JOHANNES KEPLER
UNIVERSITÄT LINZ
Netzwerk für Forschung, Lehre und Praxis

Advanced Multilevel Techniques to Topology Optimization

Dissertation

zur Erlangung des akademischen Grades

Doktor der Technischen Wissenschaften

Angefertigt am *SpezialForschungsBereich F013 - Numerical and Symbolic Scientific Computing*

Begutachter:

o.Univ.-Prof. Dr. Ulrich Langer
Prof. Dr. Martin P. Bendsøe, Technical University of Denmark

Eingereicht von:

Dipl.-Ing. Roman Stainko

Linz, February 2006

Abstract

Computer-aided design is nowadays the basis of successful product planning and production control. In almost all sectors of industry faster and faster changing and adapting product specifications demand shorter and shorter production times for competative products. In order to cope with this situation and to realize short development times, enterprises rely on computer simulations. During the last two decades a new field of applied mathematics has reached a state of maturity to enter the world of computer-aided engineering, namely the topology optimization techniques.

This work deals with mathematical methods for topology optimization problems. In particular, we focus on two specific design - constraint combinations, namely the maximization of material stiffness at given mass and the minimization of mass while keeping a certain stiffness. Both problems show different properties and will also be treated with different approaches in this work. One characteristics of topology optimization problems is that the set of feasible designs is constrained by a partial differential equation. Moreover, these optimization problems are not well-posed, so regularization techniques have to be applied.

The first combination, also known as the minimal compliance problem, is treated in the framework of an adaptive multi-level approach. Well-posedness of the problem is achieved by applying filter methods to the problem. Such a filter method is used for adaptive mesh refinement along the interface between void and material, i.e. along the boundary of the structure. The resulting optimization problems on each level are solved by the method of moving asymptotes, a well-known optimization technique in the field of topology optimization. In order to ensure an efficient solution of the linear systems, raising from the finite element discretization of the partial differential equations, a multigrid method is applied.

The treatment of the second combination is by far less understood as the minimal compliance problem. The main source of difficulties is a lack of constraint qualifications for the set of feasible designs, defined by local stress constraints. To overcome these difficulties the set of constraints is reformulated, involving only linear and 0–1 constraints. These are finally relaxed by a Phase–Field approach, which also regularizes the problem. This relaxation scheme results in large-scale optimization problems, which finally solved by an interior-point optimization method.

Most of the computing time of these optimization routines is actually spent in solving linear saddle point problems. In order to speed up computations an efficient solver with optimal complexity for these system is of high importance. Multigrid methods certainly belong to the most efficient methods for solving large-scale systems, e.g., arising from discretized partial differential equations. One of the most important ingredients of an efficient multigrid method is a proper smoother. In this work a multiplicative Schwarz-type smoother is considered, that consists of the solution of several small local saddle point problems, and that leads to an KKT-solver with linear complexity.

Zusammenfassung

In der heutigen Zeit bildet die computerunterstützte Simulation das Fundament für erfolgreiche Produktplanungen und effiziente Produktionsvorgänge. Dies trifft auf verschiedenste Bereiche der Industrie und Wirtschaft zu. Die sich immer schneller ändernde Nachfrage und härtere Marktsituation verlangt kurze Entwicklungszeiten und konkurenzfähige Produkte. Um diese kostengünstig und in kurzer Zeit herstellen zu können, wird nicht nur mit rechnergestützten Entwicklungszyklen gearbeitet, sondern auch die Produktoptimierung schon in den frühen Plangungsphasen eingesetzt. Viele Entscheidungen in diesen Vorgängen beruhen auf langjährigen Erfahrungswerten, die durch computerunterstützte Simulationen bestätigt und erweitert werden. In den letzten Jahren wuchs das akademische und wirtschaftliche Interesse für das Gebiet der Topologieoptimierung enorm. Hier werden durch mathematische Methoden gute Ausgangsmodelle für den Designprozess in unterschiedlichen Produktionsvorgängen geschaffen. So wird zum Beispiel mittels Materialeinsparung das Gewicht eines Bauteiles oder einer Maschine reduziert, ohne dass sich dadurch Funktionsfähigkeit und Leistung verringern, hingegen eventuell sogar verstärken. Da diese Optimierungsvorgänge auf dem Computer simuliert werden, anstatt durch aufwendige Versuche mit Prototypen, helfen sie teure Entwicklungszeiten und Entwicklungskosten zu sparen.

Diese Arbeit beschäftigt sich mit mathematischen Methoden der Topologieoptimierung. Hier wird die Frage nach der Optimalität eines Ausgangsentwurfes durch ein mathematisches Optimierungsproblem modelliert. Diese Optimierungsprobleme bestehen aus einer Kostenfunktion, die jedem Design einen gewissen Wert zuordnet, und einem Zulaessigkeitsbereich, der die Menge der akzeptierbaren Entwürfe beschreibt. Dieser Zulässigkeitsbereich wird unter anderem auch durch partielle Differentialgleichungen beschrieben. Daher ist für das effiziente Lösen solch komplexer Optimierungsprobleme eine erfolgreiche Kombination mehrerer mathematischer Bereiche notwendig. Diese umfassen neben der numerischen Optimierung auch die Analyse und Numerik partieller Differentialgleichungen, effiziente Lösungverfahren für Gleichungsysteme und das Modellieren physikalischer Vorgänge, zum Beispiel der Festkörpermechanik.

Nach einer kurzen Einführung in die oben genannten mathematischen Disziplinen werden zwei typische Topologieoptimierungsprobleme betrachtet. Beim ersten Modellbeispiel wird ein möglichst steifes Bauteil mit beschränktem Volumen bezüglich angreifenden Kräften gesucht. Das zweite Problem befasst sich mit der Fragestellung einer möglichst leichten Konstruktion eines Bauteiles, sodass keine Materialschäden, wie Risse oder Brüche, unter Belastung auftreten. Beide Probleme unterscheiden sich nicht nur in ihrer Fragestellung, sondern auch in ihrer effizienten mathematischen Behandlung und dem Fortschritt ihrer theoretischen Untersuchungen. Es existieren zwei grundsätzliche Ansätze, um Optimierungsprobleme mit partiellen Differentialgleichungen als Nebenbedingungen zu behandeln. Dazu wird die Menge der zu bestimmenden Parameter in zwei Gruppen unterteilt. Die Zustandsparameter repräsentieren

den Zustand der Differentialgleichung und die Designparameter beschreiben das Design. Neben dem Ansatz beide Gruppen von Unbekannten gleichzeitig im Optimierungsproblem zu behandeln, gibt es auch die Methode, die Zustandsparameter aus dem Optimierungsproblem zu eliminieren. Beide Verfahren werden vorgestellt und ihre Eigenschaften diskutiert.

Für das Problem der maximalen Steifigkeit wird ein adaptives Multilevel-Verfahren vorgestellt, das die Elimination der Zustandsparameter benützt. Das Problem wird auf einer feiner werdenden Hirarchie von Diskretisierungen gelöst, wobei die zugrundeliegenden Netze entlang des Randes der Struktur verfeinert werden. Um eine effiziente Behandlung der Probleme auf den jeweiligen Netzen zu ermöglichen, wird die 'Method of Moving Asymptotes' zum Lösen der Optimierungsproblem benutzt. Weiters wird ein Mehrgitterverfahren zur Lösung der linearen Gleichungssysteme, resultierend aus der Finiten Elemente Diskretisierung der partiellen Differentialgleichung, herangezogen.

Das Problem der minimalen Masse wird anders behandelt. Hier liefert die Beibehaltung beider Variablengruppen die Möglichkeit das Optimierungsproblem umzuschreiben. Durch diese Umformulierung können ernsthafte Probleme umgangen werden, die bei der ursprünglichen Schreibweise des Problems auftreten. Diese beinhalten unter anderem nicht konvexe Zulaessigkeitsbereiche, die gewisse Regularitätsbedingungen nicht erfüllen. Neben der Umformulierung wird eine Phase–Field Relaxierung angewandt, die schließlich die Lösung des Problems mit gängigen Optimierungsmethoden ermöglicht. Durch die große Anzahl von Spannungsnebenbedingungen und durch die Einführung zusätzlicher Variablen im Rahmen der Umformulierung resultiert dieser Ansatz in großdimensionerten Optimierungsproblemen.

Die Optimalitätsbedingungen für Lösungen von restingierten Optimierungsproblemen führen auf groß-dimensionierte lineare indefinite Gleichungssysteme, sogenannte Sattelpunktprobleme, insbesondere wenn die Anzahl der Unbekannten hoch ist. Ein effizientes Behandeln dieser Gleichungssysteme kann das Bestimmen von Lösungen solcher Optimierungsprobleme enorm beschleunigen. Eines der effektivsten Lösungsverfahren für linear Gleichungssysteme ist das Mehrgitterverfahren, dessen erfolgreiche Anwendung mehrere Zutaten erfordert. Ein wichtiger Bestandteil ist eine passende Glättungsmethode. Um die oben genannten Sattelpunktprobleme effizient zu lösen, wird ein lokaler Patch-Glätter nach der multiplikativen Schwarz Technik angewandt.

Acknowledgement

First of all I am very grateful to my supervisor Professor U. Langer for employing me as a doctorate student at the SFB F013, for guiding my work, for his enthusiasm in times of success and for all his support over the years. At the same time I am greatly indebted to Professor M. Bendsøe for co-refereeing this thesis.

Moreover, I want to thank all my colleagues of the Special Research Program SFB F013 'Numerical and Symbolic Scientific Computing' for the nice working climate, the warm social surrounding and for all the time of joy and laughter. Special thanks go to W. Mühlhuber for guiding me and caring for me in the very beginning, to J. Schöberl for his patience with my numerous questions concerning his software package NETGEN/NGSolve, and to M. Burger for sharing his ideas and inspiring my work. I also would like to thank the faculty staff of the Institute of Computational Mathematics for support and advice whenever I asked for.

I want to thank Professor M. Bendsøe for inviting me twice to DTU (Technical University of Denmark), and for the fruitful discussions with him, Professor O. Sigmund and M. Stolpe. I really enjoyed being a part-time Dane.

Last but not least, all this would not have been possible without the financial support of the Austrian Science Fund 'Fonds zur Förderung der wissenschaftlichen Forschung' FWF under the grant SFB F013/F1309.

Notation and Symbols

$\mathbb{R}, \mathbb{R}^d$	–	Set of real numbers and set of vectors $\mathbf{x} = (x_i)_{i=1,\dots,d}^T$, $x_i \in \mathbb{R}$, $i = 1, \dots, d$.
x, $\mathbf{x}$	–	Scalar $x \in \mathbb{R}$ and vector $\mathbf{x} \in \mathbb{R}^d$.
Ω, Γ	–	Bounded domain (open and connected subset of $\mathbb{R}^d$, $d = 1, 2, 3$) with sufficiently smooth boundary $\Gamma = \partial\Omega$.
$u, \mathbf{u}$	–	Scalar valued function u, vector valued function $\mathbf{u}$.
$\mathbf{x} < \mathbf{y}$	–	Binary symbols like $=$, $<$, $\leq$, etc. in combination with vectors are always ment by components.
$\hookrightarrow$	–	Compact embedding.
$\overset{*}{\rightharpoonup}$	–	weak* convergence.
χ_A	–	Characteristic function of the set A.
supp f	–	supp $f = \{x \in \Omega \mid f(x) \neq 0\}$.
$\langle \cdot, \cdot \rangle$	–	Duality or inner product in an Hilbert space.
∇	–	Gradient operator, $\nabla u(\mathbf{x}) = \left(\dfrac{\partial u(\mathbf{x})}{\partial x_1}, \dots, \dfrac{\partial u(\mathbf{x})}{\partial x_d} \right)^T$ for $\mathbf{x} \in \mathbb{R}^d$.
$\triangle$	–	Laplace operator, $\triangle u(\mathbf{x}) = \displaystyle\sum_{i=1}^{d} \frac{\partial^2 u(\mathbf{x})}{\partial x_i^2}$ for $\mathbf{x} \in \mathbb{R}^d$.
div	–	Divergence operator, $\operatorname{div} \mathbf{u}(\mathbf{x}) = \displaystyle\sum_{i=1}^{d} \frac{\partial u_i(\mathbf{x})}{\partial x_i}$ for a vector valued function $\mathbf{u}(\mathbf{x}) = \big(u_1(\mathbf{x}), \dots, u_d(\mathbf{x})\big)^T$ and for $\mathbf{x} \in \mathbb{R}^d$.
$C(\Omega; R^d)$	–	$C(\Omega; \mathbb{R}^d) = \big\{ \mathbf{u} : \Omega \to \mathbb{R}^d \;\big|\; \mathbf{u} \text{ is continuous} \big\}$. If $d = 1$ $C(\Omega)$ will be used instead of $C(\Omega; R)$.
$C^k(\Omega)$	–	$C(\Omega) = \big\{ \mathbf{u} : \Omega \to \mathbb{R}^d \;\big|\; \mathbf{u} \text{ is } k\text{-times continuous differentiable} \big\}$.
$C^\infty(\Omega)$	–	$C^\infty(\Omega) = \big\{ \mathbf{u} : \Omega \to \mathbb{R}^d \;\big|\; \mathbf{u} \text{ is infinitely differentiable} \big\}$.
$C_0^k(\Omega), C_0^\infty(\Omega)$	–	$C_0^k(\Omega), C_0^\infty(\Omega)$, etc., denote these functions in $C^k(\Omega), C^\infty(\Omega)$, etc., with compact support.
$L_p(\Omega)$	–	$L_p(\Omega) = \big\{ u : \Omega \to \mathbb{R} \;\big|\; u \text{ is Lebesgue measurable}, \|u\|_{L_p(\Omega)} < \infty \big\}$.

$\|u\|_{L_p(\Omega)}$	–	$\|u\|_{L_p(\Omega)} = \left(\int_\Omega \|u(x)\|^p \, dx \right)^{\frac{1}{p}}$, $(1 \leq p < \infty)$.
$L_2(\Omega)$	–	Space of scalar square-integrable functions on Ω.
$L_{2,0}(\Omega)$	–	$L_{2,0}(\Omega) = \left\{ u \in L_2(\Omega) \;\middle|\; \int_\Omega u(\mathbf{x}) \, d\mathbf{x} = 0 \right\}$.
$\|u\|_{L_2(\Omega)}$, $\|u\|_0$	–	$\|u\|_{L_2(\Omega)} = \|u\|_0 = (u,u)^{\frac{1}{2}}_{L_2(\Omega)}$.
$(u,v)_{L_2(\Omega)}$, $(u,v)_0$	–	$(u,v)_{L_2(\Omega)} = (u,v)_0 = \int_\Omega u(x)v(x) \, dx$.
$L_\infty(\Omega)$	–	$L_\infty(\Omega) =$ $= \left\{ u : \Omega \to \mathbb{R} \;\middle|\; u \text{ is Lebesgue measurable}, \|u\|_{L_\infty(\Omega)} < \infty \right\}$.
$\|u\|_{L_\infty(\Omega)}$	–	$\|u\|_{L_\infty(\Omega)} =$ ess $\sup_\Omega \|u\|$.
$W^k_p(\Omega)$	–	$W^k_p(\Omega) = \left\{ u \in L_p(\Omega) \;\middle|\; \text{There exists a weak derivative } \partial^\alpha u \in L_p(\Omega), \; \forall \; 0 \leq \|\alpha\| \leq k \right\}$, $(1 \leq p < \infty)$.
$W^k_\infty(\Omega)$	–	$W^k_\infty(\Omega) = \left\{ u \in L_\infty(\Omega) \;\middle|\; \text{There exists a weak derivative } \partial^\alpha u \in L_\infty(\Omega), \; \forall \; 0 \leq \|\alpha\| \leq k \right\}$.
$H^1(\Omega)$	–	$H^1(\Omega) = \left\{ u \in L_2(\Omega) \;\middle|\; \nabla u \in L_2(\Omega; \mathbb{R}^d) \right\}$.
$\|u\|_{H^1(\Omega)}$, $\|u\|_1$	–	$\|u\|_{H^1(\Omega)} = \|u\|_1 = (u,u)^{\frac{1}{2}}_{H^1(\Omega)}$.
$(u,v)_{H^1(\Omega)}$, $(u,v)_1$	–	$(u,v)_{H^1(\Omega)} = (u,v)_1 = \int_\Omega u(x)v(x) \, dx + \int_\Omega \nabla u(x)^T \nabla v(x) \, dx$.
$H^1_{\Gamma_u}(\Omega)$	–	$H^1_{\Gamma_u}(\Omega) = \left\{ u \in H^1(\Omega) \;\middle|\; u = \mathbf{0} \text{ on } \Gamma_u \right\}$
$BV(\Omega; \{0,1\})$	–	$BV(\Omega; \{0,1\}) = \left\{ \mathbf{u} \in L_1(\Omega; \{0,1\}) \;\middle|\; \|\mathbf{u}\|_{BV} < \infty \right\}$.
$\|\mathbf{u}\|_{BV}$	–	$\|\mathbf{u}\|_{BV} =$ $= \sup \left\{ \int_\Omega \operatorname{div} \boldsymbol{\phi}(\mathbf{x}) \; \mathbf{u}(\mathbf{x}) \, d\mathbf{x} \;\middle|\; \boldsymbol{\phi} \in C^\infty_0(\Omega; \mathbb{R}^d), \; \|\boldsymbol{\phi}\|_\infty \leq 1 \right\}$.

a.e.	–	Almost everywhere.
BVP	–	Boundary value problem.
CAD	–	Computer aided design.
CG	–	Conjugate gradients.
FEM	–	Finite element method.
KKT	–	Karush-Kuhn-Tucker.
MMA	–	Method of moving asymptotes.
MG	–	Mulitgrid.
PCG	–	Proconditioned conjugate gradients.
PDE	–	Partial differential equation.
SCP	–	Sequential convex programming.
SLP	–	Sequential linear programming.
SQP	–	Sequential quadratic programming.
spd	–	Symmetric and positive definite.

Contents

Chapter 1

Introduction

1.1 State of the Art in Topology Optimization

Nowadays in almost all areas and sectors of industry and business faster and faster changing and adapting product specifications demand shorter and shorter production times. On the other hand, more and more competitive selling conditions call urgently for high quality products. In order to cope with this situation and to realize short development times, enterprises and business companies rely on computer simulations. These simulations reduce time consuming and costly experiments with prototypes. In this way they speed up precious development time and significantly help to reduce development expenses. The development loop, consisting of simulations and modifications of the model with respect to the results of the simulations, is normally repeated more than once. If the feedback of the simulations indicates only changes in the detailed design, these loops are rather cheap in comparison to the situation if changes in the basic layout of the design are enforced. Then it could happen that the whole development process is relocated to its conceptual stage, which endangers to delay the whole schedule. In a broad sense computer-aided engineering can be seen as the use of computer software and technology to assist the engineers in their design and development tasks. It is long since computer-aided design and finite element analysis became a basis of computer-aided engineering, which can't be imagined without nowadays. To avoid the situation described above, where the basic design has to be modified, already this initial design should be almost optimal in some sense.

During the last decade another field of applied mathematics reached a state of maturity to enter the world of computer-aided engineering, namely the *structural optimization* techniques. Structural optimization is a discipline dealing mostly with optimal designs of load-carrying structures, but also with other problems, like electro-magnetical tasks. For instance, consider the construction of new bridges, cars, airplanes and even satellites, or just parts of them. These are all examples where structural optimization can support engineers in their task to construct an initial design that fulfills some given requirements. Of course engineers don't start from scratch. They start the design process with already known designs and use their experience and knowledge to adapt recent layouts, where several, up to hundreds of man-years of labour and experience are hidden, to meet given requirements. We find typical settings for instance in the automotive and aerospace industry. Consider the body of a car. On the one hand it should show sufficient structural strength to guarantee safety in crash situations, which would indicate a strong and stiff frame. But on the other hand, it should have as little

weight as possible in order to lower fuel consumption. So it should be as light as possible but as stiff as necessary. The same situation occurs in the aircraft construction, where the overall weight of the plane is to be minimized without violating safety regulations. When constructing a satellite its support structure should not use more than a certain amount of material, but should be stiff enough to carry all its devices.

In all these examples we see the importance of the optimization of the geometry and the lay-out of structures. Structural optimization in general can be divided into four main areas: *Sizing optimization*, *shape optimization*, *material optimization*, and *topology optimization*. In all of them a physical quantity is optimized while equilibrium of forces and other constraints on the design are satisfied.

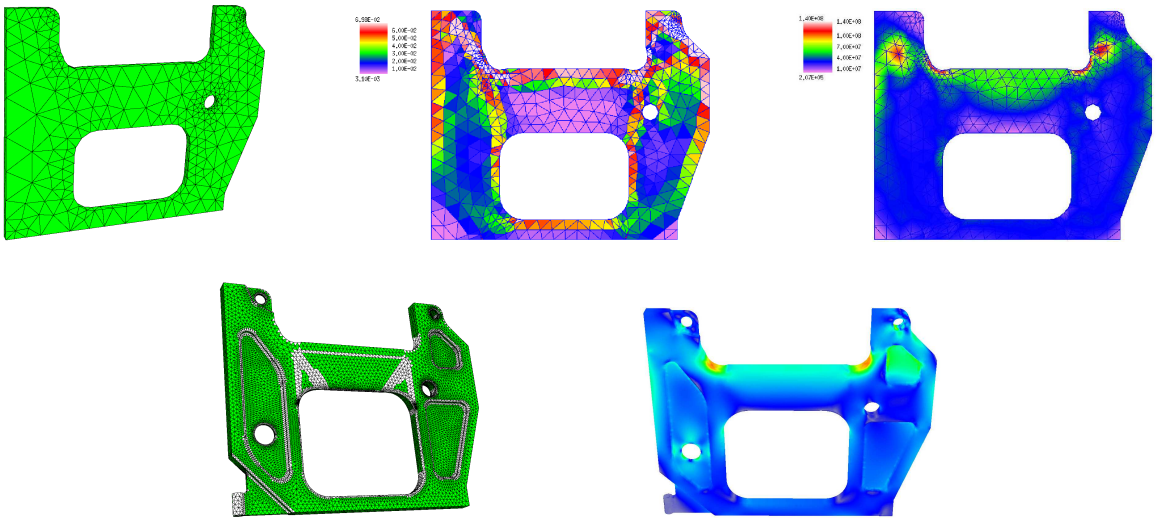

Figure 1.1: An optimal sizing problem of an industrial frame. Top Left: The original ground structure with an uniform thickness distribution, Top Mid: The optimal thickness distribution amongst others w.r.t. a global von Mises stress constraint, and Top Right: The optimal von Mises stress distribution. Below Left: A 3D meshed CAD model based on the result of the optimization. Below Right: The actual von Mises stress distribution in the 3D model. By courtesy of Engel Austria Gmbh.

Figure 1.2: Sizing optimization of a truss structure. Left: Ground structure with supports and load case, Right: Optimal design.

Sizing optimization problems can be seen as the simplest structural optimization problems, where the geometry of the design is fixed throughout of the optimization process. Optimal sizing is a $2\frac{1}{2}$-dimensional optimization where the design parameter is the thickness over a constant cross section, e.g., the thickness distribution over a elastic body or the volume of bars in a truss structure. Truss topology design problems are a subfield of optimal sizing problems. See Figure 1.1 for an example of an industrial optimal sizing problem and Figure

1.2 for an example of a typical truss optimization problem. For truss optimization with local stability considerations we refer e.g. to Achtziger [1, 2].

In *shape optimization* the design parameter is some kind of parameterization of (a part of) the boundary of the design, but the basic topology of the design is still pre-described. For instance shape optimization can be used as a post-processing tool after topology optimization in order to smooth a rough boundary, to remove stress concentrations along the boundary or to reduce e.g. the drag of a wing of an airplane. But it is a huge area on its own with lots of successful applications to real life problems, see e.g. Figure 1.3 for an application in magnetostatics. The measurements of the behavior of the electromagnet with the optimized pole-heads showed an increased performance by a factor of 4.5, see Lukáš et al. [88].

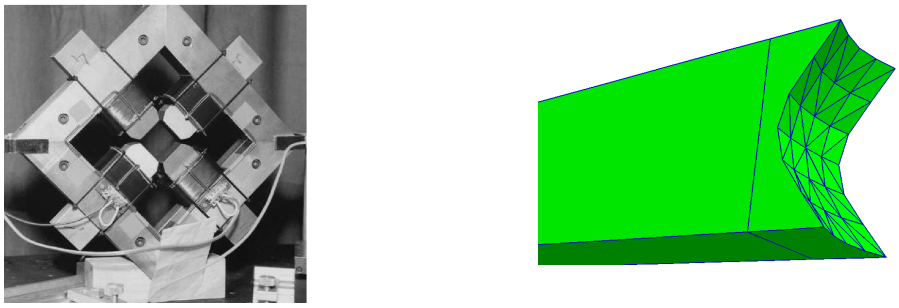

Figure 1.3: An optimal shape design of the pole-heads of an electromagnet. Left: The Maltese Cross with the original pole-heads, manufactured by the Institute of Physics, VŠB-Technical University of Ostrava. Right: The 3D optimized design of the pole-heads (from Lukáš [87]).

Material optimization is concerned with the design of materials with improved properties. The philosophy behind it can be described as: "Any material is a structure if you look at it through a sufficiently strong microscope." (taken from Bendsøe and Sigmund [22]). If we look through a microscope at any material, it will show a certain microstructure, e.g. like a honeycomb, where two phases, material and void, are arranged in a periodic way. In material optimization methods of topology optimization are now applied to those microstructures to create new materials with extreme or counterintuitive properties. See e.g. Sigmund [121] for a material with negative Poisson's ratio, i.e. a material consisting of microstructures that elongates transversely to an applied tensile load. A corresponding structure has also been manufactured in micro-scale, see Larson, Sigmund, and Bouwstra [84].

Figure 1.4: The Messerschmidt-Bölkow-Blohm (MBB) beam in topology optimization. Left: Ground structure with supports and load case, Right: A solution with a volume fraction of 50%.

In *topology optimization* no a-priori assumption on the topology of the structure is made. It is the most general area of structural optimization where in an ideal setting for every point in space it is determined, whether there should be material or not. The only problem

specifications are, e.g. for mechanical problems, the load cases, potential supports, possible restriction on the used volume or appearing stresses and so on. Topology optimization can be seen as a generalization of the areas above, since it determines the shape of the boundary of the structure, the number and shape of holes in the structure and even, if not prescribed, the optimal layout of the microstructures of the used material. Topology optimization has become an important tool in computer-aided engineering, because it helps designers to gain insight into alternative topological possibilities. In Figure 1.4 we see a typical example of a topology optimization problem. The MBB-beam has the function of carrying the floor in the fuselage of an Airbus passenger carrier.

So it turned out that for the past two decades the field of topology optimization, despite being relatively new, has been rapidly expanding with an enormous development in terms of theory, computational methods and applications including commercial applications (e.g. Optistruct by Altair). One can regard the work of Michell [93] as the first systematic contribution to the field, although the basic principles of topology optimization have been known for centuries. In 1904 Michell developed a theory for designs with very low volume fractions, resulting in thin-bar trusses that are optimal with regard to weight. Topology optimization for higher volume fractions started with the *homogenization theory* for the computation of effective properties for materials with periodic *microstructures*. These microstructures are constructed from a unit cell that consists, at a macroscopic level, of laminations of two or more materials. The composite material is then made up of an infinite number of such cells, now infinitely small and repeated periodically. Cheng and Olhoff [55] showed 1981 with optimal thickness distribution for elastic plates that composite materials appear naturally in structural optimization problems. Based on the homogenization theory the kickoff of finite element based topology optimization was caused by Bendsøe and Kikuchi [20] in 1988. Their *homogenization approach to topology optimization* yields a mathematically well-founded theory, where they simulate holes by allowing the stiffness of the composite tend to zero. Only one year later Bendsøe [16] offered in 1989 a different approach using the SIMP method (*Solid Isotropic Material with Penalization*). The naming refers to the limiting case of microstructures that are entirely occupied by one material, cf. Rozvany and Zhou [110]. So in comparison to the homogenization approach to topology optimization, the SIMP method usually models isotropic materials. These pioneering publications caused a vast development in the field of topology optimization using the microstructure and the SIMP approach. Since a detailed history of the evolution of topology optimization would go beyond the scope of this introductive chapter we refer to the monographs of Bendsøe [17] and Bendsøe and Sigmund [22] that describe the state of the art of topology optimization in the years 1995 and 2003. Moreover we would like to refer to survey papers, like, e.g. Eschenauer and Olhoff [62] and Rozvany [108]. Moreover, e.g. the recent monograph Allaire [5] covers the subject of homogenization in detail.

In the following we will continue with short introductions to significant issues of topology optimization. Besides the literature mentioned below, Bendsøe and Sigmund [22] always serves as a valuable reference. In this spirit we define the most common topology optimization problem, the minimal compliance problem, in an abstract way:

$$\begin{aligned} & J(\rho) \;\rightarrow\; \min_{\rho} \\ \text{subject to} \quad & \int_\Omega \rho(\mathbf{x})\, d\mathbf{x} \;\leq\; V, \\ & \rho(\mathbf{x}) \;\in\; \{0,1\}, \qquad \text{a.e. in } \Omega, \end{aligned} \tag{1.1}$$

where the aim is to find the stiffest design with limited volume V. So the objective $J(\rho)$ describes the compliance of the structure and the density function $\rho(\mathbf{x})$ specifies the structure as $\rho(\mathbf{x}) = 1$ if $\mathbf{x} \in \Omega$ is occupied with material and $\rho(\mathbf{x}) = 0$ otherwise. A well known example of the minimal compliance problem is shown in Figure 1.4. We will present a more detailed discussion of the minimal compliance problem in Chapter 4. Moreover the problem (1.1) is written in its nested version, see Chapter 3 for a comparison between the nested and the simultaneous approach.

