u \in H^s(\mathcal{T}_h)$ such that

$$\sum_T \int_T \nu\nabla u \cdot \nabla v \; \mathrm{d}\mathbf{x} - \sum_{e\in\Gamma_h^0\cup\Gamma_h^D} \int_e \{\nu\nabla u \cdot \mathbf{n}_r\}_e [\![v]\!]_e \, \mathrm{d}s = \int_\Omega f v \; \mathrm{d}\mathbf{x} + \sum_{e\in\Gamma_h^N} \int_e \nu\nabla u \cdot \mathbf{n}\, v \, \mathrm{d}s.$$

4. **Incorporate natural boundary conditions**
Incorporating Neumann boundary conditions is of course done by replacing the conormal derivative along Neumann boundary edges by the corresponding Neumann data g_N:

$$\sum_{e\in\Gamma_h^N} \int_e \nu\nabla u \cdot \mathbf{n}\, v \, \mathrm{d}s = \sum_{e\in\Gamma_h^N} \int_e g_N\, v \, \mathrm{d}s.$$

5. **Add IPG terms**

- Since the exact solution of (3.8) is smooth on Ω and satisfies the Dirichlet boundary conditions, we have $[\![u]\!]_e = 0$ for interior edges and $[\![u - g_D]\!]_e = 0$ for Dirichlet boundary edges. Thus, adding the term

$$\beta \left(\sum_{e\in\Gamma_h^0} \int_e \{\nu \nabla v \cdot n\} [\![u]\!] \mathrm{d}s + \sum_{e\in\Gamma_h^D} \int_e \{\nu \nabla v\} [\![u - g_D]\!] \mathrm{d}s \right) \tag{3.12}$$

to the bilinear form does not destroy consistency. The parameter β defines the different IPG methods.

- In order to penalize discontinuities of the solution $u \in H^s(\mathcal{T}_h)$, the additional penalty term

$$\sum_{e \in \Gamma_h^0} \frac{\sigma_e}{|e|} \int_e [\![u]\!]_e [\![v]\!]_e \mathrm{d}s + \sum_{e \in \Gamma_h^D} \frac{\sigma_e}{|e|} \int_e (u - g_D) v \ \mathrm{d}s \tag{3.13}$$

is added to the bilinear form. As we will see in Section 3.4, the penalty parameter σ_e can be used to make the bilinear form coercive. Note that if the mesh size h goes to zero, so do the lengths of the edges, $|e|$, and the penalty term (3.13) becomes more and more dominant, making discontinuities of the solution across edges more and more "expensive".

For different choices of β we obtain different IPG methods:
$\beta = -1$ *Symmetric Interior Penalty Galerkin* (SIPG)
$\beta = 0$ *Incomplete Interior Penalty Galerkin* (IIPG)
$\beta = 1$ *Non-symmetric Interior Penalty Galerkin* (NIPG)

Remark 3.3. *Let us make some remarks on the derivation of the IPG methods:*

- *The trace of a function $v \in H^1(\mathcal{T}_h)$ is well-defined along any edge of the mesh because of a local version of Theorem 2.12.*
- *From now on, we will omit the subscripts in $\{\cdot\}_e$ and $[\![\cdot]\!]_e$ if there is no danger of confusion.*
- *Note that we do not require the subdivision to be admissible.*

In the following, we will restrict ourselves to the case where $\beta = -1$ (SIPG). Thus, after discretization, the resulting stiffness matrix is symmetric. The SIPG formulation of problem (3.8) looks as follows (recall that we require $s > 3/2$):

$$\text{Find } u \in H^s(\mathcal{T}_h): \quad a_h(u, v) = \langle F, v \rangle \quad \forall v \in H^s(\mathcal{T}_h), \tag{3.14}$$

with the SIPG bilinear form

$$\begin{aligned} a_h(u, v) = &\sum_T \int_T \nu \nabla u \cdot \nabla v \ \mathrm{d}\mathbf{x} - \sum_{e \in \Gamma_h^0 \cup \Gamma_h^D} \int_e \{\nu \nabla u \cdot n\} [\![v]\!] \ \mathrm{d}s \\ &- \sum_{e \in \Gamma_h^0 \cup \Gamma_h^D} \int_e \{\nu \nabla v \cdot n\} [\![u]\!] \ \mathrm{d}s + \sum_{e \in \Gamma_h^0 \cup \Gamma_h^D} \frac{\sigma_e}{|e|} \int_e [\![u]\!][\![v]\!] \ \mathrm{d}s \end{aligned} \tag{3.15}$$

and the functional F defined by

$$\langle F, v \rangle = \int_\Omega f \ v \ \mathrm{d}x + \sum_{e \in \Gamma_h^N} \int_e g_N \ v \ \mathrm{d}s + \sum_{e \in \Gamma_h^D} \int_e \nu \nabla v \cdot \mathbf{n} \ g_D \ \mathrm{d}s + \sum_{e \in \Gamma_h^D} \frac{\sigma_e}{|e|} \int_e g_D \ v \ \mathrm{d}s. \tag{3.16}$$

Remark 3.4. *Note that problem* (3.14) *does not depend on the choice of the normal* $\mathbf{n}_e$*. Indeed, let* e *be one edge shared by two elements* T_i *and* T_j *and let* $\mathbf{n}_{ij}$ *be the unit normal vector pointing from* T_i *to* T_j*. If* $\mathbf{n}_e$ *coincides with* $\mathbf{n}_{ij}$*, we have*

$$\int_e \{\nu \nabla u \cdot n\} [\![v]\!]_e ds = \int_e \{\nu \nabla u \cdot \mathbf{n}_{ij}\}(v|_{T_i} - v|_{T_j}) ds.$$

If $\mathbf{n}_e$ *has the opposite direction to* $\mathbf{n}_{ij}$ *then the jump* $[\![w]\!]$ *has a different sign and*

$$\int_e \{\nu \nabla u \cdot n\} [\![v]\!]_e ds = \int_e \{\nu \nabla u \cdot (-\mathbf{n}_{ij})\}(v|_{T_j} - v|_{T_i}) ds$$

which gives the same expression as above.

Proposition 3.5 (Consistency, [27])**.** *Let* $s > 3/2$ *and assume that the weak solution* u *of problem* (3.8) *belongs to* $H^s(\mathcal{T}_h)$*. Then* u *satisfies the variational problem* (3.14)*.*
Conversely, if $u \in H^1(\Omega) \cap H^s(\mathcal{T}_h)$ *satisfies* (3.14)*, then* u *is the solution of problem* (3.8)*.*

The first statement immediately follows from the derivation of (3.14). A proof of this proposition can be found in Rivière [27]. The proposition is valid for any fixed $\beta \in \mathbb{R}$.

3.3 Finite Element Approximation

For the discretization of problem (3.14) we are going to use a subspace of $H^1(\mathcal{T}_h)$ consisting of piecewise polynomials: For a given subdivision $\mathcal{T}_h$ of the computational domain Ω and a given polynomial degree $k \in \mathbb{N}$, we define the finite element space

$$\mathcal{D}_k(\mathcal{T}_h) = \{v_h \in L^2(\Omega) : v_h|_T \in P_k(T) \; \forall \, T \in \mathcal{T}_h\} \tag{3.17}$$

where $P_k(T)$ denotes the space of polynomials of total degree less than or equal to k on the element T. Note that functions in $\mathcal{D}_k(\mathcal{T}_h)$ are in general discontinuous across element interfaces. Without any difficulty we can extend the above definition and allow for different polynomial degrees k_T on each element T.

The discretized version of problem (3.14) gives the following DG variational problem on $\mathcal{D}_k(\mathcal{T}_h)$:

$$\text{Find } u_h \; \in \mathcal{D}_k(\mathcal{T}_h) : \quad a_h(u_h, v_h) = \langle F, v_h \rangle \quad \forall v_h \in \mathcal{D}_k(\mathcal{T}_h) \tag{3.18}$$

with $a_h(\cdot,\cdot)$ and $\langle F, \cdot \rangle$ as in (3.15) and (3.16), respectively.

Next we will treat the question of existence and uniqueness of a solution to the discretized problem (3.18).

3.4 Existence and Uniqueness

Next, we will check the properties of coercivity and boundedness of the bilinear form (3.15) on the finite element space $\mathcal{D}_k(\mathcal{T}_h)$. For a given subdivision $\mathcal{T}_h$ of our computational domain Ω, we define the following *energy norm* on $\mathcal{D}_k(\mathcal{T}_h)$, sometimes also called *DG-norm*:

$$\|v\|_{\mathcal{E}} = \|v\|_{DG} = \left(\sum_{T \in \mathcal{T}_h} \int_T \nu \nabla v \cdot \nabla v \; \mathrm{d}x + \sum_{e \in \Gamma_h \cup \Gamma_D} \frac{\sigma_e}{|e|} \int_e [\![v]\!]_e^2 \; \mathrm{d}s \right)^{1/2}. \tag{3.19}$$

Note that this is a norm if and only if $meas(\Gamma_D) > 0$. Next, we will state and prove coercivity of the bilinear form $a_h(\cdot,\cdot)$ for the SIPG method (Rivière [27]):

Lemma 3.6. *Let $\mathcal{T}_h$ be a subdivision of the computational domain Ω. Let ν_0, ν_1 be as in (3.9), C_t as in (3.4) and n_0 be the maximum number of neighbors which an element T can have, i.e., $n_0 = 3$ for a conforming triangular mesh. Furthermore, choose $\sigma_e \geq \frac{2C_t^2\nu_1^2 n_0}{\nu_0}$ for all interior and Dirichlet boundary edges.*
Then the bilinear form $a_h(\cdot,\cdot)$ defined in (3.15) is coercive with coercivity constant $\mu_1 = \frac{1}{2}$ on the space $\mathcal{D}_k(\mathcal{T}_h)$ equipped with the norm $\|\cdot\|_{\mathcal{E}}$ defined in (3.19).

Proof. Our aim is to estimate the DG bilinear form

$$a_h(v,v) = \sum_T \int_T \nu\nabla v\cdot\nabla v\mathrm{d}\mathbf{x} - 2\sum_{e\in\Gamma_h^0\cup\Gamma_h^D}\int_e \{\nu\nabla v\cdot n\}[\![v]\!]\mathrm{d}s + \sum_{e\in\Gamma_h^0\cup\Gamma_h^D}\frac{\sigma_e}{|e|}\int_e [\![v]\!]_e[\![v]\!]\mathrm{d}s$$

on $\mathcal{D}_k(\mathcal{T}_h)$ from below by $\mu_1\|v\|_{\mathcal{E}}^2$ for some constant μ_1. Therefore, we take a closer look at the second sum. Using Cauchy-Schwarz's inequality we get the following estimate:

$$\begin{aligned}\sum_{e\in\Gamma_h\cup\Gamma_D}\int_e \{\nu\nabla v\cdot\mathbf{n}\}[\![v]\!]\,\mathrm{d}s &\leq \sum_{e\in\Gamma_h\cup\Gamma_D}\|\{\nu\nabla v\cdot\mathbf{n}\}\|_{L^2(e)}\|[\![v]\!]\|_{L^2(e)}\\ &= \sum_{e\in\Gamma_h\cup\Gamma_D}\|\{\nu\nabla v\cdot\mathbf{n}\}\|_{L^2(e)}\left(\frac{1}{|e|}\right)^{1/2-1/2}\|[\![v]\!]\|_{L^2(e)}. \qquad (3.20)\end{aligned}$$

Using the definition of the average (3.10), where we denote by T_1^e and T_2^e the two elements adjacent to an interior edge e, and the triangluar inequality as well as the boundedness of ν (3.9) and the trace inequality (3.4), we can further estimate the term $\|\{\nu\nabla v\cdot\mathbf{n}\}\|_{L^2(e)}$ for an interior edge e:

$$\begin{aligned}\|\{\nu\nabla v\cdot\mathbf{n}\}\|_{L^2(e)} &\leq \frac{1}{2}\|(\nu\nabla v\cdot\mathbf{n})|_{T_1^e}\|_{L^2(e)} + \frac{1}{2}\|(\nu\nabla v\cdot\mathbf{n})|_{T_2^e}\|_{L^2(e)}\\ &\leq \frac{\nu_1}{2}\|(\nabla v\cdot\mathbf{n})|_{T_1^e}\|_{L^2(e)} + \frac{\nu_1}{2}\|(\nabla v\cdot\mathbf{n})|_{T_2^e}\|_{L^2(e)}\\ &\leq \frac{C_t\nu_1}{2}h_{T_1^e}^{-1/2}\|\nabla v\|_{L^2(T_1^e)} + \frac{C_t\nu_1}{2}h_{T_2^e}^{-1/2}\|\nabla v\|_{L^2(T_2^e)}.\end{aligned}$$

Using the obvious inequality $|e| \leq h_T$ where h_T denotes the diameter of element T, we obtain

$$\begin{aligned}\int_e \{\nu\nabla v\cdot\mathbf{n}\}[\![v]\!] &\leq \frac{C_t\nu_1}{2}\left(\frac{1}{|e|}\right)^{1/2}|e|^{1/2}\left(h_{T_1^e}^{-1/2}\|\nabla v\|_{L^2(T_1^e)} + h_{T_2^e}^{-1/2}\|\nabla v\|_{L^2(T_2^e)}\right)\|[\![v]\!]\|_{L^2(e)}\\ &\leq \frac{C_t\nu_1}{2}\left(\frac{1}{|e|}\right)^{1/2}\left(h_{T_1^e}^{1/2}h_{T_1^e}^{-1/2}\|\nabla v\|_{L^2(T_1^e)} + h_{T_2^e}^{1/2}h_{T_2^e}^{-1/2}\|\nabla v\|_{L^2(T_2^e)}\right)\|[\![v]\!]\|_{L^2(e)}\\ &= \frac{C_t\nu_1}{2}\left(\frac{1}{|e|}\right)^{1/2}\left(\|\nabla v\|_{L^2(T_1^e)} + \|\nabla v\|_{L^2(T_2^e)}\right)\|[\![v]\!]\|_{L^2(e)}\\ &\leq \frac{\sqrt{2}C_t\nu_1}{2}\left(\frac{1}{|e|}\right)^{1/2}\left(\|\nabla v\|_{L^2(T_1^e)}^2 + \|\nabla v\|_{L^2(T_2^e)}^2\right)^{1/2}\|[\![v]\!]\|_{L^2(e)}\\ &\leq C_t\nu_1\left(\frac{1}{|e|}\right)^{1/2}\left(\|\nabla v\|_{L^2(T_1^e)}^2 + \|\nabla v\|_{L^2(T_2^e)}^2\right)^{1/2}\|[\![v]\!]\|_{L^2(e)}\end{aligned}$$

where we have used the algebraic inequality $a+b \leq \sqrt{2a^2+2b^2}$ that is valid for non-negative a, b. For a boundary edge e with adjacent element T_1^e we can proceed in an analogous way to obtain

$$\int_e \{\nu \nabla v \cdot \mathbf{n}\}[\![v]\!] \leq C_t \nu_1 \left(\frac{1}{|e|}\right)^{1/2} \|\nabla v\|^2_{L^2(T_1^e)} \|[\![v]\!]\|_{L^2(e)}.$$

Considering the sum over all interior and boundary edges and applying Cauchy-Schwarz's inequality in the Euclidean space, $\sum a_i b_i \leq \left(\sum a_i\right)^{1/2} \left(\sum b_i\right)^{1/2}$, we get the following:

$$\begin{aligned}
\sum_{e \in \Gamma_h \cup \Gamma_h^D} \int_e \{\nu\ \nabla v \cdot \mathbf{n}\}[\![v]\!] &\leq C_t \nu_1 \left(\sum_{e \in \Gamma_h \cup \Gamma_h^D} \frac{1}{|e|} \|[\![v]\!]\|^2_{L^2(e)} \right)^{1/2} \\
&\quad \times \left(\sum_{e \in \Gamma_h} \|\nabla v\|^2_{L^2(T_1^e)} + \|\nabla v\|^2_{L^2(T_2^e)} + \sum_{e \in \Gamma_h^D} \|\nabla v\|^2_{L^2(T_1^e)} \right)^{1/2} \\
&\leq C_t \nu_1 \sqrt{n_0} \left(\sum_{e \in \Gamma_h \cup \Gamma_h^D} \frac{1}{|e|} \|[\![v]\!]\|^2_{L^2(e)} \right)^{1/2} \left(\sum_{T \in \mathcal{T}_h} \|\nabla v\|^2_{L^2(T)} \right)^{1/2}.
\end{aligned}$$

Using the lower bound in (3.9) and Young's inequality, $ab \leq \frac{\delta}{2} a^2 + \frac{1}{2\delta} b^2$ for $\delta > 0$, we obtain

$$\begin{aligned}
\sum_{e \in \Gamma_h \cup \Gamma_h^D} \int_e \{\nu\ \nabla v \cdot \mathbf{n}\}[\![v]\!] &\leq C_t \nu_1 \sqrt{(}n0) \left(\sum_{e \in \Gamma_h \cup \Gamma_h^D} \frac{1}{|e|} \|[\![v]\!]\|^2_{L^2(e)} \right)^{1/2} \left(\sum_{T \in \mathcal{T}_h} \frac{1}{\sqrt{\nu_0}} \|\nu^{1/2} \nabla v\|^2_{L^2(T)} \right)^{1/2} \\
&\leq \frac{\delta}{2} \sum_{T \in \mathcal{T}_h} \int_T \nu |\nabla v|^2 \, \mathrm{d}x + \frac{C_t^2 \nu_1^2 n_0}{2\delta \nu_0} \sum_{e \in \Gamma_h \cup \Gamma_h^D} \frac{1}{|e|} \|[\![v]\!]\|^2_{L^2(e)}.
\end{aligned}$$

Now we get a lower bound for $a_h(v,v)$,

$$a_h(v,v) \geq (1-\delta) \sum_{T \in \mathcal{T}_h} \int_T \nu \nabla v \cdot \nabla v \, \mathrm{d}x + \sum_{e \in \Gamma_h \cup \Gamma_h^D} \frac{1}{|e|} \left(\sigma_e - \frac{C_t^2 \nu_1^2 n_0}{\delta \nu_0} \right) \|[\![v]\!]\|^2_{L^2(e)},$$

and obtain coercivity with coercivity constant $\mu_1 = 1/2$ with the choice $\delta = 1/2$ provided that $\sigma_e \geq \frac{2 C_t^2 K_1^2 n_0}{K_0}$. □

Remark 3.7.

- *The IIPG bilinear form ($\beta = 0$) is coercive with $\mu_1 = \frac{1}{2}$ if we choose $\sigma_e \geq \frac{C_t^2 K_1^2 n_0}{K_0}$ for all interior end Dirichlet boundary edges. The proof is identical to the proof above.*
- *It is easily seen that the NIPG bilinear form ($\beta = 1$) is coercive with coercivity constant $\mu_1 = 1$ if $\sigma_e > 0$ is chosen on each edge.*

Remark 3.8. *In Lemma 3.6 we showed the coercivity of the bilinear form* (3.15) *on $\mathcal{D}_k(\mathcal{T}_h)$ provided that the penalty parameter σ_e is larger than a threshold value which involves the*

constant C_t defined in (3.4)*, which is in general unknown. In this finite-dimensional framework, we can now make use of the trace inequality* (3.5) *to obtain a more precise threshold value. For a given triangle T, let θ^T denote the smallest angle in T, let ν_0^T, ν_1^T the lower and upper bound of ν on T, respectively, and k^T the polynomial degree of the approximation on T. Furthermore, for an interior edge e let T_1 and T_2 be the two adjacent elements, and for a boundary edge e we denote the neighboring element by T. Then the limiting value of the penalty is given by*

$$\sigma_e^* = \frac{3(\nu_1^{T_1})^2}{2\nu_0^{T_1}}(k^{T_1})(k^{T_1}+1)\ cot(\theta^{T_1}) + \frac{3(\nu_1^{T_2})^2}{2\nu_0^{T_2}}(k^{T_2})(k^{T_2}+1)\ cot(\theta^{T_2}) \qquad \forall e \in \Gamma_h, \quad (3.21)$$

$$\sigma_e^* = \frac{6(\nu_1^{T})^2}{\nu_0^{T}}(k^{T})(k^{T}+1)\ cot(\theta^{T}) \qquad \forall e \in \Gamma_h^D. \quad (3.22)$$

A proof can be found in Epshteyn and Rivière *[9].*

If $\sigma_e > 0$ for all edges e, it can easily be shown that the bilinear form $a_h(\cdot,\cdot)$ defined in (3.15) is bounded on $\mathcal{D}_k(\mathcal{T}_h)$ equipped with the energy norm $\|\cdot\|_{\mathcal{E}}$:

$$a_h(v,w) \le \mu_2 \|v\|_{\mathcal{E}} \|w\|_{\mathcal{E}} \qquad \forall v, w \in \mathcal{D}_k(\mathcal{T}_h). \quad (3.23)$$

Corollary 3.9 (Existence and Uniqueness)**.** *Let the assumptions of Lemma 3.6 hold. Then, the DG problem* (3.18) *has a unique solution $u_h \in \mathcal{D}_k(\mathcal{T}_h)$.*

Proof. The statement follows immediately from Lemma 2.18 with $V = \mathcal{D}_k(\mathcal{T}_h)$ and $\|\cdot\|_V = \|\cdot\|_{\mathcal{E}}$ together with Lemma 3.6 and property (3.23). □

Remark 3.10. *Note that the bilinear form $a_h(\cdot,\cdot)$ is in general not continuous on the broken space $H^2(\mathcal{T}_h)$ with respect to the energy norm (*Rivière *[27]). Therefore, we cannot apply the lemma of Lax-Milgram to problem* (3.14)*.*

3.5 Error Estimates

In this section we will again assume that the solution u belongs to $H^s(\mathcal{T}_h)$ for some $s > 3/2$ and state a priori error estimates in both the energy norm (3.19) and the $L^2(\Omega)$ norm. In the following, $u_h \in \mathcal{D}_k(\mathcal{T}_h)$ denotes the solution of problem (3.18).

Theorem 3.11 (Energy Error Estimate)**.** *Assume that the exact solution to* (3.14) *belongs to $H^s(\mathcal{T}_h)$ for $s > 3/2$. Assume also that the penalty parameter σ_e is chosen according to Lemma 3.6. Then, there exists a constant C independent of h such that the following optimal a priori error estimate holds:*

$$\|u - u_h\|_{\mathcal{E}} \le C\, h^{min(k+1,s)-1} |||u|||_{H^s(\mathcal{T}_h)} \quad (3.24)$$

Theorem 3.12 (L^2 Error Estimate)**.** *Assume that the assumptions of Theorem 3.11 hold. Then, there exists a constant C independent of h such that*

$$\|u - u_h\|_{L^2(\Omega)} \le C\, h^{min(k+1,s)} |||u|||_{H^s(\mathcal{T}_h)}. \quad (3.25)$$

Proofs to both theorems can be found in Rivière [27].

	linear ($k = 1$)		quadratic ($k = 2$)	
	$\sigma_I = 8$, $\sigma_D = 14$		$\sigma_I = 20$, $\sigma_D = 38$	
N_h	$L^2(\Omega)$ error	energy error	$L^2(\Omega)$ error	energy error
128	0.16448088	12.216117	0.13579264	11.605123
512	0.14367020	11.830577	0.01214983	2.4160455
2048	0.040351376	5.8590522	0.0014725878	0.62405851
8192	0.010499697	2.8966216	0.00018207063	0.15694465
32768	0.0026601091	1.4397248	$2.2699724 \cdot 10^{-5}$	0.039246521

Table 3.1: Errors in the $L^2(\Omega)$ norm and the DG-energy norm of SIPG solution to (3.8) with the above-mentioned data for linear and quadratic ansatz functions for different mesh sizes

3.6 Numerical Experiments

In this section we will present the numerical results we obtained by applying the SIPG method to two problems of the type (3.8). First we will present our results for the Poisson equation, then we will solve a problem from 2D magnetostatics.

3.6.1 Application to Poisson Equation

Consider problem (3.8) with $\nu(x) \equiv 1$, $\Omega = (0,1)^2$ and $\Gamma_D = \partial\Omega$. We want to approximate the exact solution

$$u(x,y) = \cos(8\pi x) + \cos(8\pi y), \quad (x,y) \in \Omega. \tag{3.26}$$

Thus, we choose $f(x) = 64\pi^2(\cos(8\pi x)+\cos(8\pi y))$ and $g_D(x,y) = u|_{\Gamma_D}(x,y)$. We computed the SIPG solution for piecewise linear and piecewise quadratic ansatz functions. Table 3.1 shows the error in the $L^2(\Omega)$ norm and in the energy norm (3.19), where we chose the penalization value σ_e according to (3.21) and (3.22). The computation was performed on a structured mesh as shown in the left picture in Figure 3.2, where the minimum angle of each triangle is $\theta^T = \pi/4$. This gives the threshold values $\sigma_I = 6$ on interior edges and $\sigma_D = 12$ on Dirichlet boundary edges for linear finite element functions ($k = 1$), as well as $\sigma_I = 18$ and $\sigma_D = 36$ for $k = 2$. We chose values slightly above those threshold values.

In each of the four columns of Table 3.1 one can observe the convergence rates established in Section 3.5. The middle and right picture in Figure 3.2 show the SIPG solution and the exact solution of (3.8) with the above-mentioned data using linear ansatz functions on a structured mesh of 32768 elements, respectively.

3.6.2 Application to Equations of linear 2D Magnetostatics

Now we will consider the linear 2D magnetostatics problem (2.29). We want to compute the magnetic field $u = u(x,y)$ generated by a direct current electromagnet as depicted in Figure 5.1 on page 42. These electromagnets are used for measurements of magneto-optic effects and have been developed at the Technical University of Ostrava, Czech Republic, see Lukáš [20] and the references therein. The left picture in Figure 3.3 shows a cross-section of the electromagnet. We assume that only two of the four coils at the four poles are active (indicated by '+' and '-' in the figure), which yields symmetry with respect to the x-axis and

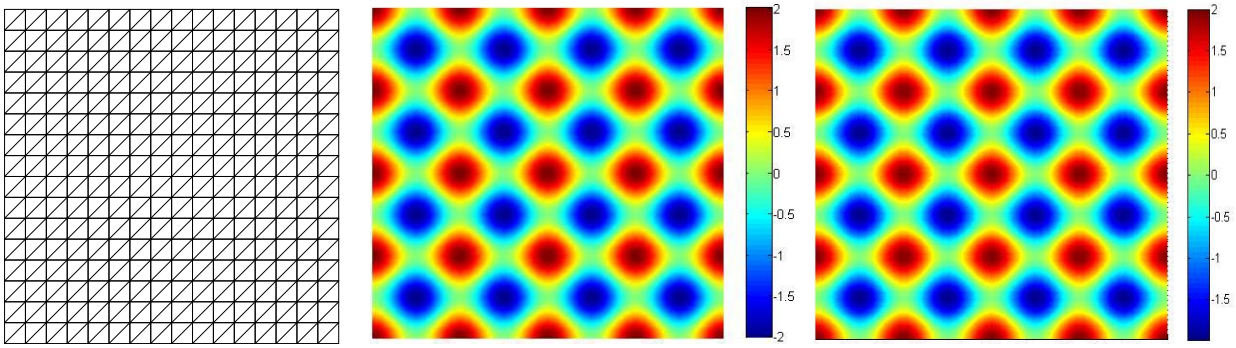

Figure 3.2: left: structured mesh consisting of $N_h = 512$ elements; center: solution obtained by the SIPG method on a mesh consisting of $N_h = 32768$ elements; right: exact solution (3.26)

antisymmetry with respect to the y-axis. Due to these symmetry properties, it suffices to consider only a quarter of the electromagnet. The antisymmetry with respect to the y-axis is modeled by imposing homogeneous Dirichlet boundary conditions, and the symmetry with respect to the x-axis by imposing homogeneous Neumann conditions on the respective parts of $\partial\Omega$. We assume homogeneous Dirichlet boundary conditions on the remaining part of the boundary, which corresponds to so-called conduction boundary conditions. The right picture in Figure 3.3 shows the computational domain Ω with the part occupied by ferromagnetic material (blue) and the coil (red). The magnetic reluctivity ν jumps from $\nu_0 = 10^7/(4\pi)$ in areas of air to $\nu_1 = \nu_0/5100$ in the ferromagnetic material. Note that the magnetic reluctivity of air can be assumed to be the same as the reluctivity of vacuum. The right hand side f corresponds to the impressed currents and is zero outside the coil (red area) and -10^6 inside.

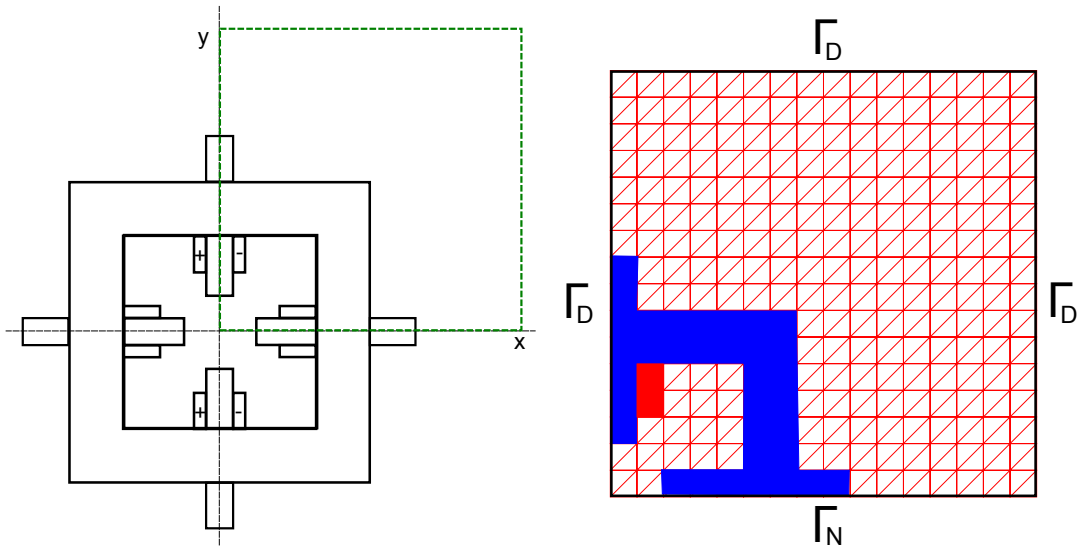

Figure 3.3: left: cross section of Maltese cross electromagnet; right: right upper section of left picture with the area of ferromagnetic material (blue) and the coil (red)

The solution of problem (2.29) is depicted in Figure 3.4. The left picture shows the solution we obtained by applying the SIPG method on a structured mesh with 32768 elements. As we do not know the exact solution of (2.29), we compared our solution with the solution obtained by a conforming finite element method on a mesh of the same size. The right picture in Figure 3.4 shows the conforming finite element solution obtained by a program developed by Dr. Clemens Pechstein at the Institute of Computational Mathematics at the Johannes Kepler University Linz (cf. PECHSTEIN ET AL. [25]).

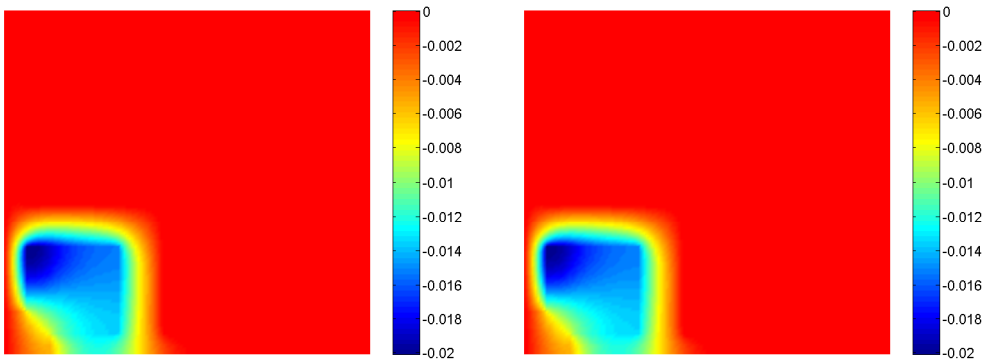

Figure 3.4: Numerical solution of the described problem obtained by SIPG (left) and by a conforming FE method (right) provided by Dr. Clemens Pechstein, Institute of Computational Mathematics, Johannes Kepler University Linz

Chapter 4

Abstract Topology Optimization

In this chapter we will give a brief overview on different ways of treating an abstract topology optimization problem. This overview is mainly based on Bendsøe and Sigmund [3], Stainko [30] and Christensen and Klarbring [7]. A topology optimization problem in its general form looks as follows:

$$\min_{\rho,u} \mathcal{J}(\rho, u) \tag{4.1a}$$

$$\text{subject to} \quad a(\rho; u, v) = \langle F, v \rangle \quad \forall v \in V_0 \tag{4.1b}$$

$$\int_\Omega \rho(\mathbf{x}) \ \mathrm{d}\mathbf{x} \leq V_{max} \tag{4.1c}$$

$$\rho(\mathbf{x}) \in \{0, 1\} \tag{4.1d}$$

Here, $\mathcal{J}$ represents the cost functional that is quantity to be minimized, u denotes the state variable and ρ the design variable which is only allowed to attain the values 1 (if point $\mathbf{x}$ should be occupied with material) or 0 (otherwise). These two variables are linked via the *state equation* (4.1b), here in variational form with V_0 being the corresponding space of test functions. Furthermore, the volume of the resulting structure is constrained by (4.1c). Due to this relation we will also refer to ρ as the *density* variable.

Remark 4.1. *For better imaginability the reader may think of the standard problem in structural mechanics, the minimal compliance problem, where the aim is to maximize the stiffness (or equivalently: to minimize the compliance $\mathcal{J}$) of a mechanical structure under given external loading. The state equations are the equations of (linearized) elasticity. We remark that in this case, in order to be able to guarantee coercivity of the bilinear form $a(\rho; \cdot, \cdot)$, the density ρ should not attain 0. Therefore, the constraint* (4.1d) *is usually replaced by $\rho(\mathbf{x}) \in \{\rho_{min}, 1\}$ with a small constant $\rho_{min} > 0$ while regions with $\rho(\mathbf{x}) = \rho_{min}$ are still interpreted as void. In contrast to elasticity, we allow ρ to be zero in electromagnetics, see Section 5.1.*

Remark 4.2. *In practical applications, we often have to impose a priori requirements on the resulting structure, i.e., some parts of the domain Ω should be occupied with material in any case or some parts should not. This can be achieved without any problem by defining the density variable ρ only on a proper subset $\Omega_d \subset \Omega$, called the design domain. In this chapter, we will, for simplicity, assume $\Omega_d = \Omega$. In Chapter 5, we will encounter an application from electromagnetics where we will have to account for such restrictions.*

Remark 4.3 (Nested and Simultaneous Formulation)**.** *There are basically two approaches to attack a PDE-constrained optimization problem like* (4.1)*. Let us briefly neglect the constraints* (4.1c) *and* (4.1d) *and consider the abstract problem*

$$\min_{\rho,u} \mathcal{J}(\rho, u)$$
$$\text{subject to } A_\rho u = F$$

where A_ρ denotes the differential operator in the state equation. When using the nested approach, the state variable u is - assuming the unique solvability of the state equation - expressed in terms of the design variable ρ via the state equation as $u = u(\rho) = A_\rho^{-1} F$. This results in an unconstrained optimization problem of the form

$$\min_{\rho} \tilde{\mathcal{J}}(\rho) := \mathcal{J}(\rho, u(\rho)).$$

On the one hand this reduces the problem's dimension as the optimization is now performed only in ρ, but on the other hand this also means that the state equation has to be solved for every evaluation of the objective functional. The nested approach is a so-called feasible path method, which means that the optimization is only performed on the manifold where the constraining PDE is satisfied. The nested approach is sometimes also called black-box method since the PDE solver realizing the operation $u = A_\rho^{-1} F$ can be inserted as black-box code into the optimization programm.

The second approach is called the simultaneous or all-at-once approach. The constrained optimization problem is solved by setting up and solving the first-order necessary optimality conditions (KKT conditions) as discussed in Section 2.1. This approach does not follow the feasible path, the state equation needs to be satisfied only by the final solution. Hence, a significant speed-up can be expected. We will follow the simultaneous approach for our benchmark problem in Chapter 5.

It is well-known that topology optimization problems of the form (4.1) are very likely to be ill-posed. More precisely, they often lack existence and uniqueness of solutions, which can be seen in the form of numerical problems of different kinds when applying straightforward solution methods (cf. Petersson and Sigmund [26]). To overcome these numerical anomalies we will investigate several different approaches.

But first we notice that we are facing a discrete-valued design problem, or a 0-1 problem, and that integer programming techniques for large-scale problems are usually rather inefficient. Therefore, the usual procedure in topology optimization is to relax condition (4.1d) by introducing a continuous density variable $\rho \in L^\infty(\Omega)$ which can take all values between 0 and 1. Problem (4.1) then becomes

$$\min_{\rho,u} \mathcal{J}(\rho, u) \tag{4.2a}$$
$$\text{subject to} \quad a(\rho; u, v) = \langle F, v \rangle \quad \forall v \in V_0 \tag{4.2b}$$
$$\int_\Omega \rho(\mathbf{x}) \ \mathrm{d}\mathbf{x} \leq V_{max} \tag{4.2c}$$
$$0 \leq \rho(\mathbf{x}) \leq 1 \tag{4.2d}$$

Problem (4.2) is often referred to as the *variable thickness sheet problem* and can also be regarded as a sizing optimization problem. A proof of existence of solutions in the case of the minimal compliance problem can be found in Stainko [30] and the references therein.

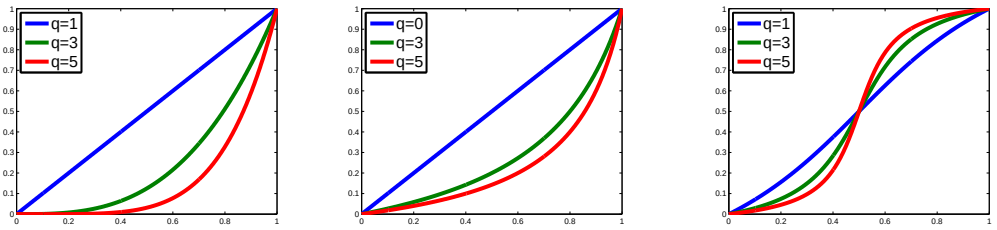

Figure 4.1: Illustration of penalizations of SIMP(left), RAMP(center) and ArcTan (right) type for different penalization values

Anyways, problem (4.2) is a different problem and leads to solutions different from the solutions of the original problem (4.1). Our goal is still to eventually obtain a black and white picture of the final structure, i.e., a solution with ρ attaining only values (close to) 0 and (close to) 1. We try to achieve this by *penalizing* intermediate values of ρ.

4.1 Penalization

The different penalization methods used in topology optimization can basically be split into two types:

In methods of the first type, the design variable ρ in the state equation (and only there) is replaced by $\eta(\rho)$ with η being a continuous, monotonously increasing function satisfying $\eta(0) = 0$ and $\eta(1) = 1$. The idea behind this approach is the following: Using $\eta(\rho) = \rho$ can be seen as a linear interpolation between the material properties of void ($\rho = 0$) and material ($\rho = 1$). Using a nonlinear function $\eta(\rho)$ means performing a nonlinear interpolation between 0 and 1. The function η is chosen in such a way that it is inefficient for the algorithm to choose intermediate values, $\rho(\mathbf{x}) \in (0, 1)$, as the obtained decrease of the objective functional is disproportionately low compared to the amount of used material. In other words, it makes it "uneconomical" to have intermediate densities in the optimal design.
The probably most popular penalization method is the SIMP method (*Solid Isotropic Material with Penalization*) where $\eta(\rho)$ is taken to be ρ^q for some $q > 1$. Other examples comprise the RAMP method (*Rational Approximation of Material Properties*) with

$$\eta(\rho(\mathbf{x})) = \frac{\rho(\mathbf{x})}{1 + q(1 - \rho(\mathbf{x}))}, \qquad q > 0 \tag{4.3}$$

or the ArcTan-choice

$$\eta(\rho(\mathbf{x})) = \frac{1}{2}\left(1 + \frac{\arctan(q(2\rho(\mathbf{x}) - 1))}{\arctan(\rho(\mathbf{x}))}\right), \qquad q > 0 \tag{4.4}$$

which - unlike the previous two approaches - penalizes 0 and 1 equally (cf. Lukáš [21]). These three penalization methods are depicted in Figure 4.1.