1.1.1 Regularization

The basic motivation for the development of the homogenization theory and microstructures was to ensure well-posed optimization problems. It is well established that the problem (1.1) lacks existence of solutions in its general continuum setting. A discretized version of the problem trivially has a solution, since the design space is finite dimensional. A physical explanation of the ill-posedness is, that given a structure with a certain volume one can improve the stiffness by introducing a lot of small holes without changing the actual volume, which will lead to an indefinite perforation of the structure. Mathematically speaking the reason for this effect is the non-closedness of the feasible design set. The indefinite perforation leads to microstructures that are typically anisotropic. Therefore they cannot be modelled with the isotropic formulation of (1.1). Hence, the set of admissible designs lacks closure. For a well-founded review about this subject we refer to Sigmund and Petersson [129].

So there is need for regularization to transform the ill-posed problem to a well-posed problem, because this *non-existence* of solutions is indeed a problem for the numerical solutions of topology optimization problems. The effect that a larger number of holes in the structure appears and that more and more fine-scaled parts yield a more detailed structure, when solving the same problem on finer and finer grids, is called *mesh-dependence*. An illustration of the mesh-dependence effect can be seen in Figure 1.5. Ideally refining the mesh should result in the same optimal design, but with a better and smoother description of the boundary. Basically there are two different ways to circumvent the ill-posedness, namely *relaxation methods* and *restriction methods*.

Relaxation methods in principle enlarge the feasible set of designs so that a closure is formed. Without going into detail, we say that the relaxed problem is well-posed and the optimal designs of the relaxed problem are limits of sequences of optimal designs of the original problem. One relaxation method is the homogenization approach to topology optimization, described in detail in e.g. Bendsøe and Sigmund [22]. Due to Allaire and Kohn [9] and Allaire et al. [6] we know that well-posedness can be achieved by using so-called ranked layered microstructures. But complete theoretical knowledge is presently only known for problems involving compliance and fundamental frequency optimization. The second relaxation method is the *free material approach to topology optimization*, see Bendsøe et al. [19] and Bendsøe et al. [18]. Here the distribution of any material with a symmetric and positive semi-definite stiffness tensor achieves well-posedness for a broad range of problems. Using relaxation methods usually results in optimal designs with large areas with perforated microstructures and composite materials. Because of this high complexity of the optimal designs, the structures will probably be expensive and complicated to manufacture. Nevertheless, design with composite material is an important area on its own, e.g. for material optimization, see also Subsection 1.1.3.

From another viewpoint it is attractive to generate 0-1 solutions on a macroscopic level.

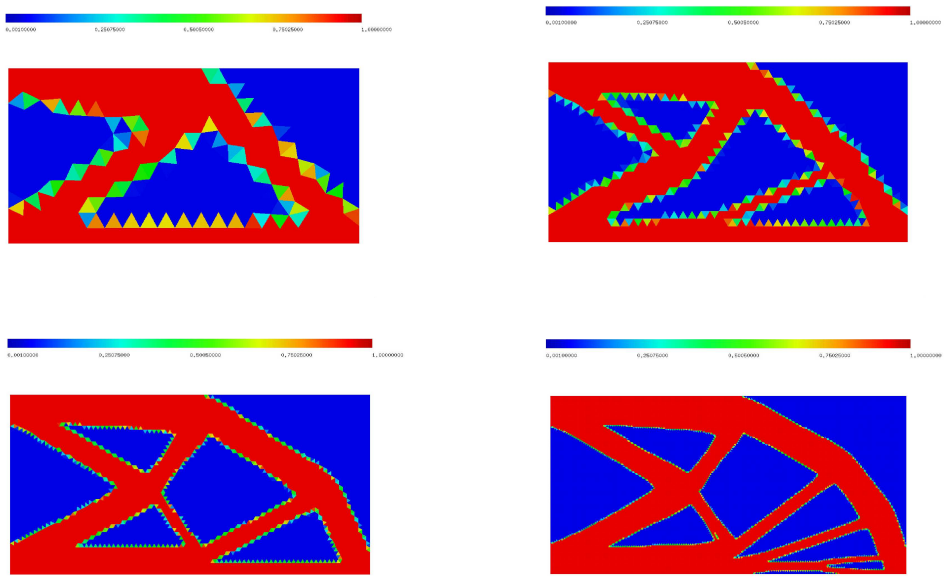

Figure 1.5: Mesh refinement without regularization. Solutions on a mesh with 449, 1839, 7319, and 29443 elements, respectively.

This can be achieved by doing quite the opposite to enlarging the design space, namely to restrict it. *Restriction methods* reduce the original feasible design set to a sufficiently compact subset by adding some local or global restriction on the variation of density. Basically there exist three kinds of restriction methods. *Perimeter* and *gradient control* can be added as a constraint to the problem or added as a penalty term to the objective. The third method is *mesh-independent filtering*. They all have in common that they rule out the possibility for fine scale structures to appear. Of course this reduces the cost and the complexity of the manufacturing process. But, unfortunately, there also drawbacks, like the uncertainty how to choose the corresponding penalty or constraint parameters. Also, since we restrict the design space, we might end up with solutions that are a trade-off due to the restrictions. The most serious drawback, however, is the fact that the optimization problem becomes non-convex. An overview and comparison of restriction methods is given e.g. in Borrvall [27]. *Perimeter* controlled restriction regularizes the problem in the sense, that the perimeter describes, vaguely speaking, the sum of the lengths or areas of all boundaries of the structure. So, limiting the perimeter obviously restricts the number of holes in the structure. Existence of solution to the perimeter constrained topology optimization problem was proved by Ambrosio and Buttazzo [10] and some convergence results can be found in Petersson [104]. When using e.g. the SIMP method and the density function is smooth enough one can simulate a bound to the perimeter as a bound on the *total variation* of the density function ρ. The total variation is then described by a L^1-bound on the gradient of ρ, like $\int_\Omega |\nabla\rho(\mathbf{x})|\, d\mathbf{x} \leq c$. Another possibility to restrict the gradient is to impose a *global gradient constraint* like the H^1-norm of the design $\left(\int_\Omega \rho(\mathbf{x})^2 + |\nabla\rho(\mathbf{x})|^2\, d\mathbf{x}\right)^{\frac{1}{2}} \leq c$. A proof of existence of solutions when using this bound (also the H^1-seminorm $\left(\int_\Omega |\nabla\rho(\mathbf{x})|^2\, d\mathbf{x}\right)^{\frac{1}{2}} \leq c$). is given in Bendsøe [17]. For the consideration of

general L^p-constraints, $\left(\int_\Omega |\rho(\mathbf{x})|^p \, d\mathbf{x}\right)^{\frac{1}{p}} \leq c$, we refer to Borrvall [27]. One can also impose *local gradient constraints* to the optimization problem. For constraining the density variation locally, Petersson and Sigmund [102] showed existence of solutions and convergence of the finite element scheme. Instead of adding extra constraints or additional penalty terms to the objective functional one can achieve existence of solutions by *filter methods*. Given a point $\mathbf{x}$, filtering the density means that the stiffness at that point depends on the density function in a neighborhood of that point. Mathematically well founded filtered techniques for the density are e.g. proposed in Bourdin [30] and Borrvall and Petersson [28]. Another way is to filter the sensitivities in a similar fashion, like proposed in Sigmund [121]. More information about regularization using filter techniques will be given in Section 4.2. In a comparison of the mentioned restriction methods we mention that they divide into two categories, global and local methods. Perimeter and global gradient restriction are global constraints and they allow thin bars to form. Just one extra constraint to the problem is added, but to determine the bound of the global constraints for new design problems is a serious problem. If this bound is too large, the constraint remains inactive, i.e. it has no regularizing effect. On the other hand, if it is to small, there might exist no optimal design. Especially for three-dimensional problem this task is tricky (cf. Fernandes, Guedes and Rodrigues [63]) and has to be solved by mostly costly experiments. Local methods like local gradient constraints or filter methods add a high number of extra constraints (in order of the number of finite elements after discretization) or an extra filter operator to the problem. But they will generally remove thin bars and, moreover, they put us in a position to control the minimum length scale of the optimal design. Which is an important issue for manufacturing considerations.

All the mentioned restriction methods eliminate not only mesh-dependence, but also the *checkerboard effect*, see Figure 1.6 for an example. The alternating arrangement of elements

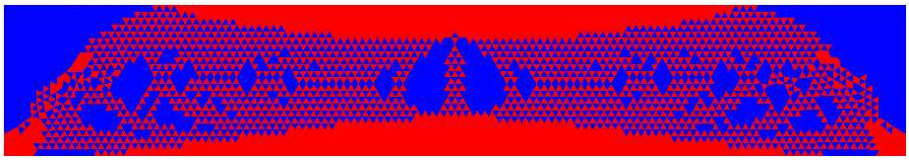

Figure 1.6: The checkerboard effect in the MBB beam example.

filled with material and void gave the effect its name. Due to bad numerical modelling these patterns are given an artificial high stiffness when analyzed in the discretized formulation. The reason for this is that the finite element discretization of the design problem constrains designs for which the finite element discretized set of admissible displacement fields is too small to give a sufficient information of the state of equilibrium. But this numerical anomaly does not only appear in topology optimization, but also e.g. in the finite element analysis of Stokes' flows. As the Stokes' flow problem, the topology optimization problem of finding an optimal design by distributing material is a two field problem. In Stokes problem (see also Subsection 2.2.2) we solve for the velocity field and the pressure and in an optimal design problem we seek, e.g. for the optimal material distribution and the corresponding displacement field. For saddle point problems like the Stokes' problem criteria have been developed to guarantee a stable finite element discretization, namely Brezzi's theory (see Brezzi and Fortin [40]) and the LBB-condition (2.17). Unfortunately the saddle point formulations arising from topology optimization, Chapter 3, don't fit into that framework.

Only under special assumptions PETERSSON [103] showed that it is possible to adopt Brezzi's theory of mixed finite element methods to the *variable thickness sheet problem*. Nevertheless theory tells us that certain combinations of finite elements for the design and the displacement field yield stable discretizations and some do not. So one suggestion to avoid checkerboard problems is to use higher order finite elements for the displacement field. This is of special interest when one does not want to use restriction methods, e.g. when analyzing the behavior of optimal structures with a very fine scale. For a summary of alternative methods to avoid the checkerboard effect we again refer to SIGMUND AND PETERSSON [129] and BENDSØE AND SIGMUND [22].

1.1.2 Material Interpolation

Let us reconsider the topology optimization problem (1.1). For each point $\mathbf{x} \in \Omega$ we have to decide whether to occupy it with material or not. So in the ideal case we treat a discrete valued design problem, or a *0-1* problem. In order to avoid (usually slow and low-scale) integer programming techniques, the constraint $\rho(\mathbf{x}) \in \{0, 1\}$ is relaxed to $0 \leq \rho(\mathbf{x}) \leq 1$ with a continuous variable $\rho(\mathbf{x})$ for $\mathbf{x} \in \Omega$. Nevertheless, it has been shown lately, that a broad class of topology optimization can be reformulated as linear or convex quadratic mixed 0–1 programs, which can be solved to global optimality using branch and bound techniques, see STOLPE AND SVANBERG [140] and STOLPE [136].

But commonly it is useful to consider a reformulation of (1.1) with a continuous design variable, having the application of gradient based mathematical programming algorithms in mind. Especially when the number of design variables after discretization is high ($\gg 1000$), which is usually the case. However, the design variable is then allowed to attain values between 0 and 1. Hence some kind of penalization of the intermediate density values is introduced to obtain again a more or less 0–1 design. In practical computations 0 is replaced by some small positive bound $\rho_{\min}$ to ensure ellipticity of the equilibrium equation, but we will omit this in this introductory section. If no penalization of the non-integer values is performed, the equilibrium equation depends linearly on the design and we call this problem the *variable thickness sheet problem*. It was first studied in RUSSOW AND TAYLOR [107] and acts as a basis for computational topology design. The linear dependence of the stiffness and volume on the density ρ yields existence of solutions for the minimal compliance problem. It can also be regarded as a sizing optimization problem.

The probably most popular penalization method is the already mentioned SIMP method, also called the *power-law approach* to topology optimization. Here a non-linear interpolation model of the form $\rho(\mathbf{x})^p$, with $p \geq 1$ is used. That means that material properties of intermediate densities are given by the properties of solid material times the element density raised to the power p. Combined with a volume constraint this approach penalizes intermediate values, since intermediate values give very little stiffness in comparison to the amount of used material. In other words, by choosing a higher value than 1 for the power p, it is inefficient for the algorithm to choose intermediate density values. When minimizing compliance the volume constraint is usually active in the optimal design and computations showed that in this case the optimal layout turns out to be an almost black and white design, if the value of p is high enough, usually $p \geq 3$ is needed. Additionally, if one wants to interpret areas with intermediate density values ('grey' areas) in the final design as a composite of materials, also $p \geq 3$ is required. This justification of the SIMP method with respect to intermediate values is given in BENDSØE AND SIGMUND [21], where also several other interpolation

schemes, e.g. based on the *Hashin-Shtrikman* bounds for two-phase materials, are discussed. Like the power-law approach there exist several other interpolation models with isotropic materials , e.g. an approach with rational function as in Stolpe and Svanberg [137]. This proposed scheme has been given the acronym RAMP for *Rational Approximation of Material Properties*. Further discussions related to SIMP and RAMP will be presented in Subsection 4.1.2.

A totally different approach to penalize intermediate density values is to choose a linear material interpolation and add an additional constraint to the optimization problem to encourage 0–1 optimal design. Such a penalty constraint could e.g. look like $\int_\Omega \rho(1-\rho)\, d\mathbf{x}$. But again, similar to the case of global regularizing constraints of the previous subsection, it is unclear how to choose a suitable bound for the constraint or a proper weighting factor when added to the objective.

1.1.3 Other objectives

Besides the classical minimal compliance problem lots of other applications in the field of topology design arose in the past. Of course we can only give a very short abstract of alternative applications; for an overview and an extensive list of references we refer again to Bendsøe and Sigmund [22], especially for the topics where we don't list any references.

First of all we would like to mention problems where the underlying equilibrium equations are not the elasticity equations (see Section 2.4) like in classical structural optimization problems. For topology optimization of fluids in Stokes flow we refer e.g. to Borrvall and Petersson [29] and Gersborg-Hansen, Sigmund and Haber [69]. As a representative of topology optimization with the Maxwell's equations we mention e.g. Hoppe, Petrova and Schulz [77] and Yoo and Hong [152]. By means of the Navier equation and the Helmholtz equation one can design structures and materials subject to wave propagation and band gaps. A band gap material does not allow wave propagation for certain frequency ranges, see e.g. Sigmund and Jensen [127]. Moreover, one can maximize the transmission through photonic crystal devices, like in Sigmund and Jensen [128].

However, one of the first topology optimization problems that appeared beside the minimal compliance problem have been topology designs for *vibration problems* and for *stability problems*. These problems in dynamics are of interest if one wants to create a certain gap between the eigenfrequencies of a structure and of an e.g. attached engine. In an optimal design with respect to maximal stiffness locally high stresses under loading can appear. Since this high stresses cause material failure, imposing stress constraints on structural optimization problems is an extremely important topic. For more informations about topology optimization with *local stress constraints* we refer to Chapter 5. A second topic that shows a weak point of the classical density based approach are *design dependent loads*, like pressure loads. Without additional parameterization the boundary between solid and void, where the pressure loads are attacking, is not well defined throughout the optimization process. Ways to resolve the problem can be found e.g. in Bourdin and Chambolle [31], using a phase–field approach (see also the Subsections 1.1.4 and 5.1.2) and in Sigmund and Clausen [126], where a mixed formulation is used (cf. Subsection 2.2.2).

Another subject in the field of topology optimization is mechanisms design. Here we distinguish between compliant mechanisms and articulated mechanisms. In contrast to articulated mechanisms with hinges, *compliant mechanisms* attain their mobility from the flexibilities of their components. A basic compliant mechanism design problem is the displacement

inverter. This and examples of crunching and gripping mechanisms can e.g. be found in SIGMUND [122]. Micro-Electro-Mechanical Systems (MEMS) are an important application of compliant mechanisms. Because of their microscale size ($\ll 1$ mm) one cannot use e.g. hinges in the production process. In SIGMUND [124, 125] we find various examples of MEMS design, like thermal, two-material thermal, and electro-thermo-mechanical actuators. We would like to mention that some of the actuators have actually been built and tested at microscale, see e.g. JONSMANN, SIGMUND AND BOUWSTRA [78]. For the design of *articulated mechanisms* trusses connected with hinges are used. Due to possible large displacement and high external loads geometrical non-linearity and buckling must be considered. Again we find typical tasks like to control the ratio between input and output displacements, like to maximize the output displacement or that the output point of the mechanisms has to follow a pre-described path. As an example of topology design of articulated mechanisms we refer to KAWAMOTO [81].

As a last application example of topology design we would like to mention *material design*. Besides the facts we already noted in Section 1.1 we want to state that the effective material properties are found by microstructure homogenization. But since we seek a microstructure with pre-described properties this material design method is called the *inverse homogenization method*. Objectives for material design are extremal elastic (see e.g. LIPTON [85]) or extremal multi physics properties like optimized thermoelastic design (see c.f. SIGMUND AND TORQUATO [130] and TURTELTAUB [145]), coupled piezoelectric-elastic design (see e.g. SIGMUND, TORQUATO AND AKSAY [131]) and combined elastic and conduction design (see SIGMUND [123] for an investigation of bone microstructure).

1.1.4 Non-classical methods

Besides the classical approach to topology optimization consisting of optimal material distribution using material interpolation, several other alternative approaches have been evolving. Here we do not want to discuss the *evolutionary* methods, where usually no sensitivity analysis is applied, but rather mention approaches like the level set method, the phase–field method and applications of the topological derivative.

The *level set* method was developed by OSHER AND SETHIAN [99] as a method for computing and analyzing the motion of an interface Γ. This interface describes a possible multiply connected set Ω. The method consists now of analyzing the motion of the interface under a velocity field, which can depend on various parameters like the position, time and geometry of Γ as well as on external physics. The interface is described by the zero level set of a sufficiently smooth function $\phi(\mathbf{x}, t)$ as $\Gamma(t) = \{\mathbf{x} \mid \phi(\mathbf{x}, t) = 0\}$. The interior of the set Ω is then defined as the set where $\phi(\mathbf{x}, t)$ is negative. For a more detailed discussion of the level set method we refer to OSHER AND FEDKIW [98] and to BURGER AND OSHER [47] for an up-to-date review. In e.g. SETHIAN AND WIEGMANN [120] the level set method was extended to capture the free boundary of a structure on a fixed mesh and to use for example the stresses to modify the design towards optimizing chosen properties. A framework for incorporating the level set method, based on the classical shape derivative, into shape optimization is given in BURGER [43], especially for adjusting the velocity field for the evolution of the interface. For examples of topology optimization using the level set method we refer e.g. to ALLAIRE, JOUVE AND TOADER [8] and to WANG, WANG AND GUO [148].

As an alternative approach the *topological derivative* is a tool that allows to quantify the sensitivity of a given objective functional with respect to the introduction of an infinitesimally small hole in the design domain. The idea is to test the optimality of a structure to topology

variations by creating a small hole. With a suitable criterion one can predict the most effective position for inserting a hole with appropriate boundary conditions. The topological derivative, based on the same idea as the *bubble method* (cf. Eschenauer and Schuhmacher [61]) was rigorously analyzed in Sokolowski and Zochowski [132]. The level set method can easily remove holes but can hardly create new ones in the middle of the design. This drawback is inconvenient mostly in two dimensions, since in three dimensions new holes can be created by pinching two boundaries. A remedy for two dimensional computations is to couple the level set method with the topological derivative as proposed e.g. in Burger, Hackl and Ring [44]. This approach was successfully tested in the field of structural optimization by e.g. Allaire Et Al [7].

The last non-classical method we will mention is the so called *phase–field* method in optimal design. Since this method is discussed in more detail in Subsection 5.1.2 we just give a brief introduction. The phase–field method consists in using a linear material interpolation and an additional Cahn-Hilliard (cf. Cahn and Hilliard [51]) type penalization functional, which is added to the objective functional. This parameter-dependent penalization functional is used to approximate the perimeter of the structure and to ensure that the material density converges pointwise to 0 or 1 as the parameter tends to 0. The phase–field method, which is closely related to level set methods, was, to the knowledge of the author, first introduced by Bourdin and Chambolle [31] to the field of topology optimization for a problem with design dependent loads, as also mentioned in the previous subsection.

1.2 Overview

The emphasis of this thesis is to present advanced multilevel methods for topology optimization problems. In particular we focus on two special design - constraint combinations, namely minimizing the compliance of a structure with respect to limited mass and the minimization of mass while keeping a certain stiffness. Both problems differ in various aspects and are also treated in different ways. Structural optimization problems are located on the interface to nonlinear numerical optimization, to the analysis and numerical treatment of partial differential equations, to solution techniques of systems of linear equations, and to modelling physical processes, e.g. in solid mechanics. Thus, as a starting point we give a brief introduction into those fields of scientific computing. The remainder of the thesis is then organized as follows.

There exist basically two approaches for solving optimization problems governed by partial differential equations. Therefor, we split the unknowns into to groups of parameters, the state parameters and the design parameters. The state parameters represent the state of the PDE and the design parameters describe the design. Beside the method to eliminate the state parameters, hence to reduce the total number on unknowns, also the approach to treat both groups of unknowns simultaneously in the optimization problem, gains more importance in topology optimization. A short introduction and a brief discussion of the two approaches, the *nested approach* and the *simultaneous approach*, is given in Chapter 3.

In Chapter 4 we discuss an adaptive multilevel approach to the minimal compliance problem in its nested formulation. It is well known that topology optimization problems are not well-posed, so regularization is needed. We discuss to *filter techniques* by Sigmund [121] and Borrvall and Peterson [28] that are used for regularization. Moreover, we use one of the methods also for mesh-refinement along the interface of material and void, i.e., the boundary of the structure. The resulting optimization problems on each level in the hierarchy of nested

meshes are solved by the *method of moving asymptotes*. The systems of linear equations, that arise from the finite element discretization from the PDEs are solved by a *multigrid method*. STAINKO [133] acts as a framework for this chapter.

A new approach to the minimal mass problem with respect to local stress constraints is presented in Chapter 5. Here a main source of difficulties is a lack of constraint qualifications for the set of feasible designs, defined by the local stress constraints. As one cornerstone of the approach serves a *reformulation* of the set of constraints in the continuous setting, due to STOLPE AND SVANBERG [140] in a discrete framework. This reformulation is possible because of the used simultaneous formulation of the problem. The reformulation results in linear and 0–1 constraints only. These are finally relaxed by a *phase–field* method, the second cornerstone of the approach. I.e., the 0-1 constraints are approximated by a *Cahn-Hillard* type penalty in the objective functional, which yields convergence of minimizers to 0-1 designs as the related penalty parameter tends to zero. A major advantage of this kind of relaxation opposed to standard approaches is a uniform constraint qualification that is satisfied for any positive value of the penalization parameter. We solve the finite-dimensional programming problems, resulting from finite element discretization, by an *interior-point* method. A shorter version of this chapter is given in the paper BURGER AND STAINKO [48].

The optimality conditions of restricted optimization problems lead to indefinite systems of linear equations, in fact saddle point problems. Most of the computing time of interior-point methods is actually spent to the solution of such saddle point problems. In Chapter 6 we derive an efficient solver with optimal complexity for these systems. *Multigrid methods* certainly belong to the most efficient methods for solving large-scale systems, e.g., arising from discretized partial differential equations. One of the most important ingredient of an multigrid method is an appropriate smoother. In this chapter we consider a *multiplicative Schwarz-type smoother*, that consists of the solution of several small local saddle point problems.

Finally, in Chapter 7 we present some conclusions and an outlook on possible related future work.

Chapter 2

Basics

2.1 Numerical Optimization

2.1.1 Basics of Constrained Optimization

In this section we list conditions for local solutions (minima) for general constrained optimization problems. For sake of simplicity we shall restrict ourselves to a finite dimensional setting. We refer the reader e.g. to the following literature Fletcher [64], Nocedal and Wright [96], and Geiger and Kanzow [68]. Lets turn our attention to the minimization of a differentiable objective function $J(\mathbf{x})$ where the variables are subjected to constraints:

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

$\mathcal{E}$ and $\mathcal{I}$ are the disjoint sets of the indices of *equality* and *inequality* constraints. Throughout this section we will assume that $J \in C^2$ and $c_i \in C^1$ map $\mathbb{R}^n$ into $\mathbb{R}$. We call a point $\mathbf{x}$ *feasible* if it satisfies the constraints in (2.1). In contrast to unconstrained optimization, where we can specify conditions for local solutions only using the objective J, we state conditions for constrained optimization using the Lagrange function or *Lagrangian* for (2.1)

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

The *Lagrange multiplier* vector $\boldsymbol{\lambda}$ is defined with the components λ_i with $i \in \mathcal{E} \cup \mathcal{I}$. At any feasible point $\mathbf{x}$ we define the *active set*

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

Before we state the first-order necessary conditions we have to take a closer look at the properties of the constraints and make sure that they don't show any degenerate behavior. We do this by assuming that so-called *constraint qualifications* hold at a suspected minimum $\overline{\mathbf{x}}$. Such constraint qualification are like the following:

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

Definition 2.2 (MFCQ)**.** *The* Mangasarian–Fromovitz constraint qualification *holds at the point* $\overline{\mathbf{x}}$ *if the set of gradients of the equality constraints* $\{\nabla c_i(\overline{\mathbf{x}}) \mid i \in \mathcal{E}\}$ *is linear independent and if there exists a vector* $\mathbf{d} \in \mathbb{R}^n$*, so that*

$$\nabla c_i(\overline{\mathbf{x}})^T \mathbf{d} < 0, \quad \forall\ i \in \mathcal{A}(\overline{\mathbf{x}}) \cap \mathcal{I} \qquad \textit{and} \qquad \nabla h_i(\overline{\mathbf{x}})^T \mathbf{d} = 0, \quad \forall\ i \in \mathcal{E}.$$

Definition 2.3 (SCQ)**.** *The* Slater constraint qualification *holds if the constraints* c_i*, for* $i \in \mathcal{E}$*, are affine linear and* c_i*, for* $i \in \mathcal{I}$*, are convex and there exists a feasible point* $\overline{\mathbf{x}}$ *so that*

$$c_i(\overline{\mathbf{x}}) < 0, \qquad \forall\ i \in \mathcal{I}.$$

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

Theorem 2.1 (First-Order Necessary Conditions)**.** *Suppose that* $\overline{\mathbf{x}}$ *is a local solution of* (2.1) *and that the LICQ holds at* $\overline{\mathbf{x}}$*. Then there exists a Lagrange multiplier* $\overline{\boldsymbol{\lambda}}$*, such that the following conditions are satisfied:*

$$\begin{aligned}
\nabla_{\mathbf{x}} \mathcal{L}(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}) &= 0, && && \text{(2.2a)}\\
c_i(\overline{\mathbf{x}}) &= 0, && \forall\ i \in \mathcal{E}, && \text{(2.2b)}\\
\overline{\lambda}_i \geq 0 \quad \textit{and} \qquad c_i(\overline{\mathbf{x}}) &\leq 0, && \forall\ i \in \mathcal{I}, && \text{(2.2c)}\\
\overline{\lambda}_i c_i(\overline{\mathbf{x}}) &= 0, && \forall\ i \in \mathcal{E} \cup \mathcal{I}. && \text{(2.2d)}
\end{aligned}$$

Proof. See e.g. Nocedal and Wright [96]. □

Any point $\overline{\mathbf{x}}$ that satisfies (2.2) under the above assumptions is said to be a *first-order critical* or a *KKT* point for the problem (2.1). The conditions (2.2d) is called the *complementarity* condition, while (2.2a) requires that the gradient of the Lagrangian vanishes at a KKT point:

$$\nabla_{\mathbf{x}} \mathcal{L}(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}) = \nabla J(\overline{\mathbf{x}}) + \sum_{i \in \mathcal{A}(\overline{\mathbf{x}})} \overline{\lambda}_i \nabla c_i(\overline{\mathbf{x}}).$$

Here we write $i \in \mathcal{A}(\overline{\mathbf{x}})$ instead of $i \in \mathcal{E} \cup \mathcal{I}$ because the complementarity condition implies that the Lagrange multipliers corresponding to the inactive inequality constraints ($i \notin \mathcal{A}(\overline{\mathbf{x}})$) are zero.

As theorem 2.1 states only necessary conditions we need more information if $\overline{\mathbf{x}}$ is a local minimum or not. With the following definition we state the necessary and sufficient conditions:

Definition 2.4 (Cone of Critical Directions)**.** *Given a point* $\overline{\mathbf{x}}$ *and the active constraint set* $\mathcal{A}(\overline{\mathbf{x}})$*, the cone* $C(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}})$ *is defined by*

$$C(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}) = \left\{ \mathbf{s} \in \mathbb{R}^n \;\middle|\; \begin{array}{ll} \nabla c_i(\overline{\mathbf{x}})^T \mathbf{s} = 0 & \forall\ i \in \mathcal{E}, \\ \nabla c_i(\overline{\mathbf{x}})^T \mathbf{s} = 0 & \forall\ i \in \mathcal{A}(\overline{\mathbf{x}}) \cap \mathcal{I} \textit{ with } \overline{\lambda}_i > 0, \\ \nabla c_i(\overline{\mathbf{x}})^T \mathbf{s} \leq 0 & \forall\ i \in \mathcal{A}(\overline{\mathbf{x}}) \cap \mathcal{I} \textit{ with } \overline{\lambda}_i = 0. \end{array} \right\}$$

Theorem 2.2 (Second-Order Necessary Conditions)**.** *Suppose that* $\overline{\mathbf{x}}$ *is a local solution of* (2.1) *and that the LICQ holds at* $\overline{\mathbf{x}}$*. Let* $\overline{\boldsymbol{\lambda}}$ *be a corresponding Lagrange multiplier such that* $(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}})$ *satisfy the KKT conditions* (2.2)*, and let* $C(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}})$ *be defined as above. Then*

$$\mathbf{s}^T \nabla^2_{\mathbf{x}} \mathcal{L}(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}) \mathbf{s} \geq 0, \qquad \forall\ \mathbf{s} \in C(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}).$$

Proof. See e.g. NOCEDAL AND WRIGHT [96]. □

Any point $\overline{\mathbf{x}}$ that satisfies these conditions is said to be a *strong second-order critical point* for the problem (2.1). In the case of unconstrained optimization $J(\mathbf{x}) \to \min_{\mathbf{x}\in\mathbb{R}^n}$ Theorem 2.2 simplifies to the condition that the Hessian of the objective is positive semidefinite ($C(\overline{\boldsymbol{\lambda}}) = \mathbb{R}^n$), the well known second-order necessary condition for unconstrained optimization.