The second type of penalization methods consists in using the linear material interpolation $\eta(\rho(\mathbf{x})) = \rho(\mathbf{x})$, but driving the intermediate density values towards 0 or 1 by a term of the form

$$P(\rho(\mathbf{x})) = \int_\Omega W(\rho(\mathbf{x}))\mathrm{d}\mathbf{x} \tag{4.5}$$

with a non-negative, lower semicontinuous function $W : [0,1] \to \mathbb{R}$ which has exactly two roots, at 0 and 1. Often used choices of $W(\cdot)$ are

$$W(\rho(\mathbf{x})) = \rho(\mathbf{x})(1-\rho(\mathbf{x})) \quad \text{or} \quad W(\rho(\mathbf{x})) = \rho(\mathbf{x})^2(1-\rho(\mathbf{x}))^2.$$

The functional P in (4.5) can be included to the optimization problem in two different ways, either by adding an additional inequality constraint of the form

$$P(\rho(\mathbf{x})) \leq \varepsilon_P,$$

to the problem or by adding it as a penalty term to the objective functional as $J(\rho, u) + \gamma_P P(\rho)$. However, in both cases it is a tricky task to find proper values for ε_P or γ_P.

4.2 An Example from Structural Mechanics

We want to motivate our further proceedings by presenting some numerical results which were obtained by applying the SIMP method to the minimal compliance problem from structural mechanics. The results were produced in the course of a project for the lecture "Structural Optimization" held at the Division of Solid Mechanics at the University of Lund, Sweden, in spring 2011. The implementation was done in Matlab using the FEM package "CALFEM" developed and provided by that same department. The task was to find the optimal design for constructing a bridge.

Problem description: The bridge to be constructed consists of five layers of 5mm thickness each that are joined together. Each layer can be considered as a two-dimensional structure (*plain stress assumption*). The task is to find the distribution of material with material parameters $E = 210$ [GPa] and $\nu = 0.3$ and density 7800 [kg/m^3] on the design domain $[0,30]\times[0,18]$ [m] such that its stiffness is maximized without exceeding the maximum allowed weight of 10000kg. External forces simulating a total weight of 20000kg, which can be assumed to be homogeneously distributed, are applied at the top of the structure. In mathematical terms, this problem can be reformulated as follows:

$$\begin{aligned} &\min_{\mathbf{u}\in V_0,\rho} f(\mathbf{u}) \\ \text{subject to} \quad &a(\rho;\mathbf{u},\mathbf{v}) = f(\mathbf{v}) \quad \forall v \in V_0 \\ &\int_\Omega \rho(\mathbf{x})\ \mathrm{d}\mathbf{x} \leq V_{max} \\ &\rho(\mathbf{x}) \in \{\rho_{min}, 1\} \end{aligned} \tag{4.6}$$

with the equations of linearized elasticity as state equation, i.e.,

$$a(\rho;\mathbf{u},\mathbf{v}) = \int_\Omega \varepsilon(\mathbf{u}) \colon\ \mathbf{C}\varepsilon(\mathbf{v})\ \mathrm{d}\mathbf{x}, \tag{4.7}$$

$$f(\mathbf{v}) = \int_{\Gamma_t} \mathbf{v}\cdot\mathbf{t}\ \mathrm{d}s, \tag{4.8}$$

where

$$\varepsilon(\mathbf{u}) = \frac{1}{2}\left(\nabla u + \nabla u^T\right) \tag{4.9}$$

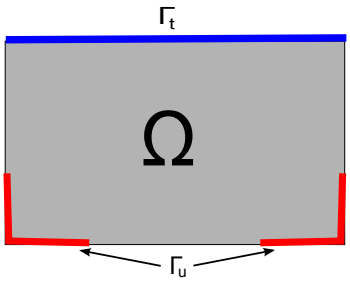

Figure 4.2: Initial design

denotes the linearized Green-St.Venant strain tensor and

$$C = \frac{E}{1-\nu^2}\begin{pmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & 1-\nu \end{pmatrix} \tag{4.10}$$

is the elasticity tensor under the plane stress assumption, i.e., the relation between the stress components $(\sigma_{11}, \sigma_{22}, \sigma_{12})^T$ and the strain components $(\varepsilon_{11}, \varepsilon_{22}, \varepsilon_{12})^T$. By Γ_t we denote the upper boundary of our computational domain Ω, see Figure 4.2, and t denotes the surface traction acting on that part of the boundary. The maximum allowed volume is denoted by $V_{max} = weight_{max}/density = 10000/7800\,[m^3]$ and $\rho_{min} > 0$ is an artificially introduced lower bound on the density ρ which assures ellipticity of the bilinear form (4.7). We set $\rho_{min} = 10^{-6}$. Furthermore, V_0 is the space of admissible displacements, defined as

$$V_0 = \{v : \overline{\Omega} \to \mathbb{R}^2 \mid v_i \in H^1(\Omega), i = 1, 2, \text{ and } v = \mathbf{0} \text{ on } \Gamma_u\} \tag{4.11}$$

with Γ_u as in Figure 4.2.

Introducing and penalizing intermediate values of ρ by the SIMP method as discussed above means to replace problem (4.6) by

$$\begin{aligned} &\min_{\mathbf{u}\in V_0, \rho \in L^\infty(\Omega)} f(\mathbf{u}) \\ \text{subject to} \quad &a(\rho^q; \mathbf{u}, \mathbf{v}) = f(\mathbf{v}) \quad \forall v \in V_0 \\ &\int_\Omega \rho(\mathbf{x})\ \mathrm{d}\mathbf{x} \le V_{max} \\ &\rho_{min} \le \rho(\mathbf{x}) \le 1 \end{aligned} \tag{4.12}$$

for some $q > 1$.

The optimization problem (4.12) was solved using the *optimality criteria* (OC) method. This method starts out from the nested formulation of (4.12), discretizes it using piecewise constant functions and linearizes the objective function in the variables

$$y_k = \rho_k^{-\alpha} \tag{4.13}$$

where ρ_k denotes the thickness of the k-th element. The resulting problem turned out to be convex and was solved using Lagrangian duality. For more details on that method we refer to Christensen and Klarbring [7].

Figure 4.3 shows the results obtained for different choices of the parameter α in the OC method. Figure 4.4 shows the result for the same parameters as in the left picture of Figure 4.3 on a mesh of half the size.

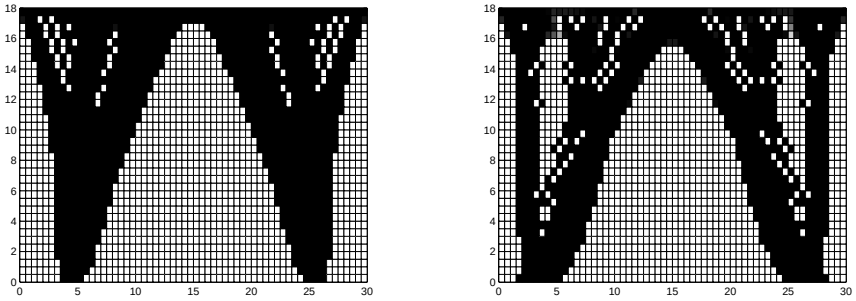

Figure 4.3: Results of problem (4.12) obtained by SIMP and OC method with penalization parameter $q = 3$ on a quadrilateral mesh with 2160 elements for different linearization parameters α; left: $\alpha = 2$, right: $\alpha = 1$

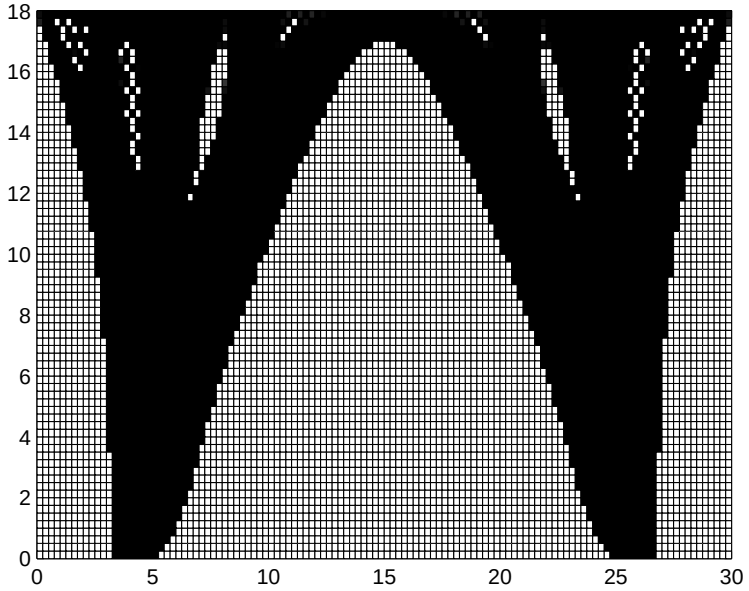

Figure 4.4: Result of problem (4.12) obtained by SIMP and OC method with penalization parameter $q = 3$, linearization parameter $\alpha = 2$ on a mesh with 8640 elements

4.3 Numerical Instabilities

A well-founded overview on numerical instabilities arising in topology optimization is given in Petersson and Sigmund [26]. The numerical problems appearing in topology optimization can basically be divided into three categories all of which we could observe in our numerical studies of the previous section:

By **mesh dependence** we understand the phenomenon that performing the same algorithm on a finer mesh yields a qualitatively different solution with more holes and finer structural elements rather than an improved picture of the same solution. This numerical anomaly is rooted in the non-existence of solutions to the underlying optimization problem. To be more precise, the reason is the non-losedness of the feasible design set. In terms of the minimal compliance problem this can be seen as follows: Inserting more and more holes while keeping the actual volume will result in stiffer and stiffer structures. This leads to an indefinite perforation and finally microstructures which are typically anisotropic and thus not in the feasible set any more, which indicates the lack of closure of the feasible design set. Comparing the left picture in Figure 4.3 and Figure 4.4 we can see that we obtain a topologically different solution on the finer grid even though the same algorithmic parameters are used.

The term **checkerboard pattern** refers to the reoccurring phenomenon of high oscillations of the density variable in the "optimal" solution of a topology optimization problem. In Figure 4.3 we can observe the density function ρ alternating between material and void - similar to the arrangement of a chess board, giving this numerical instability its name. The reason for this behaviour is usually bad numerical modeling.

Local minima: By penalizing intermediate density values, problem (4.2) is often turned from convex to non-convex. This means that performing the same optimization algorithm for different starting values or different algorithmic parameters can result in completely different "optimal" solutions. We could observe this behaviour in Figure 4.3 in terms of the algorithmic parameter α.

In the next section we will present different ways of dealing with these presented difficulties.

4.4 Regularization Methods

In this section we will shortly discuss various remedies for the mentioned numerical instabilities. For a more detailed survey we again refer to Petersson and Sigmund [26].

Let us first address the issue of non-existence of a solution. As mentioned above, the reason for non-existence of a solution is the lack of closure of the feasible design set. Now, given a non-closed set Q there are basically two ways to obtain a closed set: Either choose a closed subset $Q' \subset Q$ or choose a closed superset $Q'' \supset Q$. These approaches are called *restriction* and *relaxation*, respectively.

Remark 4.4. *Before going over to topology optimization problems we want to illustrate these methodologies in a simple example (*Christensen and Klarbring *[7]). Consider the optimization problem in* $\mathbb{R}$

$$\min_{x \in \mathcal{H}} f(x) \tag{4.14}$$

with $f(x) = \frac{1}{x}$ *and the set* $\mathcal{H} = \{x \in \mathbb{R} \mid x \geq 1\}$. *Obviously,* $\mathcal{H}$ *is not closed and the minimization problem has no solution. One possible cure for this ill-posed problem would now consist in restricting the feasible set* $\mathcal{H}$ *to a subset* $\mathcal{H}' = \{x \in \mathbb{R} \mid 1 \leq x \leq c\}$ *with some constant* $c > 1$*, yielding a well-posed problem.*
The opposite approach would be to enlarge the feasible set $\mathcal{H}$ *and to form a closure. By setting* $\mathcal{H}'' = \{x \in \mathbb{R} \mid x \geq 1\} \cup \{+\infty\}$ *and defining* $f(+\infty) = 0$ *we again obtain a solvable minimization problem.*

We will now discuss these two approaches in connection with topology optimization problems. Let us begin with relaxation:

4.4.1 Relaxation

Without going into detail, performing relaxation on a topology optimization problem would mean to form a closure of the feasible design set by including infinitely perforated structures. One relaxation method is the homogenization approach to topology optimization, described in detail in Bendsøe and Sigmund [3]. Applying relaxation to a topology optimization problem usually results in large areas with perforated microstructures and composite materials, which will probably be expensive and complicated to manufacture. Nevertheless, design with composite material is an area on its own (*material optimization*, cf. Bendsøe and Sigmund [3]).

4.4.2 Restriction

In topology optimization, restriction of the design set is realized by adding constraints to the optimization problem, again either in the form of an additional inequality constraint or by adding a penalty term to the objective functional. In order to exclude microstructures from the feasible design set, this additional constraint should somehow bound the maximum of allowed oscillation of ρ. Basically there exist three kinds of restriction methods:

In *perimeter* control, high variations of the density ρ are avoided by posing a bound on the perimeter of the structure which is - roughly speaking - the sum of lengths of all inner and outer boundaries. Of course, in presence of areas with intermediate density ($0 < \rho(\mathbf{x}) < 1$) it is not clear how to define the perimeter of design. Therefore, the usual way to simulate a bound on the perimeter is to pose a bound on the *total variation* of the density function ρ, that is $\int_\Omega |\nabla\rho(\mathbf{x})| \, \mathrm{d}\mathbf{x}$ if ρ is smooth enough. This choice makes sense since the total variation of ρ coincides with the perimeter of an area Ω_s when ρ is 1 in Ω_s and 0 elsewhere (Petersson and Sigmund [26]).

Another restriction method often used is called *gradient control* where one poses *global* or *local* bounds on the gradient of ρ, again assuming ρ to be sufficiently smooth. A global constraint can be in the form of a bound on the H^1-norm or H^1-seminorm. A proof of existence of solutions when using these bounds can be found in Bendsøe and Sigmund [3]. It is also possible to impose local constraints on the gradient, which results in a high number of additional constraints (in the order of the number of finite elements after discretization).

The third restriction method is to include *filters* in the optimization process as it is done in image processing to reduce high frequency components. Stainko [30] showed existence of solutions to the minimal compliance problem when combining a penalization like (4.5) with a filter operator.

We want to remark that for global restriction methods like perimeter or global gradient control only one constraint has to be added, whereas local methods like local gradient constraints or filter methods result in a high number of additional constraints. However, local methods will generally remove thin bars, which is not necessarily the case for global methods. Furthermore, determining a proper bound for a global constraint is a serious problem, which has to be solved by experiments. If the bound is too large the constraint will remain inactive and there is no regularizing effect. If, on the other hand, it is chosen too small there might be no optimal solution.

All the mentioned restriction methods bound the maximum of allowed variation of ρ and therefore eliminate not only the mesh-dependence, but also the checkerboard patterns in the resulting structure. However, with applying one of the restriction methods above comes one serious problem: The optimization problem is often turned from convex to non-convex. In order to cope with this, one can use *continuation methods*, i.e., one gradually changes the optimization problem from an (artificial) convex problem to the regularized (non-convex) problem one is actually interested in. For the SIMP method, this would mean to gradually increase the penalization parameter q from 1 (giving problem (4.2) which is in many cases convex) to higher values. If this change from convex to non-convex happens to abruptly, one risks getting stuck in a local minimum of the non-convex problem.

One example of a continuation method is the *phase-field* method, which we will discuss in detail in the next section.

4.5 The Phase-Field Method

Let from now on the density variable ρ be defined only on an open subset of Ω, the so-called design domain Ω_d.

The phase-field method is basically a restriction method as discussed in Section 4.4.2. More precisely, perimeter control is performed by adding the *total variation* of the density function ρ to the objective functional, weighted with a proper factor $\gamma > 0$. That means that (4.1a) is replaced by

$$\min_{\rho,u} \gamma \mathcal{J}(\rho, u) + |\rho|_{TV} \tag{4.15}$$

with the total variation of a function defined by

$$|\rho|_{TV} = \sup_{\substack{\mathbf{g} \in C_0^\infty(\Omega_d;\mathbb{R}^2) \\ \|\mathbf{g}\|_\infty \leq 1}} \int_{\Omega_d} \operatorname{div} \mathbf{g}(\mathbf{x})\, \rho(\mathbf{x})\, \mathrm{d}\mathbf{x}. \tag{4.16}$$

Note that for $\rho \in W^{1,1}(\Omega)$,

$$\int_{\Omega_d} \operatorname{div} \mathbf{g}\ \rho\ \mathrm{d}\mathbf{x} = -\int_{\Omega_d} \mathbf{g} \cdot \nabla\rho\ \mathrm{d}\mathbf{x} \quad \forall\, \mathbf{g} \in C_0^\infty(\Omega_d;\mathbb{R}^2),$$

and the supremum over all $\mathbf{g}$ with $\|\mathbf{g}\|_\infty \leq 1$ is given by

$$|\rho|_{TV} = \int_{\Omega_d} |\nabla\rho|\ \mathrm{d}\mathbf{x}, \tag{4.17}$$

cf. Caselles et al. [6]. The phase-field method now consists of two steps: First, as it is usual in computational topology optimization, the 0-1 condition (4.1d) is relaxed by introducing an $L^\infty(\Omega_d)$ function ρ that can attain any value between 0 and 1. The second step is to approximate the perimeter term in (4.15) by a functional of the form

$$P_\epsilon(\rho) = \frac{\epsilon}{2}\int_{\Omega_d} |\nabla\rho(\mathbf{x})|^2 \,\mathrm{d}\mathbf{x} + \frac{1}{\epsilon}\int_{\Omega_d} W(\rho(\mathbf{x})) \,\mathrm{d}\mathbf{x} \tag{4.18}$$

where $W : \mathbb{R} \to \mathbb{R}\cup\{+\infty\}$ is a positive lower semicontinuous function with exactly two roots, at 0 and at 1, and ϵ is a positive regularization parameter. The phase-field reformulation of the abstract topology optimization problem (4.2) then looks as follows:

$$\min_{\rho,u} \mathcal{J}_\epsilon(\rho, u) \tag{4.19a}$$

$$\text{subject to} \quad a(\rho; u, v) = \langle F, v\rangle \quad \forall v \in V_0 \tag{4.19b}$$

$$\int_{\Omega_d} \rho(\mathbf{x}) \,\mathrm{d}\mathbf{x} \le V_{max} \tag{4.19c}$$

$$0 \le \rho(\mathbf{x}) \le 1 \quad \text{a.e. in } \Omega_d \tag{4.19d}$$

with

$$\begin{aligned}\mathcal{J}_\epsilon(\rho, u) &= \gamma\,\mathcal{J}(\rho, u) + P_\epsilon(\rho)\\ &= \gamma\,\mathcal{J}(\rho, u) + \frac{\epsilon}{2}\int_{\Omega_d} |\nabla\rho(\mathbf{x})|^2 \,\mathrm{d}\mathbf{x} + \frac{1}{\epsilon}\int_{\Omega_d} W(\rho(\mathbf{x})) \,\mathrm{d}\mathbf{x}.\end{aligned} \tag{4.20}$$

The functional P_ϵ in (4.18) consists of two parts where the first term causes the regularizing effect of the method and the second term penalizes intermediate density values. They are weighted by ϵ and $1/\epsilon$, respectively. Thus, depending on the magnitude of ϵ, the first or the second part is dominant. The idea is now to perform a continuation in ϵ, i.e., to solve problem (4.19) for a decreasing sequence $\{\epsilon^{(k)}\}$. For large ϵ, the first term in (4.18) dominates, and, in many cases, the functional J_ϵ can be shown to be convex provided that ϵ is large enough (cf. Stainko [30] for the case where $J(\rho, u) = \int_\Omega \rho \,\mathrm{d}\mathbf{x}$). On the other hand, for small ϵ, the density is forced to take values close to 0 or 1, and J_ϵ becomes in general non-convex. As outlined in the previous section, continuation methods gradually change a problem from convex to non-convex. When decreasing ϵ in the phase-field method the optimal design of the previous step can be expected to be a good initial guess for the next step such that the algorithm does not get stuck in a local minimum, even though the functional J_ϵ might be non-convex.

The *theorem of Modica and Mortola* assures the convergence of minimizers $\{\overline{\rho}^{(k)}\}$ of the functional $\mathcal{J}_\epsilon$ as defined in (4.20) to a minimizer of (4.15) as $\epsilon \to 0$ in the sense of Γ-convergence. For more details we refer the reader to Stainko [30] and the references therein.

Common choices for the function W in (4.18) are

$$W(\rho) = \rho(1-\rho) \,\mathrm{d}\mathbf{x} \qquad \text{or} \qquad W(\rho) = \rho^2(1-\rho)^2 \,\mathrm{d}\mathbf{x}. \tag{4.21}$$

The solution of problem (4.19) with small ϵ can be expected to exhibit a sharp transition between areas of material and void. The thickness of the transition layer is in fact of order ϵ. Numerical results when applying this method to the minimal mass problem in structural mechanics are presented by Stainko in [30]. The minimal mass problem consists in minimizing the volume of a structure while keeping a certain stiffness to avoid material failure.

Chapter 5

Application to a Benchmark Problem from Electromagnetics

In this chapter we are going to apply the phase-field method described in Section 4.5 to a benchmark problem from electromagnetics. The task is to optimize the geometry of a direct current electromagnet such that the arising magnetic field minimizes a given functional. The benchmark problem is taken from Lukáš [20] and was treated by means of topology optimization in Lukáš [21]. In Section 5.1 we will give a description of the physical problem and set up a two-dimensional mathematical model. In Section 5.2 we will then formulate the necessary optimality conditions, perform a DG discretization as introduced in Chapter 3 and apply Newton's method to the arising system of nonlinear equations.

5.1 Problem Description

We consider a direct electric current electromagnet, the *Maltese Cross* electromagnet, as depicted in Figure 5.1, consisting of a ferromagnetic yoke and four poles that are equipped with coils which are pumped with direct electric current. This electromagnet is used for measurements of so-called *Kerr magneto-optic* effects which play a role in high capacity data storage media. In connection with this application, it is important that the magnetic field in the area around the pole heads in the center of the electromagnet is as *homogeneous*, (i.e. constant) as possible in different given directions. Our goal is to find an even better geometry for those purposes. In mathematical terms, it would be desirable to have a geometry such that the functional

$$\mathcal{J}\left(\operatorname{curl}\mathbf{u}\right) := \frac{1}{2}\int_{\Omega_m} |\operatorname{curl}\mathbf{u}(\mathbf{x}) - B_m^{avg}\mathbf{n}_m|^2 \,\mathrm{d}\mathbf{x} \tag{5.1}$$

is minimized where $\mathbf{u}$ is the magnetic potential such that $\mathbf{B} = \operatorname{curl}\mathbf{u}$ with $\mathbf{B}$ denoting the magnetic flux density as introduced in Section 2.4. The subdomain Ω_m is the magnetization area around the pole heads, see Figure 5.2, $\mathbf{n}_m$ is a given direction and B_m^{avg} denotes the average value of the magnetic field $\mathbf{B}$ in direction $\mathbf{n}_m$:

$$B_m^{avg} = B_m^{avg}(\operatorname{curl}\mathbf{u}) := \frac{1}{\operatorname{meas}\,\Omega_m}\int_{\Omega_m} |\operatorname{curl}\mathbf{u}(\mathbf{x})\cdot\mathbf{n}_m| \,\mathrm{d}\mathbf{x} \tag{5.2}$$

A Maltese Cross electromagnet is capable of performing such measurements in eight directions by just switching the currents in some coils off and on or by switching their directions. We will consider the magnetic field in only one direction which is generated only by the upper and lower coil in the right picture of Figure 5.1. Thus, we will consider the left and right coil to be inactive. Another issue is that the magnetic field $\mathbf{B}$ should also be as strong as possible. Therefore, the term

$$\xi \left(\min\{0, B_m^{avg} - B^{min}\}\right)^2 \tag{5.3}$$

is usually added to the objective functional (5.1), which penalizes magnetic fields with an average below a given minimal strength of B^{min}. The parameter ξ is the penalization parameter and is typically set to a high value, e.g., $\xi = 10^6$. However, in this thesis we will assume for simplicity that the average B_m^{avg} is a given constant, which makes the penalization term (5.3) redundant.

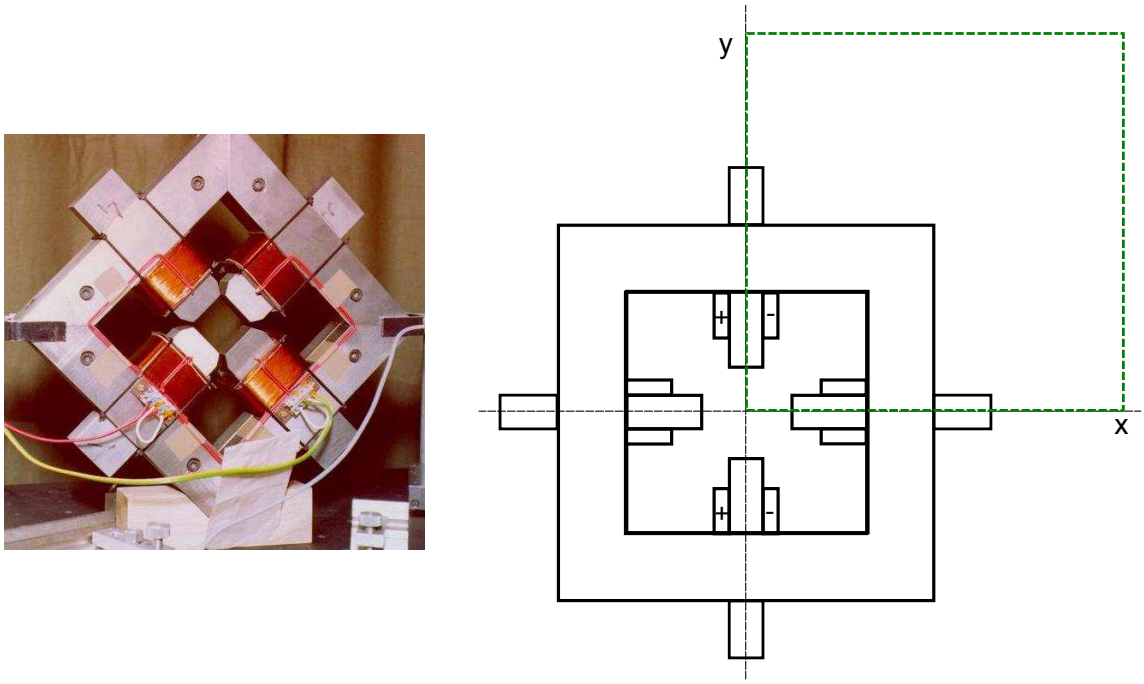

Figure 5.1: The Maltese Cross electromagnet

Let us now formulate the topology optimization problem (4.1) for the case of this benchmark problem. We will treat the problem in only two space dimensions, and use the mathematical model derived in Section 2.4.1. Then the vector-valued function $\mathbf{u}(x_1, x_2, x_3)$ can be replaced by a scalar function $u(x_1, x_2)$ and the curl operator in (5.1) and (5.2) becomes the operator $(\partial_2, -\partial_1)^T$. Due to symmetry considerations we consider the problem only on one quarter of the domain, see Figure 5.2. The anti-symmetry with respect to the y-axis in the right picture of Figure 5.1 yields homogeneous Dirichlet boundary conditions on the left boundary of Ω, and the symmetry with respect to the x-axis yields homogeneous Neumann boundary conditions on the lower boundary. Furthermore, we consider homogeneous Dirichlet boundary conditions on the upper and right part of the boundary, see Figure 5.2. By Ω_d we denote the subset of Ω where the material should be distributed. Note that we do

not allow material in the areas around the pole heads as well as in the coils. The objective functional lives on Ω_m and the state equations have to be fulfilled on the entire domain Ω. For simplicity, we assume linear behaviour of the materials involved such that our governing equations are the equations of linear 2D magnetostatics (2.41) on p. 16. As the direction in which we want to homogenize the magnetic field, we choose $\mathbf{n}_m = (0,1)^T$. As a further simplification, we will neglect the volume constraint (4.1c). Problem (4.1) then reads as follows:

$$\min_{u\in V_0,\rho} \quad \frac{1}{2}\int_{\Omega_m} \left| \begin{pmatrix} \partial_2 u \\ -\partial_1 u \end{pmatrix} - B_m^{avg}\mathbf{n}_m \right|^2 \mathrm{d}\mathbf{x} \tag{5.4a}$$

$$\text{subject to} \quad a(\rho; u, v) = \langle F, v\rangle \quad \forall v \in V_0, \tag{5.4b}$$

$$\rho(\mathbf{x}) \in \{0,1\} \quad \forall \mathbf{x} \in \Omega_d. \tag{5.4c}$$

Here, the governing equation is given in variational form with

$$a(\rho; u, v) = \int_\Omega \nu(\rho(\mathbf{x}))\nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x}) \, \mathrm{d}\mathbf{x} \qquad \text{and} \tag{5.5}$$

$$\langle F, v\rangle = \int_\Omega f(\mathbf{x}) v(\mathbf{x}) \, \mathrm{d}\mathbf{x} \tag{5.6}$$

where f denotes the third component of the impressed current density (denoted as J_3 in Section 2.4.1), and ν represents the magnetic reluctivity of air or of the ferromagnetic material,

$$f(\mathbf{x}) = \begin{cases} -10^6 & \mathbf{x} \in \Omega_c \\ 0 & \text{else} \end{cases} \qquad \nu(\mathbf{x}) = \begin{cases} \nu_1 & \text{if } \rho(\mathbf{x}) = 1 \\ \nu_0 & \text{if } \rho(\mathbf{x}) = 0 \end{cases} \in L^\infty(\Omega)$$

with $\nu_0 = \frac{1}{4\pi}10^7$, $\nu_1 = \frac{1}{5100}\nu_0$ and Ω_c denoting the area occupied by the coils. As the space of test functions V_0 we choose $V_0 = \{v \in H^1(\Omega) : v\big|_{\Gamma_D} = 0\}$. We mention that, in contrast to the corresponding topology optimization problem in elasticity (4.6), the bilinear form (5.5) is elliptic for any fixed ρ.

Remark 5.1. *We will not require the solution to be in $H^s(\Omega)$ for some $s > 3/2$ (as it would be necessary for applying Theorem 2.12), but interpret the normal derivative of a function u from $H^1(\Omega)$ as a functional from $H^{-1/2}(\partial\Omega)$, cf. Remark 2.14. As we are only treating homogeneous Neumann boundary conditions, this functional will be the zero functional.*

5.2 Application of Phase-Field Method

In this section we will apply the phase-field method as introduced in Section 4.5 to the benchmark problem (5.4) formulated in the previous section. We will derive the first-order necessary optimality conditions, discretize them by means of a discontinuous Galerkin method as introduced in Chapter 3 and finally solve the arising system of nonlinear equations by Newton's method.

As outlined in Section 4.5, applying the phase-field method to a topology optimization problem like (5.4) consists of two steps: On the one hand, the function $\rho(\mathbf{x}) \in \{0,1\}$ on Ω_d, cf. (5.4c), is replaced by a function $\rho(\cdot)$ which can attain all values between 0 and 1 on Ω_d. On the other hand, a phase-field functional of the type (4.18), which approximates the perimeter

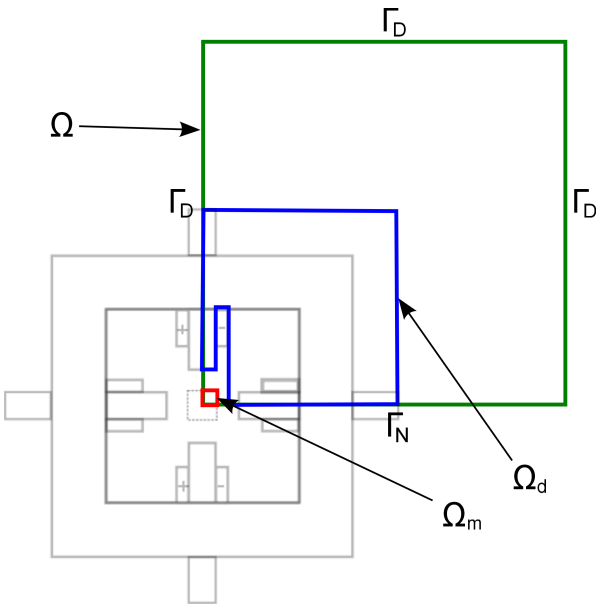

Figure 5.2: The Maltese Cross electromagnet

control term (4.16), is added to the objective functional. For the function W in (4.18), we choose $W(\rho) = \rho^2(1-\rho)^2$. Problem (5.4) then becomes

$$\begin{aligned} \min_{u\in V_0, \rho\in H^1(\Omega_d)\cap L^\infty(\Omega_d)} \mathcal{J}_\epsilon(u,\rho) &= \frac{\gamma}{2}\int_{\Omega_m} \left| \begin{pmatrix} \partial_2 u \\ -\partial_1 u \end{pmatrix} - B_m^{avg}\mathbf{n}_m \right|^2 \mathrm{d}\mathbf{x} \\ &\quad + \frac{\epsilon}{2}\int_{\Omega_d} |\nabla\rho|^2 \,\mathrm{d}\mathbf{x} + \frac{1}{\epsilon}\int_{\Omega_d} \rho^2(1-\rho)^2 \,\mathrm{d}\mathbf{x} \end{aligned} \tag{5.7a}$$

$$\text{subject to} \quad a(\rho; u, v) = \langle F, v\rangle \quad \forall v \in V_0 \tag{5.7b}$$

$$0 \le \rho(\mathbf{x}) \le 1 \qquad \text{a.e. in } \Omega_d \tag{5.7c}$$

with the notation from Section 5.1. We use the linear material interpolation

$$\nu(\rho(\mathbf{x})) = \begin{cases} \nu_0 + (\nu_1 - \nu_0)\rho(\mathbf{x}) & \text{if } \mathbf{x} \in \Omega_d \\ \nu_0 & \text{otherwise} \end{cases} \in L^\infty(\Omega) \tag{5.8}$$

that interpolates linearly between the material properties of air ($\nu = \nu_0$) and the ferromagnetic material ($\nu = \nu_1$). Note that we assume linear behaviour of the ferromagnetic material, i.e., that the reluctivity ν_1 is constant and does depend on the magnetic field. In order to guarantee that $\nu \in L^\infty(\Omega)$, we require $\rho \in H^1(\Omega_d) \cap L^\infty(\Omega_d)$. Recall that $\gamma > 0$ is a fixed weighting factor and $\epsilon > 0$ is the regularization parameter. Problem (5.7) is solved for a decreasing sequence $\{\epsilon^{(l)}\}$ where the solution for $\epsilon = \epsilon^{(l)}$ is used as an initial guess for the next iteration step with $\epsilon = \epsilon^{(l+1)}$. For simplicity we will neglect the box constraints $0 \le \rho(\mathbf{x}) \le 1$.

5.2.1 Derivation of KKT System in Variational Form

We note that the variational equation (5.7b) can also be written in operator form as

$$G(u,\rho) := A_\rho u - F = 0 \quad \text{in } V_0^* \tag{5.9}$$

with the operator $A_\rho : V_0 \to V_0^*$ defined by the identity

$$\langle A_\rho u, v\rangle = a(\rho; u, v)\, \forall u, v \in V_0.$$

Interpreting (5.7b) in terms of (5.9), the minimization problem (5.7) is of the same form as (2.4) on p. 5 and we can define the Lagrangian. Identifying $(V_0^*)^*$ with V_0 via the Riesz respresentation theorem (cf., e.g., Adams and Fournier [1]), Definition 2.3 gives
$\mathcal{L} : V_0 \times H^1(\Omega_d) \cap L^\infty(\Omega_d) \times V_0 \to \mathbb{R}$,

$$\mathcal{L}(u, \rho, \lambda) = \mathcal{J}_\epsilon(u, \rho) + \langle \lambda, A_\rho u - F\rangle_{V_0 \times V_0^*}. \tag{5.10}$$

For convenience, we will flip the arguments in the duality product and skip the indices if there is no danger of confusion, i.e., we will write $\langle A_\rho u - F, \lambda\rangle$ instead of $\langle A_\rho u - F, \lambda\rangle_{V_0^* \times V_0} = \langle \lambda, A_\rho u - F\rangle_{V_0 \times V_0^*}$. Both the objective functional J_ϵ and the operator G are Fréchet differentiable. Before we derive the first-order necessary optimality conditions in variational form we check the constraint qualification (2.5) for our operator G. Defining $Y := V_0 \times H^1(\Omega_d)$, $\overline{y} = (\overline{u}, \overline{\rho})$ and $y = (u, \rho)$, condition (2.5) reads

$$G'(\overline{y})Y = V_0^*,$$

which means that, for any functional H in V_0^*, there should exist a $y \in Y$ such that $G'(\overline{y})y = H$ where $\overline{y} \in Y$ denotes the (unknown) exact solution of (5.7). For our operator this means that there should exist (u, ρ) such that

$$\begin{aligned} H \overset{!}{=} G'(\overline{y})y = \nabla G(\overline{u}, \overline{\rho}) \begin{pmatrix} u \\ \rho \end{pmatrix} &= \nabla_u G(\overline{u}, \overline{\rho})u + \nabla_\rho G(\overline{u}, \overline{\rho})\rho = A_{\overline{\rho}}u + A_\rho^{(2)}\overline{u} \\ &= a(\overline{\rho}; u, \cdot) + a^{(2)}(\rho; \overline{u}, \cdot) \end{aligned}$$

where the operator $A_\rho^{(2)}$ is defined by the identity

$$\langle A_\rho^{(2)} u, v\rangle = a^{(2)}(\rho; u, v) := (\nu_1 - \nu_0) \int_{\Omega_d} \rho \nabla u \cdot \nabla v \, \mathrm{d}\mathbf{x} \quad \forall u, v \in V_0.$$

Note that the trilinear form $a^{(2)}(\rho; u, v)$ represents the derivative of the bilinear form (5.5) with respect to ρ in case of the linear material interpolation (5.8). In other words, for any right hand side $H \in V_0^*$ there should exist an element $u \in V_0$ and $\rho \in H^1(\Omega_d)$ such that

$$a(\overline{\rho}; u, v) + a^{(2)}(\rho; \overline{u}, v) = \langle H, v\rangle \quad \forall v \in V_0. \tag{5.11}$$

If we now simply choose $\rho = 0$ then (5.11) becomes the linear problem (2.41) for which we showed the existence of a unique solution u for any right hand side $H \in V_0^*$ in Section 2.4.3, provided that the coefficient function $\nu(\mathbf{x}) = \nu(\rho(\mathbf{x}))$ is bounded from below and above by two positive constants ν_0 and ν_1 as in (2.46). This is of course satisfied by the optimal solution of (5.7) in combination with (5.8). Note that the discussion of the constraint qualification does not involve the regularization parameter ϵ. We say the problem satisfies a *uniform constraint qualification*.