Next we state the second-order sufficient conditions for the problem (2.1):

Theorem 2.3 (Second-Order Sufficient Conditions)**.** *Suppose that for some feasible point* $\overline{\mathbf{x}}$ *there exists a corresponding Lagrange multiplier such that* $(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}})$ *satisfy the KKT conditions* (2.2)*. Suppose also that*

$$\mathbf{s}^T \nabla^2_{\mathbf{x}} \mathcal{L}\big(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}\big)\mathbf{s} > 0, \qquad \forall\ \mathbf{s} \in C\big(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}\big),\ \mathbf{s} \neq 0.$$

Then $\overline{\mathbf{x}}$ *is a strict local minimum for* (2.1)*.*

Proof. See e.g. NOCEDAL AND WRIGHT [96]. □

In the case of unconstrained optimization Theorem 2.3 states the usual second-order sufficient condition for unconstrained optimization, the positive definiteness of the Hessian of the objective.

Notes and remarks for subsection 2.1.1

- If we consider the complementary condition (2.2d) there are 3 cases that can occur for $i \in \mathcal{I}$:
  a) $c_i(\overline{\mathbf{x}}) = 0 \ \wedge\ \overline{\lambda}_i > 0$: The constraint c_i at the point $\overline{\mathbf{x}}$ is said to be *strongly active*.
  b) $c_i(\overline{\mathbf{x}}) = 0 \ \wedge\ \overline{\lambda}_i = 0$: The constraint c_i at the point $\overline{\mathbf{x}}$ is said to be *weakly active*.
  c) $c_i(\overline{\mathbf{x}}) < 0$: The constraint c_i at the point $\overline{\mathbf{x}}$ is said to be *inactive*.
- Lagrange multipliers are often known as *dual variables*.

2.1.2 The Method of Moving Asymptotes

Typically a large number of design variables appears in topology optimization problems, since for a good representation of the design we have to work with rather fine finite element meshes. So on the other hand we have for each element at least one design variable, but on the other hand we have usually a rather small number of constraints. It is a common approach of mathematical programming methods for non-linear optimization problems to formulate a local model at an iteration point. This local model approximates the original one at the given iteration point, but is easier to solve. For an overview of non-linear programming methods see e.g. NOCEDAL AND WRIGHT [96]. Classical methods like *Sequential Quadratic Programming (SQP)* (see e.g. BOGGS AND TOLLE [26] for a survey) use such local models. But with respect to the large number of design variables the use of SQP methods and solving the local models is very costly if not even impossible, due to the fact that gathering second order information for the approximation of the Hessian could be an insuperable task.

One method that turned out to be very efficient for topology optimization problems, in academical and industrial environment, is the *Method of Moving Asymptotes (MMA)* by

SVANBERG [141] (1987). As its mother method *CONLIN*, see FLEURY AND BRAIBANT [65] (1986), the MMA works with a sequence of simpler approximating subproblems (similar to *Sequential Linear Programming (SLP)* and SQP), but their approximation is based on terms of direct and reciprocal design variables. A major advantage of the MMA is that these local models are convex and separable and only require one function and gradient evaluation at the iteration point. This is an important fact since evaluations of the original problem can be very time consuming, especially if the original problem is formulated in a nested way (see Chapter 3) and the evaluation includes a finite element analysis. Separability means that the necessary optimality conditions of the subproblem do not couple the design variables. This yields that instead of one n-dimensional problem we have to solve n one-dimensional problems. Convexity means that that dual or primal-dual methods can be used to attack the subproblems. These valuable properties allow to reduce computational costs for solving the subproblems significantly. A solution of a subproblem is then used as the next iteration point.

Let us now consider a structural optimization problem of the following form:

$$\begin{aligned} J(\mathbf{x}) &\rightarrow \min_{\mathbf{x}\in R^n} \\ \text{subject to} \quad c_i(\mathbf{x}) &\leq \hat{c}_i, \qquad i \in \{1,\dots,m\}, \\ \underline{\mathbf{x}} \leq \mathbf{x} &\leq \overline{\mathbf{x}}, \end{aligned}$$

where the bound constraints are understood by components. Given an iteration point $\mathbf{x}^{(k)}$ an approximation of a given function f will look like the following: Firstly two parameters $\mathbf{L}^{(k)}$ and $\mathbf{U}^{(k)}$ are chosen, such that $\mathbf{L}^{(k)} < \mathbf{x}^{(k)} < \mathbf{U}^{(k)}$. Based on these parameters the approximation $\tilde{f}^{(k)}$ of the given function f is defined as

$$\tilde{f}^{(k)}(\mathbf{x}) = r^{(k)} + \sum_{i=1}^{n} \left(\frac{p_i^{(k)}}{U_i^{(k)} - x_i} + \frac{q_i^{(k)}}{x_i - L_i^{(k)}} \right), \tag{2.3}$$

where $r^{(k)}$ and the coefficients $\mathbf{p}^{(k)}$, $\mathbf{q}^{(k)}$ are chosen as

$$\begin{aligned} p_i^{(k)} &= \begin{cases} \left(U_i^{(k)} - x_i^{(k)}\right)^2 \frac{\partial f}{\partial x_i}\left(\mathbf{x}^{(k)}\right), & \text{if } \frac{\partial f}{\partial x_i}\left(\mathbf{x}^{(k)}\right) > 0, \\ 0, & \text{if } \frac{\partial f}{\partial x_i}\left(\mathbf{x}^{(k)}\right) \leq 0, \end{cases} \\ q_i^{(k)} &= \begin{cases} 0, & \text{if } \frac{\partial f}{\partial x_i}\left(\mathbf{x}^{(k)}\right) \geq 0 \\ -\left(x_i^{(k)} - L_i^{(k)}\right)^2 \frac{\partial f}{\partial x_i}\left(\mathbf{x}^{(k)}\right), & \text{if } \frac{\partial f}{\partial x_i}\left(\mathbf{x}^{(k)}\right) < 0, \end{cases} \\ r^{(k)} &= f\left(\mathbf{x}^{(k)}\right) - \sum_{i=1}^{n} \left(\frac{p_i^{(k)}}{U_i^{(k)} - x_i^{(k)}} + \frac{q_i^{(k)}}{x_i^{(k)} - L_i^{(k)}} \right). \end{aligned}$$

It is worth noting that $\tilde{f}^{(k)}$ is a convex function and a first order approximation of f. The parameters $\mathbf{L}^{(k)}$ and $\mathbf{U}^{(k)}$ act like asymptotes in (2.3) and control, loosely speaking, the range for which $\tilde{f}^{(k)}$ approximates f reasonably. The tighter we choose $\mathbf{L}^{(k)}$ and $\mathbf{U}^{(k)}$ around $\mathbf{x}^{(k)}$ the more curvature is given to the approximating function, the more conservative becomes the approximation of the original problem. In fact we can state the following (see SVANBERG [141]):

Remark 2.1 (Properties of The Asymptotes)**.**

a) Assume that $\tilde{f}^{(k)}$ *and* $\tilde{\tilde{f}}^{(k)}$ *are two approximating functions corresponding to* $\tilde{\mathbf{L}}^{(k)} \leq \tilde{\tilde{\mathbf{L}}}^{(k)} < \mathbf{x}^{(k)} \leq \tilde{\tilde{\mathbf{U}}}^{(k)} \leq \tilde{\mathbf{U}}^{(k)}$. *Than for* $\tilde{\tilde{\mathbf{L}}}^{(k)} < \mathbf{x} \leq \tilde{\tilde{\mathbf{U}}}^{(k)}$ *it holds that* $\tilde{f}^{(k)}(\mathbf{x}) \leq \tilde{\tilde{f}}^{(k)}(\mathbf{x})$.

b) Assume that $\mathbf{L}^{(k)}$ *and* $\mathbf{U}^{(k)}$ *are chosen far away from* $\mathbf{x}^{(k)}$, *then the approximation becomes close to linear. Let* $'\mathbf{L}^{(k)} = -\infty'$ *and* $'\mathbf{U}^{(k)} = \infty'$ *then we have that* $\tilde{f}^{(k)}(\mathbf{x}) = f(\mathbf{x}) + \sum_j \frac{\partial f}{\partial x_j}(\mathbf{x})(x_j - x_j^{(k)})$.

More details on how to choose the asymptotes and how to generate strictly conservative approximations can be found in Svanberg [141, 143] and in Bruyneel, Duysinx, and Fleury [41]. Since these update schemes for the asymptotes rely on information from previous iterations, the approximating subproblems are also based on some iteration history.

The resulting subproblem at iteration point $\mathbf{x}^{(k)}$ looks now like the following:

$$\begin{aligned} \tilde{J}^{(k)}(\mathbf{x}) \;&\rightarrow\; \min_{\mathbf{x}\in R^n} \\ \text{subject to} \qquad \tilde{c}_i^{(k)}(\mathbf{x}) \;&\leq\; \hat{c}_i, \qquad\qquad i \in \{1,\ldots,m\}, \\ \max\left\{\underline{\mathbf{x}}, \boldsymbol{\alpha}^{(k)}\right\} \;&\leq\; \mathbf{x} \;\leq\; \min\left\{\overline{\mathbf{x}}, \boldsymbol{\beta}^{(k)}\right\}, \end{aligned}$$

with $\mathbf{L}^{(k)} < \boldsymbol{\alpha}^{(k)} \leq \boldsymbol{\beta}^{(k)} < \mathbf{U}^{(k)}$. $\tilde{J}^{(k)}$ and $\tilde{c}_i^{(k)}$ for $i \in \{1,\ldots,m\}$ are the approximating functions of J and c_i respectively, constructed as in (2.3). These subproblems can be solved now using a dual methods or primal-dual methods, like an interior point approach, see e.g. Zillober [154].

Notes and remarks for subsection 2.1.2

- Although the original MMA is considered to be a reliable and fast method, it is not globally convergent. It is possible to construct problems on which it does not converge. Globally convergent versions of the MMA can be found in Zillober [155], by adding a line-search procedure, and in Svanberg [143], relying on strictly convex conservative approximations.

- In Svanberg [143] a new class of optimization methods is presented and called *conservative separable approximation (CCSA) methods*, where the MMA is one special case. Due to the conservative approximation schemes of these methods the iteration points are always feasible with respect to the original problem. Which is e.g. not the case for the linearized constraints in the QP subproblems of SQP methods.
  For more families of the MMA-like approximations we refer to Bruyneel, Duysinx, and Fleury [41], where various versions are presented, utilizing the gradient and function values at two successive design points to improve the quality of the approximation.

- Moreover, a software tool for structural optimization problems named SCPIP is described in Zillober [156]. Here SCPIP stands for *sequential convex programming* combined with a primal-dual *interior point* approach for the resulting the MMA-like subproblems combined with a line-search for global convergence.

2.1.3 Interior–Point Methods

In the last two decades interior–point algorithms have evolved to efficient methods for large scale nonlinear programming since their revival in 1984. For a survey see e.g. the related chapters in Nocedal and Wright [96] and Wright [151] and the references cited therein.

The rediscovery of interior–point methods is rooted in the desire to find algorithms with a better complexity than the *simplex method* for linear programming by Dantzig in 1947. Since then, the simplex method dominated the field of linear programming, although its worst-case complexity is exponential in the size of the problem dimension. After Karmakar's announcement in 1984 of the *projective algorithm*, a polynomial-time method for linear programs, interior–point methods have been the subject of intense research. In principle there are two ways to motivate these methods nowadays, namely minimizing a barrier function or perturbing the optimality conditions.

For a short introduction we consider again the following general optimization problem (as (2.1)):

$$\begin{aligned} & J(\mathbf{x}) \;\rightarrow\; \min_{\mathbf{x}\in\mathbb{R}^n} & & \\ \text{subject to} \quad & c_i(\mathbf{x}) \;=\; 0, & & i \in \mathcal{E}, \\ & c_i(\mathbf{x}) \;\leq\; 0, & & i \in \mathcal{I}. \end{aligned} \tag{2.4}$$

where all appearing functions should be sufficiently differentiable. For the sake of simplified notation we will denote $\mathbf{c}_\mathcal{I}(\mathbf{x})$ as $(c_i(\mathbf{x}))_{i\in\mathcal{I}}$ and $\mathbf{c}_\mathcal{E}(\mathbf{x})$ as $(c_i(\mathbf{x}))_{i\in\mathcal{E}}$. This problem is then modified such that the restricting inequality constraints are treated implicitly by adding them to the objective functional using some barrier term. Appropriate barrier functions are characterized by the following properties:

Remark 2.2 (Properties of barrier functions)**.** *Let $B_\mu(\mathbf{x})$ denote a barrier function, then:*

a) $B_\mu(\mathbf{x})$ depends only on inequality constraints $\mathbf{c}_\mathcal{I}(\mathbf{x})$ and is infinite outside the interior of the feasible region (of (2.4)*).*

b) $B_\mu(\mathbf{x})$ is smooth inside the feasible region and preserves the continuity properties of $\mathbf{c}_\mathcal{I}(\mathbf{x})$.

c) The value of $B_\mu(\mathbf{x})$ tends to ∞ as $\mathbf{x}$ approaches the boundary of the feasible region.

The predominant barrier function is the logarithmic barrier function and so the new barrier objective $J_\mu(\mathbf{x}) := J(\mathbf{x}) + B_\mu(\mathbf{x})$ is now the sum of the original one and a logarithmic interior part:

$$\begin{aligned} J(\mathbf{x}) - \mu \sum_{i\in\mathcal{I}} \ln\big(-c_i(\mathbf{x})\big) \;&\rightarrow\; \min_{\mathbf{x}\in\mathbb{R}^n} \\ \text{subject to} \quad \mathbf{c}_\mathcal{E}(\mathbf{x}) \;&=\; \mathbf{0}, \end{aligned} \tag{2.5}$$

where $\mu > 0$ is called the *barrier parameter*. A major characteristic of these methods is that all inequality constraints are (have to be) satisfied strictly, which leads to the labelling *interior-point* methods. Minimization of (2.5) for a decreasing sequence of the barrier parameter $\mu \to 0$ will result (under appropriate assumptions) in a sequence of minimizers $\overline{\mathbf{x}}_\mu \to \overline{\mathbf{x}}_0 = \overline{\mathbf{x}}$ converging to the minimizer $\overline{\mathbf{x}}$ of the original problem (2.4). The sequence $\overline{\mathbf{x}}_\mu$ also defines a path to $\overline{\mathbf{x}}$, which is either called the *central path* or the *barrier trajectory*. The central path is a path of strictly feasible points that satisfy the perturbed complementarity conditions, see below. It is the essential idea of most interior–point methods to follow this path numerically more or less exactly. Path following methods are related to *homotopy* methods for general nonlinear equations, which define a path to the solution as well.

Using the following notation we state the first order necessary optimality conditions for (2.5): $\mathbf{C}_{\mathcal{I}}(\mathbf{x}) = \operatorname{diag}(c_i(\mathbf{x}), i \in \mathcal{I})$, $\boldsymbol{\lambda}_{\mathcal{E}}$ the vector of Lagrange multipliers for the equality constraints and $\mathbf{e}$ a vector of ones in the appropriate dimension:

$$\begin{aligned} \nabla J(\mathbf{x}) + \mu \nabla \mathbf{c}_{\mathcal{I}}(\mathbf{x})^T \mathbf{C}_{\mathcal{I}}(\mathbf{x})^{-1}\mathbf{e} + \nabla \mathbf{c}_{\mathcal{E}}(\mathbf{x})^T \boldsymbol{\lambda}_{\mathcal{E}} &= \mathbf{0}, \\ \mathbf{c}_{\mathcal{E}}(\mathbf{x}) &= \mathbf{0}. \end{aligned} \tag{2.6}$$

Usually Newton's method is used to solve (2.6) and to find minimizers $\overline{\mathbf{x}}$. Unfortunately, the scaling of the objective $J_\mu(\mathbf{x})$ becomes poorer and poorer as $\mu \to 0$. The extreme behavior of the barrier function close to the boundary of the feasible set translates to ill-conditioning in the barrier Hessian $\nabla^2_{\mathbf{x}} B_\mu(\mathbf{x})$. As a consequence the quadratic Taylor series approximation, on which Newton-like methods are based, does not reflect the behavior of the original function except in a small neighbourhood of $\overline{\mathbf{x}}$. This fact was one of the major motivations for the downfall of barrier methods before 1984, since it e.g. causes poor numerical performance of unconstrained optimization methods ($\mathcal{E} = \emptyset$). Fortunately Newton's method (in a carefully implemented algorithm, see FORSGREN, GILL, AND WRIGHT [66]) is insensitive to this poor scaling.

The true reason for the inefficiency of classical barrier methods is another one. Unfortunately, it is often not possible to take a full Newton step, because this step would move the current iterate out of the feasible region, especially when the current iterate is very close to a minimizer of $B_\mu(\mathbf{x})$ with a fixed μ. Suppose the current iterate is the minimizer $\overline{\mathbf{x}}_\mu$ of (2.5) with a fixed μ and the barrier parameter μ is now reduced to $\hat{\mu}$ with $\mu > \hat{\mu}$. If the ratio $\mu/\hat{\mu}$ exceeds a certain factor and the next Newton step is computed with respect to the new barrier parameter $\hat{\mu}$, a full Newton step will move the iterate to a significant infeasible point.

There are several remedies to overcome these poor steps that occur after a reduction of the barrier parameter, but the best one is to use *primal-dual* interior methods. In primal-dual methods we treat the primal variables and the dual variables (the Lagrangian multipliers of the problem) independently. In this spirit we now create an independent variable $\boldsymbol{\lambda}_{\mathcal{I}}$ of multipliers for the inequality constraints from the relation $\boldsymbol{\lambda}_{\mathcal{I}} = -\mu \mathbf{C}_{\mathcal{I}}(\mathbf{x})^{-1}\mathbf{e}$. Furthermore, if we consider $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{\mathcal{I}}, \boldsymbol{\lambda}_{\mathcal{E}})$ and $\mathbf{c}(\mathbf{x}) = (\mathbf{c}_{\mathcal{I}}(\mathbf{x}), \mathbf{c}_{\mathcal{E}}(\mathbf{x}))$, we can rewrite (2.6) as a system in the primal variables $\mathbf{x}$ and the dual variables $\boldsymbol{\lambda}$:

$$\nabla J(\mathbf{x}) + \nabla \mathbf{c}(\mathbf{x})^T \boldsymbol{\lambda} = \mathbf{0}, \tag{2.7a}$$

$$\mathbf{C}_{\mathcal{I}}(\mathbf{x})\boldsymbol{\lambda}_{\mathcal{I}} + \mu \mathbf{e} = \mathbf{0}, \tag{2.7b}$$

$$\mathbf{c}_{\mathcal{E}}(\mathbf{x}) = \mathbf{0}. \tag{2.7c}$$

The second equation (2.7b) can be interpreted as the *perturbed complementarity condition* for the inequality constraints in the KKT conditions for (2.4). The success of primal–dual methods is now partly due to their effectiveness at following the central path, especially in steps where the barrier parameter is reduced.

The left-hand-side of (2.7) defines a function $F_\mu(\mathbf{x}, \boldsymbol{\lambda})$. Instead of minimizing (2.5) for $\mu \to 0$, we look for solutions of $F_\mu(\mathbf{x}, \boldsymbol{\lambda}) = 0$ for $\mu \to 0$. For a fixed μ (2.7) can be solved, e.g., using a modified Newton-type method such that $\mathbf{x}$ and $\boldsymbol{\lambda}_{\mathcal{I}}$ fulfill the inequality constraints $\mathbf{c}_{\mathcal{I}}(\mathbf{x}) \leq 0$ and $\boldsymbol{\lambda}_{\mathcal{I}} \geq 0$ strictly. The Newton direction $(\triangle\mathbf{x}, \triangle\boldsymbol{\lambda})$ of such a method is defined as the solution of $\nabla F_\mu(\mathbf{x}, \boldsymbol{\lambda})(\triangle\mathbf{x}, \triangle\boldsymbol{\lambda}) = -F_\mu(\mathbf{x}, \boldsymbol{\lambda})$:

$$\begin{pmatrix} \nabla^2 \mathbf{H} & -\nabla \mathbf{c}_{\mathcal{I}}^T & \nabla \mathbf{c}_{\mathcal{E}}^T \\ \boldsymbol{\Lambda}_{\mathcal{I}} \nabla \mathbf{c}_{\mathcal{I}} & \mathbf{C}_{\mathcal{I}} & \mathbf{0} \\ \nabla \mathbf{c}_{\mathcal{E}} & \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \triangle\mathbf{x} \\ \triangle\boldsymbol{\lambda}_{\mathcal{I}} \\ \triangle\boldsymbol{\lambda}_{\mathcal{E}} \end{pmatrix} = - \begin{pmatrix} \nabla J + \nabla \mathbf{c}^T \boldsymbol{\lambda} \\ \mathbf{C}_{\mathcal{I}}\boldsymbol{\lambda}_{\mathcal{I}} + \mu \mathbf{e} \\ \mathbf{c}_{\mathcal{E}} \end{pmatrix}, \tag{2.8}$$

where $\mathbf{\Lambda}_{\mathcal{I}} = \text{diag}(\boldsymbol{\lambda}_i, i \in \mathcal{I})$, $\mathbf{H}(\mathbf{x}, \boldsymbol{\lambda})$ denotes the Hessian of the Lagrangian of (2.4) and all arguments in (2.8) are omitted.

Notes and remarks for subsection 2.1.3

- The computational costs of primal–dual interior–point methods are dominated by the cost of solving the linear system (2.8). So an efficient solver with efficient preconditioning to (2.8) is desirable. We refer to Chapter 6 for an example of the construction of an optimal solver. For a comprehensive review over numerical solutions of saddle point problems we refer to Benzi, Golub and Liesen [23].
- For more information and details of implementations of primal–dual interior methods we refer to Byrd, Hribar and Nocedal [49] and Wächter and Biegler [147]. Where the algorithm *Ipopt*, described in the latter paper, is actually used in this work (see Section 5.6).

2.2 The Finite Element Method

Topology optimization problems usually contain partial differential equations, describing the state equation of some physical equilibrium. In optimization problems related to structural mechanics these equations are the (linear) elasticity equations, see Section 2.4. In case of electromagnetics the state is described by the Maxwell equations. There are several methods to compute approximations to the solutions of the partial differential equations, where each of them has its specific area of application. The most used methods are the following:

- The *Boundary Element Method*, see, e.g. Chen and Zhou [52], Steinbach [134], Sauter and Schwab [114], or Schatz, Thomée and Wendland [115].
- The *Finite Difference Method*, see, e.g. Grossmann and Roos [70] or Thomas [144].
- The *Finite Element Method*, see, e.g. Braess [32], Brenner and Scott [38], Ciarlet [57], Jung and Langer [80] or Zienkiewics [153].
- The *Finite Volume Method*, see, e.g. Grossmann and Roos [70] or Heinrich [75]

In this work only the method of finite elements is used, as it is standard in the field of computational solid mechanics. So in the following we will focus on the finite element method and introduce it with use of the two probably most common examples in the literature: the Poisson equation and the Stokes' problem.

But first of all we have to introduce some function spaces. In the following, L_p denotes the L_p spaces ($1 \leq p \leq \infty$) equipped with the norm $\|\cdot\|_{L_p}$, and H^k ($0 \leq k$) denotes the Sobolev spaces equipped with the norm $\|\cdot\|_{H^k}$. For instance the space $H^1(\Omega)$, $\Omega \subset \mathbb{R}^d$ is defined as

$$H^1(\Omega) = \left\{ u \in L_2(\Omega) \;\middle|\; \nabla u \in L_2(\Omega; \mathbb{R}^d) \right\}$$

and the space $H_0^1(\Omega)$ is defined as

$$H_0^1(\Omega) = \left\{ u \in H^1(\Omega) \;\middle|\; u = 0 \text{ on } \partial\Omega \right\}.$$

For a more details about Sobolov spaces we refer to Adams [3].

2.2.1 A Model Elliptic Boundary Value Problem

We start with the most elementary partial differential equation, the Poisson equation with homogeneous Dirichlet boundary conditions, as a first model problem:

$$\begin{aligned} -\triangle u(\mathbf{x}) &= b(\mathbf{x}), && \mathbf{x} \in \Omega, \\ u(\mathbf{x}) &= 0, && \mathbf{x} \in \partial\Omega, \end{aligned} \tag{2.9}$$

where $\Omega \subset \mathbb{R}^2$ is a bounded domain with a sufficiently smooth boundary. Let us further assume that all appearing functions are sufficiently smooth. The starting point of the finite element method is the *weak* or *variational* form of the equations (2.9). After multiplying the PDE with a test-function v and using Gauss theorem we are looking for a solution $u \in V = H_0^1(\Omega)$ that fulfills the variational problem

$$\int_\Omega \nabla u(\mathbf{x})^T \nabla v(\mathbf{x})\ d\mathbf{x} = \int_\Omega b(\mathbf{x}) v(\mathbf{x})\ d\mathbf{x} \qquad \forall\ v \in V. \tag{2.10}$$

The relationship (2.10) is called variational because the function v is allowed to vary arbitrarily in V. For a more abstract form we use the abbreviations $a(u, v)$ for the left-hand side and $f(v)$ for the right-hand side of (2.10), respectively: Find $u \in V$ such that

$$a(u, v) = f(v), \qquad \forall\ v \in V, \tag{2.11}$$

where the $f(v)$ is an element of V^*, the dual space of V. By means of the Theorem of *Lax-Milgram* there exists a unique solution $u \in V$ for the variational problem (2.11). A function u is called a *weak* solution to the problem (2.9) if it satisfies the corresponding variational problem (2.11).

Theorem 2.4 (Lax-Milgram)**.** *Given a Hilbert space V, a continuous linear functional $f(\cdot) \in V^*$ and a continuous (bounded) and elliptic (coercive) bilinear form $a(\cdot, \cdot) : V \times V \to \mathbb{R}$, i.e. that there exist constants $c_1, c_2 > 0$ such that*

$$|a(u, v)| \le c_1 \|u\|_V \|v\|_V, \qquad \forall\ u, v \in V,$$

and

$$a(v, v) \ge c_2 \|v\|_V^2, \qquad \forall\ v \in V.$$

Then there exists a unique $u \in V$ such that

$$a(u, v) = f(v), \qquad \forall\ v \in V.$$

Proof. See e.g. Brenner and Scott [38] or Braess [32]. □

Notes and remarks for subsection 2.2.1

For a *symmetric* and elliptic bilinearform $a(\cdot, \cdot)$ the variational problem (2.11) is equivalent to the related minimum problem:

$$J(u) := \frac{1}{2} a(u, u) - f(u) \ \to \ \min_{u \in V}.$$

The sufficient and necessary optimality condition for the above quadratic functional $J(u)$ is the equation (2.11).