Now Theorem 2.6 is applicable to our problem and the first-order necessary optimality conditions read as follows:

$$\begin{aligned}
\nabla_u \mathcal{L}(u,\rho,\lambda)p &= \gamma \int_{\Omega_m} \nabla u \cdot \nabla p \qquad + B_m^{avg} \begin{pmatrix} n_2 \\ -n_1 \end{pmatrix} \cdot \nabla p \, \mathrm{d}\mathbf{x} \\
&\qquad + \langle p, A_\rho^* \lambda \rangle \qquad\qquad = 0 \ \forall p \in V_0, \qquad (5.12a) \\
\nabla_\rho \mathcal{L}(u,\rho,\lambda)q &= \epsilon \int_{\Omega_d} \nabla \rho \cdot \nabla q \, \mathrm{d}\mathbf{x} \quad + \frac{1}{\epsilon} \int_{\Omega_d} (2\rho - 6\rho^2 + 4\rho^3) q \, \mathrm{d}\mathbf{x} \\
&\qquad + (\nu_1 - \nu_0) \int_{\Omega_d} q \nabla u \cdot \nabla \lambda \, \mathrm{d}\mathbf{x} \quad = 0 \ \forall q \in H^1(\Omega_d) \cap L^\infty(\Omega_d), \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (5.12b) \\
\nabla_\lambda \mathcal{L}(u,\rho,\lambda)v &= \langle A_\rho u - F, v \rangle \qquad\qquad = 0 \ \forall v \in V_0, \qquad (5.12c)
\end{aligned}$$

where $u \in V_0$, $\rho \in H^1(\Omega_d) \cap L^\infty(\Omega_d)$ and $\lambda \in V_0$. Note that in the derivation of the third term in (5.12b) we used

$$\begin{aligned}
&\left(\frac{\partial}{\partial \rho} \langle A_\rho u - F, \lambda \rangle \right) q = \left(\frac{\partial}{\partial \rho} \left(a(\rho; u, \lambda) - \langle F, \lambda \rangle \right) \right) q \\
&= \left(\frac{\partial}{\partial \rho} \left(\int_\Omega \nu_0 \nabla u \cdot \nabla \lambda \, \mathrm{d}\mathbf{x} + (\nu_1 - \nu_0) \int_{\Omega_d} \rho \, \nabla u \cdot \nabla \lambda \, \mathrm{d}\mathbf{x} - \langle F, \lambda \rangle \right) \right) q \\
&= \left(\frac{\partial}{\partial \rho} \left((\nu_1 - \nu_0) \int_{\Omega_d} \rho \, \nabla u \cdot \nabla \lambda \, \mathrm{d}\mathbf{x} \right) \right) q \\
&= (\nu_1 - \nu_0) \lim_{t \to 0} \frac{1}{t} \left(\int_{\Omega_d} (\rho + tq) \nabla u \cdot \nabla \lambda \, \mathrm{d}\mathbf{x} - \int_{\Omega_d} \rho \nabla u \cdot \nabla \lambda \, \mathrm{d}\mathbf{x} \right) \\
&= (\nu_1 - \nu_0) \int_{\Omega_d} q \, \nabla u \cdot \nabla \lambda \, \mathrm{d}\mathbf{x}.
\end{aligned}$$

Now we want to perform an IPG discretization of system (5.12). Since any DG discretization starts out from the *differential* (or *classical*) form of a PDE rather than the variational form, we first need to reformulate (5.12) in its classical form.

5.2.2 Derivation of KKT System in Differential Form

To formulate system (5.12) in differential form, we need to impose smoothness requirements on the functions u, ρ and λ in the classical sense:

$$u \in C^2(\Omega) \cap C^1(\Omega \cup \Gamma_N) \cap C^0(\Omega \cup \Gamma_D) \qquad (5.13)$$

$$\rho \in C^2(\Omega_d) \cap C^1(\overline{\Omega}_d) \qquad (5.14)$$

$$\lambda \in C^2(\Omega) \cap C^1(\Omega \cup \Gamma_N) \cap C^0(\Omega \cup \Gamma_D) \qquad (5.15)$$

A procedure that is essential when deriving the differential form of a PDE from the variational form is commonly refered to as "Euler's trick":

Remark 5.2 (Euler's trick)**.** *Let $\Omega \subset \mathbb{R}^d$. If*

$$\int_\Omega g(\mathbf{x}) v(\mathbf{x}) \ \mathrm{d}\mathbf{x} = 0 \quad \forall v \in C^2\left(\overline{\Omega}\right) \ \textit{with} \ v|_{\partial\Omega} = 0 \qquad (5.16)$$

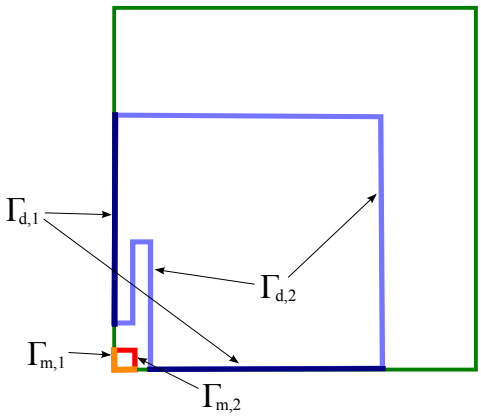

Figure 5.3: Computational domain for benchmark problem with boundaries and interfaces

and if g is continuous on $\overline{\Omega}$, then $g(\mathbf{x}) = 0$ on $\overline{\Omega}$.

This can be seen easily: Suppose that g is not the zero function, so there exists a point $\mathbf{x}^0$ with $g(\mathbf{x}^0) \neq 0$, w.l.o.g. $g(\mathbf{x}^0) > 0$. Due to the continuity of g there exists a neighborhood $U_\varepsilon(\mathbf{x}^0)$ of $\mathbf{x}^0$ such that $g(\mathbf{y}) > 0$ for all $\mathbf{y} \in U_\varepsilon(\mathbf{x}^0)$. If we now choose a certain non-negative function v that is zero outside $\overline{U}_\varepsilon(\mathbf{x}^0)$, positive in the interior and is smooth enough (such a function can be constructed easily, see, e.g., Fomin and Gelfand *[10]) then the integral over Ω reduces to an integral over $U_\varepsilon(\mathbf{x}^0)$ which is obviously positive, in contradiction to* (5.16)*. Hence, the assumption that g is not the zero function was false.*

This trick is also known as the fundamental lemma of calculus of variations (see, e.g., Fomin and Gelfand *[10]). We will apply this procedure to integrals over a two-dimensional domain Ω and to integrals over (a part of) the boundary $\partial\Omega$.*

A further essential ingredient for deriving the differential form of a PDE is the integration by parts formula

$$-\int_\Omega \operatorname{div}(a\nabla u)\, v \,\mathrm{d}\mathbf{x} = \int_\Omega a\nabla u \cdot \nabla v \,\mathrm{d}\mathbf{x} - \int_{\partial\Omega} a\nabla u \cdot \mathbf{n}\, v \,\mathrm{d}s \tag{5.17}$$

which is valid if a, u, v are sufficiently smooth such that the above integrals are defined.

Let us begin with the reformulation of equation (5.12a). We define $\mathbf{b}(\mathbf{x}) := B_m^{avg}\binom{n_2}{-n_1}$ and treat it as a general continuously differentiable function depending on $\mathbf{x}$ even though it is a constant vector in our case. Recall Figure 5.2 for the structure of the computational domain Ω with the boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$ and the subsets Ω_d and Ω_m which we assume to be open. We further define particular parts of the boundary and interfaces (see Fiugre 5.3):

$$\begin{aligned}\Gamma_{m,1} &= \partial\Omega_m \cap \partial\Omega \\ \Gamma_{m,2} &= \partial\Omega_m \backslash \Gamma_{m,1} \\ \Gamma_{d,1} &= \partial\Omega_d \cap \partial\Omega \\ \Gamma_{d,2} &= \partial\Omega_d \backslash \Gamma_{d,1}\end{aligned}$$

For better readability we will skip the arguments of the occurring functions. Note that the

test function p vanishes on the Dirichlet boundary Γ_D. From (5.12a) we obtain

$$0 = \gamma \int_{\Omega_m} \nabla u \cdot \nabla p + \mathbf{b} \cdot \nabla p \, \mathrm{d}\mathbf{x} + \int_{\Omega} \nu(\rho) \nabla p \cdot \nabla \lambda \, \mathrm{d}\mathbf{x} \tag{5.18}$$

$$\begin{aligned} &\overset{Ibp}{=} -\gamma \int_{\Omega_m} (\Delta u + \operatorname{div} \mathbf{b}) p \, \mathrm{d}x + \gamma \int_{\partial\Omega_m \backslash \Gamma_D} (\nabla u + \mathbf{b}) \cdot \mathbf{n} \, p \, \mathrm{d}s \\ &\quad - \int_{\Omega} \operatorname{div} \left(\nu(\rho) \nabla \lambda\right) p \, \mathrm{d}\mathbf{x} + \int_{\partial\Omega \backslash \Gamma_D} (\nu(\rho) \nabla \lambda) \cdot \mathbf{n} \, p \, \mathrm{d}s \qquad \forall \, p \in V_0. \end{aligned} \tag{5.19}$$

Since this equality must hold for all functions $p \in V_0$, it must also hold for all those functions p that vanish outside Ω_m:

$$\int_{\Omega_m} \left(-\gamma \Delta u - \gamma \operatorname{div} \mathbf{b} - \operatorname{div} \left(\nu(\rho) \nabla \lambda\right)\right) p \, \mathrm{d}\mathbf{x} = 0 \quad \forall p \in V_0 : p|_{\overline{\Omega} \backslash \Omega_m} = 0. \tag{5.20}$$

Since the involved functions are sufficiently smooth, Remark 5.2 yields

$$-\gamma \Delta u - \gamma \operatorname{div} \mathbf{b} - \operatorname{div} \left(\nu(\rho) \nabla \lambda\right) = 0 \qquad \forall \mathbf{x} \in \Omega_m \tag{5.21}$$

If we now choose a function $p \in V_0$ that vanishes on $\partial\Omega$ and on $\Gamma_{m,2}$ and if we plug in (5.21) into (5.19), we obtain

$$-\int_{\Omega \backslash \Omega_m} \operatorname{div}(\nu(\rho) \nabla \lambda) p \, \mathrm{d}\mathbf{x} = 0 \quad \forall p \in V_0 : p|_{\partial\Omega \cup \Gamma_{m,2}} = 0. \tag{5.22}$$

Again, applying Euler's trick from Remark 5.2 gives

$$-\operatorname{div}(\nu(\rho) \nabla \lambda) = 0 \qquad \forall \mathbf{x} \in \Omega \backslash \Omega_m. \tag{5.23}$$

We again plug in our findings into (5.19) and see that only the boundary integrals

$$\gamma \int_{\partial\Omega_m \backslash \Gamma_D} (\nabla u + \mathbf{b}) \cdot \mathbf{n} \, p \, \mathrm{d}s + \int_{\partial\Omega \backslash \Gamma_D} (\nu(\rho) \nabla \lambda) \cdot \mathbf{n} \, p \, \mathrm{d}s$$

remain. Next we now choose p to vanish on $\Gamma_{m,2} \cup (\partial\Omega \backslash \Gamma_{m,1})$, which after again applying Euler's trick yields the boundary condition

$$\gamma (\nabla u + \mathbf{b}) \cdot \mathbf{n} + \nu(\rho) \nabla \lambda \cdot \mathbf{n} = 0 \quad \forall \mathbf{x} \in \Gamma_{m,1} \backslash \Gamma_D. \tag{5.24}$$

Repeating the described procedure with functions $p \in V_0$ that vanish on $\partial\Omega$ and $\partial\Omega_m$ yield the interface condition

$$\gamma (\nabla u + \mathbf{b}) \cdot \mathbf{n} = 0 \quad \forall \mathbf{x} \in \Gamma_{m,2} \tag{5.25}$$

and the boundary condition

$$\nu(\rho) \nabla \lambda \cdot \mathbf{n} = 0 \quad \forall \mathbf{x} \in \partial\Omega \backslash (\Gamma_{m,1} \cup \Gamma_D), \tag{5.26}$$

respectively.

So far, we have not treated the interface $\Gamma_{d,2}$ between the design domain Ω_d and the rest of Ω. We will need a condition on that interface in Section 5.2.4 where we will solve

the DG reformulation of problem (5.12) by Newton's method. Note the obvious relation $\Omega = (\Omega \backslash \overline{\Omega}_d) \cup \Omega_d$. Let us start out from equation (5.12a) once again:

$$(5.12\text{a}) = \gamma \int_{\Omega_m} (\nabla u + \mathbf{b}) \cdot \nabla p + \int_{\Omega \backslash \Omega_d} \nu(\rho) \nabla p \cdot \nabla \lambda \, \mathrm{d}\mathbf{x} \tag{5.27}$$

$$+ \int_{\Omega_d} \nu(\rho) \nabla p \cdot \nabla \lambda \, \mathrm{d}\mathbf{x} \tag{5.28}$$

Now we perform integration by parts "in the other direction" in the second and third integral, which gives

$$(5.12\text{a}) = \ldots = -\gamma \int_{\Omega_m} \operatorname{div}(\nabla u + \mathbf{b})\, p \, \mathrm{d}x + \int_{\partial\Omega_m} (\nabla u + \mathbf{b}) \cdot \mathbf{n}\, p \, \mathrm{d}s \tag{5.29}$$

$$- \int_{\Omega \backslash \Omega_d} \operatorname{div}(\nu(\rho)\nabla p)\, \lambda \, \mathrm{d}\mathbf{x} + \int_{\partial(\Omega \backslash \Omega_d)} (\nu(\rho)\nabla p) \cdot \mathbf{n}\, \lambda \, \mathrm{d}s \tag{5.30}$$

$$- \int_{\Omega_d} \operatorname{div}(\nu(\rho)\nabla p)\, \lambda \, \mathrm{d}\mathbf{x} + \int_{\partial\Omega_d} (\nu(\rho)\nabla p) \cdot \mathbf{n}\, \lambda \, \mathrm{d}s. \tag{5.31}$$

Further noting $\partial(\Omega \backslash \Omega_d) \cup \partial\Omega_d = \partial\Omega \cup \Gamma_{d,2}$ and that λ vanishes on Γ_D due to $\lambda \in V_0$, the above expression reduces to

$$(5.12\text{a}) = \ldots = -\gamma \int_{\Omega_m} \operatorname{div}(\nabla u + \mathbf{b})\, p \, \mathrm{d}x + \int_{\partial\Omega_m} (\nabla u + \mathbf{b}) \cdot \mathbf{n}\, p \, \mathrm{d}s \tag{5.32}$$

$$- \int_{\Omega} \operatorname{div}(\nu(\rho)\nabla p)\, \lambda \, \mathrm{d}\mathbf{x} + \int_{\Gamma_N} (\nu(\rho)\nabla p) \cdot \mathbf{n}\, \lambda \, \mathrm{d}s \tag{5.33}$$

$$+ \int_{\Gamma_{d,2}} (\nu(\rho)\nabla p \cdot \mathbf{n}\, \lambda)_{\Omega_d} + (\underbrace{\nu(\rho)}_{=\nu_0} \nabla p \cdot \mathbf{n}\, \lambda)_{\neg\Omega_d} \, \mathrm{d}s \tag{5.34}$$

where the subscripts in the last integral correspond to the evaluation inside or outside the subdomain Ω_d. Now, we perform integration by parts over the whole of Ω twice:

$$-\int_{\Omega} \operatorname{div}(\nu(\rho)\nabla p)\, \lambda \, \mathrm{d}\mathbf{x} = \int_{\Omega} \nu(\rho)\nabla p \cdot \nabla \lambda \, \mathrm{d}\mathbf{x} - \int_{\partial\Omega} (\nu(\rho)\nabla p) \cdot \mathbf{n}\, \lambda \, \mathrm{d}s \tag{5.35}$$

$$= -\int_{\Omega} \operatorname{div}(\nu(\rho)\nabla \lambda)\, p \, \mathrm{d}x + \int_{\partial\Omega} \nu(\rho)\nabla\lambda \cdot \mathbf{n}\, p \, \mathrm{d}s - \int_{\partial\Omega} (\nu(\rho)\nabla p) \cdot \mathbf{n}\, \lambda \, \mathrm{d}s \tag{5.36}$$

Inserting this into (5.32) and again noting that λ vanishes on Γ_D gives

$$(5.12\text{a}) = \ldots = -\gamma \int_{\Omega_m} \operatorname{div}(\nabla u + \mathbf{b})\, p \, \mathrm{d}x + \int_{\partial\Omega_m} (\nabla u + \mathbf{b}) \cdot \mathbf{n}\, p \, \mathrm{d}s \tag{5.37}$$

$$- \int_{\Omega} \operatorname{div}(\nu(\rho)\nabla \lambda)\, p \, \mathrm{d}x + \int_{\partial\Omega} \nu(\rho)\nabla\lambda \cdot \mathbf{n}\, p \, \mathrm{d}s \tag{5.38}$$

$$- \int_{\partial\Omega} (\nu(\rho)\nabla p) \cdot \mathbf{n}\, \lambda \, \mathrm{d}s + \int_{\Gamma_N} (\nu(\rho)\nabla p) \cdot \mathbf{n}\, \lambda \, \mathrm{d}s \tag{5.39}$$

$$+ \int_{\Gamma_{d,2}} (\nu(\rho)\nabla p \cdot \mathbf{n}\, \lambda)_{\Omega_d} + (\nu_0 \nabla p \cdot \mathbf{n}\, \lambda)_{\neg\Omega_d} \, \mathrm{d}s, \tag{5.40}$$

which, by (5.21), (5.23), (5.24) and (5.26), reduces to the interface condition

$$\int_{\Gamma_{d,2}} (\nu(\rho)\nabla p \cdot \mathbf{n}\,\lambda)_{\Omega_d} + (\nu_0 \nabla p \cdot \mathbf{n}\,\lambda)_{\neg\Omega_d}\ \mathrm{d}s = 0 \qquad \forall p \in V_0, \tag{5.41}$$

which will become important in Section 5.2.4.