2.2.2 A Model Mixed Boundary Value Problem

Saddle point problems appear in optimization problems (e.g. like in Section 3.1) and in the context of *mixed* variational problems, which have more than one approximation space. Although the Poisson problem can also be written in a mixed way (cf. Braess [32]), we will demonstrate this formulation by means of the Stokes' problem: Find the velocity $\mathbf{u}$ and the pressure p such that

$$\begin{aligned} -\nu\triangle\mathbf{u}(\mathbf{x}) \;-\; \nabla p(\mathbf{x}) &= \mathbf{b}(\mathbf{x}), && \mathbf{x}\in\Omega, \\ \operatorname{div}\mathbf{u}(\mathbf{x}) &= 0, && \mathbf{x}\in\Omega, \\ \mathbf{u}(\mathbf{x}) &= 0, && \mathbf{x}\in\partial\Omega. \end{aligned} \tag{2.12}$$

The scalar parameter ν describes the viscosity. As for the Poisson problem in the previous section we define appropriate function spaces for the weak formulation of (2.12). Note, that the pressure p in (2.12) is only defined up to a additive constant. With the Hilbert spaces $V = H_0^1(\Omega)^2$ and $Q = L_{2,0}$ we search for $\mathbf{u}\in V$ and $p\in Q$ such that

$$\begin{aligned} \nu\left(\nabla\mathbf{u},\nabla\mathbf{v}\right)_0 \;+\; \left(\operatorname{div}\mathbf{v},p\right)_0 &= \left(\mathbf{b},\mathbf{v}\right)_0, && \forall\ \mathbf{v}\in V, \\ \left(\operatorname{div}\mathbf{u},q\right)_0 &= 0, && \forall\ q\in Q. \end{aligned} \tag{2.13}$$

Again we define bilinear forms $a(\cdot,\cdot): V\times V\to\mathbb{R}$ and $b(\cdot,\cdot): V\times Q\to\mathbb{R}$ based on the weak form (2.13):

$$\begin{aligned} a(\mathbf{u},\mathbf{v}) = \nu\int_\Omega \nabla\mathbf{u}\cdot\nabla\mathbf{v}\;d\mathbf{x} &= \left(\nabla\mathbf{u},\nabla\mathbf{v}\right)_0, \\ b(\mathbf{u},q) = \int_\Omega \operatorname{div}\mathbf{u}\;q\;d\mathbf{x} &= \left(\operatorname{div}\mathbf{u},q\right)_0. \end{aligned}$$

Together with the linear form $f(\cdot): V\to\mathbb{R}$

$$f(\mathbf{v}) = \int_\Omega \mathbf{b}\cdot\mathbf{v}\;d\mathbf{x} \;=\; \left(\mathbf{b},\mathbf{v}\right)_0,$$

we write the mixed variational problem in the following way: Find $\mathbf{u}\in V$ and $p\in Q$ such that

$$\begin{aligned} a(\mathbf{u},\mathbf{v}) \;+\; b(\mathbf{v},p) &= f(\mathbf{v}), && \forall\ \mathbf{v}\in V, \\ b(\mathbf{u},q) &= 0, && \forall\ q\in Q. \end{aligned} \tag{2.14}$$

Similar to the conditions in the Lax-Milgram Theorem 2.4 we assume that the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are continuous:

$$\begin{aligned} |a(\mathbf{u},\mathbf{v})| &\le c_a\|\mathbf{u}\|_V\|\mathbf{v}\|_V, && \forall\ \mathbf{u},\mathbf{v}\in V, \\ |b(\mathbf{u},q)| &\le c_b\|\mathbf{u}\|_V\|\mathbf{q}\|_Q, && \forall\ \mathbf{u}\in V,\ q\in Q, \end{aligned} \tag{2.15}$$

with constants $c_a, c_b > 0$. The discussion of coercivity is a bit more sophisticated. For this purpose we first define the kernel V_0 of the bilinearform $b(\cdot,\cdot)$ by

$$V_0 = \left\{\mathbf{v}\in V \mid b(\mathbf{v},q) = 0,\ \forall\ q\in Q\right\}.$$

Now, due to Brezzi's theory (see Brezzi [39] and Brezzi and Fortin [40]), we can list two conditions on $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ to obtain stability for (2.14). Firstly, the so called V_0-ellipticity or kernel ellipticity of $a(\cdot,\cdot)$:

$$a(\mathbf{u},\mathbf{u}) \ge \alpha\|\mathbf{u}\|_V^2, \qquad \forall\ \mathbf{u}\in V_0, \tag{2.16}$$

and, secondly, the *inf-sub* or the *LBB* condition:

$$\inf_{q\in Q}\sup_{\mathbf{v}\in V}\frac{b(\mathbf{v},q)}{\|\mathbf{v}\|_V\|q\|_Q}\geq\beta, \tag{2.17}$$

with constants $\alpha,\beta>0$. Finally we state the following theorem:

Theorem 2.5. *Suppose that the conditions* (2.15)*,* (2.16) *and* (2.17) *are fulfilled. Then there exists a unique solution* $(\mathbf{u},p)\in V\times Q$ *to the variational problem* (2.14)*.*

Proof. See e.g. Brezzi and Fortin [40] or Brenner and Scott [38]. □

Notes and remarks for subsection 2.2.2

- The first component of a solution $(\mathbf{u},p)\in V\times Q$ to (2.14) is also a solution to the related minimum problem:

$$\begin{aligned} J(\mathbf{u}) \;&=\; \frac{1}{2}a(\mathbf{u},\mathbf{u})-f(\mathbf{u}) \;\rightarrow\; \min_{\mathbf{u}\in V} \\ \text{subject to}\qquad & b(\mathbf{u},q) \;=\; 0, \qquad\qquad \forall\, q\in Q. \end{aligned}$$

  If we consider the Lagrange functional $\mathcal{L}(\mathbf{u},q) := J(\mathbf{u})+b(\mathbf{u},q)$ we see that every solution $(\mathbf{u},p)$ to (2.14) fulfills the saddle point property:

$$\mathcal{L}(\mathbf{u},q)\leq\mathcal{L}(\mathbf{u},p)\leq\mathcal{L}(\mathbf{v},q),\qquad\forall\,(\mathbf{v},q)\in V\times Q.$$

- For a rigorous analysis of the theory and construction of mixed finite element methods we refer to Brezzi and Fortin [40].

2.2.3 Finite Element Discretization

The finite element method is a technique to compute an discrete approximation to the solution of variational forms like (2.11) or (2.14). It is a special Galerkin projection method that provides a natural procedure to compute numerical solutions to these problems. We replace the infinite dimensional space V by some finite dimensional subspace V^h. Here h denotes the actual discretization parameter, indicating that there should be convergence to the continuous problem for $h\rightarrow 0$. Now we do not search for a solution $u\in V$, but a approximate solution $\tilde{u}\in V^h$. The corresponding finite dimensional problem is the reduction of the variational ones: Find $\tilde{u}\in V^h$ such that

$$a(\tilde{u},\tilde{v})=f(\tilde{v}),\qquad\forall\,\tilde{v}\in V^h. \tag{2.18}$$

In order to construct the finite dimensional spaces V_h we partition the domain Ω into geometrical primitives, $\Omega=\bigcup_{i=1}^n\overline{\tau}_i$, like triangles and quadrilaterals in 2D, tetrahedrons, hexahedron, and prisms in 3D. For a detailed description of a regular triangulation $\mathcal{T}_h=\{\tau_i\mid i=1,\ldots,n\}$ we refer to Ciarlet [57]. Using a regular triangulation we introduce the finite elements by defining shape functions (usually polynomials) with local support over the geometrical elements. Thus, the spaces V_h contain usually piecewise linear, bilinear or quadratic functions. But in the last two decades it became more and more popular to use polynomials with higher degree, the so called *hp*- or *p*-version, see e.g. Schwab [119]. This increasing interest in high

order finite element methods is due to the higher convergence rates and their robustness with respect to locking effects and element distortion.

The discretization parameter h is usually related to the mesh size of a given triangulation of Ω into n finite elements. We have $n = O(h^{-d})$. Further we assume that whenever mesh refinement is performed, it is done in such a way that $V^h \supset V^H$ if $h < H$.

To prepare the FEM for computing we choose a basis $\mathbf{\Phi} := (\phi_1, \ldots, \phi_N)^T$ for the finite element space V^h, with $N = \dim V^h$. Then we can write (2.18) equivalently as

$$a(\tilde{u}, \phi_i) = f(\phi_i), \qquad i = 1, \ldots, N.$$

Using the basis $\mathbf{\Phi}$ we can write

$$\tilde{u} = {\mathbf{u}^h}^T \mathbf{\Phi} = \sum_{i=1}^{N} {u_i^h}^T \phi_i$$

with a coefficient vector $\mathbf{u}^h \in \mathbb{R}^N$. This representation leads to the following system of equations:

$$\sum_{j=1}^{N} a(\phi_j, \phi_i) u_j = f(\phi_i), \qquad i = 1, \ldots, N,$$

which we can write in matrix-vector notation as

$$\mathbf{K}\mathbf{u}^h = \mathbf{f}^h. \tag{2.19}$$

The *stiffness matrix* $\mathbf{K}$ is defined as $K_{ij} = a(\phi_i, \phi_j)$ and the *load vector* $\mathbf{f}^h$ as $f_i^h = f(\phi_i)$ for $i = 1, \ldots, N$.

Notes and remarks for subsection 2.2.3

- The names stiffness matrix and load vector have their origin in the evolutionary history of the finite element method in computational mechanics.
- The finite element discretization of mixed problems, like in Subsection 2.2.2, needs two finite elements spaces, namely $V^h \subset V$ and $Q^h \subset Q$. But different to elliptic problems, like in Subsection 2.2.1, additional conditions have to be fulfilled. Those are the discrete counterparts to the kernel ellipticity (2.16) and to the LBB condition (2.17). Again we refer the reader to Brezzi and Fortin [40].

2.3 Iterative Solvers for Linear Systems

In this section we give a introduction and a brief discussion of efficient solution methods for the linear systems, like (2.19), arising from finite element discretizations. We will restrict ourselves to the case of elliptic problems. Linear systems resulting from saddle point problems lead to indefinite systems (cf. Benzi, Golub, and Liesen [23]) and for sake of space we will omit a description of related solution techniques. Some information about solving saddle point problems will be given in Section 3.2 and in Section 6.2.

The method of finite elements with a small mesh size h leads to very large systems ($n \geq 10^6$). But the system matrices show some structure. Due to the local support of the shape

functions there are only a few non-zero entries per row in $\mathbf{K}$, i.e. the system matrix is sparse. Since the stiffness matrix $\mathbf{K}$ is defined on the basis of the bilinear form $a(\cdot,\cdot)$, properties like coercivity and symmetry are transferred from the bilinear form to the system matrix $\mathbf{K}$. In the previous subsections we did not assume the symmetry of the bilinear form $a(\cdot,\cdot)$. Nevertheless, let us assume that $\mathbf{K}$ is symmetric, since in many applications this fact holds for the bilinearform. Furthermore, the coercivity of $a(\cdot,\cdot)$ yields that $\mathbf{K}$ is positive definite. This means that the smallest eigenvalues stays positive, but still the spectral condition number

$$\kappa(\mathbf{K}) = \frac{\lambda_{\max}(\mathbf{K})}{\lambda_{\min}(\mathbf{K})}$$

becomes very large as $h \to 0$. In fact, the condition number usually behaves like $\mathcal{O}(h^{-2m})$, as $h \to 0$, where $2m$ denotes the order of an elliptic BVP. So we have to deal with large scale linear systems, with a symmetric, positive definite, and sparse system matrix, that has a large condition number.

Despite the sparsity pattern of $\mathbf{K}$, direct methods are inappropriate due to their large memory consumption, especially for 3D problems. When performing a Gauss elimination the number of non-zero entries (the fill-in) is increasing very fast, which results in an increasing complexity. Hence, iterative methods are usually the methods of choice for large problems. They are able to exploit the sparsity structure to a high extent and the needed matrix-vector applications of the system matrix can be accomplished with a complexity that is proportional to the number of unknowns. But still, for 2D problems with moderate size direct methods are an alternative to iterative methods. Direct methods for sparse linear systems carry out a Gauss or Cholesky factorization of the system matrix. To realize this factorization efficiently, a renumbering of rows and columns is performed in order to minimize the fill-in.

For a detailed discussion of iterative methods we refer to e.g. Axelsson [12], Hackbusch [71], and Meurant [91].

2.3.1 The Richardson Iteration and Preconditioning

We will now present the basic procedure of an iterative method by means of one of the simplest methods, the Richardson iteration. A nice motivation for solving the equation $\mathbf{Ku} = \mathbf{f}$, $\mathbf{K}$ spd, iteratively is that the solution of the equation describes the minimum of the quadratic function

$$q(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T\mathbf{K}\mathbf{u} - \mathbf{f}^T\mathbf{u}, \tag{2.20}$$

since $\mathbf{K}$ is assumed to be symmetric and positive definite. The simplest method is the steepest descent method with a fixed steplength, also known as the *Richardson* iteration:

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \tau\big(\mathbf{f} - \mathbf{K}\mathbf{u}^k\big), \tag{2.21}$$

with a positive damping parameter τ. Let us consider the spectral equivalence inequalities

$$\gamma_1\langle\mathbf{v},\mathbf{v}\rangle \le \langle\mathbf{K}\mathbf{v},\mathbf{v}\rangle \le \gamma_2\langle\mathbf{v},\mathbf{v}\rangle, \qquad \forall\ \mathbf{v} \in \mathbb{R}^n$$

with the spectral equivalence constants $\gamma_1, \gamma_2 > 0$. For the optimal choice $\tau = 2/(\gamma_1 + \gamma_2)$ we get the best possible convergence rate $(\gamma_2 - \gamma_1)/(\gamma_2 + \gamma_1)$. The convergence speed of the iteration method (2.21) depends on the condition number $\kappa(\mathbf{K}) = \mathcal{O}(h^{-2})$. Hence the number of iterations is proportional to the condition number and is therefore growing too

quickly. An idea to reduce the condition number, and hence the number of iterations, is to use a symmetric and positive definite preconditioner $\mathbf{C}$. Furthermore, let us assume that the preconditioner $\mathbf{C}$ is spectral equivalent to the system matrix $\mathbf{K}$, where we use the same notation γ_1, γ_2 for simplicity:

$$\gamma_1\langle \mathbf{Cv}, \mathbf{v}\rangle \leq \langle \mathbf{Kv}, \mathbf{v}\rangle \leq \gamma_2\langle \mathbf{Cv}, \mathbf{v}\rangle, \qquad \forall\ \mathbf{v} \in \mathbb{R}^n. \tag{2.22}$$

A preconditioned version of the Richardson method (2.21) is presented in Algorithm 2.1. In case of the preconditioned Richardson method the convergence factor is bounded by

$$\frac{\kappa(\mathbf{C}^{-1}\mathbf{K}) - 1}{\kappa(\mathbf{C}^{-1}\mathbf{K}) + 1}. \tag{2.23}$$

There are two requirements that we demand on a preconditioner:

Algorithm 2.1 Preconditioned Richardson iteration

Choose a damping parameter τ, $0 < \tau < \frac{2}{\gamma_2}$.
Choose a relative error bound $\varepsilon > 0$.
Initialize start value $\mathbf{u}^0$.
$k = 0$;

while not converged **do**
  $\mathbf{u}^{k+1} = \mathbf{u}^k + \tau\mathbf{C}^{-1}(\mathbf{f} - \mathbf{K}\mathbf{u}^k)$;
  $k = k + 1$;
end while

Remark 2.3 (Properties of a preconditioner)**.**

a) The preconditioning operation $\mathbf{C}^{-1}\mathbf{r}$ *or the preconditioning system*

$$\mathbf{Cw} = \mathbf{r} \tag{2.24}$$

should be efficiently evaluable, since it is applied at each iteration. The arithmetic costs of applying $\mathbf{C}^{-1}$ *should be about the same as applying* $\mathbf{K}$*. Also the memory requirement for the preconditioner should be about the same as the system matrix.*

b) The preconditioner should be constructed such that the quotient $\frac{\gamma_2}{\gamma_1}$ *is close to 1, where* γ_1 *and* γ_2 *are the spectral equivalence constants from* (2.22)*. In other words the condition number of* $\kappa(\mathbf{C}^{-1}\mathbf{K})$ *should be as close to 1 as possible, but at least independent of the mesh parameter* h*, i.e.,* $\kappa(\mathbf{C}^{-1}\mathbf{K}) = \mathcal{O}(1)$ *as* $h \to 0$*.*

For $\mathbf{C} = \mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix, we perfectly fulfill the first requirement, but not the second one. On the other hand, for $\mathbf{C} = \mathbf{K}$ we satisfy the condition $\kappa(\mathbf{C}^{-1}\mathbf{K}) = 1$ and would finish only after 1 iteration, but requirement *a*) is not fulfilled at all. So usually a good tradeoff between the two qualifications is needed for an optimal preconditioner.

Notes and remarks for Subsection 2.3.1

The disadvantages of the Richardson iteration are that we need the constants γ_1 and γ_2 close to $\lambda_{\min}(\mathbf{C}^{-1}\mathbf{K})$ and $\lambda_{\max}(\mathbf{C}^{-1}\mathbf{K})$ and that the number of iterations is proportional to $\kappa(\mathbf{C}^{-1}\mathbf{K})$. A slightly improved method is the gradient method where we don't need γ_1 and γ_2. Here the optimal damping parameter τ is deduced from the sufficient and necessary optimality condition for the quadratic function (2.20)

$$\frac{dq\big(\mathbf{u}^k + \tau\mathbf{w}^k\big)}{d\tau} = 0,$$

where $\mathbf{w}^k$ is the preconditioned residuum of the k-th iteration $\mathbf{w}^k = \mathbf{C}^{-1}\mathbf{r}^k$. We list the method in Algorithm 2.2.

Algorithm 2.2 Preconditioned gradient method

Choose a relative error bound $\varepsilon > 0$.
Initialize start value $\mathbf{u}_0$.
$k = 0$;

while not converged **do**
  $\mathbf{r}^k = \mathbf{f} - \mathbf{K}\mathbf{u}^k$;
  $\mathbf{w}^k = \mathbf{C}^{-1}\mathbf{r}^k$;
  $\tau^k = \frac{\langle \mathbf{r}^k, \mathbf{w}^k \rangle}{\langle \mathbf{w}^k, \mathbf{K}\mathbf{w}^k \rangle}$;
  $\mathbf{u}^{k+1} = \mathbf{u}^k + \tau^k \mathbf{w}^k$;
  $k = k + 1$;
end while

2.3.2 The Conjugate Gradient Method

The *conjugate gradient* (CG) method was developed by Hestenes and Stiefel in 1952. The breakthrough was accomplished in 1971, when first simple preconditioning techniques became available. Since then the *CG* method has belonged to the most efficient methods for solving sparse symmetric positive definite linear systems. The CG method was developed by adapting the gradient method. Firstly, we keep the advantage that we don't need the spectral equivalence constants γ_1 and γ_2. Secondly, we use orthogonal search directions $\mathbf{s}^k$ instead of the preconditioned residuum $\mathbf{w}^k$, which leads to a significant speed up. A preconditioned version can be found in Algorithm 2.3.

The method of conjugate gradients yields for exact computation a solution to (2.19) in at most n iterations. But this fact should not be overestimated. Since we usually treat large scale problems with a high number of unknowns and because of the influence of numerical rounding errors, we use the CG method until we get a sufficiently good approximation of the solution. It is much more important that we get good approximations after a number of iterations that is much smaller than n. The error in the k^{th} iteration step of the preconditioned CG method in the $\mathbf{K}$-energy norm is bounded by

$$\|\mathbf{u}^k - \mathbf{u}^*\|_{\mathbf{K}} \leq \eta_k \|\mathbf{u}^0 - \mathbf{u}^*\|_{\mathbf{K}}$$

Algorithm 2.3 Preconditioned conjugate gradient iteration

Choose a relative error bound $\varepsilon > 0$.
Initialize start value $\mathbf{u}^0$;
$\mathbf{r}^0 = \mathbf{f} - \mathbf{K}\mathbf{u}^0$;
$\mathbf{w}^0 = \mathbf{C}^{-1}\mathbf{r}^0$;
$\mathbf{s}_0 = \mathbf{w}_0$;
$k = 0$;

while not converged **do**
$\gamma^k = \langle \mathbf{s}^k, \mathbf{r}^k \rangle$;
$\alpha^k = \frac{\gamma^k}{\langle \mathbf{K}\mathbf{w}^k, \mathbf{w}^k \rangle}$;
$\mathbf{u}^{k+1} = \mathbf{u}^k + \alpha^k \mathbf{w}^k$;
$\mathbf{r}^{k+1} = \mathbf{r}^k - \alpha^k \mathbf{K}\mathbf{w}^k$;
$\mathbf{s}^{k+1} = \mathbf{C}^{-1}\mathbf{r}^{k+1}$;
$\beta^k = \frac{\gamma^k}{\langle \mathbf{s}^{k+1}, \mathbf{r}^k \rangle}$;
$\mathbf{w}^{k+1} = \mathbf{s}^{k+1} + \beta^k \mathbf{w}^k$;
$k = k + 1$;
end while

with

$$\eta_k = \frac{2c^k}{1 + c^{2k}} \qquad \text{and} \qquad c = \frac{\sqrt{\kappa(\mathbf{C}^{-1}\mathbf{K})} - 1}{\sqrt{\kappa(\mathbf{C}^{-1}\mathbf{K})} + 1}$$

(see e.g. JUNG AND LANGER [80]). Here let $\mathbf{u}^*$ denote the exact solution of (2.19). In comparison to the convergence factor of the Richardson iteration (2.23) an acceleration by the square root is given.

Notes and remarks for subsection 2.3.2

- In fact, the CG method belongs to the class of *Krylov subspace projection* methods. A Krylov subspace method is a method where we minimize the residuum in the Krylov subspace of the k^{th} iteration

$$\mathcal{K}_k(\mathbf{K}, \mathbf{r}^0) = \operatorname{span}\left\{\mathbf{r}^0, \mathbf{K}\mathbf{r}^0, \mathbf{K}^k\mathbf{r}^0, \ldots, \mathbf{K}^{k-1}\mathbf{r}^0\right\},$$

  where $\mathbf{r}^0 = \mathbf{f} - \mathbf{K}\mathbf{u}^0$. From an approximation theory point of view we see that the approximations obtained are of the form

$$\mathbf{K}^{-1}\mathbf{f} \approx \mathbf{u}^k = \mathbf{u}^0 + p_{k-1}(\mathbf{K})\mathbf{r}^0,$$

  in which p_{k-1} is a certain polynomial of degree $k-1$. In other words, $\mathbf{K}^{-1}\mathbf{f}$ is approximated by $p_{k-1}(\mathbf{K})\mathbf{f}$, if we assume $\mathbf{u}^0 = \mathbf{0}$. In case of the preconditioned CG method the generating matrix is $\mathbf{C}^{-1/2}\mathbf{K}\mathbf{C}^{-1/2}$. For more information about Krylov subspace methods we refer to e.g. SAAD [111].

- In comparison to the Richardson method, the CG method as well as the gradient method is a non-linear iteration scheme.

2.3.3 The Multigrid Method

The multigrid method provides an optimal order algorithm for solving linear systems arising from finite element discretizations, as well as other discretization techniques. The number of iterations of the previously mentioned iteration methods are increasing as $h \to 0$ if no proper preconditioning is applied. When using multigrid methods we get numbers of iterations that are independent from the mesh parameter h. In other words the convergence speed does not deteriorate when the discretization is refined, whereas classical iterative methods slow down for decreasing mesh size. A fundamental attribute of the multigrid method is that it is working on a hierarchy of meshes and related discretizations of a boundary value problem. We recommend e.g. the books BRAMBLE [35] and HACKBUSCH [73] for detailed reading.

The multigrid method has two main features: smoothing on a fine grid and error correction on a coarser grid. The starting point for this idea is the observation that classical iteration methods have smoothing properties, i.e. they remove the high oscillating parts of the error very fast. The smooth part of the error can already be represented and corrected on coarser grids. Hence, combining these two approaches makes the multigrid method to the most efficient solvers. For a short introduction let us consider a hierarchy of l meshes (e.g. like in

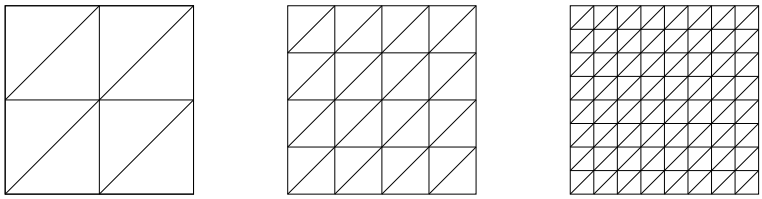

Figure 2.1: A hierarchy of 3 meshes: $\mathcal{T}_0 \subset \mathcal{T}_1 \subset \mathcal{T}_2$.

Figure 2.1)

$$\mathcal{T}_0 \subset \mathcal{T}_1 \subset \ldots \subset \mathcal{T}_l,$$

with corresponding finite element spaces $V_0 \subset \ldots \subset V_l$, mesh sizes $h_0 \geq \ldots \geq h_l$, and number of unknowns $n_0 \leq \ldots \leq n_l$. One of the ingredients of a successful multigrid method are the *intergrid transfer* operators:

Definition 2.5 (Intergrid Operators)**.** *The coarse-to-fine operator*

$$\mathcal{I}_{l-1}^{l} \; : \; V_{l-1} \to V_l$$

is called the (prolongation) *operator and the the fine-to-coarse operator*

$$\mathcal{I}_{l}^{l-1} \; : \; V_l \to V_{l-1}$$

is called the (restriction) *operator.*

Remark 2.4. *If we have a sequence $V_0 \subset \ldots \subset V_l$ of spaces, the prolongation operator $\mathcal{I}_{l-1}^{l}$ can be taken to be the natural injection. In other words, $\mathcal{I}_{l-1}^{l} v = v, \quad \forall\, v \in V_{l-1}$. Then the restriction operator is defined to be the adjoint of $\mathcal{I}_{l-1}^{l}$ with respect to $(\cdot,\cdot)_{l-1}$ and $(\cdot,\cdot)_l$ inner products. In other words, $(\mathcal{I}_{l}^{l-1} w, v)_{l-1} = (w, \mathcal{I}_{l-1}^{l} v)_l = (w, v)_l, \quad \forall v \in V_{l-1}, w \in V_l$.*

Algorithm 2.4 Two grid method

Choose a relative error bound $\varepsilon > 0$.
Choose a number ν_1 of pre-smoothing and a number ν_2 of post-smoothing steps.
Initialize start value $\mathbf{u}_h^{(0,0)}$.
$k = 0$;

while not converged **do**
/* Pre-smoothing: */
$\mathbf{u}_h^{(k,1)} = \mathcal{S}^{\nu_1}\mathbf{u}_h^{(k,0)}$;
/* Coarse grid correction: */
/* Defect calculation: */
$\mathbf{d}_h^k = \mathbf{f}_h - \mathbf{K}_h\mathbf{u}_h^{(k,1)}$;
/* Restriction onto coarse grid: */
$\mathbf{d}_H^k = \mathbf{I}_h^H\mathbf{d}_h^k$;
/* Solve coarse grid system: */
$\mathbf{K}_H\mathbf{w}_H^k = \mathbf{d}_H^k$;
/* Prolongation onto the fine grid: */
$\mathbf{w}_h^k = \mathbf{I}_H^h\mathbf{w}_H^k$;
/* Add coarse grid correction: */
$\mathbf{u}_h^{(k,2)} = \mathbf{u}_h^{(k,1)} + \mathbf{w}_h^k$;
/* Post-smoothing: */
$\mathbf{u}_h^{(k+1,0)} = \mathcal{S}^{\nu_2}\mathbf{u}_h^{(k,2)}$;
$k = k + 1$;
end while

A proper choice of the intergrid operators influences the convergence speed considerably, and may even be necessary for convergence.

In addition to the intergrid operators we need an iteration method *(smoother)* for the smoothing iterations on the fine grids. For instance we choose the smoothing operator $\mathcal{S}$ to realize the Jacobi-relaxation with a damping parameter $\tau > 0$ (cf. Algorithm 6.1):

$$\mathbf{u} \longmapsto \mathcal{S}\mathbf{u} \;=\; \mathbf{u} - \tau(\mathbf{K}\mathbf{u} - \mathbf{f}).$$

Since it reduces the high frequency error components the smoothing operator $\mathcal{S}$ is an essential part in multigrid methods. Typically, a proper smoother for a problem takes the special structure of the system matrix into account. Beside the point Jacobi smoother also the point Gauß-Seidel (cf. Algorithm 6.2), the block Jacobi and block Gauß-Seidel smoother are suitable for a large class of finite element discretized problems. E.g. in Section 6.2 we will discuss a local patch smoother. Beside the smoothing operation we also need a coarse grid correction. Let us assume that we have the corresponding system matrices $\mathbf{K}_0, \ldots, \mathbf{K}_l$ and load vectors $\mathbf{f}_0, \ldots, \mathbf{f}_l$ for each level at hand. These can either be generated by assembling on each level or can be constructed by Galerkin's method, i.e.

$$\mathbf{K}_{l-1} = \mathbf{I}_l^{l-1}\mathbf{K}_l\mathbf{I}_{l-1}^l.$$

After these preliminaries we are ready to state a two level method, as in Algorithm 2.4, where we assume that $l = 1$ and we use the following notation $h_0 = 2h_1 = H$, $\mathbf{I}_0^1 = \mathbf{I}_H^h$, and $\mathbf{I}_1^0 = \mathbf{I}_h^H$

for better readability. The parameters ν_1 and ν_2 control the number of smoothing iterations. For benign problems like the Poisson equation (2.9) it usually does not pay off to use more than two smoothing steps. In the case of more complex problems, e.g. saddle point problems like (2.12), it can be necessary to use more smoothing iterations.

The restricted system onto the coarse grid is by far easier to solve than the one on the finer grid. When switching to a mesh from mesh size h to $2h$ by uniform refinement, the number of unknowns decreases about to a quarter. But still, the complexity of solving the coarse grid system may be regarded to high. The idea to advance from a two grid method to a multigrid method is now to repeat this procedure recursively. That is to coarsen the grid until the coarsest grid yields a sufficiently small system, that is easy to solve. The linear system on the coarsest grid is usually solved directly, e.g. by some Cholesky factorization. So, instead of solving the coarse grid system, one or two multigrid steps are called, resulting in a V-cycle or a W-cycle. The patterns in Fig. 2.2 will explain the naming, where $\circ$ denotes smoothing, $\bullet$ marks the solution of the system on the coarsest grid, $\searrow$ and $\nearrow$ stand for restriction and interpolation between the grids, respectively. In the early days the common choice was a

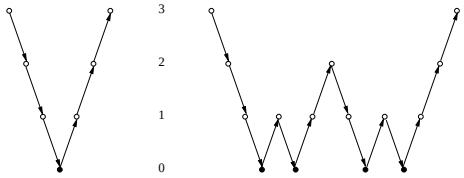

Figure 2.2: V-cycle and W-cycle on a hierarchy of 4 grids.

W-cycle to ensure that the error is not increasing too much when cycling between several grids. But most of the problems are so benign, that a V-cycle is more efficient. The following Algorithm 2.5 sketches the operations of the k^{th} multigrid iteration on level i with $1 \leq i \leq l$. For sake of readability we drop the iteration index k.

Notes and remarks for subsection 2.3.3

- The multigrid method also defines optimal preconditioners for the preconditioned CG method. As an a-priori preconditioner we choose the system matrix $\mathbf{K}$ and solve the preconditioning system (2.24) in each iteration of the CG method approximately by m iterations of the multigrid method. In this way a preconditioner is implicitly defined that fulfills the requirements of Remark 2.3. For more details we refer e.g. to Jung and Langer [79].

- The above coarsening idea can be used if we have a nested sequence of finite element spaces. This kind of multigrid approach is called *geometric* multigrid. But there are cases when geometric multigrid cannot be applied. For instance if there is no hierarchy of finite element spaces or if the system of the coarsest grid is still too large to be solved efficiently. Then *algebraic* multigrid methods are of special interest. They construct the matrix hierarchy and intergrid operators only by using the system matrix $\mathbf{K}$. An overview of algebraic multigrid methods can be found in, e.g., Reitzinger [105] and in the references therein.

Algorithm 2.5 One multigrid method iteration MGM($\mathbf{K}_i, \mathbf{u}_i, \mathbf{f}_i, i$)

Let μ describe the number of MGM calls per level i.
Let ν_1 and ν_2 denote the number of pre- and post-smoothing steps.
Initialize start value $\mathbf{u}_i^0 = \mathbf{u}_i$;

if $i == 0$ **then**
  Solve the coarsest grid system, i.e. $\mathbf{u}_0 = \mathbf{K}_0^{-1}\mathbf{f}_0$;
  return;
else
  /* Pre-smoothing: */
  $\mathbf{u}_i^1 = \mathcal{S}^{\nu_1}\mathbf{u}_i^0$;
  /* Coarse grid correction: */
  /* Defect calculation: */
  $\mathbf{d}_i = \mathbf{f}_i - \mathbf{K}_i\mathbf{u}_i^1$;
  /* Restriction onto coarse grid: */
  $\mathbf{d}_{i-1} = \mathbf{I}_i^{i-1}\mathbf{d}_i$;
  /* Recursively call MGM for coarse grid approximation: */
  $\mathbf{w}_{i-1} = \mathbf{0}$;
  for $j = 1, \ldots, \mu$ **do**
    MGM($\mathbf{K}_{i-1}, \mathbf{w}_{i-1}, \mathbf{d}_{i-1}, i-1$);
  end for
  /* Prolongation onto the fine grid: */
  $\mathbf{w}_i = \mathbf{I}_{i-1}^i\mathbf{w}_{i-1}$;
  /* Add coarse grid correction: */
  $\mathbf{u}_i^2 = \mathbf{u}_i^1 + \mathbf{w}_i$;
  /* Post-smoothing: */
  $\mathbf{u}_i^3 = \mathcal{S}^{\nu_2}\mathbf{u}_i^2$;
end if

- If all steps of the two grid method are assembled, Algorithm 2.4 can be written as

$$\mathbf{u}_h^{k+1} = \mathbf{M}_h\mathbf{u}_h^k + \mathbf{m}_h, \qquad \text{for } k = 1, 2, \ldots$$

  with

$$\mathbf{M}_h = \mathcal{S}^{\nu_1}\big(\mathbf{I} - \mathbf{I}_H^h\mathbf{K}_H^{-1}\mathbf{I}_h^H\mathbf{K}_h\big)\mathcal{S}^{\nu_2} \tag{2.25}$$

  and $\mathbf{m}_h = (\mathbf{I} - \mathbf{M}_h)\mathbf{K}_h^{-1}\mathbf{f}_h \in \mathbb{R}^{n_h}$.

2.4 Linear Elasticity

The topology optimization problems in this work treat problems from the field of solid mechanics. This means that the underlying partial differential equations describing the equilibrium of forces are the equations of linear elasticity. In this section we will give a short introduction to linear elasticity.

The deformation of elastic bodies under loading as well as the appearing stresses are usually determined using the method of finite elements. Most of the characteristical properties already show up using linear elasticity, i.e. under the assumption of small deformations. Also

the problems in this work are solved under the assumption of the linear theory. For problems with non-linear elasticity see e.g. BUHL, PEDERSON AND SIGMUND [42] and BENDSOE AND SIGMUND [22] and the references cited therein. Many materials allow only very small deformations (nearly incompressible materials). To avoid the appearing *locking* effect we need suitable variational mixed formulations. A rigorous analysis can be found in CIARLET [56] and in HAN AND REDDY [74].

2.4.1 A short insight to the theory of linear elasticity

In the theory of elasticity we consider the state of a body under loading. Of special interest are the displacements $\mathbf{u}(\mathbf{x})$, the strains $\boldsymbol{\varepsilon}(\mathbf{x})$, and the stresses $\boldsymbol{\sigma}(\mathbf{x})$ at a point $\mathbf{x}$ of the deformed body. We start with the assumption that the body without deformations and in stress free state occupies a domain $\Omega \subset \mathbb{R}^3$. The actual position of the body under loading is now described using a mapping $\mathbf{y} : \Omega \to \mathbb{R}^3$, which indicates the new position of the point $\mathbf{x} \in \Omega$. Moreover we write

$$\mathbf{y}(\mathbf{x}) = \mathbf{x} + \mathbf{u}(\mathbf{x})$$

with the *displacement* field $\mathbf{u}$. Let us now consider rigid body motions, in which the body is moved to a new position without any deformation. The displacement field alone does not provide enough information to see whether a body was deformed or not. To measure the deformation of a body we introduce a strain tensor $\boldsymbol{\eta}$ as

$$\boldsymbol{\eta}(\mathbf{u}) = \frac{1}{2}\left(\nabla\mathbf{u} + (\nabla\mathbf{u})^T + (\nabla\mathbf{u})^T(\nabla\mathbf{u})\right). \tag{2.26}$$

In the theory of linear elasticity we assume now that the deformations are sufficiently small, and so is $\nabla\mathbf{u}$, and therefore the quadratical term in (2.26) is neglected. So the above strain tensor simplifies to the *infinitesimal strain* tensor $\boldsymbol{\varepsilon}$, defined as:

$$\boldsymbol{\varepsilon}(\mathbf{u}) = \frac{1}{2}\big(\nabla\mathbf{u} + (\nabla\mathbf{u})^T\big),$$

or written in components as

$$\varepsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right).$$

It is assumed that the force-interaction can be entirely traced back to two kind of forces. The *body force* $\mathbf{b} : \Omega \to \mathbb{R}^3$ represents the force $\mathbf{b}dV$ per unit reference volume dV, e.g. the gravitational acceleration. The second kind of force acting on the body is the *surface traction* $\mathbf{t} : \Omega \times S^2 \to \mathbb{R}^3$, where S^2 denotes the unit sphere in $\mathbb{R}^3$. Let dA be a regular unit surface element of Ω with a given unit normal $\mathbf{n}$. Then $\mathbf{t}(\mathbf{x}, \mathbf{n})dA$, $\mathbf{x} \in dA$ is the force by the surface element dA exerted by the portion of the body Ω on the side of dA towards which $\mathbf{n}$ points, on the portion of dA that lies on the other side. The vector $\mathbf{t}(\mathbf{x}, \mathbf{n})$ is called *Cauchy stress vector*. Now we can state the laws of balance of linear and angular static momentum:

Postulate 2.1 (Static equilibrium)**.** *Let the body Ω be in state of equilibrium with respect to the body forces $\mathbf{b}$. Then there exists a vector field $\mathbf{t}$ on $\Omega \times S^2$ such that for every subset $V \subset \Omega$ the following holds:*

$$\int_V \mathbf{b}(\mathbf{x})\ d\mathbf{x}\ + \int_{\partial V} \mathbf{t}(\mathbf{x}, \mathbf{n})\ d\mathbf{s} = 0, \tag{2.27a}$$

$$\int_V \mathbf{x} \times \mathbf{b}(\mathbf{x})\ d\mathbf{x}\ + \int_{\partial V} \mathbf{x} \times \mathbf{t}(\mathbf{x}, \mathbf{n})\ d\mathbf{s} = 0. \tag{2.27b}$$

The equilibrium (2.27a) is called *balance of linear momentum* and (2.27b) is known as the *balance of angular momentum*. Now that we know the existence of the Cauchy stress vectors $\mathbf{t}$ we can state the following fundamental theorem:

Theorem 2.6 (Cauchy's theorem)**.** *Let* $\mathbf{t}(\cdot,\mathbf{n}) \in C^1(\Omega,\mathbb{R}^3)$, $\mathbf{t}(\mathbf{x},\cdot) \in C^0(S^2,\mathbb{R}^3)$ *and* $\mathbf{b} \in C(\Omega,\mathbb{R}^3)$ *in equilibrium* (2.27)*. Then there exists a symmetric tensor field* $\boldsymbol{\sigma} \in C^1(\Omega,\mathcal{S}^3)$ *with the following properties:*

$$\mathbf{t}(\mathbf{x},\mathbf{n}) = \boldsymbol{\sigma}(\mathbf{x})\mathbf{n}, \qquad \mathbf{x}\in\Omega,\ \mathbf{n}\in S^2, \tag{2.28a}$$

$$-\operatorname{div}\boldsymbol{\sigma}(\mathbf{x}) = \mathbf{b}(\mathbf{x}), \qquad \mathbf{x}\in\Omega, \tag{2.28b}$$

$$\boldsymbol{\sigma}(\mathbf{x}) = \boldsymbol{\sigma}^T(\mathbf{x}), \qquad \mathbf{x}\in\Omega. \tag{2.28c}$$

The tensor $\boldsymbol{\sigma}$ *is called the* Cauchy stress tensor*.*

Here $\mathcal{S}^3$ denotes the set of all symmetric 3×3 matrices. From the relation (2.27a) and a variant of the Green-Gauss theorem we obtain

$$\int_V \mathbf{b}(\mathbf{x})\ d\mathbf{x} + \int_{\partial V} \boldsymbol{\sigma}(\mathbf{x})\mathbf{n}\ d\mathbf{s} = \int_V \mathbf{b}(\mathbf{x}) + \operatorname{div}\boldsymbol{\sigma}(\mathbf{x})\ d\mathbf{x} = 0,$$

what from (2.28b) follows immediately. Moreover, the equilibrium (2.27b) yields the symmetry (2.28c). For a proof of the existence of the stress tensor (2.28a) we refer e.g. to Ciarlet [56].

An important question is now how to determine the stresses caused by some given external forces. The equilibrium equation (2.28b) only results in 3 equations, hence 6 components of the symmetric stress tensor are undefined. We gain the missing information from the particular material behavior. To start with, we call a body *linear elastic* if the stress depends linearly on the infinitesimal strain

$$\boldsymbol{\sigma} = \mathbf{C}\boldsymbol{\varepsilon}, \tag{2.29}$$

where $\mathbf{C}$ is called the *elasticity tensor*. We call the body Ω *homogeneous* if its density and the elasticity tensor $\mathbf{C}$ do not depend on the position $\mathbf{x}\in\Omega$. The mapping $\mathbf{C}$ can be represented as a fourth-order tensor in the following way

$$\sigma_{ij} = \mathbf{e}_i\big(\mathbf{C}(\mathbf{e}_k\times\mathbf{e}_l)\big)\mathbf{e}_j = C_{ijkl}\varepsilon_{kl},$$

where $\mathbf{e}_i$ denotes the vector of all zeros with an 1 at position i. Furthermore, the elastic tensor is symmetric, i.e. $C_{ijkl} = C_{jikl} = C_{ijlk} = C_{klij}$ and positive definite, that is $\boldsymbol{\varepsilon} : \mathbf{C}\boldsymbol{\varepsilon} > 0$ for all nonzero symmetric second-order tensors $\boldsymbol{\varepsilon}$. Another material property is *isotropy*, which means that the material does not possess any preferred directions or symmetries. In other words, the response of the material with respect to external forces does not depend on its orientation. Isotropy does not for hold for layered materials and e.g. wood. The most important mathematical effect of isotropy is that it reduces the number of independent components to 2. For an isotropic linearly elastic material the components of the stress tensor are then given by

$$C_{ijkl} = \lambda\delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}),$$

with δ_{ij} the Kronecker delta. The constants λ and μ are called the *Lamé*-constants. We can write the stress-strain relation (2.29) as

$$\boldsymbol{\sigma} = \lambda(\operatorname{tr}\boldsymbol{\varepsilon})\mathbf{I} + 2\mu\boldsymbol{\varepsilon}, \tag{2.30}$$

which is also known as *Hooke's law*. The constant μ is also known as the *shear* modulus and the material coefficient $K = \lambda + \frac{2}{3}\mu$ is called the *bulk* modulus. So an alternative pair of elastic material coefficients to the Lamé-constants is $\{\mu, K\}$. To an other important pair of material coefficients, namely the *Young's modulus* E and the *Poisson's ratio* ν, the following relation holds:

$$\nu = \frac{\lambda}{2(\lambda+\mu)} \qquad \text{and} \qquad E = \frac{\mu(3\lambda+2\mu)}{\lambda+\mu}. \tag{2.31}$$

Due to physical reasons we conclude that $\lambda > 0$, $\mu > 0$ and $E > 0$, $0 < \nu < \frac{1}{2}$ respectively. For many materials it holds that $\nu \approx \frac{1}{3}$ and for *nearly incompressible* materials ($\lambda \gg \mu$) ν is close to $\frac{1}{2}$.

Using the material coefficients $\{E, \nu\}$ we can rewrite Hooke's law (2.30) as

$$\boldsymbol{\sigma} = \frac{E\nu}{(1+\nu)(1-2\nu)}(\mathrm{tr}\boldsymbol{\varepsilon})\mathbf{I} + \frac{E}{1+\nu}\boldsymbol{\varepsilon}$$

and can rewrite (2.29) in components as:

$$\begin{pmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{33} \\ \sigma_{12} \\ \sigma_{13} \\ \sigma_{23} \end{pmatrix} = \frac{E}{(1+\nu)(1-2\nu)} \begin{pmatrix} 1-\nu & \nu & \nu & & & \\ \nu & 1-\nu & \nu & & 0 & \\ \nu & \nu & 1-\nu & & & \\ & & & 1-2\nu & & \\ & 0 & & & 1-2\nu & \\ & & & & & 1-2\nu \end{pmatrix} \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \varepsilon_{33} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{23} \end{pmatrix},$$

where we write, due to symmetry, $\boldsymbol{\sigma}$ and $\boldsymbol{\varepsilon}$ as vectors with 6 components.

Notes and remarks for Subsection 2.4.1

In many problems the extension of one spatial dimension is very small in comparison to the others. In these cases it is advisable to solve a simplified two dimensional problem rather than the full three dimensional problem. An example is a plate which is very thin in the x_3-direction in comparison to the other spatial directions or a dam. Here we distinguish between two cases:

- **State of plane stress:** Let us assume that only forces depending on x_1 and x_2 with zero x_3-component act on the plate. Under the further assumption that deformation in x_3-direction is possible we obtain the state of plane stress:

$$\begin{aligned} \sigma_{ij}(\mathbf{x}) &= \sigma_{ij}(x_1, x_2), && i,j = 1,2, \\ \sigma_{i3} &= \sigma_{3i} = 0, && i = 1,2,3, \\ \varepsilon_{33} &= -\frac{\nu}{1-\nu}(\varepsilon_{11} + \varepsilon_{22}). \end{aligned}$$

  The condition on ε_{33} results from $\sigma_{33} = 0$. It follows that $\varepsilon_{i3} = \varepsilon_{3i} = 0$ for $i = 1, 2$. After eliminating the strain ε_{33} we end up with the following stress-strain relationship:

$$\begin{pmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{12} \end{pmatrix} = \frac{E}{1-\nu^2} \begin{pmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & 1-\nu \end{pmatrix} \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \varepsilon_{12} \end{pmatrix}.$$

- **State of plane strain:** If we otherwise assume that, e.g. because of boundary conditions, that there are no deformations in the x_3-direction (e.g. like for a dam) we obtain the the state of plane strain:

$$\begin{aligned} u_i(\mathbf{x}) &= u_i(x_1, x_2), \quad i = 1, 2, \\ \varepsilon_{ij}(\mathbf{x}) &= \varepsilon_{ij}(x_1, x_2), \quad i, j = 1, 2, \\ \varepsilon_{i3} &= \varepsilon_{3i} = 0, \qquad i = 1, 2, 3. \end{aligned}$$

  Further we have that $u_3 = 0$ and $\sigma_{i3} = \sigma_{3i} = 0$ for $i = 1, 2$. From Hooke's law (2.30) and (2.31) we obtain $\sigma_{33} = \nu(\sigma_{11} + \sigma_{22})$. σ_{33} can be eliminated and we get the resulting stress-strain relationship:

$$\begin{pmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{12} \end{pmatrix} = \frac{E}{(1+\nu)(1-2\nu)} \begin{pmatrix} 1-\nu & \nu & 0 \\ \nu & 1-\nu & 0 \\ 0 & 0 & 1-2\nu \end{pmatrix} \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \varepsilon_{12} \end{pmatrix}.$$

2.4.2 Variational formulations for linear elasticity boundary value problems

As a starting point let us reconsider the problem of the deformation of a linearly elastic body. Let the body be denoted by an open, bounded and connected domain $\Omega \subset \mathbb{R}^3$ with a Lipschitz boundary Γ. We assume that the boundary is divided into two complementary parts Γ_u and Γ_t such that $\Gamma_u \cap \Gamma_t = \emptyset$, $\overline{\Gamma} = \overline{\Gamma}_u \cup \overline{\Gamma}_t$, and $|\Gamma_u| > 0$. On these parts of the boundary we define the boundary conditions as the following:

$$\begin{aligned} \mathbf{u} &= \mathbf{0}, \qquad \text{on } \Gamma_u, \\ \boldsymbol{\sigma} \cdot \mathbf{n} &= \mathbf{t}, \qquad \text{on } \Gamma_t. \end{aligned} \tag{2.32}$$

The governing equations for the static behavior of the body under the above boundary conditions are stated as follows:

$$-\operatorname{div} \boldsymbol{\sigma} = \mathbf{b}, \qquad \text{in } \Omega, \tag{2.33a}$$

$$\boldsymbol{\sigma} = \mathbf{C}\boldsymbol{\varepsilon}(\mathbf{u}), \qquad \text{in } \Omega, \tag{2.33b}$$

$$\boldsymbol{\varepsilon}(\mathbf{u}) = \frac{1}{2}\Big(\nabla \mathbf{u} + \big(\nabla \mathbf{u}\big)^T\Big), \qquad \text{in } \Omega, \tag{2.33c}$$

where (2.33a) is the equilibrium equation, (2.33b) is the constitutive law, and (2.33c) denotes the strain-displacement equation. Obviously (2.33) leads to a mixed formulation in $\mathbf{u}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\sigma}$. But it is possible to eliminate one or two unknown quantities. In the following we will list now two common variational formulations of (2.33).

Standard elliptic formulation

In this approach the strains $\boldsymbol{\varepsilon}$ and the stresses $\boldsymbol{\sigma}$ are eliminated and the problem is stated in the displacement field $\mathbf{u}$ only. At first we eliminate $\boldsymbol{\sigma}$ from (2.33b) to obtain

$$-\operatorname{div}\big(\mathbf{C}\boldsymbol{\varepsilon}\big(\mathbf{u}\big)\big) = \mathbf{b}, \qquad \text{in } \Omega, \tag{2.34}$$

as the equation of equilibrium. As the next step towards the weak formulation we introduce the space $V = H^1_{\Gamma_u}(\Omega)$ of admissible displacements. Then multiplication of (2.34) with an

arbitrary $\mathbf{v} \in V$ and integration by parts using the boundary conditions (2.32) yields the following variational equations:

$$a(\mathbf{u}, \mathbf{v}) = f(\mathbf{v}), \qquad \forall\ \mathbf{v} \in V, \tag{2.35}$$

with the abbreviations

$$a(\mathbf{u}, \mathbf{v}) = \int_\Omega \boldsymbol{\varepsilon}(\mathbf{u}) : \mathbf{C}\boldsymbol{\varepsilon}(\mathbf{v})\ d\mathbf{x} \tag{2.36}$$

and

$$f(\mathbf{v}) = \int_\Omega \mathbf{b} \cdot \mathbf{v}\ d\mathbf{x} + \int_{\Gamma_t} \mathbf{t} \cdot \mathbf{v}\ d\mathbf{s}. \tag{2.37}$$

As in Subsection 2.2.1 the Theorem 2.4 of Lax and Milgram answers the question of well-posedness of the problem (2.35). For the proof of the ellipticity of the bilinearform (2.36) we need *Korn's* inequality.

Lemma 2.1 (Korn's inequality)**.** *Let $\Omega \subset \mathbb{R}^3$ be an open and bounded set with a piecewise smooth boundary Γ. Furthermore, let $\Gamma_u \subset \Gamma$ with $|\Gamma_u| > 0$, $\mathbf{u} \in H^1_{\Gamma_u}(\Omega)$ and let the linearized strain tensor be defined as in* (2.33c)*. Then there exists a constant $c > 0$ depending on Ω, such that*

$$\|\mathbf{u}\|_1^2 \le c \int_\Omega \left|\boldsymbol{\varepsilon}(\mathbf{u})\right|^2\ d\mathbf{x}, \qquad \forall\ \mathbf{u} \in H^1_{\Gamma_u}(\Omega).$$

Proof. See e.g. Duvant and Lions [58]. □

With these preliminaries we can state now the following existence result (cf. Han and Reddy [74]):

Theorem 2.7. *Under the stated assumptions the problem* (2.35) *has a unique solution $\mathbf{u} \in H^1_{\Gamma_u}(\Omega)$. Furthermore, there exists a constant $c > 0$ such that*

$$\|\mathbf{u}\|_1 \le c\big(\|\mathbf{b}\|_0 + \|\mathbf{t}\|_{L_2(\Gamma_u)^3}\big).$$

The Hellinger and Reissner principle

In this method, named after *Hellinger* and *Reissner*, only the strains $\boldsymbol{\varepsilon}$ are eliminated. Thus the original problem (2.33) is stated as

$$\begin{aligned} \boldsymbol{\sigma} - \mathbf{C}\boldsymbol{\varepsilon}(\mathbf{u}) &= \mathbf{0}, && \text{in } \Omega, && (2.38a)\\ \operatorname{div} \boldsymbol{\sigma} &= -\mathbf{b}, && \text{in } \Omega, && (2.38b) \end{aligned}$$

with the boundary conditions (2.32). With the spaces $V = H^1_{\Gamma_u}(\Omega)$ and $Q = L_2(\Omega, \mathcal{S}^3)$ we define now the variational equations for (2.38) as:

$$\begin{aligned} \big(\mathbf{C}^{-1}\boldsymbol{\sigma}, \mathbf{q}\big)_0 \ -\ \big(\mathbf{q}, \boldsymbol{\varepsilon}(\mathbf{u})\big)_0 &= 0, && \forall\ \mathbf{q} \in Q,\\ -\big(\boldsymbol{\sigma}, \boldsymbol{\varepsilon}(\mathbf{v})\big)_0 &= -(\mathbf{b}, \mathbf{v})_0 + \int_{\Gamma_t} \mathbf{t} \cdot \mathbf{v}\ d\mathbf{x}, && \forall\ \mathbf{v} \in V, \end{aligned} \tag{2.39}$$

with the bilinear forms $a(\boldsymbol{\sigma}, \mathbf{q}) = (\mathbf{C}^{-1}\boldsymbol{\sigma}, \mathbf{q})_0$ and $b(\mathbf{q}, \mathbf{u}) = (\mathbf{q}, \boldsymbol{\varepsilon}(\mathbf{u}))_0$ (compare to Subsection 2.2.2). To ensure the existence of a solution $(\boldsymbol{\sigma}, \mathbf{u}) \in V \times Q$ the assumptions of Theorem 2.5 have to be fulfilled. The bilinarform $a(\cdot, \cdot)$ is Q-elliptic, since $\nu < 1/2$ and $\mathbf{C}$ positive definite. The LBB-condition follows from the following theorem.

Theorem 2.8. *Let $\Omega \subset \mathbb{R}^3$ be an open and bounded set with a piecewise smooth boundary Γ. Furthermore, let $\Gamma_u \subset \Gamma$ with $|\Gamma_u| > 0$ and $\mathbf{u} \in H^1_{\Gamma_u}(\Omega)$. Then there exists a constant $c > 0$ depending on Γ_u and Ω, such that*

$$\sup_{\mathbf{q} \in L_2(\Omega, \mathcal{S}^3)} \frac{b(\mathbf{q}, \mathbf{v})}{\|\mathbf{q}\|_0} \geq c \|\mathbf{v}\|_1, \qquad \forall\ \mathbf{v} \in H^1_{\Gamma_u}(\Omega).$$

Proof. See e.g. Braess [32]. □

The approach to create the variational equations (2.39) from (2.38) was a weak formulation of (2.38a) and a weak formulation of (2.38b) using partial integration. If we proceed the other way around, namely a weak formulation of (2.38a) using partial integration and a weak formulation of (2.38b), we obtain a different mixed formulation:

$$\begin{aligned} \left(\mathbf{C}^{-1}\boldsymbol{\sigma}, \mathbf{q}\right)_0 \ + \ (\operatorname{div} \mathbf{q}, \mathbf{u})_0 \ &= \ 0, && \forall\ \mathbf{q} \in Q_0, \\ (\operatorname{div} \boldsymbol{\sigma}, \mathbf{v})_0 \ &= \ -(\mathbf{b}, \mathbf{v})_0 - (\operatorname{div} \boldsymbol{\sigma}_t, \mathbf{v})_0\,, && \forall\ \mathbf{v} \in V, \end{aligned} \tag{2.40}$$

where we homogenized the boundary condition $\boldsymbol{\sigma}\mathbf{n} = \mathbf{t}$ with the approach $\boldsymbol{\sigma} = \boldsymbol{\sigma}_0 + \boldsymbol{\sigma}_t$ to $\boldsymbol{\sigma}_0 \cdot \mathbf{n} = \mathbf{0}$ and $\boldsymbol{\sigma}_t \cdot \mathbf{n} = \mathbf{t}$ on Γ_t. For sake of simplicity we have used $\boldsymbol{\sigma}$ instead of $\boldsymbol{\sigma}_0$ in (2.40), where we look for a solution $(\boldsymbol{\sigma}, \mathbf{u}) \in Q_0 \times V$ with the spaces $Q_0 = \{\mathbf{q} \in H(\operatorname{div}, \Omega)^{3\times 3} \mid \mathbf{q}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) = \mathbf{0} \text{ for } \mathbf{x} \in \Gamma_t\}$ and $V = L_2(\Omega, \mathcal{S}^3)$.

Which version of the Hellinger and Reissner principle is more reasonable depends amongst others on the stated boundary conditions.

Notes and remarks for Subsection 2.4.2

- In literature exists a third variational formulation for the elasticity problem, namely the *Hu-Washizu* principle:

$$\begin{aligned} (\mathbf{C}\boldsymbol{\varepsilon}, \mathbf{q})_0 \qquad\qquad - \ (\mathbf{q}, \boldsymbol{\sigma})_0 \ &= \ 0, && \forall\ \mathbf{q} \in L_2\big(\Omega, \mathcal{S}^3\big), \\ \big(\boldsymbol{\varepsilon}(\mathbf{v}), \boldsymbol{\sigma}\big)_0 \ &= \ (\mathbf{b}, \mathbf{v})_0 + \int_{\Gamma_u} \mathbf{t} \cdot \mathbf{v}\ d\mathbf{s}, && \forall\ \mathbf{v} \in H^1_{\Gamma_u}(\Omega), \\ -(\boldsymbol{\varepsilon}, \mathbf{r})_0 \ + \ \big(\boldsymbol{\varepsilon}(\mathbf{u}), \mathbf{r}\big)_0 \ &= \ 0, && \forall\ \mathbf{r} \in L_2\big(\Omega, \mathcal{S}^3\big). \end{aligned}$$

  Here we seek a solution $(\boldsymbol{\varepsilon}, \mathbf{u}, \boldsymbol{\sigma}) \in L_2(\Omega, \mathcal{S}^3) \times H^1_{\Gamma_u}(\Omega) \times L_2(\Omega, \mathcal{S}^3)$. With $a(\boldsymbol{\varepsilon}, \mathbf{u}; \mathbf{q}, \mathbf{v}) = (\mathbf{C}\boldsymbol{\varepsilon}, \mathbf{q})_0$, $b(\mathbf{q}, \mathbf{v}; \boldsymbol{\sigma}) = -(\mathbf{q}, \boldsymbol{\sigma})_0 + (\boldsymbol{\varepsilon}(\mathbf{v}), \boldsymbol{\sigma})_0$, and $f(\mathbf{q}, \mathbf{v}) = (\mathbf{b}, \mathbf{v})_0 + \int_{\Gamma_u} \mathbf{t} \cdot \mathbf{v}\ d\mathbf{s}$ the above formulation fits in the framework of mixed formulations of Subsection 2.2.2.

- As the Hu-Washizu principle is hardly used in practice, the Hellinger-Reissner principle plays an important role if the stresses $\boldsymbol{\sigma}$ have to be computed directly and not as a post-processing step, like in the pure displacement based formulation.

- A quite natural way to avoid numerical problems in the case of nearly incompressible materials and *locking* effects is to use mixed formulations with a penalty term (cf. Braess [32]).

Chapter 3

Two Approaches - Nested and Simultaneous Formulation

Structural optimization problems and many other optimization problems, like optimal control problems, are governed by a partial differential equations or by a system of PDEs. If we consider optimal design problems, the variables can be partitioned into the state and design variable, denoted by u and ρ, respectively. Of course, the properties of the objective depend strongly on the treated problems, and additionally, other constraints on the design ρ, the state u, or on both variables my be present. But in general we try to solve an optimization problem like

$$\mathcal{J}(\rho, u) \;\rightarrow\; \min_{\rho \in Q, u \in U} \tag{3.1a}$$

$$\text{subject to} \qquad e(\rho, u) \;=\; 0, \tag{3.1b}$$

where the equality constraint (3.1b) denotes the constraining PDE, also called the state equation. As mentioned above, the objective functional (3.1a) can be of various kinds. For the topology optimization problems presented in this work it is linear, once with respect to the state (the minimal compliance problem, see Chapter 4), and once with respect to the design (the minimal mass problem, see Chapter 5). For other structural optimization problems, e.g. shape optimization problems, the objective can be quite general. For optimal control and inverse problems it is usually of a quadratically least-square type with an additional regularization term. But also the state equation varies. In this work we will only treat the system of linear elasticity equations as constraining state equations. Beside elliptic equations, nonlinear and time-dependent state equations are of particular interest in many applications. The state may even consist of various state variables, related to different physical quantities, like e.g. a mechanical displacement field and an electromagnetical field. This is e.g. the case for problems from multidisciplinary structural optimization and multi-physics problems (cf. MEMS in Subsection 1.1.3).

There are basically two approaches for problems like (3.1). Under proper conditions (see the subsection below) the state equation can be solved for each design ρ to obtain a state $u(\rho)$ that depends uniquely on the design ρ. Thus, equation (3.1b) can be eliminated and formally hidden in the objective functional. Then we end up with the unconstrained optimization problem

$$\mathcal{J}\big(\rho, u(\rho)\big) \;\rightarrow\; \min_{\rho \in Q},$$

the so called *nested* approach or *black-box* approach. Alternatively, if no variables are eliminated and the state equation is treated as a constraint, see (3.1), the approach is called the non-nested approach or *simultaneous analysis and design* (SAND). In relevant literature there are also different names like, *simultaneous* optimization, *all-at-once* approach or *one-shot* methods.

In the following two sections of this chapter, we will briefly discuss the advantages and disadvantages of both approaches.