Let us now turn to equation (5.12b). Integration by parts in the first term gives

$$\begin{aligned} 0 &= \epsilon \int_{\Omega_d} \nabla\rho \cdot \nabla q\ \mathrm{d}\mathbf{x} + \frac{1}{\epsilon}\int_{\Omega_d} (2\rho - 6\rho^2 + 4\rho^3)\, q\ \mathrm{d}\mathbf{x} + (\nu_1 - \nu_0)\int_{\Omega_d} q\,\nabla u \cdot \nabla\lambda\ \mathrm{d}\mathbf{x} && (5.42)\\ &= -\epsilon \int_{\Omega_d} \Delta\rho\, q\ \mathrm{d}\mathbf{x} + \epsilon \int_{\partial\Omega_d} \nabla\rho \cdot \mathbf{n}\, q\ \mathrm{d}s \\ &\quad + \frac{1}{\epsilon}\int_{\Omega_d} (2\rho - 6\rho^2 + 4\rho^3)\, q\ \mathrm{d}\mathbf{x} + (\nu_1 - \nu_0)\int_{\Omega_d} q\,\nabla u \cdot \nabla\lambda\ \mathrm{d}\mathbf{x} \qquad \forall q \in H^1(\Omega_d) && (5.43)\end{aligned}$$

Proceeding as above, we choose a test function $q \in H^1(\Omega_d)$ that vanishes on the boundary $\partial\Omega_d$ and obtain

$$\int_{\Omega_d} \left(-\epsilon\Delta\rho + \frac{1}{\epsilon}(2\rho - 6\rho^2 + 4\rho^3) + (\nu_1 - \nu_0)\nabla u \cdot \nabla\lambda\right) q\ \mathrm{d}\mathbf{x} = 0 \quad \forall q \in H^1(\Omega_d) : q|_{\partial\Omega_d} = 0, \tag{5.44}$$

which by Euler's trick yields the partial differential equation

$$-\epsilon\Delta\rho + \frac{1}{\epsilon}(2\rho - 6\rho^2 + 4\rho^3) + (\nu_1 - \nu_0)\nabla u \cdot \nabla\lambda = 0 \qquad \forall \mathbf{x} \in \Omega_d. \tag{5.45}$$

If we plug in (5.45) into (5.43) we see that only

$$\epsilon \int_{\partial\Omega_d} \nabla\rho \cdot \mathbf{n}\, q\ \mathrm{d}s = 0 \qquad \forall q \in H^1(\Omega_d)$$

remains, which by Euler's trick yields the boundary condition for ρ

$$\epsilon\nabla\rho \cdot \mathbf{n} = 0 \qquad \forall \mathbf{x} \in \partial\Omega_d. \tag{5.46}$$

If we follow the same considerations for the third equation (5.12c) we obtain (as it was to be expected) the equations of linear 2D magnetostatics in differential form (2.29). Furthermore, the same considerations as in the derivation of the interface condition (5.41) yield:

$$\begin{aligned} 0 &= \int_\Omega \nu\nabla u \cdot \nabla v\ \mathrm{d}\mathbf{x} = \int_{\Omega\setminus\Omega_d} \nu\nabla u \cdot \nabla v\ \mathrm{d}\mathbf{x} + \int_{\Omega_d} \nu\nabla u \cdot \nabla v\ \mathrm{d}\mathbf{x} \\ &= -\int_{\Omega\setminus\Omega_d} \mathrm{div}\,(\nu\nabla u)\, v\ \mathrm{d}\mathbf{x} + \int_{\partial(\Omega\setminus\Omega_d)} \nu\nabla u \cdot \mathbf{n}\, v\ \mathrm{d}s - \int_{\Omega_d} \mathrm{div}\,(\nu\nabla u)\, v\ \mathrm{d}\mathbf{x} + \int_{\partial\Omega_d} \nu\nabla u \cdot \mathbf{n}\, v\ \mathrm{d}s \\ &= -\int_\Omega \mathrm{div}\,(\nu\nabla u)\, v\ \mathrm{d}\mathbf{x} + \int_{\partial\Omega} \nu\nabla u \cdot \mathbf{n}\, v\ \mathrm{d}s + \int_{\Gamma_{d,2}} (\nu\nabla u \cdot \mathbf{n}\, v)_{\Omega_d} + (\nu\nabla u \cdot \mathbf{n}\, v)_{\neg\Omega_d}\ \mathrm{d}s \\ &= \int_{\Gamma_{d,2}} (\nu\nabla u \cdot \mathbf{n}\, v)_{\Omega_d} + (\nu\nabla u \cdot \mathbf{n}\, v)_{\neg\Omega_d}\ \mathrm{d}s \qquad \forall v \in V_0 && (5.47)\end{aligned}$$

In the last line we used the PDE and the Neumann boundary condition from (2.29) as well as the fact that the test function vanishes on the Dirichlet boundary due to $v \in V_0$. Again, the subscripts correspond to the evaluation inside or outside the subdomain Ω_d. Applying Euler's trick gives

$$0 = (\nu \nabla u \cdot \mathbf{n})_{\Omega_d} + (\nu \nabla u \cdot \mathbf{n})_{\neg\Omega_d} =: [\nu \nabla u \cdot \mathbf{n}]_{\Gamma_{d,2}}, \tag{5.48}$$

which means that the co-normal derivative of u must be continuous across that interface.

Combining our findings, the KKT system in differential form reads

$$\chi_{\Omega_m}(-\gamma \Delta u - \gamma \mathrm{div}\, \mathbf{b}) - \mathrm{div}\,(\nu(\rho)\nabla\lambda) = 0 \quad \text{in } \Omega \tag{5.49a}$$

$$-\epsilon\Delta\rho + \frac{1}{\epsilon}(2\rho - 6\rho^2 + 4\rho^3) + (\nu_1 - \nu_0)\nabla u \cdot \nabla\lambda = 0 \quad \text{in } \Omega_d \tag{5.49b}$$

$$-\mathrm{div}\,(\nu(\rho)\nabla u) = f \quad \text{in } \Omega \tag{5.49c}$$

together with the boundary conditions for u, λ, ρ

$$u = 0 \quad \text{on } \Gamma_D \tag{5.49d}$$

$$\nu(\rho)\nabla u \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_N \tag{5.49e}$$

$$\lambda = 0 \quad \text{on } \Gamma_D \tag{5.49f}$$

$$\nu(\rho)\nabla\lambda \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_N \backslash \Gamma_{m,1} \tag{5.49g}$$

$$\gamma(\nabla u + \mathbf{b}) \cdot \mathbf{n} + \nu(\rho)\nabla\lambda \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_N \cap \Gamma_{m,1} \tag{5.49h}$$

$$\epsilon\nabla\rho \cdot \mathbf{n} = 0 \quad \text{on } \partial\Omega_d \tag{5.49i}$$

and the interface conditions

$$\gamma(\nabla u + \mathbf{b}) \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_{m,2} \tag{5.49j}$$

$$\int_{\Gamma_{d,2}} (\nu(\rho)\nabla p \cdot \mathbf{n}\,\lambda)_{\Omega_d} + (\nu_0 \nabla p \cdot \mathbf{n}\,\lambda)_{\neg\Omega_d}\, \mathrm{d}s = 0 \quad \forall p \in V_0, \tag{5.49k}$$

$$[\nu\nabla u \cdot \mathbf{n}]_{\Gamma_{d,2}} = 0 \tag{5.49l}$$

where χ_S denotes the characteristic function of a set S,

$$\chi_S(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in S \\ 0 & \text{otherwise.} \end{cases} \tag{5.50}$$

5.2.3 DG Discretization of KKT System

Now we are ready to derive the DG formulation of system (5.49). We start out from a given (not necessarily admissible) subdivision of Ω which we assume to be consistent with the interfaces $\Gamma_{m,2}$ and $\Gamma_{d,2}$, meaning that no element T is allowed to lie across an interface. A similar condition applies to a change of boundary conditions: Each edge of an element can lie only on either Γ_N or Γ_D, see Figure 5.4.

Let us specify some notation:

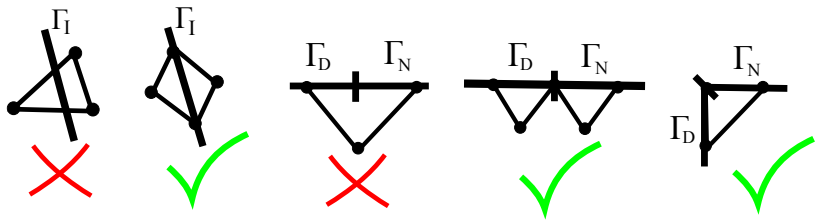

Figure 5.4: Examples of subdivisions that are non-consistent (×) and consistent (✓) with respect to interfaces and changes of boundary conditions

$\mathcal{T}_h$		triangulation of Ω
$\mathcal{T}_h^d$	$:= \{T \in \mathcal{T}_h : T \subset \Omega_d\}$	triangulation of Ω_d
$\mathcal{T}_h^m$	$:= \{T \in \mathcal{T}_h : T \subset \Omega_m\}$	triangulation of Ω_m
Γ_h	$:= \{e : e \text{ is edge of } T, T \in \mathcal{T}_h\}$	set of edges
Γ_h^D	$:= \{e \in \Gamma_h : e \subset \Gamma_D\}$	set of Dirichlet edges
Γ_h^N	$:= \{e \in \Gamma_h : e \subset \Gamma_N\}$	set of Neumann edges
$\Gamma_h^{m,0}$	$:= \{e \in \Gamma_h : \text{int}(e) \subset \text{int}(\Omega_m)\}$	set of edges in the interior of Ω_m
$\Gamma_h^{m,1}$	$:= \{e \in \Gamma_h : e \subset \Gamma_{m,1}\}$	set of edges on $\Gamma_{m,1} = \partial\Omega_m \cap \partial\Omega$
$\Gamma_h^{m,2}$	$:= \{e \in \Gamma_h : e \subset \Gamma_{m,2}\}$	set of edges on $\Gamma_{m,2} = \partial\Omega_m \backslash \partial\Omega$
$\Gamma_h^{d,0}$	$:= \{e \in \Gamma_h : \text{int}(e) \subset \text{int}(\Omega_d)\}$	set of edges in the interior of Ω_d
$\Gamma_h^{d,1}$	$:= \{e \in \Gamma_h : e \subset \Gamma_{d,1}\}$	set of edges on $\Gamma_{d,1} = \partial\Omega_d \cap \partial\Omega$
$\Gamma_h^{d,2}$	$:= \{e \in \Gamma_h : e \subset \Gamma_{d,2}\}$	set of edges on $\Gamma_{d,2} = \partial\Omega_d \backslash \partial\Omega$

We will derive the DG formulation of (5.49a) together with the corresponding boundary conditions (5.49d), (5.49f), (5.49g) and (5.49h) as in Section 3.2.1. Recall the definition of the broken Sobolev space (3.6) on p. 19 for $H^s(\mathcal{T}_h)$ and $H^s(\mathcal{T}_h^d)$. Since a DG method naturally involves normal derivatives on the edges of the mesh, we require all functions involved to be locally H^2 such that Theorem 2.12 is applicable and we can think of the normal derivatives as functions from $L^2(e)$ (or actually from $H^{1/2}(e) \subset L^2(e)$) for all edges e in Γ_h. Let us look for $u \in H^s(\mathcal{T}_h)$, $\rho \in H^s(\mathcal{T}_h^d) \cap L^\infty(\Omega_d)$ and $\lambda \in H^s(\mathcal{T}_h)$ with $s = 2$ for simplicity (in general: $s > 3/2$), and let us multiply (5.49a) with a test function $p \in H^2(\mathcal{T}_h)$ on each element T in the subdivision $\mathcal{T}_h$ and sum up over all elements. It is important to note that if a non-symmetric discretization technique as the NIPG or the IIPG is used to discretize system (5.49), the terms $-\text{div}\,(\nu(\rho)\nabla\lambda)$ in (5.49a) and $-\text{div}\,(\nu(\rho)\nabla u)$ in (5.49c) have to be treated in different ways. Even though they have the same form in the differential form, we must keep in mind that the Lagrange multiplier λ is actually the dual variable in (5.12a). Thus, when deriving a DG reformulation of (5.49a), we start out from

$$\int_T -\text{div}\,(\nu\nabla p)\lambda \, \mathrm{d}\mathbf{x} \tag{5.51}$$

rather than

$$\int_T -\text{div}\,(\nu\nabla\lambda)p \, \mathrm{d}\mathbf{x}. \tag{5.52}$$

The reason for proceeding like this will become clear in Section 5.2.4 when the Jacobian of the KKT system after discretization will turn out to be symmetric again. Now we will follow

the procedure described in Section 3.2.1:

$$\begin{aligned}&\sum_{T\in\mathcal{T}_h}\int_T (\chi_{\Omega_m}(-\gamma\Delta u-\gamma\text{div }\mathbf{b}))\,p-\text{div }(\nu(\rho)\nabla p)\,\lambda\,\mathrm{d}\mathbf{x} \\ &=-\sum_{T\in\mathcal{T}_h^m}\int_T(\gamma\Delta u+\gamma\text{div }\mathbf{b})p\,\mathrm{d}\mathbf{x}-\sum_{T\in\mathcal{T}_h}\int_T\text{div }(\nu(\rho)\nabla p)\,\lambda\,\mathrm{d}\mathbf{x}.\end{aligned} \tag{5.53}$$

Performing integration by parts on each element leads to

$$\begin{aligned}(5.53)=&-\sum_{T\in\mathcal{T}_h^m}\int_T\gamma(\nabla u+\mathbf{b})\cdot\nabla p\,\mathrm{d}\mathbf{x}-\int_{\partial T}\gamma(\nabla u+\mathbf{b})\cdot\mathbf{n}\,p\,\mathrm{d}s\\ &-\sum_{T\in\mathcal{T}_h}\int_T\nu(\rho)\nabla p\cdot\nabla\lambda\,\mathrm{d}\mathbf{x}-\int_{\partial T}\nu(\rho)\nabla p\cdot\mathbf{n}\,\lambda\,\mathrm{d}s.\end{aligned} \tag{5.54}$$

Following the procedure in Section 3.2.1, the first sum in (5.54) is replaced by

$$\begin{aligned}&\sum_{T\in\mathcal{T}_h^m}\int_T\gamma(\nabla u+\mathbf{b})\cdot\nabla p\,\mathrm{d}\mathbf{x}-\sum_{e\in\Gamma_h^{m,0}\cup(\Gamma_h^{m,1}\cap\Gamma_h^D)}\int_e\{\gamma(\nabla u+\mathbf{b})\cdot\mathbf{n}\}[\![p]\!]\,\mathrm{d}s\\ &-\sum_{e\in\Gamma_h^{m,1}\cap\Gamma_h^N}\int_e\gamma(\nabla u+\mathbf{b})\cdot\mathbf{n}\,p\,\mathrm{d}s-\sum_{e\in\Gamma_h^{m,2}}\int_e\underbrace{\gamma(\nabla u+\mathbf{b})\cdot\mathbf{n}}_{=0}p\,\mathrm{d}s\end{aligned} \tag{5.55}$$

with the average over an edge, $\{\cdot\}$, as defined in (3.10) and the jump operator, $[\![\cdot]\!]$, defined in (3.11). The integral over the edges on the interface $\Gamma_{m,2}$ vanishes due to the interface condition (5.49j). However, we need to be careful when treating the second sum in (5.54), in particular on the interface between the subdomain Ω_d and the rest of the domain, $\Omega\backslash\overline{\Omega}_d$, i.e. on $\Gamma_{d,2}$. As the density variable ρ is not defined outside this subdomain we cannot simply take a mean value of $\nu(\rho)$ across this interface. At a first glance this might seem to be not a big deal since we could just extend the definition of ρ to all elements adjacent to Ω_d and set the value of ρ to zero there. This is true in the case where we want to apply that type of method "only" to solve a partial differential equation like (5.49a). On the other hand, in Section 5.2.4 we want to solve the resulting DG reformulation of (5.49) by Newton's method which involves derivatives of the DG forms. When dealing with the derivative with respect to ρ of the resulting DG form we cannot simply extend the definition in a reasonable way. Therefore, we do not introduce averages on those mentioned interface edges, but stick to the edge integrals as they arise from the integration by parts. Of course, proceeding like this will not harm consistency. The described procedure looks as follows: We have a closer look on the second sum in (5.54),

$$-\sum_{T\in\mathcal{T}_h}\int_T\nu(\rho)\nabla p\cdot\nabla\lambda\,\mathrm{d}\mathbf{x}-\int_{\partial T}\nu(\rho)\nabla p\cdot\mathbf{n}\,\lambda\,\mathrm{d}s \tag{5.56}$$

$$\begin{aligned}=&\sum_{T\in\mathcal{T}_h}\int_T\nu(\rho)\nabla p\cdot\nabla\lambda\,\mathrm{d}\mathbf{x}-\sum_{e\in(\Gamma_h^0\backslash\Gamma_h^{d,2})\cup\Gamma_h^D}\int_e\{\nu(\rho)\nabla p\cdot\mathbf{n}\}[\![\lambda]\!]\,\mathrm{d}s\\ &-\sum_{e\in\Gamma_h^N}\int_e\{\nu(\rho)\nabla p\cdot\mathbf{n}\}[\![\lambda]\!]\,\mathrm{d}s-\sum_{e\in\Gamma_h^{d,2}}\int_e(\nu(\rho)\nabla p\cdot\mathbf{n}\,\lambda))_{\Omega_d}+(\nu(\rho)\nabla p\cdot\mathbf{n}\,\lambda))_{\neg\Omega_d}\,\mathrm{d}s\end{aligned} \tag{5.57}$$

where the subscripts $(\cdot)_{\Omega_d}$ and $(\cdot)_{\neg\Omega_d}$ correspond to the restriction to element adjacent to edge e that is inside or outside Ω_d, respectively. Note that these restrictions have to be interpreted in the sense of Theorem 2.12 and are well-defined because all functions are locally H^2. If we add up (5.55) and (5.57) we see that the integrals over the Neumann boundary edges vanish due to the boundary conditions (5.49g) and (5.49g). Adding the corresponding IPG terms as in (3.12) and (3.13) on p. 22 completes our DG reformulation of (5.49a):

$$\begin{aligned}
0 = &\sum_{T\in\mathcal{T}_h^m}\int_T \gamma(\nabla u + \mathbf{b})\cdot\nabla p \,\mathrm{d}\mathbf{x} - \sum_{e\in\Gamma_h^{m,0}\cup(\Gamma_h^{m,1}\cap\Gamma_h^D)}\int_e \gamma\{(\nabla u+\mathbf{b})\cdot\mathbf{n}\}[\![p]\!]\,\mathrm{d}s\\
&+\beta\sum_{e\in\Gamma_h^{m,0}\cup(\Gamma_h^{m,1}\cap\Gamma_h^D)}\int_e \gamma\{\nabla p\cdot\mathbf{n}\}[\![u]\!]\,\mathrm{d}s + \sum_{e\in\Gamma_h^{m,0}\cup(\Gamma_h^{m,1}\cap\Gamma_h^D)}\int_e \frac{\sigma_e}{|e|}[\![u]\!][\![p]\!]\,\mathrm{d}s\\
&+\sum_{T\in\mathcal{T}_h}\int_T \nu(\rho)\nabla p\cdot\nabla\lambda\,\mathrm{d}\mathbf{x} - \sum_{e\in(\Gamma_h^0\setminus\Gamma_h^{d,2})\cup\Gamma_h^D}\int_e \{\nu(\rho)\nabla p\cdot\mathbf{n}\}[\![\lambda]\!]\,\mathrm{d}s\\
&+\beta\sum_{e\in(\Gamma_h^0\setminus\Gamma_h^{d,2})\cup\Gamma_h^D}\int_e \{\nu(\rho)\nabla\lambda\cdot\mathbf{n}\}[\![p]\!]\,\mathrm{d}s + \sum_{e\in\Gamma_h^0\cup\Gamma_h^D}\int_e \frac{\sigma_e}{|e|}[\![p]\!][\![\lambda]\!]\,\mathrm{d}s\\
&-\sum_{e\in\Gamma_h^{d,2}}\int_e (\nu(\rho)\nabla p\cdot\mathbf{n}\,\lambda))_{\Omega_d} + (\nu(\rho)\nabla p\cdot\mathbf{n}\,\lambda))_{\neg\Omega_d}\,\mathrm{d}s
\end{aligned} \tag{5.58}$$

The derivations of the DG reformulations of (5.49b) and (5.49c) together with the remaining boundary conditions are analogous to Section 3.2.1.