3.1 Nested Formulation

In this section we will list some properties and facts about the *nested* formulation of the problem (3.1):

$$\tilde{\mathcal{J}}(\rho) = \mathcal{J}\big(\rho, u(\rho)\big) \;\to\; \min_{\rho \in Q}. \tag{3.2}$$

As a starting point we state conditions (cf. Arian, Battermann und Sachs [11]), that admit a unique solution of the state equation (3.1b) with respect to the state.

Remark 3.1. *Let $e : Q \times U \to Z$ be twice continuously Fréchet differentiable, where U, Q, and Z are proper Hilbert spaces. Furthermore, let the partial Fréchet derivative e_u of e with respect to the state u be bijective and continuous, and let the partial Fréchet derivative e_ρ of e with respect to the design ρ be continuous.*

Then, the state equation, respectively its linearization, admits a unique solution with respect to the state:

$$e_u^{-1} : Z \to U \text{ exists and is a continuous linear operator for all } (u, \rho) \in U \times Q. \tag{3.3}$$

Under the assumption (3.3) and using the implicit function theorem, we can define a solution operator S for the state equation.

Lemma 3.1 (Solution Operator)**.** *Let the assumptions in Remark 3.1 hold. Moreover, let $\overline{\mathcal{Q}}$ be an open neighbourhood of $\overline{\rho} \in Q$ with $(\overline{\rho}, \overline{u}) \in Q \times U$ and $e(\overline{\rho}, \overline{u}) = 0$. Then, there exists a unique* solution operator *$S : \overline{\mathcal{Q}} \to U$ that is twice continuously Fréchet differentiable in $\overline{\mathcal{Q}}$ and that satisfies the identity*

$$e\big(\rho, S(\rho)\big) = 0, \qquad \forall\ \rho \in \overline{\mathcal{Q}}.$$

Furthermore, the derivative of S with respect to ρ is given by

$$S'(\rho) = -e_u^{-1}\big(\rho, S(\rho)\big) e_\rho\big(\rho, S(\rho)\big), \qquad \forall\ \rho \in \overline{\mathcal{Q}}.$$

In structural design it is quite common to use the nested problem formulation (3.2). Let us now list some general facts about the nested approach:

- The reduction of the problem from the product space $U \times Q$ to the design space Q reduces the problem's dimension. This has an advantageous effect if the dimension of the design space is small, in particular, if it is much smaller than the dimension of the state space U.

- Because of the introduction of the solution operator S, the state constraint is fulfilled (up to a discretization error) at every iteration step of the optimization routine. Thus, the nested approach can be regarded as a feasible path method. Even if the optimization algorithm is interrupted ahead of termination, the actual design has a physical meaning, since $e(\rho^{(k)}, u^{(k)}) = 0$ for all iteration numbers k.

- But on the other hand, due to the introduction of the solution operator S, it is necessary to solve the state equation (3.1b) for each evaluation of the objective functional and of it's gradient, respectively. Additionally, for each gradient evaluation, the adjoint problem to the state equation has to be solved. This can be rather time consuming, even when solution methods of optimal order are applied, like a multigrid method or a multigrid preconditioned conjugate gradient algorithm, especially if the state equation is nonlinear.

- Moreover, the reduction of variables will change the original properties of the objective functional. Even if $\mathcal{J}(\rho, u)$ is linear or convex, $\tilde{\mathcal{J}}(\rho)$ may become severely nonlinear and nonconvex. For the minimal compliance problem (see next Chapter 4) we e.g. have that $\mathcal{J}(\boldsymbol{\rho}^h, \mathbf{u}^h) = {\mathbf{f}^h}^T \mathbf{u}^h$, but $\tilde{\mathcal{J}}(\boldsymbol{\rho}^h) = {\mathbf{f}^h}^T \mathbf{K}^{-1}(\boldsymbol{\rho}^h)\mathbf{f}^h$ due to $\mathbf{S}(\boldsymbol{\rho}^h) = \mathbf{K}^{-1}(\boldsymbol{\rho}^h)\mathbf{f}^h$.

- In topology optimization the nested approach is usually used together with the method of moving asymptotes, see Subsection 2.1.2. If the design space is small, also SQP-type methods are possible.

3.2 Simultaneous Optimization

Let us take now a closer look at the simultaneous approach. As a reminder, let us state the original model problem again:

$$\mathcal{J}(\rho, u) \;\rightarrow\; \min_{\rho \in Q, u \in U} \tag{3.4a}$$

$$\text{subject to} \qquad e(\rho, u) \;=\; 0. \tag{3.4b}$$

In comparison to the nested approach, no solution operator is required. The state equation is not eliminated, but treated as an equality constraint. Thus, optimization is carried out in the product space $Q \times U$. In this spirit let us formulate the first-order necessary conditions (Theorem 2.1) for problem (3.4). The Lagrangian functional for problem (3.4) is given by

$$\mathcal{L}(\rho, u, \lambda) = \mathcal{J}(\rho, u) + \big\langle \lambda, e(\rho, u) \big\rangle, \tag{3.5}$$

where $\lambda \in Z^*$ denotes the Lagrangian multiplier of the state constraint. Then the first-order necessary conditions are given by the following equations:

$$\nabla_\rho \mathcal{L} \;=\; \mathcal{J}_\rho + \langle \lambda, e_\rho \rangle \;=\; 0, \tag{3.6a}$$

$$\nabla_u \mathcal{L} \;=\; \mathcal{J}_u + \langle \lambda, e_u \rangle \;=\; 0, \tag{3.6b}$$

$$\nabla_\lambda \mathcal{L} \;=\; e(\rho, u) \qquad\; =\; 0. \tag{3.6c}$$

Here, (3.6a) is the design equation, (3.6b) is usually called the adjoint or costate equation, and (3.6c) is the state equation. The equations (3.6) can be combined to a (nonlinear) mapping

$F(\rho, u, \lambda) = \mathbf{0}$, that can be solved using Newton's method. Thus, we end up with the following linear system, which has to be solved at each Newton iteration:

$$\begin{pmatrix} \mathcal{L}_{\rho\rho} & \mathcal{L}_{\rho u} & e_\rho^* \\ \mathcal{L}_{u\rho} & \mathcal{L}_{uu} & e_u^* \\ e_\rho & e_u & 0 \end{pmatrix} \begin{pmatrix} \triangle\rho \\ \triangle u \\ \triangle\lambda \end{pmatrix} = - \begin{pmatrix} \mathcal{J}_\rho + \langle \lambda, e_\rho \rangle \\ \mathcal{J}_u + \langle \lambda, e_u \rangle \\ e(\rho, u) \end{pmatrix}. \tag{3.7}$$

Note, that assumption (3.3) implies that e_u is a regular operator, which must not hold for e_q. The fact that this *saddle point* system arises from the KKT-conditions, gave it the name *KKT-system*. After a suitable discretization, e.g. using the finite element method, we can rewrite the system (3.7) as a discrete linear system

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{B}_1^T \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{B}_2^T \\ \mathbf{B}_1 & \mathbf{B}_2 & \mathbf{0} \end{pmatrix} \begin{pmatrix} \triangle\boldsymbol{\rho}^h \\ \triangle\mathbf{u}^h \\ \triangle\boldsymbol{\lambda}^h \end{pmatrix} = \begin{pmatrix} \mathbf{f}_1^h \\ \mathbf{f}_2^h \\ \mathbf{g}^h \end{pmatrix} \tag{3.8}$$

In the sequel we will also use the following abbreviation of the KKT-system (3.8):

$$\begin{pmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix}.$$

We use the abbreviation $\mathbf{A}$ for the discrete analogon of the Hessian of the Lagrangian, thus $\mathbf{A}$ is symmetric. Let us shortly list some properties in comparison to the nested problem formulation:

- Potential comfortable properties of the problem are not destroyed, e.g. like linearity of the objective or sparsity of the KKT system.
- The simultaneous approach does not follow the feasible path by the state equation. The state equation has only to be fulfilled by the final state $\overline{\mathbf{u}}$, $\mathbf{u}^{(k)} \to \overline{\mathbf{u}}$, and the final optimal design $\overline{\boldsymbol{\rho}}$, $\boldsymbol{\rho}^{(k)} \to \overline{\boldsymbol{\rho}}$. Hence, a significant speedup can be expected. See, e.g. Burger and Mühlhuber [45, 46], for an application of the simultaneous approach to inverse problems (parameter identification problems).
- The simultaneous approach contains a solution to the sparse, symmetric, but indefinite linear system (3.8), at each Newton step.
- Up to now the simultaneous analysis and design method is hardly used in continuous topology optimization. As examples we refer to Maar and Schulz [89] and Hoppe and Petrova [76]. In both cases, interior-point methods (cf. Subsection 2.1.3) are used to solve the arising optimization problem.
- It has been shown that various examples of nonlinear mixed 0-1 topology optimization problems can be modelled in an equivalent way as linear mixed 0-1 problems (cf. Stolpe and Svanberg [140] and Stolpe [136]). The simultaneous approach is the cornerstone of these reformulations, that allow to solve these problems to global optimality.

The keystone to a simultaneous approach is now an efficient solution technique for the KKT-system (3.8). Since the actual properties of the system matrix in (3.8), and in particular of the block matrices, change from each problem to the other, it is not possible to advise a

general solution approach. There exist several ways how to apply iterative solution methods and how to construct efficient preconditioners.

Reduced approaches consist of an a-priori elimination of the equality constraint $\mathbf{B}_1 \triangle \boldsymbol{\rho}^h + \mathbf{B}_2 \triangle \mathbf{u}^h = \mathbf{g}^h$, which results in an elimination of the state $\mathbf{u}^h$ and Lagrangian multiplier $\boldsymbol{\lambda}^h$. After some calculations this yields the reduced system

$$\mathbf{K}_r \triangle \boldsymbol{\rho}^h = \mathbf{f}_r^h,$$

with the Schur complement

$$\mathbf{K}_r = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{B}_2^{-1} + \mathbf{B}_1^T \mathbf{B}_2^{-T}(\mathbf{A}_{21} - \mathbf{A}_{22}\mathbf{B}_2^{-1}\mathbf{B}_1)$$

and the corresponding right-hand side

$$\mathbf{f}_r^h = \mathbf{f}_1^h - \mathbf{A}_{22}\mathbf{B}_2^{-1}\mathbf{g}^h - \mathbf{B}_1^T \mathbf{B}_2^{-T}(\mathbf{f}_2^h - \mathbf{A}_{22}\mathbf{B}_2^{-1}\mathbf{g}^h).$$

Note, that due to assumption (3.3), $\mathbf{B}_2$ is regular and the above computations are possible. The reduced approach is of particular interest if the size of the design space is much smaller than the size of the state space. Nevertheless, the application of $\mathbf{K}_r$ is likely to be more expensive than an application of the KKT-matrix, since it involves two solutions of systems with the matrix $\mathbf{B}_2$. This drawback is usually overcome by using Broyden-type update schemes for the reduced system matrix. This strategy is frequently used in inverse problems and optimal control. For applications we refer e.g. to Sachs [113] and to Schulz and Bock [118].

An alternative to reduced approaches, namely the *simultaneous solution* of the KKT-system has been recently investigated, especially in connection with optimal control problems, see e.g. Battermann and Heinkenschloss [14], Biros and Ghattas [24], and Battermann and Sachs [15]. A reason for this is, that the assembling and application of the reduced system matrix is more expensive then the assembling and application of the KKT-matrix, even if it is larger and indefinite. In the following we will discuss briefly some iterative methods to solve the KKT-system (3.8) and address the question of preconditioning. A survey on solution methods for saddle point problems is given by Benzi, Golub, and Liesen [23].

Since (3.8) is a linear indefinite system, it seems quite natural to use variants of the conjugate gradient algorithm that are applicable to indefinite systems. Those methods belong the class of *Krylov subspace* methods (cf. Saad [111] and Subsection 2.3.2). Suitable methods would be GMRES (cf. Saad and Schultz [112]), QMR (cf. Freund and Nachtigal [67]), and MINRES (cf. Paige and Saunders [100]). Thus, appropriate preconditioning is essential. Another possibility is a positive definite reformulation of the saddle point system and applying a CG method with a proper inner product (cf. Bramble and Pasciak [36]).

Another class of iterative methods for indefinite linear systems are *inexact UZAWA* algorithms. Convergence properties have been investigated, e.g. by Langer and Queck [82, 83] and more recently by Bramble, Pasciak and Vassilev [37] and Zulehner [158]. Inexact Uzawa methods have been developed for the iterative solution of the Stokes' problem (cf. Subsection 2.2.2) and similar mixed problems. Due to this original motivation, these methods rely on the positive semi-definiteness of the upper left block matrix $\mathbf{A}$. Unfortunately, this property cannot be fulfilled by most topology optimization problems. An inexact Uzawa method for the system (3.8) writes as

$$\begin{aligned} \hat{\mathbf{A}}\big(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\big) &= \mathbf{f} - \mathbf{A}\mathbf{x}^{(k)} - \mathbf{B}^T\mathbf{y}^{(k)}, \\ \hat{\mathbf{C}}\big(\mathbf{y}^{(k+1)} - \mathbf{y}^{(k)}\big) &= \mathbf{B}\mathbf{x}^{(k+1)} - \mathbf{g}, \end{aligned}$$

where $\hat{\mathbf{A}}$ and $\hat{\mathbf{C}}$ are symmetric positive definite preconditioners for $\mathbf{A}$ and for the (negative) Schur complement $\mathbf{C} = \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T$. In order to construct a efficient solver in this way, good preconditioners $\hat{\mathbf{A}}$ and $\hat{\mathbf{C}}$ to $\mathbf{A}$ and $\mathbf{C}$ respectively, are of high importance.

As a third class of iterative solution methods for the KKT system we mention *multigrid methods* with proper smoothers. Two classes of smoothers are of special interest, namely local patch smoothers, e.g. see SCHÖBERL AND ZULEHNER [116], and global block smoothers, e.g. see BRAESS AND SARAZIN [33] and ZULEHNER [157]. In SCHÖBERL AND ZULEHNER [116] a Schwarz-type iteration method as smoothers in a multigrid method for saddle point problems is considered and rigorously analyzed. A multigrid convergence proof is given for the additive case, numerical examples are presented for both, the additive and the multiplicative Schwarz-type smoother. For more information we refer to Chapter 6, where an optimality system arising from an interior-point formulation is solved by the use of a multigrid method with a multiplicative patch smoother. Another class of smoothers is analyzed in ZULEHNER [157], where a symmetric positive definite preconditioner $\hat{\mathbf{A}}$ for $\mathbf{A}$ and a symmetric positive definite preconditioner $\hat{\mathbf{S}}$ for the (negative) inexact Schur complement $\mathbf{B}\hat{\mathbf{A}}^{-1}\mathbf{B}^T$ are needed. Again, both classes of smoothers depend on the semi-positive definiteness of $\mathbf{A}$. A class of smoothers that do not rely on this property, are the so-called transforming smoothers (cf. WITTUM [150]). They are used in MAAR AND SCHULZ [89] for their simultaneous optimization approach to the minimal compliance problem. However no regularization is applied and no convergence analysis is given.

Chapter 4

An Adaptive Multilevel Approach to the Minimal Compliance Problem

In this chapter we present an adaptive multilevel approach to the minimal compliance problem. Here we search for an optimal material distribution with respect to maximal stiffness under a given loading and restriction of the total volume used. This problem contains the system of linear elasticity partial differential equations as constraints, resulting in a large scaled optimization problem after the finite element discretization. Due to the repeated solution of the direct field problem given by the PDE constraints, efficient solution techniques are required. Next to adaptive mesh-refinement we use a multigrid approach for the direct problem. Minimizing compliance turned out to be a standard problem in topology optimization. However, it already contains the most basic, but non-trivial difficulties like mesh-dependent solutions, local minima and checkerboard phenomena. Due to this ill-posedness we need regularization. In our algorithm we combine two filter methods, such that their disadvantages are eliminated and only their positive properties remain. Numerical examples are performed with several benchmark problems, where our adaptive multilevel approach turns out to be quite efficient. For solving the optimization problem arising in each iteration step, the method of moving asymptotes is used.

We start with an introduction to the minimal compliance problem and briefly discuss material interpolation methods. In the next section we treat the aspect of regularization using filter methods, which is a cornerstone of our adaptive multilevel approach. Afterwards, the approach itself is discussed and finally successful numerical examples are presented in the last section.

4.1 Preliminaries

4.1.1 The Minimal Compliance Problem

This section is devoted to the problem of minimizing the compliance of a structure subject to a weight constraint. For this let us consider Figure 4.1. Let $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ be an open, bounded connected domain with a Lipschitz boundary Γ, the so called *ground structure*. Moreover, let $\Gamma_u \subset \Gamma$, $|\Gamma_u| > 0$ be the part of the boundary where the displacements are

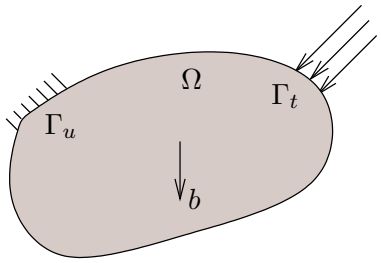

Figure 4.1: The reference domain and applied forces in a minimal compliance problem.

fixed, and $\Gamma_t = \Gamma \setminus \Gamma_u$ the part where boundary tractions are prescribed (cf. (2.32)). Later on, the optimal design is generated referring to this ground structure. Our aim is now to distribute in Ω a certain amount of material, so that the resulting structure is as stiff as possible under loading. For sake of simplicity we will restrict ourselves to the case of isotropic material and to the case where no body forces are applied ($\mathbf{b} = \mathbf{0}$). A first ideal formulation of the minimal compliance problem looks now like the following:

$$\ell(\mathbf{u}) = \int_{\Gamma_t} \mathbf{t} \cdot \mathbf{u} \, d\mathbf{s} \;\rightarrow\; \min_{\rho,\mathbf{u}} \tag{4.1a}$$

$$\text{subject to} \qquad a(\rho; \mathbf{u}, \mathbf{v}) \;=\; \ell(\mathbf{v}), \qquad \forall\, \mathbf{v} \in V_0, \tag{4.1b}$$

$$\int_\Omega \rho(\mathbf{x}) \, d\mathbf{x} \;\leq\; m_0, \tag{4.1c}$$

$$\rho(\mathbf{x}) \;\in\; \{0,1\}, \qquad \text{a.e. in } \Omega. \tag{4.1d}$$

The compliance is given by the objective functional (4.1a), which is also the right hand side of the equilibrium constraints (4.1b). These equilibrium constraints contain the linear elasticity equations (cf. Subsection 2.4.1) in a weak formulation, where $V_0 = H^1_{\Gamma_u}(\Omega; \mathbb{R}^d)$ denotes the set of kinematically admissible displacement fields. Constraint (4.1d) tells that each point $\mathbf{x} \in \Omega$ should be occupied with material ($\rho(\mathbf{x}) = 1$) or void ($\rho(\mathbf{x}) = 0$). Let $\mathbf{C}^0$ describe an elasticity tensor of fourth order, satisfying the usual symmetry, ellipticity and boundedness assumption. Then, the material tensor related to $\rho(\mathbf{x}) = 1$ is $\mathbf{C}(1) = \mathbf{C}^0$ and the one that corresponds to $\rho(\mathbf{x}) = 0$, is $\mathbf{C}(0) = \mathbf{0}$. Moreover, constraint (4.1c) limits the amount of available material.

In order to use gradient based optimization methods and to avoid e.g. branch and bound techniques to solve the 0-1 problem, the constraint (4.1d) is relaxed. So the discrete valued constraint is replaced by a continuous version $\rho(\mathbf{x}) \in [0,1]$ with $\rho \in L_\infty(\Omega)$. For methods to still obtain a 0-1 solution, we refer to Subsection 1.1.2 and to the two following subsections. Below we replace the lower bound 0 by a small value $\rho_{\min}$, $0 < \rho_{\min} \ll 1$, to still ensure the ellipticity of the bilinearform $a(\rho; \mathbf{u}, \mathbf{v})$. Furthermore, let

$$\eta : [\rho_{\min}, 1] \rightarrow (0, 1]$$

be a continuous, monotonously increasing function. This *material interpolation* function η describes how the actual density $\rho(\mathbf{x})$ influences the elasticity tensor $\mathbf{C}$ (e.g. to enforce 0-1 designs) at a given point $\mathbf{x} \in \Omega$. Then the actually used elasticity tensor is variable over the ground structure and is defined as

$$\mathbf{C}\big(\rho(\mathbf{x})\big) = \eta\big(\rho(\mathbf{x})\big)\mathbf{C}^0, \qquad \text{a.e. in } \Omega. \tag{4.2}$$

The energy bilinearform on $V_0 \times V_0$ is then given by

$$a(\rho; \mathbf{u}, \mathbf{v}) = \int_\Omega \boldsymbol{\varepsilon}\big(\mathbf{u}(\mathbf{x})\big) : \mathbf{C}\big(\rho(\mathbf{x})\big)\boldsymbol{\varepsilon}\big(\mathbf{v}(\mathbf{x})\big)\ d\mathbf{x}.$$

Summarizing all this steps we end up at the following formulation of the minimal compliance problem:

$$\begin{aligned} \ell(\mathbf{u}) \ &\rightarrow \min_{\rho \in L_\infty(\Omega), \mathbf{u} \in V_0} && && (4.3a)\\ \text{subject to} \quad a(\rho; \mathbf{u}, \mathbf{v}) \ &= \ \ell(\mathbf{v}), && \forall\ \mathbf{v} \in V_0, && (4.3b)\\ \int_\Omega \rho(\mathbf{x})\ d\mathbf{x} \ &\leq \ m_0, && && (4.3c)\\ \rho_{\min} \leq \rho(\mathbf{x}) \ &\leq \ 1, && \text{a.e. in } \Omega. && (4.3d) \end{aligned}$$

In the above formulation (4.3) the problem is stated in a simultaneous formulation. But usually, and also in this chapter, the state variable $\mathbf{u}$ is eliminated through the state equation, resulting in a nested formulation. For a given admissible ρ the solution of (4.3b) is ensured and denoted by $\mathbf{u}(\rho)$, arriving at the nested formulation of the minimal compliance problem:

$$\begin{aligned} \ell\big(\mathbf{u}(\rho)\big) \ &\rightarrow \min_{\rho \in L_\infty(\Omega)} && && (4.4a)\\ \text{subject to} \quad \int_\Omega \rho(\mathbf{x})\ d\mathbf{x} \ &\leq \ m_0, && && (4.4b)\\ \rho_{\min} \leq \rho(\mathbf{x}) \ &\leq \ 1, && \text{a.e. in } \Omega, && (4.4c) \end{aligned}$$

Now the state constraint (4.3b) is hidden in the objective (4.4a), which means that for every evaluation of the objective functional, or of it's gradient, the state equation (4.3b) has to be solved.

Notes and remarks for Subsection 4.1.1

There are several extensions to the minimal compliance problem, which are all omitted in this section for sake of simplicity.

- Include multiple load cases to the problem. This is done rather easily, since the optimization with respect to multiple load conditions is formulated as an optimization of a weighted average of the compliances for each of the load cases.

- Include the self-weight of the structure in the optimization process. From a modeling viewpoint this means that body forces have to be considered ($\mathbf{b} \neq \mathbf{0}$) and the load vector $\ell(\rho, \mathbf{v})$ becomes design dependent.

- Consider areas in the ground structure Ω that have to be filled with material or that have to be free of material in the optimal design.

4.1.2 Material Interpolation

In the previous Subsection we replaced the integer valued constraint (4.1d) by the continuous relaxation (4.4c). But in order to obtain 0-1 optimal designs the intermediate density values have to be be penalized.

One way to penalize intermediate values is to use a proper non-linear material interpolation function that penalizes intermediate values indirectly. This is done, e.g. if intermediate densities give little stiffness in comparison to the amount of used material. The most known penalization method is the SIMP (*Solid Isotropic Material with Penalization*) approach (cf. Subsection 1.1.2). Here a non-linear interpolation model of the form $\eta(\rho(\mathbf{x})) = \rho(\mathbf{x})^p$ with $p \geq 1$ is used. Then the relation (4.2) between the density and the material tensor in the state equation is given by

$$\mathbf{C}_p\big(\rho(\mathbf{x})\big) = \rho(\mathbf{x})^p \mathbf{C}^0, \qquad \forall\, \mathbf{x} \in \Omega. \tag{4.5}$$

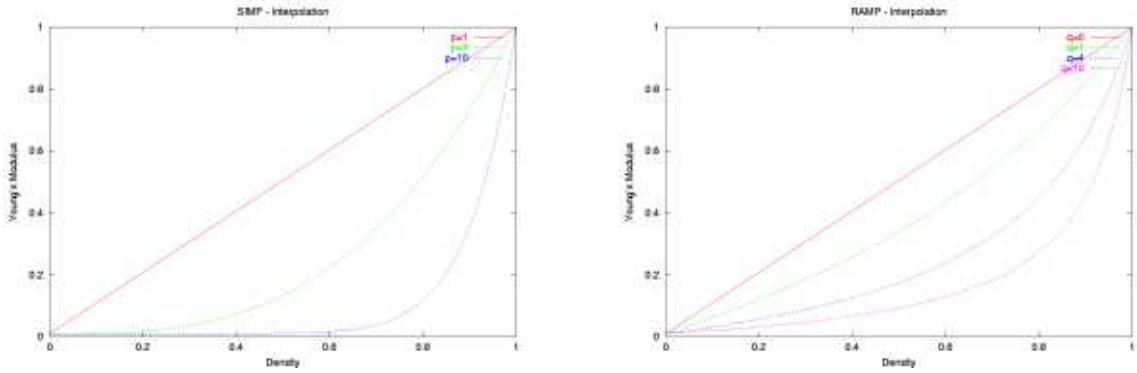

Figure 4.2: Left: SIMP-Interpolation scheme with various values for p. Right: RAMP-Interpolation with various values for q.

An alternative approach to the SIMP method is the following interpolation model:

$$\mathbf{C}_q\big(\rho(\mathbf{x})\big) = \frac{\rho(\mathbf{x})}{1 + q\big(1 - \rho(\mathbf{x})\big)} \mathbf{C}^0, \tag{4.6}$$

which is called RAMP (*Rational Approximation of Material Properties*) and was mentioned in Rietz [106] and treated thoroughly in Stolpe and Svanberg [137]. The two interpolation models satisfy that

$$\begin{aligned} \mathbf{C}_p\big(\rho_{\min}\big) &= \rho_{\min}^p \mathbf{C}^0, \\ \mathbf{C}_p(1) &= \mathbf{C}^0, \end{aligned} \qquad \text{and} \qquad \begin{aligned} \mathbf{C}_q\big(\rho_{\min}\big) &= \frac{\rho_{\min}}{1 + q\big(1 - \rho_{\min}\big)} \mathbf{C}^0, \\ \mathbf{C}_q(1) &= \mathbf{C}^0, \end{aligned}$$

with $0 < \rho_{\min}^p \ll 1$ and $0 < \frac{\rho_{\min}}{1+q(1-\rho_{\min})} \ll 1$ for $0 < \rho_{\min} \ll 1$, $p \geq 1$, and $q \geq 0$. So $\mathbf{C}_p(\rho_{\min})$ and $\mathbf{C}_q(\rho_{\min})$ can be regarded as two compliant materials (in comparison to $\mathbf{C}^0$), pretending to be void. Above the lower bound $\rho_{\min} > 0$ was introduced to guarantee the ellipticity of the bilinearform $a(\rho; \mathbf{u}, \mathbf{v})$, which resulted in a very compliant material for the

regions of Ω occupied with void. Following this idea we could see the used material over the ground structure as a composite of two materials, whereas one of them is interpreted as void. Consider $\mathbf{C}_0$ and $\mathbf{C}_1$ as two material tensors with the same Poisson's ratios ν_0 and ν_1 but with different Young moduli, e.g. $0 < E_0 \ll E_1$. The material tensor of the composite is then given by

$$\mathbf{C}\big(\rho(\mathbf{x})\big) = \mathbf{C}_0 + \eta\big(\rho(\mathbf{x})\big)\triangle\mathbf{C}, \qquad \text{a.e. in } \Omega,$$

with $\triangle\mathbf{C} = \mathbf{C}_1 - \mathbf{C}_0$. Now it is possible to set $\rho_{\min} = 0$ and, because of $E_0 > 0$, to still ensure the V_0-ellipticity of the bilinearform. One of the advantages of the RAMP model with respect to the SIMP model is the behavior of the derivative of the material model at $\rho(\mathbf{x}) = 0$. Comparing the derivatives yields the following:

$$\mathbf{C}'_p(0) = \begin{cases} \mathbf{C}^0 & \text{if } p = 1, \\ \mathbf{0} & \text{if } p > 1, \end{cases} \qquad \text{vs.} \qquad \mathbf{C}'_q(0) = \frac{1}{1+q}\mathbf{C}^0,$$

where it is worth noticing that $\mathbf{C}'_p(0)$ is discontinuous in the parameter p, while $\mathbf{C}'_q(0)$ is continuous in the parameter $q \geq 0$. The fact that $\mathbf{C}'_p(0) = \mathbf{0}$ for $p > 1$ makes it hard to move material, which is not the case when using the RAMP model. But the primary motivation for the RAMP model is the reason that problems with penalization of intermediate values become in general non-convex. Without any penalization ($\eta(\rho) = \rho$) compliance is a convex functional (cf. Svanberg [142]). So a common suggestion is to use a *continuation* on the penalization parameter to shift the objective cautiously from convexity to non-convexity, hopefully concavity, when the penalization parameter is high enough. This increases the possibility to obtain a global optimal minimum to the 0-1 problem (4.1). For a sufficiently high value of q, the compliance turns into a concave functional for the RAMP scheme, which needs not to happen when increasing p in the SIMP model (cf. Stolpe and Svanberg [137]).

A totally different approach to penalize intermediate density values is use the material interpolation function $\eta(\rho(\mathbf{x})) = \rho(\mathbf{x})$ and to add an additional constraint to the optimization problem to encourage 0-1 optimal designs. Such a penalty constraint could e.g. look like the following:

$$P\big(\rho(\mathbf{x})\big) = \int_\Omega \big(1 - \rho(\mathbf{x})\big)\big(\rho(\mathbf{x}) - \rho_{\min}\big)\, d\mathbf{x} \;\leq\; \varepsilon_P. \tag{4.7}$$

Of course such a penalty function can also be added as a penalty term to the objective as $\mathcal{J}(\rho, \mathbf{u}) + w_P P(\rho)$, with a weighting factor $w_P > 0$. But in both cases it is a tricky task to choose proper values for ε_P or w_P.

Notes and remarks for Subsection 4.1.2

- An other disadvantage of SIMP versus RAMP is that the mass depends linearly on the density ρ and the stiffness depends on a power of ρ, which results in a non finite ratio of mass to stiffness when ρ attends zero.

- Since problems with a penalizing material interpolation function are in general non-convex, continuation methods have been proposed, see above. But it is shown in Stolpe and Svanberg [138] that the trajectory, defined as the path followed by the solutions to the penalized problems, as the penalization is intensified, may be discontinuous, no matter how gently the penalization parameter is increased.

4.2 Regularization using Filter Methods

A naive formulation of topology optimization tasks like the minimal compliance problem will lead to difficulties in the sense that there are no optimal solutions (cf. Subsection 1.1.1). In this section we will discuss two filter methods to obtain a well-posed version of the problem (4.4).

An optimization problem is said to be well-posed when the two following conditions are valid: The objective functional has to be lower semi-continuous and the feasible set has to be compact, and both properties have to be fulfilled with respect to the same topology. We begin with the definition of lower semi-continuity.

Definition 4.1 (Lower Semi-continuity)**.** *Suppose X is a topological space, $\overline{x} \in X$ and $f : X \to \mathbb{R}$ is a real-valued mapping. The mapping f is called* lower semi-continuous *at $\overline{x}$ if for every $\varepsilon > 0$ there exists a neighborhood $\overline{N}_\varepsilon$ of $\overline{x}$ such that $f(x) > f(\overline{x}) - \varepsilon$ for all $x \in \overline{N}_\varepsilon$. Equivalently, this can be expressed as*

$$\liminf_{x \to \overline{x}} f(x) \geq f(\overline{x}).$$

The mapping f is called lower semi-continuous if it is lower semi-continuous for all $x \in X$. Then $\{x \in X \mid f(x) > a\}$ is an open set for every $a \in \mathbb{R}$. The mapping f is called upper semi-continuous *if $(-f)$ is lower semi-continuous.*

Note, that the combination of lower and upper semi-continuity yields continuity.

Let us now investigate the existence of solutions to the minimal compliance problem for the variable thickness sheet, i.e. problem (4.4) with $\eta(\rho) = \rho$.

Lemma 4.1. *Let $\mathcal{F}$ denote the set of all feasible designs of the minimal compliance problem (4.4), i.e. $\mathcal{F} = \left\{\rho \in L_\infty(\Omega) \;\middle|\; \int_\Omega \rho(\mathbf{x})\, d\mathbf{x} \leq m_0,\ \rho_{\min} \leq \rho(\mathbf{x}) \leq 1 \text{ a.e. in } \Omega\right\}$. Then $\mathcal{F}$ is weakly* closed in $L_\infty(\Omega)$.*

Proof. Let $\{\rho^{(k)}\} \subset \mathcal{F}$ be a sequence such that $\rho^{(k)} \overset{*}{\rightharpoonup} \overline{\rho}$ in $L_\infty(\Omega)$. Since Ω is bounded, the weak* convergence yields that

$$\int_\Omega \overline{\rho}(\mathbf{x})\, d\mathbf{x} = \lim_{k \to \infty} \int_\Omega \rho^{(k)}(\mathbf{x})\, d\mathbf{x} \leq m_0.$$

Moreover, choose an arbitrary measurable subset $\omega \subset \Omega$ and $\varepsilon > 0$ arbitrary but fixed. Then there exists an integer K, such that

$$\int_\omega 1 - \overline{\rho}(\mathbf{x})\, d\mathbf{x} = \int_\omega 1 - \rho^{(k)}(\mathbf{x})\, d\mathbf{x} + \int_\omega \rho^{(k)}(\mathbf{x}) - \overline{\rho}(\mathbf{x})\, d\mathbf{x} \geq 0 - \varepsilon = -\varepsilon,$$

and

$$\int_\omega \overline{\rho}(\mathbf{x}) - \rho_{\min}\, d\mathbf{x} = \int_\omega \overline{\rho}(\mathbf{x}) - \rho^{(k)}(\mathbf{x})\, d\mathbf{x} + \int_\omega \rho^{(k)}(\mathbf{x}) - \rho_{\min}\, d\mathbf{x} \geq -\varepsilon + 0 = -\varepsilon,$$

holds for all $k > K$. Now, since ω and ε were arbitrarily and $\overline{\rho}$ satisfies the constraints (4.4b) and (4.4c), the feasible set $\mathcal{F}$ is weakly* closed in $L_\infty(\Omega)$. □

Lemma 4.2 (Lower Semi-continuity of Compliance)**.** *Let $\ell(\mathbf{u}(\cdot))$ denote compliance like in (4.4a) with $\eta(\rho) = \rho$. Then $\ell(\mathbf{u}(\cdot))$ is lower semi-continuous in the weak* topology of $L_\infty(\Omega)$.*

Proof. See e.g. BENDSOE [17]. □

Theorem 4.1. *Let the objective functional $\ell(\mathbf{u}(\cdot))$ be lower semi-continuous with $\eta(\rho) = \rho$ and the feasible design set $\mathcal{F}$ weakly* closed in $L_\infty(\Omega)$. Moreover, let $\mathcal{F}$ be non-empty. Then there exists a solution $\overline{\rho}$ to the minimal compliance problem* (4.4).

Proof. For admissible designs $\rho \in \mathcal{F}$ the objective functional is bounded below by $\ell(\mathbf{u}(\rho_{\min}))$. Hence, we can find a minimizing sequence $\{\rho^{(k)}\} \subset \mathcal{F}$ for problem (4.4). Due to the boundedness of $\{\rho^{(k)}\}$ in $L_\infty(\Omega)$ we can extract a subsequence (again denoted by the superscript k) so that $\rho^{(k)} \overset{*}{\rightharpoonup} \overline{\rho}$ for a $\overline{\rho} \in L_\infty(\Omega)$. Lemma 4.1 yields now $\overline{\rho} \in \mathcal{F}$.
By reason of Lemma 4.2, $\ell(\mathbf{u}(\cdot))$ is lower semi-continuous and thus it holds for arbitrary $\rho \in \mathcal{F}$ that

$$\ell\big(\mathbf{u}(\overline{\rho})\big) \leq \liminf_{k\to\infty} \ell\big(\mathbf{u}(\rho^{(k)})\big) \leq \ell\big(\mathbf{u}(\rho)\big).$$

□

We know now that for the minimal compliance problem (4.4) the feasible set $\mathcal{F}$ is weakly* compact in $L_\infty(\Omega)$. However, in contrast to the choice $\eta(\rho) = \rho$, the objective functional $\ell(\mathbf{u}(\cdot))$ is not weakly* lower semi-continuous anymore when the material interpolation function is chosen according to the SIMP or RAMP scheme. A possible remedy would be to use $\eta(\rho) = \rho$ and to penalize intermediate density values with an additional constraint like (4.7). But, (4.7) only defines a set that is closed in a strong sense, which is in fact weaker than weakly* closed, and then the feasible set $\mathcal{F}$ is not weakly* closed anymore (cf. BORRVALL AND PETERSSON [28]).

Below we will present two different regularizing filter methods to restrict the design space. The first method one is called *Regularized Intermediate Density Control* (*RIDC*) and is discussed in detail in BORRVALL AND PETERSSON [28]. Here the penalization constraint (4.7) is modified to become weakly* closed.

Theorem 4.2. *Let $\ell(\mathbf{u}(\cdot))$ with $\eta(\rho) = \eta$ be lower semi-continuous in the weak* topology of $L_\infty(\Omega)$, and let $P_S = P \circ S$ where $P : L_p(\Omega) \to \mathbb{R}$ is strongly semi-continuous and $S : L_p(\Omega) \to L_p(\Omega)$, $1 < p < \infty$, is compact and linear. Then there exists at least one solution to the minimal compliance problem* (4.4) *with the additional constraint $P_S(\rho) \leq \varepsilon_P$.*

Proof. See BORRVALL AND PETERSSON [28]. □

The motivation of the operator S is to convert weakly* convergent sequences into strongly convergent ones. An example of such a compact and linear operator is an integral operator.

Theorem 4.3. *Let p and q be conjugate exponents, $1 < p, q < \infty$. Moreover, let $\phi : \Omega\times\Omega \to \mathbb{R}$ be positive, measurable and such that $\|\phi(\mathbf{x}, \cdot)\|_q < \infty$ for all $\mathbf{x} \in \Omega$. Define $S : L_p(\Omega) \to L_p(\Omega)$ as*

$$S(f) = \int_\Omega \phi(\cdot, \mathbf{y}) f(\mathbf{y})\ d\mathbf{y}.$$

Then S is a linear and compact operator.

Proof. See BORRVALL AND PETERSSON [28]. □

Corollary 4.1. *Let $\ell(\mathbf{u}(\cdot))$ with $\eta(\rho) = \eta$ be lower semi-continuous in the weak* topology of $L_\infty(\Omega)$, and let $P_S = P \circ S$ where $P : L_p(\Omega) \to \mathbb{R}$, $1 < p < \infty$ is strongly semi-continuous and S is the integral operator described in Theorem 4.3. Then there exists at least one solution to the following optimization problem:*

$$\begin{aligned} & \ell\big(\mathbf{u}(\rho)\big) \;\to\; \min_{\rho \in L_\infty(\Omega)} \\ \textit{subject to} \qquad & \int_\Omega \rho(\mathbf{x})\, d\mathbf{x} \;\le\; m_0, \\ & P_S(\rho) \;\le\; \varepsilon_P, \\ & \rho_{\min} \le \rho(\mathbf{x}) \;\le\; 1, \qquad \textit{a.e. in } \Omega. \end{aligned}$$

We know now how to modify the penalization constraint P by the use of a linear compact operator S to obtain a regularized penalization function P_S. Let $S : L_2(\Omega) \to L_2(\Omega)$ be an integral operator defined as

$$S(\rho) = \int_\Omega \phi(\mathbf{x}, \mathbf{y}) \rho(\mathbf{y})\, d\mathbf{y}, \quad \forall\, \mathbf{x} \in \Omega, \tag{4.8}$$

with the kernel

$$\phi(\mathbf{x}, \mathbf{y}) = C(\mathbf{x}) \max\left(0, 1 - \frac{|\mathbf{x} - \mathbf{y}|}{R}\right), \tag{4.9}$$

where ϕ fulfills the requirements in Theorem 4.3 and additionally $\int_\Omega \phi(\mathbf{x}, \mathbf{y})\, d\mathbf{y} = 1$. $C(\mathbf{x})$ is chosen such that the latter is satisfied, i.e. $C(\mathbf{x}) = \big(\int_\Omega \phi(\mathbf{x}, \mathbf{y})\, d\mathbf{y}\big)^{-1}$. Basically this means a linear convolution with a cone of base radius R. Using Fubini's theorem and the symmetry of the kernel ϕ it is possible to show (cf. Borrvall and Petersson [28]) that a penalization of $S(\rho)$ implies a penalization of ρ:

$$0 \le P(\rho) \le (P \circ S)(\rho) \le \varepsilon_P.$$

The filter constraint now looks as the following:

$$P_S(\rho) = \int_\Omega \Big(1 - S\big(\rho(\mathbf{x})\big)\Big)\Big(S\big(\rho(\mathbf{x})\big) - \underline{\rho}\Big)\, d\mathbf{x} \;\le\; \varepsilon_P, \tag{4.10}$$

where a suitable value for ε_P must be found by experiments. This procedure is mostly very expansive. This is a serious disadvantage of the filter method. But on the other hand for problems like minimal compliance it is mathematically well defined.

The second filter technique is used together with the RAMP interpolation scheme (4.6) and was first proposed in Sigmund [121]. Here not the density, but the discrete element sensitivities of an discrete objective $J^h(\boldsymbol{\rho}^h)$ are modified as follows:

$$\widehat{\frac{\partial J^h}{\partial \rho_k^h}} = \frac{1}{\rho_k^h \sum_{i=1}^n H_{i,k}} \sum_{i=1}^n H_{i,k} \rho_i^h \frac{\partial J^h}{\partial \rho_i^h}, \tag{4.11}$$

where the convolution operator $H_{i,k}$ with filter radius R is defined as

$$H_{i,k} = \max\big\{0, R - \operatorname{dist}(i,k)\big\}, \quad \text{for } i, k = 1, \ldots, n.$$

The operator $\mathrm{dist}(i,k)$ represents the distance of the geometrical centroids of finite element k and element i. Roughly speaking, this filter replaces the original derivative by a weighted average of the derivatives of the surrounding area. The advantage of this filter approach is that it is very easy to implement and it turned out to work very well in various different topology optimization problems in 2D and in 3D. Moreover it is very robust with respect to coarse grids. But it must be pointed out that this filter is purely heuristic and it is not quite understood which problem is actually solved. In the following we will call this filter the *mesh-independence filter* and refer to it as *MIF*.

Both filter techniques are able to control the minimal length scale of the components in the optimal design. The larger the filter radius R, the larger is the minimal length scale, or, the thicker are the occurring components, which is important, e.g. to ensure that the optimal structure is not too complicated to be manufactured. This influence of the filter radius can be seen in Figure 4.3.

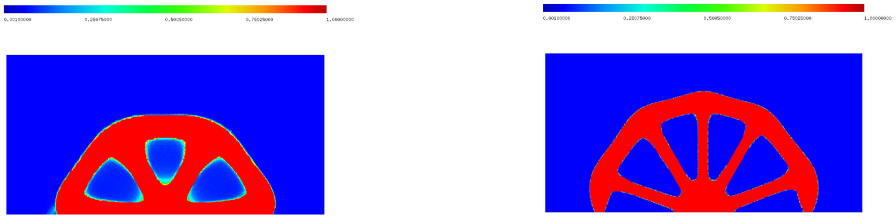

Figure 4.3: Different optimal designs for the same optimization problem w.r.t. different filter radii (RIDC): Left $R = 0.15$ and right $R = 0.05$.

Notes and remarks for Section 4.2

- Before the MIF approach was introduced to topology optimization by Sigmund [121], similar ideas to ensure mesh-independence have been used, e.g. in bone-modeling (cf. Mullender, Huiskes and Weinans [95]).

- In Bourdin [30] another filter method on the density is presented and analyzed. Here no additional constraint is added to the optimization problem, but the dependence of the material properties on the density is replaced by the filtered density.

4.3 An Adaptive Multilevel Approach

Our basic motivation for a multilevel algorithm is to solve the problem efficiently and to save computational costs. This is achieved by solving the problem firstly on a coarse grid to get a first coarse design for rather cheap computational costs. Then we will use this first coarse design as an initial design on a finer grid and repeat the optimization on the finer grid, and so on. As the first coarse optimal design is mostly close to the succeeding finer optimal designs, this procedure will help us to avoid unnecessary long and expansive computations on very fine meshes.

4.3.1 Discretization

When solving problems like (4.4) numerically they are usually discretized using finite elements on a triangulation $\mathcal{T}_h = \{\tau_i \mid i = 1, \dots, n\}$ (cf. Subsection 2.2.3). Following a standard finite element procedure the ground structure Ω is partitioned into $n = \mathcal{O}(h^{-d})$ ($n = n_{el} = n_\rho$) triangles τ_i (or tetrahedrons for $d = 3$), where h is the discretization parameter. It is worth noticing that there are two different variables, the displacements $\mathbf{u}$ and the density ρ. For both variables the same finite element mesh is used, but not the same finite elements.

The density ρ is approximated by a piecewise constant finite element function $\tilde{\rho}$, i.e. $\tilde{\rho}$ is constant over every triangle τ_i. The displacement field $\mathbf{u}$ is approximated using continuous element-wise quadratic functions. The finite element function $\tilde{\mathbf{u}} \in V_0^h$ is now the unique solution of the finite element equations for given feasible $\tilde{\rho} \in Q^h$

$$a(\tilde{\rho}; \tilde{\mathbf{u}}, \tilde{\mathbf{v}}) = \ell(\tilde{\mathbf{v}}), \qquad \forall\ \tilde{\mathbf{v}} \in V_0^h, \tag{4.12}$$

where $V_0^h = \{\tilde{\mathbf{v}} \in \mathcal{P}_2(\mathcal{T}_h) \mid \tilde{\mathbf{v}} \text{ continuous and } \tilde{\mathbf{v}} = 0 \text{ on } \Gamma_u\}$ denotes the finite dimensional subspace of V_0 and $\tilde{\rho} \in Q^h = \mathcal{P}_0(\mathcal{T}_h)$. Here $\mathcal{P}_k(\mathcal{T}_h)$ denotes the space of polynomials of maximal degree k over the triangles τ_i. Whenever mesh refinement is performed, it is done in such a way that $V_0^h \supset V_0^H$ for $h \leq H$. Let the vectors $\mathbf{u}^h \in \mathbb{R}^{n_\mathbf{u}}$ and $\boldsymbol{\rho}^h \in \mathbb{R}^n$ contain the coefficients of the finite element functions $\tilde{\mathbf{u}} \in V_0^h$ and $\tilde{\rho} \in Q^h$, respectively. Then, the discrete analogon of the state equations (4.3b) turns from (4.12) to the following system of linear equations:

$$\mathbf{K}(\boldsymbol{\rho}^h)\mathbf{u}^h = \mathbf{f}^h, \tag{4.13}$$

where $\mathbf{f}^h \in R^{n_\mathbf{u}}$ denotes the load vector. The stiffness matrix $\mathbf{K}(\boldsymbol{\rho}^h)$ depends on the design vector $\boldsymbol{\rho}^h = (\rho_i^h)_{i=1,\dots,n}$ as follows:

$$\mathbf{K}(\boldsymbol{\rho}^h) = \sum_{i=1}^{n} \eta(\rho_i^h)\mathbf{K}_i,$$

where $\mathbf{K}_i$ are the element stiffness matrices extended to $n_u \times n_u$ matrices, which are weighted with the values of the material interpolation function η evaluated at the element densities.

Now, with $\mathbf{u}^h(\boldsymbol{\rho}^h)$ referring to the unique solution of (4.13) for a given feasible design $\boldsymbol{\rho}^h$, the discrete analogon of the objective (4.4a) is ${\mathbf{f}^h}^T\mathbf{K}^{-1}(\boldsymbol{\rho}^h)\mathbf{f}^h = {\mathbf{f}^h}^T(\mathbf{u}^h(\boldsymbol{\rho}^h))$. Furthermore, let the vector $\mathbf{m}^h = (m_i^h)_{i=1,\dots,n}$ represent the volumes the finite elements so that $m_i^h = |\tau_i|$. Then the discrete version of the minimal compliance problem (4.4) can be posed as follows:

$$\begin{aligned} {\mathbf{f}^h}^T\big(\mathbf{u}^h(\boldsymbol{\rho}^h)\big) \ &\to \ \min_{\boldsymbol{\rho}^h \in \mathbb{R}^n} \\ \text{subject to} \qquad {\mathbf{m}^h}^T\boldsymbol{\rho}^h \ &\leq \ m_0, \\ \rho_{\min} \ \leq \ \rho_i^h \ &\leq \ 1, \qquad\qquad i = 1, \dots, n. \end{aligned} \tag{4.14}$$

4.3.2 Adaptive Mesh-Refinement

Elements inside a region, solely occupied by material or void, far away from the structure's boundary, are very unlikely to be affected by the optimization on finer levels. Far more interesting is the interface between material and void, i.e. the boundary of the structure. It is much more efficient to identify this interface and only refine elements along this interface,

instead of an uniform refinement. For identifying the interface the filter operator S, defined in (4.8), turns out to be a useful tool. Consider an arbitrary but fixed point $\overline{\mathbf{x}} \in \Omega$. If the function ρ is locally constant inside the filter region of $\overline{\mathbf{x}}$ (the support $\mathrm{supp}_\phi(\overline{\mathbf{x}})$ of the integral kernel $\phi(\overline{\mathbf{x}}, \cdot)$ in (4.9)), then $S(\rho)(\overline{\mathbf{x}}) = \rho(\overline{\mathbf{x}})$ holds. This is exactly the case for regions where the design is material, i.e. $\rho(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathrm{supp}_\phi(\overline{\mathbf{x}})$, or void, i.e. $\rho(\mathbf{x}) = \rho_{\min}$ for all $\mathbf{x} \in \mathrm{supp}_\phi(\overline{\mathbf{x}})$. If ρ is not constant inside the filter region, $|S(\rho)(\overline{\mathbf{x}}) - \rho(\overline{\mathbf{x}})|$ will have values different from 0. In fact $|S(\rho)(\overline{\mathbf{x}}) - \rho(\overline{\mathbf{x}})| \in [0, 1 - \rho_{\min}]$, since $S(\rho)(\mathbf{x}) \in [\rho_{\min}, 1]$ for $\mathbf{x} \in \Omega$ (cf. Borrvall and Petersson [28]). So, we mark the element τ_i to be refined, if

$$\left| \left(\mathbf{\Phi} \boldsymbol{\rho}^h \right)_i - \rho_i^h \right| \geq \delta_1 > 0, \tag{4.15}$$

for some δ_1 with $1 \gg \delta_1 > 0$. In (4.15) $\mathbf{\Phi} \in \mathbb{R}^{n \times n}$ denotes the convolution matrix corresponding to the integral kernel ϕ. In Figure 4.4 we see an example of the application of this

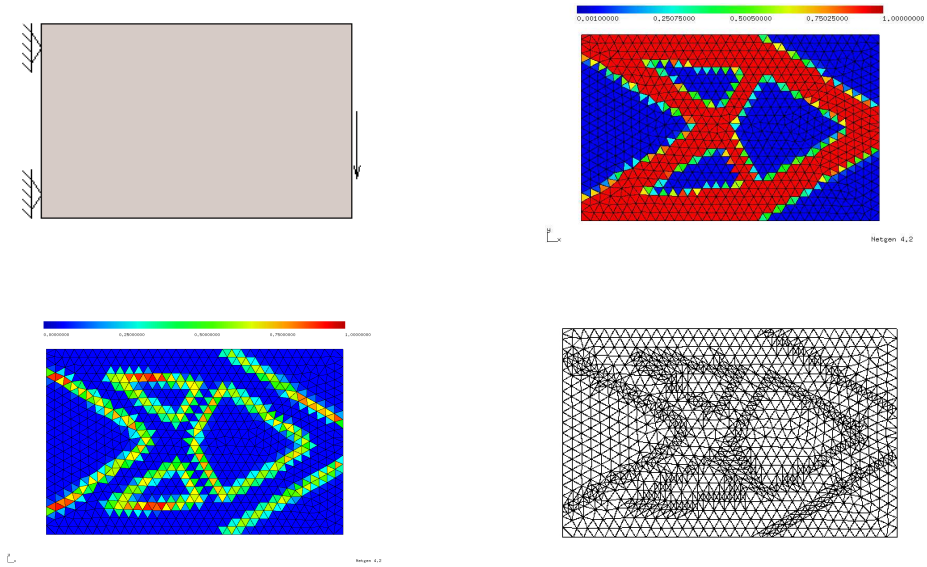

Figure 4.4: Sketch, coarse solution, identified boundary and refined mesh of the cantilever problem in 2D.

refinement idea to the cantilever example in 2D. The lower left picture shows the identified interface using the refinement indicator (4.15) (scaled to [0,1]). Moreover, in Figure 4.5, we see the refinement indicator for the cantilever problem in 3D. Varying the size of the filter radius R_{ref} for the refinement indicator we can control the sensitivity of the indicator with respect to the interface. The larger R_{ref} is chosen, the more elements around the interface will be refined. But of course R_{ref} should be at least greater than the distance of all elements centroids to their at most $d + 1$ adjacent neighboring elements centroids:

$$R_{\text{ref}} = \delta_2 \cdot \max_{i=1,\ldots,n} \left\{ \, \mathrm{dist}(\tau_i, \tau_k) \mid \tau_k \in NH(\tau_i) \, \right\}, \qquad \text{with } \delta_2 > 1,$$

where the set $NH(\tau_i)$ represents the set of the adjacent neighboring elements of the element τ_i. In Figure 4.6 two different refined meshes are shown, resulting from two different refinement filter radii R_{ref}.

For other publications dealing with adaptivity in topology optimization we refer e.g. to Maute, Schwarz and Ramm [92].

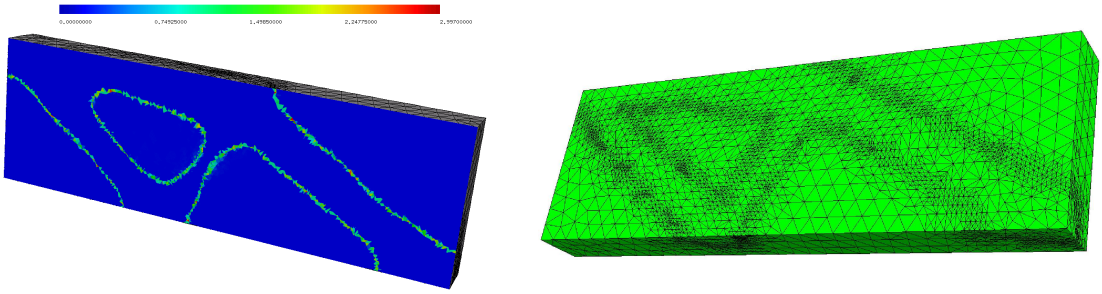

Figure 4.5: Left: The refinement indicator for the 3D cantilever problem. Right: Finally refined mesh of the 3D cantilever problem.

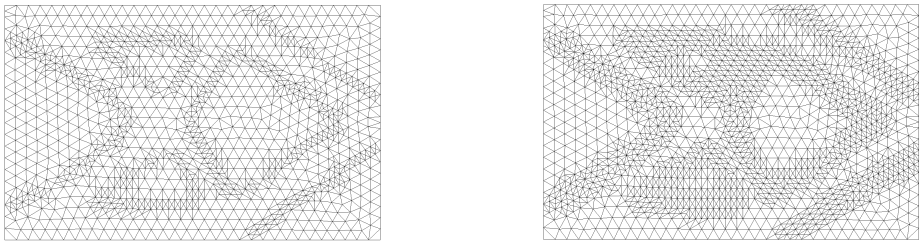

Figure 4.6: Influence of the refinement filter radius R_{ref} on the refined mesh. Original mesh: 1463 elements. Left: Refined mesh with 2625 elements. Right: Refined mesh with 3741 elements.

4.3.3 A Multilevel Approach

In our multilevel approach we basically tried to combine the two filter methods, see Section 4.2, so that their disadvantages are eliminated and only their advantages remain.

Assume that we have a hierarchy of adaptively refined meshes $\mathcal{T}_0 \subset \mathcal{T}_1 \subset \ldots \subset \mathcal{T}_l$ at hand. At the beginning the problem is solved on the coarsest grid $\mathcal{T}_0$. Here, at the first level, we use the mesh-independency filter MIF for regularization together with the RAMP interpolation scheme to penalize intermediate density values, combined with a continuation method. The latter means that the RAMP- parameter q is slowly raised through the optimization progress. In the first few iterations $q = q_0$ is chosen, the for the next ones some higher value, and so on, until a wanted value $q^{\max}$ is reached where the design is then finally fully optimized. The advantage of such a continuation method is that it avoids to get stuck early in an unwanted local minima. That may happen if the calculation is done only with one value of q, which is chosen too large. There are two major reasons why we use the MIF method combined with the RAMP scheme on the coarsest grid. On the one hand we can use coarser grids than with the RIDC method. On the other hand we can use the optimal design $\boldsymbol{\rho}^H$ of the coarsest grid to get a realistic value for ε_P in (4.7), setting $\varepsilon_P = P_S^H(\boldsymbol{\rho}^H)$, where P_S^H denotes the discretization of P_S. This saves costly experiments to find a proper value for ε_P. Although we adapted the MIF formula (4.11) to work also on adaptively refined grids, see (4.16), the filter lost its regularizing properties.

$$\widehat{\frac{\partial J}{\partial \rho_k^h}} = \frac{1}{\rho_k^h \sum_{i=1}^n H_{i,k}|\tau_i|} \sum_{i=1}^n H_{i,k}\rho_i^h \frac{\partial J}{\partial \rho_i^h}|\tau_i|. \tag{4.16}$$

Thus, we continue on the refined grids with the RIDC method, which works fine on unstructured grids and is mathematically well-founded. Moreover, since the effective density in $a(\rho;\cdot,\cdot)$ and the original density ρ are the same ($\eta(\rho) = \rho$), there are no doubts which density is the one to plot. Consequently, we solve, on the coarse level $l = 0$ problem (4.14), and, on higher levels $l \geq 1$, problem (4.14) with the additional penalty constraint (see (4.17)) by an MMA-like algorithm. In each iteration step k of the MMA algorithm we solve the following optimization problem

$$\begin{aligned}
\mathbf{f}^{h^T}\big(\mathbf{u}^h(\boldsymbol{\rho}^h)\big) \;&\rightarrow\; \min_{\boldsymbol{\rho}^h \in \mathbb{R}^n} && \text{(4.17a)}\\
\text{subject to} \qquad \mathbf{m}^{h^T}\boldsymbol{\rho}^h \;&\leq\; m_0, && \text{(4.17b)}\\
\mathbf{p}_k^{h^T}\boldsymbol{\rho}^h + q_k \;&\leq\; \varepsilon_P^l, && \text{(4.17c)}\\
\rho_{\min} \;\leq\; \rho_i^h \;&\leq\; 1, \qquad i = 1,\ldots,n, && \text{(4.17d)}
\end{aligned}$$

for determining $\boldsymbol{\rho}^{h(k+1)}$, where (4.17c) is a linearization of $P_S(\rho)$ with respect to the design variables $\boldsymbol{\rho}^{h(k)}$.