Our modified IPG reformulation of the KKT system (5.49) (or (5.12)) formally reads as follows:

Find $(u,\rho,\lambda) \in\in H^2(\mathcal{T}_h)\times(H^2(\mathcal{T}_h^d)\cap L^\infty(\Omega_d))\times H^2(\mathcal{T}_h)$ such that

$$b_h(u,p) + a_h(\rho;p,\lambda) = \langle F_1, p\rangle \qquad \forall p\in H^2(\mathcal{T}_h) \tag{5.59a}$$

$$c_h(\rho,q) + a_h^{(2)}(q,u,\lambda) = 0 \qquad \forall q\in H^2(\mathcal{T}_h^d)\cap L^\infty(\Omega_d) \tag{5.59b}$$

$$a_h(\rho;u,v) = \langle F_3, v\rangle \qquad \forall v\in H^2(\mathcal{T}_h) \tag{5.59c}$$

with the bilinear forms defined by

$$\begin{aligned}
b_h(u,p) = &\sum_{T\in\mathcal{T}_h^m}\gamma\int_T \nabla u\cdot\nabla p\,\mathrm{d}\mathbf{x} - \sum_{e\in\Gamma_h^{m,0}\cup(\Gamma_h^{m,1}\cap\Gamma_h^D)}\gamma\int_e \{\nabla u\cdot\mathbf{n}\}[\![p]\!]\,\mathrm{d}s\\
&+\beta\sum_{e\in\Gamma_h^{m,0}\cup(\Gamma_h^{m,1}\cap\Gamma_h^D)}\gamma\int_e \{\nabla p\cdot\mathbf{n}\}[\![u]\!]\,\mathrm{d}s + \sum_{e\in\Gamma_h^{m,0}\cup(\Gamma_h^{m,1}\cap\Gamma_h^D)}\int_e \frac{\sigma_e}{|e|}[\![u]\!][\![p]\!]\,\mathrm{d}s,
\end{aligned} \tag{5.60}$$

$$
\begin{aligned}
a_h(\rho;u,v) = &\sum_{T\in\mathcal{T}_h}\int_T \nu(\rho)\nabla u\cdot\nabla v\ \mathrm{d}\mathbf{x} - \sum_{e\in(\Gamma_h^0\setminus\Gamma_h^{d,2})\cup\Gamma_h^D}\int_e \{\nu(\rho)\nabla u\cdot\mathbf{n}\}[\![v]\!]\ \mathrm{d}s \\
&+\beta\sum_{e\in(\Gamma_h^0\setminus\Gamma_h^{d,2})\cup\Gamma_h^D}\int_e \{\nu(\rho)\nabla v\cdot\mathbf{n}\}[\![u]\!]\ \mathrm{d}s + \sum_{e\in\Gamma_h^0\cup\Gamma_h^D}\int_e \frac{\sigma_e}{|e|}[\![u]\!][\![v]\!]\ \mathrm{d}s \\
&-\sum_{e\in\Gamma_h^{d,2}}\int_e (\nu(\rho)\nabla u\cdot\mathbf{n}\,v))_{\Omega_d} + (\nu(\rho)\nabla u\cdot\mathbf{n}\,v))_{\neg\Omega_d}\ \mathrm{d}s, \text{ and}
\end{aligned}
\tag{5.61}
$$

$$
\begin{aligned}
c_h(\rho,q) = &\sum_{T\in\mathcal{T}_h^d}\epsilon\int_T \nabla\rho\cdot\nabla q\ \mathrm{d}\mathbf{x} + \frac{1}{\epsilon}\int_T (2\rho-6\rho^2+4\rho^3)q\ \mathrm{d}\mathbf{x} - \sum_{e\in\Gamma_h^{d,0}}\int_e \epsilon\{\nabla\rho\cdot\mathbf{n}\}[\![q]\!]\ \mathrm{d}s \\
&+\beta\sum_{e\in\Gamma_h^{d,0}}\int_e \epsilon\{\nabla q\cdot\mathbf{n}\}[\![\rho]\!]\ \mathrm{d}s + \sum_{e\in\Gamma_h^{d,0}}\int_e \frac{\sigma_e}{|e|}[\![\rho]\!][\![q]\!]\ \mathrm{d}s,
\end{aligned}
\tag{5.62}
$$

the *tri*linear form

$$
a_h^{(2)}(\rho,u,\lambda) = \sum_{T\in\mathcal{T}_h^d}(\nu_1-\nu_0)\int_T \rho\nabla u\cdot\nabla\lambda\ \mathrm{d}\mathbf{x} \tag{5.63}
$$

and the linear forms

$$
\langle F_1,p\rangle = -\sum_{T\in\mathcal{T}_h^m}\gamma\int_T \mathbf{b}\cdot\nabla p\ \mathrm{d}\mathbf{x} + \sum_{e\in\Gamma_h^{m,0}\cup(\Gamma_h^{m,1}\cap\Gamma_h^D)}\gamma\int_e \{\mathbf{b}\cdot\mathbf{n}\}[\![p]\!]\ \mathrm{d}s \tag{5.64}
$$

$$
\langle F_3,v\rangle = \sum_{T\in\mathcal{T}_h}\int_T f\,v\ \mathrm{d}\mathbf{x}. \tag{5.65}
$$

Note that since the third term in (5.49b) does not involve a derivative of q, no integration by parts is performed in that term such that its DG reformulation (5.63) does not involve any edge integrals. However, like in the derivation of IPG methods in Section 3.2.1, we have the freedom to add any term that does not destroy consistency. We note that for the exact solution $(\overline{u},\overline{\rho},\overline{\lambda})^T$ of the KKT system (5.12), the components $\overline{u}\in V_0$ and $\overline{\lambda}\in V_0$ have continuous traces across element interfaces. Hence, the jump over any edge vanishes for all edges of the mesh $\mathcal{T}_h$,

$$
\begin{aligned}
[\![\overline{u}]\!]_e &= 0 \quad \text{a.e. on } e \quad \forall e\in\Gamma_h,\\
[\![\overline{\lambda}]\!]_e &= 0 \quad \text{a.e. on } e \quad \forall e\in\Gamma_h.
\end{aligned}
$$

Thus, adding the terms

$$
-\sum_{e\in(\Gamma_h^0\setminus\Gamma_h^{d,2})\cup\Gamma_h^D}(\nu_1-\nu_0)\int_e \{\rho\nabla u\cdot\mathbf{n}\}[\![\lambda]\!]\ \mathrm{d}s + \beta\sum_{e\in(\Gamma_h^0\setminus\Gamma_h^{d,2})\cup\Gamma_h^D}(\nu_1-\nu_0)\int_e \{\rho\nabla\lambda\cdot\mathbf{n}\}[\![u]\!]\ \mathrm{d}s \tag{5.66}
$$

to $a^{(2)}(\rho,u,\lambda)$ does not destroy consistency of the method. Furthermore, differentiating the interface condition (5.49k) with respect to ρ yields the condition

$$
(\nu_1-\nu_0)\int_{\Gamma_{d,2}}(\rho\nabla p\cdot\mathbf{n}\,\lambda)_{\Omega_d}\ \mathrm{d}s = 0 \quad \forall p\in V_0, \tag{5.67}
$$

which is also a necessary condition for the exact solution. Thus, choosing $p = \overline{u}$ gives

$$(\nu_1 - \nu_0) \int_{\Gamma_{d,2}} \left(\overline{\rho}\nabla\overline{u} \cdot \mathbf{n}\,\overline{\lambda}\right)_{\Omega_d} \,\mathrm{d}s = 0 \tag{5.68}$$

for the exact solution $(\overline{u}, \overline{\rho}, \overline{\lambda})^T$. Therefore, also this term is consistent and subtracting it from the trilinear form $a^{(2)}(\cdot,\cdot,\cdot)$ does not cause any troubles. We will replace (5.63) by

$$a_h^{(2)}(\rho, u, \lambda) = \sum_{T\in\mathcal{T}_h^d} (\nu_1 - \nu_0) \int_T \rho\nabla u \cdot \nabla\lambda \,\mathrm{d}\mathbf{x} - \sum_{e\in(\Gamma_h^0\setminus\Gamma_h^{d,2})\cup\Gamma_h^D} (\nu_1 - \nu_0) \int_e \{\rho\nabla u \cdot \mathbf{n}\}[\![\lambda]\!] \,\mathrm{d}s \tag{5.69}$$

$$+ \beta \sum_{e\in(\Gamma_h^0\setminus\Gamma_h^{d,2})\cup\Gamma_h^D} (\nu_1 - \nu_0) \int_e \{\rho\nabla\lambda \cdot \mathbf{n}\}[\![u]\!] \,\mathrm{d}s - \sum_{e\in\Gamma_h^{d,2}} (\nu_1 - \nu_0) \int_e \left(\rho\nabla u \cdot \mathbf{n}\,\lambda\right)_{\Omega_d} \,\mathrm{d}s, \tag{5.70}$$

which makes the Jacobian of the discretized KKT system symmetric, as we will in see Section 5.2.4.

For a discretization of (5.59) we use ansatz and test functions from the (non-conforming) finite element spaces as defined in (3.17),

$$V_{h,k} := \mathcal{D}_k(\mathcal{T}_h), \tag{5.71}$$

$$V_{h,k}^d := \mathcal{D}_k(\mathcal{T}_h^d), \tag{5.72}$$

where $k \in \mathbb{N}$ denotes the polynomial degree. Let N be the dimension of $V_{h,k}$ and $M(\leq N)$ the dimension of $V_{h,k}^d$. Let $\{\varphi_1, \ldots, \varphi_N\}$ be a basis of $V_{h,k}$, i.e.,

$$V_{h,k} = \operatorname{span}\{\varphi_1, \ldots, \varphi_N\} \tag{5.73}$$

and assume w.l.o.g. that the φ_i are ordered in such a way that the first M basis functions form a basis of $V_{h,k}^d$,

$$V_{h,k}^d = \operatorname{span}\{\varphi_1, \ldots, \varphi_M\}. \tag{5.74}$$

The discretization yields the system

Find $(u_h, \rho_h, \lambda_h) \in V_{h,k} \times \in V_{h,k}^d \times \in V_{h,k}$ such that

$$b(u_h, \varphi_i) + a(\rho_h; \varphi_i, \lambda_h) = \langle F_1, \varphi_i\rangle \qquad \forall i = 1, \ldots, N \tag{5.75a}$$

$$c(\rho_h, \varphi_i) + a^{(2)}(\varphi_i, u_h, \lambda_h) = 0 \qquad \forall i = 1, \ldots, M \tag{5.75b}$$

$$a(\rho_h; u_h, \varphi_i) = \langle F_3, \varphi_i\rangle \qquad \forall i = 1, \ldots, N \tag{5.75c}$$

Note that, due to the trilinear form $a^{(2)}(\cdot,\cdot,\cdot)$ and to the nonlinear term in $c_h(\cdot,\cdot)$, system (5.75) is **nonlinear** ! In the next subsection, we will apply Newton's method, presented in Section 2.2 to problem (5.75).

5.2.4 Solving KKT System using Newton's Method

Newton's method is applicable to a general system of nonlinear equations of the form (2.10) with continuously differentiable G with continuously invertible Jacobian G' and the method converges locally q-quadratically if G is twice continuously differentiable. In order to apply Newton's method to (5.75) we have to rewrite this system in the form (2.10) with a suitable mapping G. Let us first note that by plugging in the basis representations

$$u_h(\mathbf{x}) = \sum_{j=1}^{N} u_j \varphi_j(\mathbf{x}) \tag{5.76a}$$

$$\rho_h(\mathbf{x}) = \sum_{j=1}^{M} \rho_j \varphi_j(\mathbf{x}) \tag{5.76b}$$

$$\lambda_h(\mathbf{x}) = \sum_{j=1}^{N} \lambda_j \varphi_j(\mathbf{x}), \tag{5.76c}$$

with the basis functions from (5.73) and (5.74), problem (5.75) can be rewritten in the form

Find $(u_1, \dots, u_N, \rho_1, \dots, \rho_M, \lambda_1, \dots \lambda_N) \in \mathbb{R}^{N+M+N}$ such that

$$b(u_h, \varphi_i) + a(\rho_h; \varphi_i, \lambda_h) - \langle F_1, \varphi_i \rangle = 0 \qquad \forall i = 1, \dots, N \tag{5.77a}$$

$$c(\rho_h, \varphi_i) + a^{(2)}(\varphi_i, u_h, \lambda_h) = 0 \qquad \forall i = 1, \dots, M \tag{5.77b}$$

$$a(\rho_h; u_h, \varphi_i) - \langle F_3, \varphi_i \rangle = 0 \qquad \forall i = 1, \dots, N, \tag{5.77c}$$

which has the required form $G(z) = 0$ with $z := (u_1, \dots, u_N, \rho_1, \dots, \rho_M, \lambda_1, \dots \lambda_N)^T$ and $G : \mathbb{R}^{N+M+N} \to \mathbb{R}^{N+M+N}$ defined as

$$G(z) = \begin{pmatrix} \left[b(\sum_{j=1}^{N} u_j \varphi_j, \varphi_i) + a(\sum_{j=1}^{M} \rho_j \varphi_j; \varphi_i, \sum_{j=1}^{N} \lambda_j \varphi_j) - \langle F_1, \varphi_i \rangle \right]_{i=1,\dots,N} \\ \left[c(\sum_{j=1}^{M} \rho_j \varphi_j, \varphi_i) + a^{(2)}(\varphi_i, \sum_{j=1}^{N} u_j \varphi_j, \sum_{k=1}^{N} \lambda_k \varphi_k) \right]_{i=1,\dots,M} \\ \left[a(\sum_{j=1}^{M} \rho_j \varphi_j; \sum_{j=1}^{N} u_j \varphi_j, \varphi_i) - \langle F_3, \varphi_i \rangle \right]_{i=1,\dots,N} \end{pmatrix}. \tag{5.78}$$

For applying Newton's method we now have to compute the Jacobian of the function G in (5.78) which turns out to be

$$G'(z) = \left(\frac{\partial G_i}{\partial z_k} \right)_{i,k=1,\dots,N+M+N} = \begin{pmatrix} K_{uu} & K_{u\rho} & K_{u\lambda} \\ K_{\rho u} & K_{\rho\rho} & K_{\rho\lambda} \\ K_{\lambda u} & K_{\lambda\rho} & \mathbf{0} \end{pmatrix} \tag{5.79}$$

with the matrices

$$(K_{uu})_{i,k} = b_h(\varphi_k, \varphi_i) \qquad \text{for } i,k = 1,\dots,N, \tag{5.80}$$

$$\begin{aligned}(K_{u\rho})_{i,k} &= \sum_{T\in\mathcal{T}_h^d} (\nu_1-\nu_0)\int_T \varphi_k \nabla\varphi_i\cdot\nabla\lambda_h \,\mathrm{d}\mathbf{x} - \sum_{e\in(\Gamma_h^0\setminus\Gamma_h^{d,2})\cup\Gamma_h^D} (\nu_1-\nu_0)\int_e \{\varphi_k\nabla\varphi_i\cdot\mathbf{n}\}[\![\lambda_h]\!]\,\mathrm{d}s \\ &\quad + \beta \sum_{e\in(\Gamma_h^0\setminus\Gamma_h^{d,2})\cup\Gamma_h^D} (\nu_1-\nu_0)\int_e \{\varphi_k\nabla\lambda_h\cdot\mathbf{n}\}[\![\varphi_i]\!]\,\mathrm{d}s \\ &\quad - \sum_{e\in\Gamma_h^{d,2}} (\nu_1-\nu_0)\int_e (\varphi_k\nabla\varphi_i\cdot\mathbf{n}\,\lambda_h)_{\Omega_d}\,\mathrm{d}s \\ &= a^{(2)}(\phi_k,\phi_i,\lambda_h) \qquad \text{for } i,k=1,\dots,M,\end{aligned} \tag{5.81}$$

where the subscript in the last integral denotes the evaluation of the integral on the neighboring element of the interface edge e that is *inside* Ω_d,

$$(K_{u\lambda})_{i,k} = a_h(\rho_h;\varphi_i,\varphi_k) \qquad \text{for } i,k=1,\dots,N, \tag{5.82}$$

$$(K_{\rho u})_{i,k} = a_h^{(2)}(\varphi_i;\varphi_k,\lambda_h) \qquad \text{for } i,k=1,\dots,M, \tag{5.83}$$

$$\begin{aligned}(K_{\rho\rho})_{i,k} &= \sum_{T\in\mathcal{T}_h^d} \epsilon\int_T \nabla\varphi_k\cdot\nabla\varphi_i\,\mathrm{d}\mathbf{x} + \frac{1}{\epsilon}\int_T (2-12\varphi_k+12\varphi_k^2)\varphi_i\,\mathrm{d}\mathbf{x} \\ &\quad - \sum_{e\in\Gamma_h^{d,0}}\int_e \epsilon\{\nabla\varphi_k\cdot\mathbf{n}\}[\![\varphi_i]\!]\,\mathrm{d}s + \beta\sum_{e\in\Gamma_h^{d,0}}\int_e \epsilon\{\nabla\varphi_i\cdot\mathbf{n}\}[\![\varphi_k]\!]\,\mathrm{d}s \\ &\quad + \sum_{e\in\Gamma_h^{d,0}}\int_e \frac{\sigma_e}{|e|}[\![\varphi_k]\!][\![\varphi_i]\!]\,\mathrm{d}s \\ &= a^{(2)}(\phi_k,u_h,\phi_i) \qquad \text{for } i,k=1,\dots,M,\end{aligned} \tag{5.84}$$

$$(K_{\rho\lambda})_{i,k} = a_h^{(2)}(\varphi_i;u_h,\varphi_k) \qquad \text{for } i,k=1,\dots,M, \tag{5.85}$$

$$(K_{\lambda u})_{i,k} = a_h(\rho_h;\varphi_k,\varphi_i) \qquad \text{for } i,k=1,\dots,N, \tag{5.86}$$

$$\begin{aligned}(K_{\lambda\rho})_{i,k} &= \sum_{T\in\mathcal{T}_h^d} (\nu_1-\nu_0)\int_T \varphi_k\nabla u_h\cdot\nabla\varphi_i\,\mathrm{d}\mathbf{x} - \sum_{e\in(\Gamma_h^0\setminus\Gamma_h^{d,2})\cup\Gamma_h^D} (\nu_1-\nu_0)\int_e \{\varphi_k\nabla u_h\cdot\mathbf{n}\}[\![\varphi_i]\!]\,\mathrm{d}s \\ &\quad + \beta \sum_{e\in(\Gamma_h^0\setminus\Gamma_h^{d,2})\cup\Gamma_h^D} (\nu_1-\nu_0)\int_e \{\varphi_k\nabla\varphi_i\cdot\mathbf{n}\}[\![u_h]\!]\,\mathrm{d}s \\ &\quad - \sum_{e\in\Gamma_h^{d,2}} (\nu_1-\nu_0)\int_e (\varphi_k\nabla u_h\cdot\mathbf{n}\,\varphi_i)_{\Omega_d}\,\mathrm{d}s \qquad \text{for } i,k=1,\dots,M.\end{aligned} \tag{5.87}$$

5.2.5 Summary

Let us summarize the proceedings of this section in an abstract algorithm: The phase-field method stipulates that problem (5.7) is solved for a decreasing sequence $\{\epsilon^{(l)}\}$ whereas the weighting factor γ is fixed. We choose to decrease ϵ as $\epsilon^{(l+1)} = \delta\epsilon^{(l)}$ with a proper factor $0<\delta<1$, e.g. $\delta = 1/2$. For each ϵ the KKT system is discretized using a DG method and

the arising system of nonlinear equations is solved by the (damped) Newton method. The initial value for the very first iteration where $\epsilon = \epsilon^{(0)}$ has to be provided. In the subsequent iterations (i.e. for $\epsilon = \epsilon^{(l)}$, $l = 1, 2, \dots$) the optimal design of the previous step is used as an initial guess for Newton's method. The algorithm terminates after a certain number of iterations, when $\epsilon \leq \epsilon^{min}$. The resulting final design is expected to consist of almost only black and white areas, meaning areas where the density ρ is 1 or 0, respectively.

Algorithm 2. Phase-Field Method for Benchmark Problem (5.4)

Choose γ, ϵ^0, $z^0 = (u_1^0, \dots, u_N^0, \rho_1^0, \dots, \rho_M^0, \lambda_1^0, \dots, \lambda_N^0)^T$, δ, ϵ^{min}
While $\epsilon > \epsilon^{min}$
    Solve KKT system (5.49) using DG discretization by Newton's method:
    For $k = 0$ until convergence do
        set up system matrix $G'(z^k)$ in (5.79)
        solve $G'(z^k)w^k = -G(z^k)$
        choose $\tau^k \in \{1, 1/2, 1/4, 1/8, \dots\}$ maximal such that $\|z^k + \tau^k w^k\| < \|z^k\|$
        set $z^{k+1} = z^k + \tau^k w^k$
    end
set $\epsilon = \delta\, \epsilon$

Chapter 6

Numerical Experiments

In this chapter we will shortly present the observations and numerical results we obtained from the numerical realization of Algorithm 2. Recall that this algorithm summarizes the procedure of solving our benchmark problem (5.4), which corresponds to the topology optimization of an electromagnet, cf. Section 5.1. The goal is to find a geometry represented by the density variable ρ such that the resulting magnetic field minimizes the functional

$$\int_{\Omega_m} \left| \begin{pmatrix} \partial_2 u \\ -\partial_1 u \end{pmatrix} - B_m^{avg} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right|^2 \, \mathrm{d}\mathbf{x} \tag{6.1}$$

where B_m^{avg} is a prescribed average value, which we choose as $B_m^{avg} = \frac{1}{10}$. We chose a DG discretization with globally discontinuous, piecewise linear ansatz functions u_h, ρ_h, λ_h. The numerical experiments were performed on a PC using a 1.8 GHz Intel CPU and 2 GB memory.

Recall that Algorithm 2 states that, for each value $\epsilon = \epsilon^{(j)}$ of a decreasing sequence $\{\epsilon^{(j)}\}$, the system of nonlinear equations (5.77) has to be solved using Newton's method, where in iteration k the Newton correction equation

$$\begin{pmatrix} K_{uu} & K_{u\rho} & K_{u\lambda} \\ K_{\rho u} & K_{\rho\rho} & K_{\rho\lambda} \\ K_{\lambda u} & K_{\lambda\rho} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \underline{w}_u^k \\ \underline{w}_\rho^k \\ \underline{w}_\lambda^k \end{pmatrix} = -G(\underline{z}^{(k)}) \tag{6.2}$$

with $\underline{z}^{(k)} = (\underline{u}^k, \underline{\rho}^k, \underline{\lambda}^k)^T = (u_1^k, \ldots, u_N^k, \rho_1^k, \ldots, \rho_M^k, \lambda_1^k, \ldots \lambda_N^k)^T$ and the block matrices defined in (5.80) - (5.87) has to be solved. Note that also these matrices depend on the current iterates $u_h = u_h^{(k)}$, $\rho_h = \rho_h^{(k)}$ and $\lambda_h = \lambda_h^{(k)}$ and that the vectors $\underline{u}^k$, $\underline{\rho}^k$, $\underline{\lambda}^k$ are related to the functions $u_h^{(k)}$, $\rho_h^{(k)}$ and $\lambda_h^{(k)}$ via (5.76). In our implementation, this large-scale, sparse system of linear equations was solved using the software package PARDISO (*Parallel Sparse Direct Solver*, see Schenk and Gärtner[12]). The computational domain $\Omega = [0,1]^2$ was divided into 638 elements using the structured mesh depicted in Figure 6.1. Here, the magnetization area $\Omega_m = [0, 1/16]^2$ was subdivided into 128 elements.

Throughout all testing it could be observed that the convergence behaviour of the presented method is heavily dependent on the choice of the parameters γ, $\{\epsilon^{(j)}\}$ and the starting value $\underline{z}^0 = (\underline{u}^0, \underline{\rho}^0, \underline{\lambda}^0)^T$. For us, choosing the initial value $\underline{z}^0$ reduced to choosing an initial design $\underline{\rho}^0$ and computing $\underline{u}^0 = \underline{u}^0(\underline{\rho}^0)$ from (5.77c) and furtheron $\underline{\lambda}^0 = \underline{\lambda}^0(\underline{u}^0, \underline{\rho}^0)$ via (5.77a).

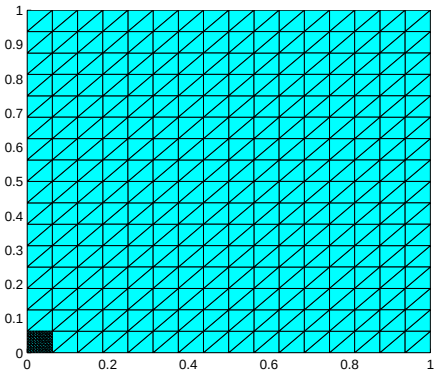

Figure 6.1: Subdivision of the computational domain $\Omega = [0,1]^2$ used for numerical experiments

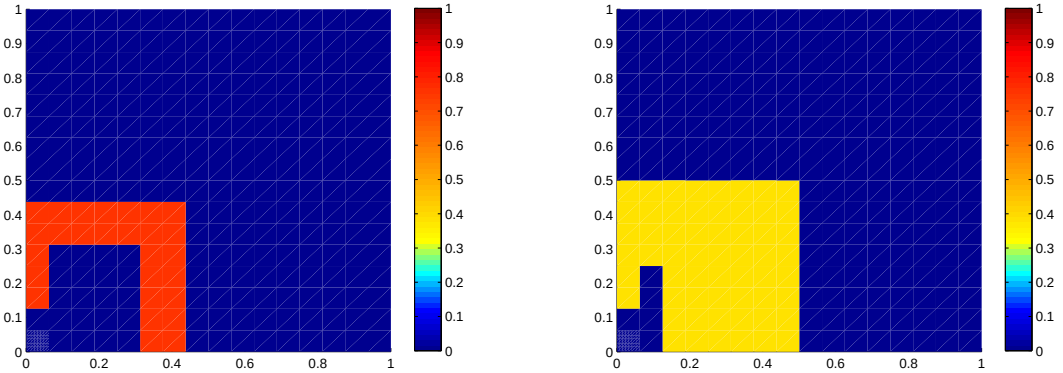

Figure 6.2: Left picture: initial design (geometry of Maltese cross electromagnet). Right picture: design after 17 iterations of the damped Newton method.

As a first initial design we chose the geometry of the Maltese Cross electromagnet, see Figure 5.1 on p. 42. However, for many choices of γ and $\epsilon^{(0)}$, fast convergence towards a constant density value on the whole of Ω_d could be observed. See Figure 6.2 for the initial density ρ(left) and the density after 17 iterations (right). The parameters chosen in this test were $\epsilon = 3$ and $\gamma = 0.2/(|\Omega_m||B_m^{avg}|) = 512$. It is common to scale the objective functional (6.1) by the factor $1/|\Omega_m||B_m^{avg}$. We incorporated this factor in the parameter γ. Due to these observations, we restricted ourselves to constant initial values ρ^0.

We pointed out in Chapter 4 that the penalization of intermediate density values can cause a convex problem to become non-convex. This non-convexity could be observed in numerous numerical tests as the convergence behaviour turned out to be very sensitive with respect to the choice of the parameters and initial values. If the initial value is too far from the exact solution, the (damped) Newton method converges very slowly. For the choice $\gamma = 0.02 \cdot 2560 = 51.2$ and the initial value $\rho^0 \equiv 0.75 = const$ on the entire design domain Ω_d we did not obtain convergence for $\epsilon^0 = 0.05$ and $\epsilon^0 = 0.005$ and aborted each of the two tests after around 370 iterations. The damping parameter τ of the damped Newton method (cf. (2.14)) typically ranged in the order of 10^{-4} and hence the method did not entail any major

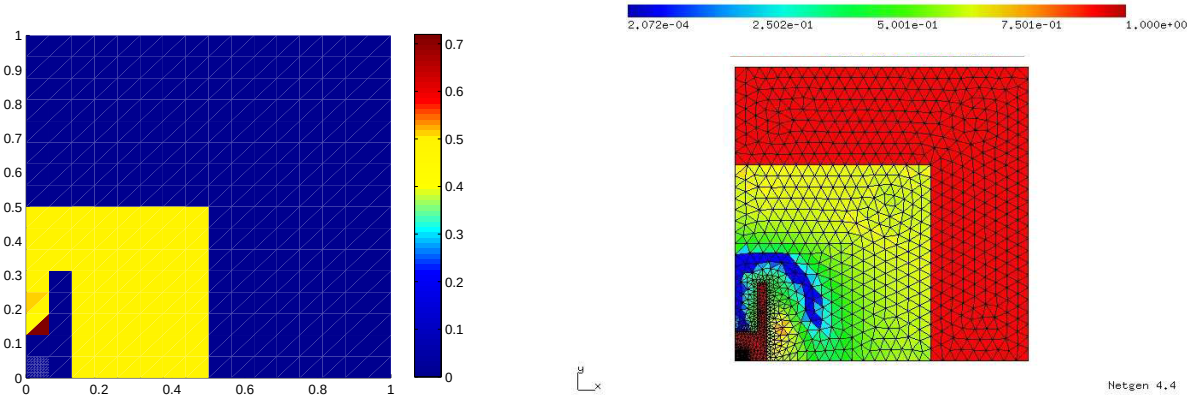

Figure 6.3: Left picture: Density values after 160 iterations of damped Newton method (converged to local maximum $\rho \equiv 0.5$). Right picture: Design development after 800 iterations of the phase-field method by Lukáš [21].

changes of the initial geometry. For the choice $\epsilon^0 = 0.0025$ we obtained quadratic convergence after a total of 160 iterations of the damped Newton method. However, the density was driven close to $\rho \equiv 0.5$ which is a local maximum of the function $W(\rho) = \rho^2(1-\rho)^2$ that we used for penalizing intermediate values of ρ. The result we obtained is depicted in Figure 6.3. This poor convergence behaviour was already observed by D. Lukáš in Lukáš [21] where the phase-field approach was applied to that same model problem and did not yield satisfactory results even after 800 iterations. The intermediate results by Lukáš are depicted in the right picture of Figure 6.3.

However, for the choice $\gamma = 2 \cdot 2560 = 5120$, $\epsilon = 3$ and $\rho_h^0 \equiv 0.75$ we were able to observe first signs of a tendency towards the optimal result when we compared intermediate results with the initial design. The left picture in Figure 6.4 shows the difference $\rho_h^{(26)} - \rho_h^{(0)}$. Even though the density on the whole design domain Ω_d is decreasing, one can see a certain tendency towards the optimal solution depicted in the right picture of Figure 6.4. This optimal solution is taken from Lukáš [21] and was obtained by applying the penalization method presented in (4.4) on p. 33.

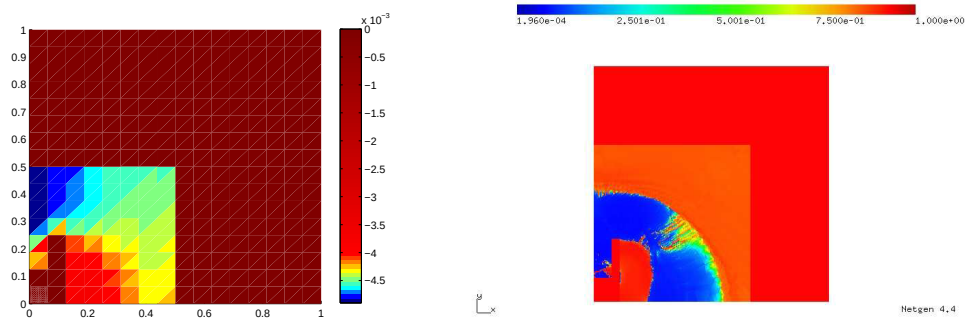

Figure 6.4: Left picture: Difference between density values of intermediate result after 26 iterations of the damped Newton method and the initial design. Right picture: Optimal solution computed by D. Lukáš in [21].

Chapter 7

Conclusion and Outlook

In this thesis we gave an overview on the method of topology optimization and discussed various possible regularization approaches. We investigated in more detail the phase-field method and applied it to a real world problem from electromagnetics in combination with a discontinuous Galerkin discretization.

In Chapter 3 we introduced a class of DG methods, the interior penalty Galerkin methods, derived the variational formulation, discussed the existence and uniquness issue, provided error estimates and verified them in numerical tests.

In Chapter 4 we presented different aspects of topology optimization in an abstract framework. We discussed different penalization schemes and motivated the use of regularization methods by pointing out possible numerical instabilities. After a brief discussion of possible remedies, we focused on one regularization method, the phase-field method.

In Chapter 5 we gave a detailed description of a benchmark problem from electromagnetics, which we regularized by the phase-field method. We derived the first-order necessary optimality conditions for the resulting minimization problem and discretized them by a discontinuous Galerkin method. The discretized system turned out to be nonlinear and we applied Newton's method to it.

In Chapter 6 numerical experiments were presented. The phase-field method turned out to be very sensitive with respect to the involved parameters and the initial design. For inappropriate choices, the convergence was either very slow so that no major improvement of the design was achieved, or the method converged to a local maximum.

This work can be continued in the following directions:

- *Different Regularization Methods:*
  As indicated in Chapter 4, the regularization term in the phase-field functional (4.18) could just as well be replaced by any other regularization term. Likewise, the penalization term could be exchanged.

- *Efficient Solver:*
  The presented method turned out to be dependent on the choice of the involved regularization parameters. A next step would be to solve the linear system in each Newton step by an iterative method such as the minimal residual (MINRES) method. For that purpose, it would be desirable to have a preconditioner that is robust with respect to all involved parameters.

- *Include Inequality Constraints:*
  In this thesis, for simplicity we did neither account for the box constraints on the density variable nor for the volume constraint in (4.2). For a more realistic setting, these constraints should be included in the optimization problem.

- *Nonlinear Materials:*
  For simplicity we assumed linear behaviour of the material where the magnetic reluctivity does not depend on the magnetic field itself and we could use model (2.29). The ferromagnetic material that is distributed in the design domain behaves nonlinearly and the model (2.28) would be more realistic.

- *Domain Decomposition DG-FEM:*
  A discretization with discontinuous ansatz functions is reasonable on and around the interface between material and void in the final design, as a jump of the density function of any desired height can be forced just by a simple modification of the DG bilinear form. However, in the parts of the computational domain where no jumps appear, a discretization by a classical FE method would be more efficient as it would mean a notable decrease in the number of global degrees of freedom. Hence, it might be a good idea to make a domain decomposition and to perform a DG discretization on the (unknown) interface and a classical FE discretization on the rest of the computational domain Ω. A further issue is the question of how to determine the location of the interface between material and void.

Bibliography

[1] R. A. Adams and J. J. F. Fournier. *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics*. Academic Press, Amsterdam, Boston, second edition, 2003.

[2] M. P. Bendsøe. *Optimization of Structural Topology, Shape and Material*. Springer, 1995.

[3] M. P. Bendsøe and O. Sigmund. *Topology Optimization: Theory, Methods and Applications*. Springer, Berlin, 2003.

[4] J. Bergh and J. Löfström. *Interpolation Spaces: An Introduction*. Springer, Berlin, Heidelberg, New York, 1976.

[5] M. Burger and R. Stainko. Phase-field relaxation of topology optimization with local stress constraints. *SIAM Journal on Control and Optimization*, 45(4):1447–1466, 2006.

[6] A. Chambolle, V. Caselles, M. Novaga, D. Cremers, and T. Pock. An introduction to total variation for image analysis. *Radon Series on Computational and Applied Mathematics*, 9:1–78, 2010. `http://www.math.unipd.it/~novaga/papers/chambolle/notes/Fornasier.pdf`.

[7] P. W. Christensen and A. Klarbring. *An Introduction to Structural Optimization*. Springer, 2009.

[8] P. Deuflhard. *Newton Methods for Nonlinear Problems*. Springer, 2004.

[9] Y. Epshteyn and B. Rivière. Estimation of penalty parameters for symmetric interior penalty Galerkin methods. *Journal of Computational and Applied Mathematics*, 206:843–872, 2007.

[10] S. V. Fomin and I. M. Gelfand. *Calculus of Variations*. Dover, 1963.

[11] P. Gangl. Exact and inexact semismooth Newton methods for elliptic optimal control problems. Bachelor Thesis, Johannes Kepler University Linz, Institute of Computational Mathematics, 2010. `http://www.numa.uni-linz.ac.at/Teaching/Bachelor/gangl-bakk.pdf`.

[12] K. Gärtner and O. Schenk. On fast factorization pivoting methods for symmetric indefinite systems. Elec. Trans. Numer. Anal, 23:158-179, 2006. `http://www.pardiso-project.org`.

[13] J.S. Hesthaven and T. Warburton. On the constants in hp-finite element trace inverse inequalities. *Computer methods in applied mechanics and engineering*, 192:2765–2773, 2003.

[14] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM Journal on Optimization*, 13(3):865–888, 2003.

[15] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints*. Springer, 2009.

[16] M. Jung and U. Langer. *Methode der finiten Elemente für Ingenieure: Eine Einführung in die numerischen Grundlagen und Computersimulation*. B.G. Teubner GmbH, 2010.

[17] M. Kaltenbacher. *Numerical Simulation of Mechatronic Sensors and Actuators, second edition*. Springer, Berlin-Heidelberg, 2007.

[18] U. Langer. Numerik 1: Operatorgleichungen. Lecture Notes, 1995/96. `http://www.numa.uni-linz.ac.at/Teaching/Notes/`.

[19] U. Langer. Lecture on Computational Electromagnetics. 2010.

[20] D. Lukáš. *Optimal Shape Design in Magnetostatics*. PhD thesis, Technical University of Ostrava, 2003.

[21] D. Lukáš. Progress report on topology and shape optimization in magnetostatics. Technical Report 2004-40, Special Research Programme SFB F013, Johannes Kepler University Linz, 2004. `http://www.sfb013.uni-linz.ac.at/reports/2004/pdf-files/rep_04-40_lukas.pdf`.

[22] P. Monk. *Finite Element Methods for Maxwell's Equations*. Oxford University Press, 2003.

[23] J. Nocedal and S. Wright. *Numerical Optimization*. Springer, New York, 1999.

[24] C. Pechstein. Multigrid-Newton-methods for nonlinear magnetostatic problems. Master's thesis, Johannes Kepler University Linz, Institute of Computational Mathematics, 2004. `http://www.numa.uni-linz.ac.at/Teaching/Diplom/Finished/pechstein_dipl.pdf`.

[25] C. Pechstein and Institute of Computational Mathematics at Johannes Kepler University Linz. Data-sparse boundary and finite element domain decomposition methods in electromagnetics. FWF-Project P19255. `http://www.numa.uni-linz.ac.at/P19255/`.

[26] J. Petersson and O. Sigmund. Numerical instabilities in topology optimization: A survey on procedures dealing with checkerboards, mesh-dependencies and local minima. *Structural Optimization*, 16:68–75, 1998.

[27] B. Rivière. *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations: Theory and Implementation*. SIAM, 2008.

[28] J. Schöberl. Numerical methods for Maxwell equations. Lecture Notes, 2005. `http://www.asc.tuwien.ac.at/~schoeberl/wiki/lva/notes/maxwell.pdf`.

[29] J. Schöberl. Numerical methods for partial differential equations. Lecture Notes, 2009. `http://www.asc.tuwien.ac.at/~schoeberl/wiki/lva/notes/numpde.pdf`.

[30] R. Stainko. *Advanced Multilevel Techniques to Topology Optimization.* PhD thesis, Johannes Kepler University of Linz, Institute of Computational Mathematics, 2006. `http://www.numa.uni-linz.ac.at/Teaching/PhD/Finished/stainko-diss.pdf`.

[31] F. Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen: Theorie, Verfahren und Anwendungen.* Vieweg, Wiesbaden, 2005.

[32] S. Zaglmayr. *High Order Finite Element Methods for Electromagnetic Field Computation.* PhD thesis, Johannes Kepler University of Linz, Institute of Computational Mathematics, 2006. `http://www.numa.uni-linz.ac.at/Teaching/PhD/Finished/zaglmayr-diss.pdf`.

[33] W. Zulehner. *Numerische Mathematik: Eine Einführung anhand von Differentialgleichungsproblemen, Band1: Stationäre Probleme.* Birkhäuser, 2008.

Eidesstattliche Erklärung

Ich, Peter Gangl, erkläre an Eides statt, dass ich die vorliegende Masterarbeit selbständig 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.

Linz, Februar 2012

_____

Peter Gangl

Curriculum Vitae

Name: Peter Gangl

Nationality: Austria

Date of Birth: 22 February 1988

Place of Birth: Schärding, Austria

Education:

1998 - 2006	Bundesgymnasium (secondary comprehensive school) Schärding
2007 - 2010	Bachelor Studies in Technical Mathematics, Johannes Kepler University Linz
2010 - 2012	Master Studies in Industrial Mathematics, Johannes Kepler University Linz

Special Activities:

September 2010	24th ECMI Modeling Week, Milano, Italy
January 2011 - June 2011	Exchange Semester, Lund, Sweden

Work Experience:

August 2009 - January 2011	Siemens Transformers Austria Linz