In the optimal coarse grid design the interface $I = \{\mathbf{x} \in \Omega \mid \rho_{\min} < \rho(\mathbf{x}) < 1\}$ between void and material might have a quite significant width (a fuzzy interface). In order to minimize these zones of intermediate densities we reduce ε_P from level to level like $\varepsilon_P^{i+1} = \delta_3 \varepsilon_P^i$ with $0 \leq i < l$ and $0 < \delta_3 \leq 1$. So the initial diffuse interface turns, as i increases, into a sharp interface. Unfortunately, if δ_3 is chosen too small, it may happen that the optimization algorithm is unable to find a feasible design at the next level $i+1$. The following choice of δ_3 turned out to work quite well, in fact it was successful with all test examples:

$$\delta_3 = 1 - \frac{1}{2}\frac{\int_\Omega \chi_I(\mathbf{x})\rho(\mathbf{x})\; d\mathbf{x}}{\int_\Omega \rho(\mathbf{x})\; d\mathbf{x}} \in \big[\tfrac{1}{2}, 1\big],$$

where χ_I denotes the characteristic function of the interface I. The fraction on the right hand side describes the ratio of the mass of the interfacial region and the mass of the overall structure.

Algorithm 4.1 An adaptive multilevel approach

Initialize start value $\boldsymbol{\rho}_0^H$, e.g. like $\boldsymbol{\rho}_0^H = m_0/|\Omega|$
Choose the parameters δ_1 and δ_2 with $0 < \delta_1 \ll 1$ and $1 < \delta_2$ respectively.
$l = 0$;

Coarse grid solution $\boldsymbol{\rho}^H$ with MIF and RAMP;
Determine $\boldsymbol{\varepsilon}_\mathbf{P}^\mathbf{0}$ by $\varepsilon_P^0 = P_S^H(\boldsymbol{\rho}^H)$;
while design not satisfactory **do**
  Mesh-refinement along the interface of void and material. Possible reduction of ε_P: $\varepsilon_P^{l+1} = \delta_3 \varepsilon_P^l$, $0 < \delta_3 \leq 1$;
  Fine grid solution $\boldsymbol{\rho}^h$ using RIDC;
  $l = l + 1$;
end while

4.4 Numerical Experiments

The approach described above was tested with several benchmark examples and we got very good results from all of them. All computations were performed on a computer with a 2.4 GHz CPU and 2 GB memory. Moreover, we used the following parameters for all our numerical experiments: $\delta_1 = 0.1$ and $\delta_2 = 1.1$. For solving the discrete optimization problems (4.14) and (4.17) the method of moving asymptotes was used (cf. Subsection 2.1.2). The finite element assembling and the meshing part was realized with the software package *NETGEN/NGSolve* by SCHÖBERL ET AL. [117], a powerful meshing and finite element software tool. The additional code for the optimization routine including MMA was written in C++ and coupled with NETGEN/NGSolve. In Table 4.1 we list the computational data gained

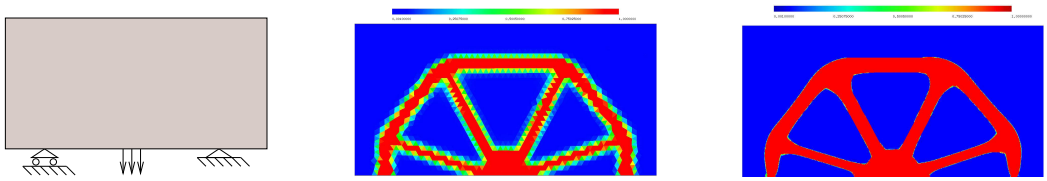

Figure 4.7: Sketch, coarse grid and fine grid solution of the 'wheel' example.

from the 'wheel' example (measurements of the ground structure Ω: 4×2, filter radius $R = 0.1$, volume fraction: $0.25|\Omega|$). In Figure 4.7 we see the sketch, coarse grid solution and the final fine grid solution of the wheel example. The columns n_{el} and $n_{\mathbf{u}}$ contain the number of finite elements and the degrees of freedom with respect to the displacements. The other columns t_{state}, t_{∇}, t_{opt}, t_{fil} and t_{it} show the time used for one evaluation of the state equation, of the derivatives, for the solution of the MMA subproblem, for applying the filter and the overall time per iteration. In the last column the number of needed iterations is listed. The algorithm was stopped at each level when the maximum norm of the difference between two successive designs is less then 0.1 and the relative difference of two successive objective values was less then 10^{-5}. It turned out that this is a sufficiently tight convergence criteria for good design results. For solving the state equation (4.13) we used a multigrid preconditioned conjugate gradient method, where per each iteration a V-cycle with one pre- and post-smoothing step with a Gauß-Seidel smoother is done. Moreover, we computed the 2D examples in the framework of plane strain (cf. Subsection 2.4.1). The adaptive multilevel algorithm was also

l	N_{el}	$N_{\mathbf{u}}$	t_{state}	t_{∇}	t_{opt}	t_{fil}	t_{it}	*Iter.*
level 0:	3334	13666	0.8	0.1	0.1	0.0	1.0	84
level 1:	7654	30974	2.5	0.3	0.2	0.1	3.1	18
level 2:	16877	67920	6.1	0.6	0.5	0.4	8.2	16
level 3:	31280	125554	12.8	1.2	1.2	2.1	19.9	9
level 4:	60833	243796	29.0	2.3	3.1	12.0	59.3	9
level 5:	111397	446072	56.1	4.3	7.9	58.2	186.3	8

Table 4.1: Computational features from the 2D wheel example.

applied to other known 2D examples, where the solutions are shown in Figure 4.8. In all

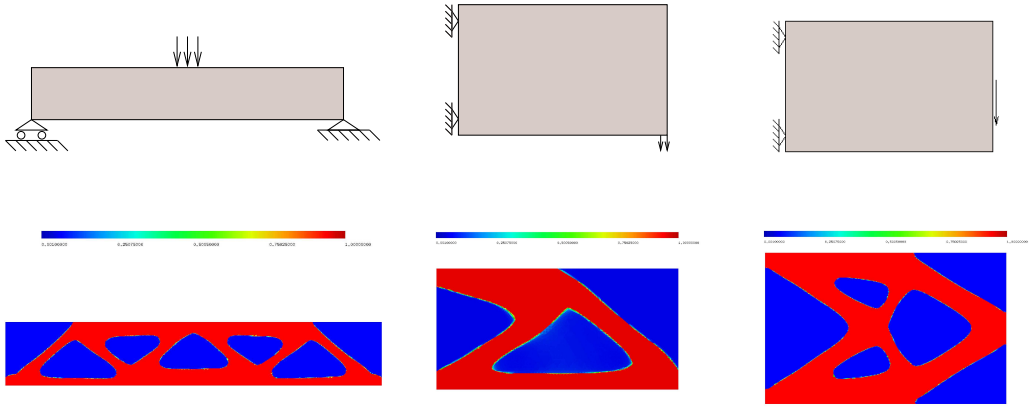

Figure 4.8: Sketches and fine grid solutions of other 2D examples.

these examples the available volume was restricted to $0.5|\Omega|$ and the filter radius was chosen as $R = 0.1$. The measurements of the ground structures of the examples are 6×1, 3.2×2, and 2×1, respectively. The time tables are basically the same as for the wheel example, hence we omit them.

Taking a close look at Table 4.1 it is apparent that the time used for applying the filter is growing significantly (with quadratic order) for very fine meshes, which is still a bottleneck of this approach. So far no sophisticated methods are used to speed up the (sparse) matrix operation. Here, a more efficient way has yet to be investigated for very fine resolutions. Approaches using $\mathcal{H}$-matrices (see e.g. HACKBUSCH [72]), other data-sparse representations techniques, or multipol/multilevel (see e.g. OF AND STEINBACH [97]) techniques are good candidates to overcome this bottleneck.

The same effect appears of course when calculating 3D examples, like the cantilever beam example (measurements of Ω: $16 \times 10 \times 3$, filter radius R = 0.5, volume fraction $0.25|\Omega|$). Due to symmetry, the actual computation was performed only in a quarter of the domain Ω. Figure 4.9 shows sketch, coarse grid solution with 8100 elements and fine grid solution with 1410880 elements. For visualization purposes for the 3D examples (Figures 4.9 and 4.10) we blanked all elements with a density lower than 0.9. Again, in Table 4.2, we list the

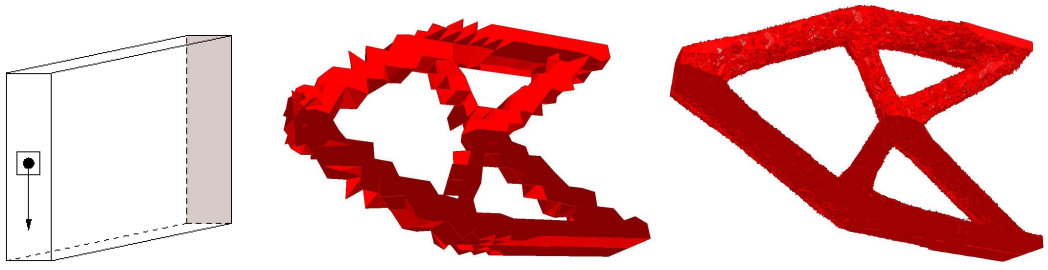

Figure 4.9: The cantilever beam in 3D: Sketch, coarse grid solution and fine grid solution.

computational data of the 3D example. In Figure 4.10 we present another 3D example. As before, the computation was just done in a quarter of the domain. Since the computational

l	N_{el}	$N_{\mathbf{u}}$	t_{state}	t_{∇}	t_{opt}	t_{fil}	t_{it}	*Iter.*
level 0:	1725	9774	1.4	0.2	0.0	0.0	1.7	96
level 1:	8857	41307	11.4	1.2	0.2	0.2	13.4	93
level 2:	49437	214374	76.6	6.7	1.0	10.2	108.5	55
level 3:	189288	794628	330.5	26.1	3.9	160.7	691.8	45

Table 4.2: Computational features from the 3D cantilever beam example.

data is similar to Table 4.2, it is omitted again. Here the coarse grid solution is computed with 30236 (7559) elements and the fine grid solution with 608340 elements corresponding to 608340 (152085) design unknowns and 953964 (238491) displacement unknowns, respectively.

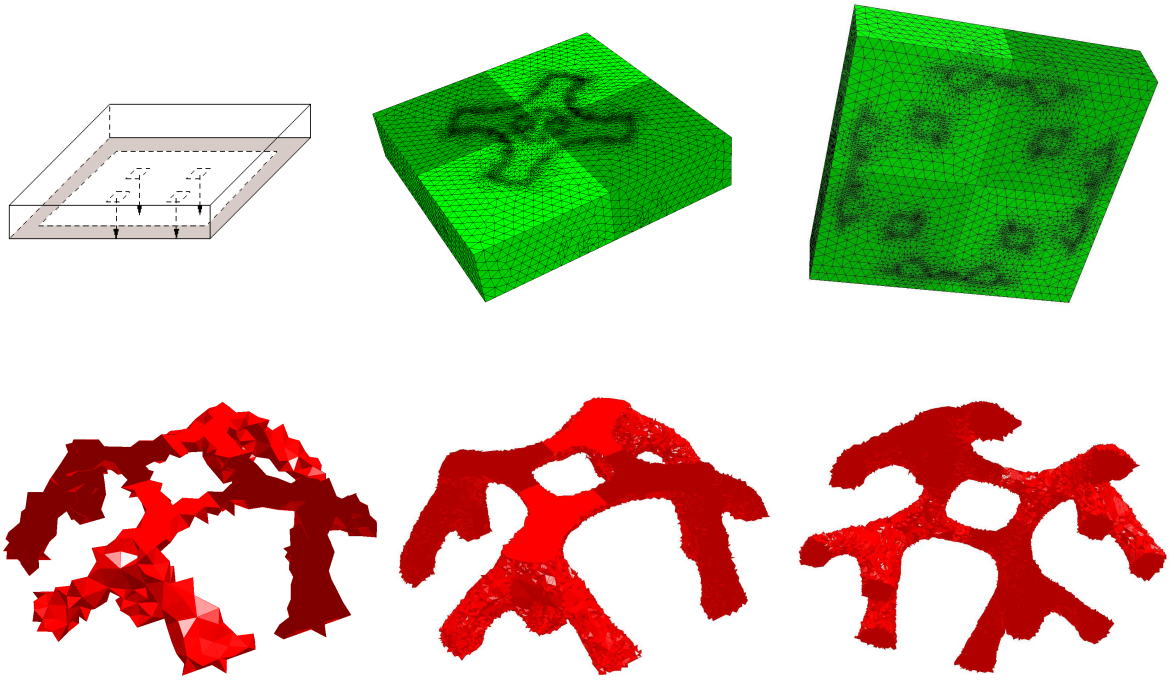

Figure 4.10: The 'roof' example, first line: Sketch and the finally refined mesh, once from above and below. Second line: coarse grid solution and fine grid solution, once from above and once from below.

Chapter 5

Phase–Field Relaxation to Topology Optimization with Local Stress Constraints

We introduce a new relaxation scheme for topology optimization problems with local stress constraints based on a *phase–field* approach. The starting point of the relaxation is a reformulation of the material distribution problem involving linear and 0-1 constraints only. The 0-1 constraints are then relaxed and approximated by a Cahn-Hillard type penalty in the objective functional, which yields convergence of minimizers to 0-1 designs as the related penalty parameter decreases to zero. A major advantage of this kind of relaxation opposed to standard approaches is a uniform constraint qualification that is satisfied for any positive value of the penalization parameter.

The relaxation scheme yields a large-scale optimization problem with a high number of linear inequality constraints. We discretize the problem by finite elements and solve the arising finite-dimensional programming problems by a primal-dual interior–point method. Numerical experiments for problems with local stress constraints based on different stress criteria indicate the success and robustness of the new approach.

The reminder of the chapter is organized as follows: In the first section we give an introduction to the field of topology optimization with local stress constraints and motivate the phase–field method. Then, we consider the reformulation of the constraints, which is the first fundamental of our approach. The second one, the phase–field relaxation, is introduced in the next section, where we also analyze its basic properties. The finite element discretization yielding linearly constrained programming problems is discussed in the section afterwards. Finally, we present numerical results obtained for different stress criteria.

5.1 Introduction

In structural optimization there are two design - constraint combinations of particular importance, namely the maximization of material stiffness (minimizing the compliance) at given mass and the minimization of mass while keeping a certain stiffness. The first combination, also known as the minimal compliance problem, seems to be mathematically well understood and various successful numerical techniques to solve the problem have been proposed. (see also Chapter 4). The treatment of the second problem is by far less understood and until now

there seems to be no approach that is capable of computing reliable (global) optimal designs within reasonable computational effort. The main source of difficulties in this problem is a lack of constraint qualifications for the set of feasible designs, defined by the local stress constraints.

Starting point of our analysis is a reformulation of the equality constraints describing the elastic equilibrium and the local inequality constraints for the stresses into a system of linear inequality constraints as recently proposed by Stolpe and Svanberg [140]. A remaining difficulty is that the arising problem also involves 0-1 constraints in addition to the linear inequalities. The computational effort of methods for the global minimization of these mixed linear programming problems grows fast with the number of degrees of freedom in the discretization, so that the problem could be solved only for very coarse discretizations so far (cf. Stope and Svanberg [140] and Stolpe [135, 136]). Instead of solving mixed linear programming problems we propose to use a phase–field relaxation of the reformulated problem. Due to the well-known ill-posedness of topology optimization problems we might add a perimeter penalization to the objective functional. The phase–field relaxation consists in using a material interpolation function $\eta(\rho) = \rho$, and additionally, a Cahn-Hillard type penalization functional is used to approximate the perimeter.

Let $\Omega_{\text{mat}} = \{\mathbf{x} \in \Omega \mid \rho(\mathbf{x}) = 1\} \subset \Omega \subset \mathbb{R}^d$ $(d = 2, 3)$, denote the optimal design, which is of course initially unknown. Furthermore, let $\Gamma_{t_0} \subset \Gamma_t$ describe the part of the boundary Γ_t where the traction forces are zero, i.e. $\mathbf{t} = \mathbf{0}$. Then, the stress constrained topology optimization problem that we are going to investigate in this chapter states as follows:

$$
\begin{aligned}
J(\rho) = \int_\Omega \rho(\mathbf{x})\, d\mathbf{x} \;&\to\; \min_{\rho,\mathbf{u}} && && (5.1a)\\
\text{subject to} \qquad \operatorname{div} \boldsymbol{\sigma} \;&=\; 0, && \text{in } \Omega_{\text{mat}}, && (5.1b)\\
\boldsymbol{\sigma} - \mathbf{C}\boldsymbol{\varepsilon}(u) \;&=\; \mathbf{0}, && \text{in } \Omega, && (5.1c)\\
\mathbf{u} \;&=\; \mathbf{0}, && \text{on } \Gamma_u, && (5.1d)\\
\boldsymbol{\sigma} \cdot \mathbf{n} \;&=\; \mathbf{t}, && \text{on } \Gamma_t, && (5.1e)\\
\boldsymbol{\sigma} \cdot \mathbf{n} \;&=\; \mathbf{0}, && \text{on } \big(\partial\Omega_{\text{mat}} \setminus \Gamma_t\big) \cup \Gamma_{t_0}, && (5.1f)\\
\rho(\mathbf{x}) \;&\in\; \{0, 1\}, && \text{a.e. in } \Omega, && (5.1g)\\
\Phi^{\min} \;\le\; \Phi\big(\boldsymbol{\sigma}(\mathbf{x})\big) \;&\le\; \Phi^{\max}, && \text{a.e. in } \Omega_{\text{mat}}, && (5.1h)\\
\mathbf{u}^{\min} \;\le\; \mathbf{u}(\mathbf{x}) \;&\le\; \mathbf{u}^{\max}, && \text{a.e. in } \Omega. && (5.1i)
\end{aligned}
$$

Thus, in a first formulation, the objective functional (5.1a) only consists of a mass term. Note, that we only optimize with respect to the design ρ and the displacements $\mathbf{u}$, because the stresses $\boldsymbol{\sigma}$ can be eliminated using the stress-strain relation (5.1c). But for sake of better readability we will keep the stresses in the formulation. The constraints (5.1b) - (5.1f) describe the elasticity equations with corresponding boundary conditions (cf. Section 2.4), where we again neglect bodyforces for sake of simplicity. In an ideal case, the material density ρ only attains two values, 1 for material and 0 for void, see the 0-1 constraint (5.1g). Moreover, the vectors $\mathbf{u}^{\min}$ and $\mathbf{u}^{\max}$ in the bound constraint (5.1i) are lower and upper bounds for the displacements $\mathbf{u}$. In the bound constraints (5.1h), Φ denotes a proper stress criterion. For $\Phi(\boldsymbol{\sigma}) = \boldsymbol{\sigma}$ we have that $\boldsymbol{\sigma}^{\min} \le \boldsymbol{\sigma} \le \boldsymbol{\sigma}^{\max}$ and we shall call this criterion total stress. Alternatively, if the case of von Mises stress constraints is of interest, the local constraints on

$\boldsymbol{\sigma}$ are replaced by

$$\Phi\big(\boldsymbol{\sigma}(\mathbf{x})\big) \leq \Phi^{\max}, \qquad \text{a.e. in } \Omega_{\text{mat}}. \tag{5.2}$$

Since von Mises stress is always non-negative, the lower bound $\Phi^{\min}$ can be omitted. Here the von Mises stress is denoted via the functional $\Phi : \mathbb{R}^{d\times d} \to \mathbb{R}$ given by

$$\Phi(\boldsymbol{\sigma}) = \sqrt{\frac{\sum_{i,j=1}^{d}(\lambda_i - \lambda_j)^2}{2}},$$

where λ_i, $i = 1, \ldots, d$ are the principal stresses (the eigenvalues of $\boldsymbol{\sigma}$) (cf. Han and Reddy [74]). Note that for $d = 2$, the case we are focusing on, we simply have

$$\Phi(\boldsymbol{\sigma}) = |\lambda_2 - \lambda_1| = \sqrt{(\sigma_{11} - \sigma_{22})^2 + 4\sigma_{12}^2}.$$

The two following subsections we will give a short introduction to topology optimization with local stress constraints and to the phase–field method.

5.1.1 Topology Optimization with Local Stress Constraints

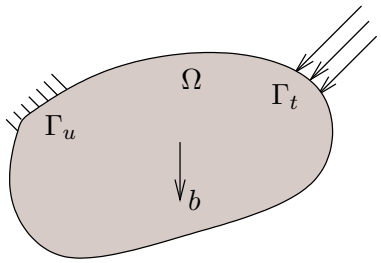

Figure 5.1: The reference domain and applied forces in a minimal mass problem.

Let us again consider a sketch of a usual structural optimization problem in solid mechanics, e.g. Figure 5.1. In difference to the minimal compliance problem, we aim for the lightest structure that operates without material failure under loading. The most abstract way to describe this problem in mathematical terms is probably the following:

$$\begin{aligned} &\text{mass} \;\to\; \min \\ \text{subject to} \qquad &\Phi\big(\boldsymbol{\sigma}(\mathbf{x})\big) \;\leq\; \Phi^{\max}, \qquad \forall\, \mathbf{x} \in \Omega. \end{aligned}$$

But in order to solve this problem efficiently, several challenges must be overcome:

- For a 0-1 formulation, like e.g. (5.1), the stress constraints are well defined. But for intermediate density values, the form of the stress criterion is not a-priori defined. For instance, for a stress criterion for the SIMP model, we refer to Duysinx and Bendsøe [59].
- Treating structural optimization problems with local stress constraints usually results in a large scale optimization problem with a large number of constraints, e.g. two local stress constraints per finite element after discretization.

- The design domain (i.e. the feasible set defined by the constraints in (5.1)) may be nonconvex (even nonconnected) and contain degenerated appendices with lower measure. (Global) optima are very likely to be located in this lower dimensional regions (cf. ROZVANY [109]), where constraint qualifications are lacking. In literature, this effect is often called the *singularity* problem.

- In order to avoid mesh-dependent solution, a proper regularization has to be applied.

Under all these difficulties, the singularity problem is the most severe one. An explanation for this phenomenon is the discontinuous nature of the stress constraints at zero density (cf. CHENG AND JIANG [54]). If the density of an related finite element tends to zero, the corresponding stresses of that element tend to finite values. But the stresses of an element with no density should not taken into account. Since stress constraints should only be imposed if material is present, a reformulation, e.g. for the case of von Mises stress, of the original stress constraints would be

$$\Phi\big(\boldsymbol{\sigma}(\mathbf{x})\big) \leq \Phi^{\max}, \qquad \text{a.e. in } \Omega, \text{ if } \rho(\mathbf{x}) > 0.$$

To avoid the condition $\rho(\mathbf{x}) > 0$ in the constraints (which would yield a non-constant number of constraints throughout the optimization process), the constraints are modified as:

$$\rho(\mathbf{x})\big(\Phi\big(\boldsymbol{\sigma}(\mathbf{x})\big) - \Phi^{\max}\big) \leq 0, \qquad \text{a.e. in } \Omega.$$

But unfortunately, the above reformulation does not eliminate the singularity phenomenon, which is actually rooted in the lack of constraint qualifications (the Slater condition, see e.g. Definition 2.3). If no constraint qualifications are valid, gradient based optimization algorithms, based on the necessary first-order conditions (see Theorem 2.1), cannot reach the optima in the degenerated parts of the design space. In other words, the algorithm is not able to totally remove some low density regions and, as a consequence, to come up with optimal designs.

A remedy is the so called ϵ-relaxation approach proposed in CHENG AND GOU [53]. Here the original constraints are replaced by the following perturbations

$$\rho(\mathbf{x})\big(\Phi\big(\boldsymbol{\sigma}(\mathbf{x})\big) - \Phi^{\max}\big) \leq \epsilon\big(1 - \rho(\mathbf{x})\big), \qquad \epsilon^2 = \rho^{\min} \leq \rho(\mathbf{x}) \leq 1, \text{ a.e. in } \Omega,$$

where the 0-1 constraint (5.1g) is relaxed to $\rho(\mathbf{x}) \in [\rho^{\min}, 1]$. The ϵ-relaxed problems now have a regular design space for all $\epsilon > 0$. The resulting parameter dependent problem is then solved using a continuation approach in ϵ, i.e. the problem is solved for a decreasing sequence $\epsilon \to 0$. Then the related sequence of design domains and of optimal designs converge to the original (degenerated) design domain and to the corresponding optimal design, respectively. For applications of the ϵ-relaxation in continuous topology optimization we refer e.g. to DUYSINX AND BENDSØE [59] and PEREIRA, FANCELLO AND BARCELLOS [101]. But the ϵ-relaxation approach may fail, especially if the relaxed problem has many local minima, as shown in STOLPE AND SVANBERG [139] for a simple truss optimization problem.

Another remedy would be to use one global stress constraint, like a global L_p constraint $\|\Phi(\boldsymbol{\sigma})\|_p \leq \Phi^{\max}$, in contrast to the high number of local stress constraints, see e.g. DUYSINX AND SIGMUND [60]. The computational complexity is then of course much lower, however the global constraint cannot assure that local stress values are below the given limit for all areas in the optimal design. In Figure 1.1, we see the application of a global stress constraint in

an industrial project. In order to ensure that the stresses in the optimal design stay below the a-priori given limit, the 3D elasticity problem for the CAD prototype was solved, and the actual stresses are re-checked.

Notes and remarks for Subsection 5.1.1

- Above, we mentioned references related to truss optimization and continuous topology optimization. As an example for stress constrained material optimization we refer e.g. to Lipton and Stuebner [86].
- "However, the best way to solve stress constrained problems has probably yet to be suggested" (taken from Bendsøe and Sigmund [22]).

5.1.2 The Phase–Field Method

It is well known that optimal design problems are very likely to be ill-posed. Adding a perimeter term to the objective functional regularizes the problem, since it prevents highly oscillating optimal designs, i.e. we minimize

$$\gamma \int_\Omega \rho(\mathbf{x})\ d\mathbf{x} + |\rho|_{BV}, \tag{5.3}$$

with a proper parameter $\gamma > 0$ and for $\rho(\mathbf{x}) \in \{0,1\}$, where the definition of the BV-seminorm is given as follows:

$$|\rho|_{BV} = \sup_{\substack{\mathbf{g}\in C_0^\infty(\Omega;\mathbb{R}^d)\\ \|\mathbf{g}\|_\infty \le 1}} \int_\Omega \operatorname{div} \mathbf{g}(\mathbf{x})\ \rho(\mathbf{x})\ d\mathbf{x}.$$

The *phase–field* method consists now of introducing a continuous density $\rho(\mathbf{x}) \in [0,1]$ and of approximating the perimeter term in (5.3) by a Cahn-Hilliard type penalization functional (cf. Cahn and Hilliard [51]) of the form

$$P_\epsilon(\rho) = \frac{\epsilon}{2} \int_\Omega \left|\nabla \rho(\mathbf{x})\right|^2\ d\mathbf{x} + \frac{1}{\epsilon} \int_\Omega W\big(\rho(\mathbf{x})\big)\ d\mathbf{x}, \tag{5.4}$$

where $W : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is a positive lower semicontinuous function with exactly two roots at 0 and 1. We would like to mention that the penalization term P_ϵ takes only into account the part of the perimeter of Ω_{mat} that lies in the interior of Ω. The composed objective functional then looks as

$$\gamma \int_\Omega \rho(\mathbf{x})\ d\mathbf{x} + P_\epsilon(\rho). \tag{5.5}$$

The first term of the penalty functional P_ϵ controls the perimeter of the level sets of ρ, while the second term ensures that the values of the material density ρ converge to 0 or 1 as $\epsilon \to 0$. The latter means that we use a continuation method in ϵ in order to compute optimal designs $\{\overline{\rho}^{(k)}\}$ with respect to a decreasing sequence $\{\epsilon^{(k)}\}$. Due to a famous result by Modica and Mortola [94] (cf. also Alberti [4]) there exists a subsequence $\{\overline{\rho}^{(k)}\}$ (again denoted by k) of the sequence of minimizers of (5.5) and a subset $\Omega_{\text{mat}} \subset \Omega$, so that $\overline{\rho}^{(k)} \to \chi_{\Omega_{\text{mat}}}$ almost everywhere in Ω. Then $\{\overline{\rho}^{(k)}\}$ converges to a solution of (5.3). In other words, minimizers of P_ϵ with fixed volume $\int_\Omega \rho(\mathbf{x})\ d\mathbf{x}$ converge to minimizers $\overline{\rho}$ of the perimeter at fixed volume over functions satisfying $\rho(\mathbf{x}) \in \{0,1\}$ almost everywhere. This convergence arises in the

framework of Γ-convergence (cf. Braides [34] and the references cited therein), which ensures in particular convergence of minimizers.

Of course, the phase–field method is not the only possible relaxation, one could e.g. use standard material interpolation schemes or level set methods, which are closely related to phase–field methods. However the phase–field approach incorporates some advantages with respect to such approaches:

- In contrast to a direct relaxation to a continuous density variable and in contrast to material interpolation schemes, the phase–field method still provides geometric information. In particular, one can expect for sufficiently small ϵ the set $\{\mathbf{x} \in \Omega \mid \rho(\mathbf{x}) > \delta\}$, with $\delta \ll \frac{1}{2}$, to be a superset of the limit Ω_{mat} and the set $\{\mathbf{x} \in \Omega \mid \rho(\mathbf{x}) < 1 - \delta\}$ to be a subset of Ω_{mat}.

- With the phase–field relaxation one can still use the density linearly ($\eta(\rho) = \rho$), which is not true for material interpolation schemes, like e.g. SIMP and RAMP, or for the level set method. In the latter case, the unknown is a signed distance function to some boundary, and in the relaxation usually an application of a smoothed Heaviside function is used. The additional nonlinearity does not only complicate the problem, but might also destroy constraint qualifications.

- The parameter ϵ can be used for continuation. For ϵ being large, the functional P_ϵ is strictly convex (cf. Theorem 5.2), so that global optima can be computed for arbitrary initial designs. When decreasing ϵ, the optimal design of the previous step can be expected to provide a good initial guess for the next step carried out with a smaller ϵ.

To the knowledge of the author the phase–field method was first introduced in topology optimization by Bourdin and Chambolle [31] for a design problem with design-dependent loads, another type of problem where classical methods encounter difficulties. The method has recently been applied to minimal compliance problems by Wang and Zhou [149].

5.2 Reformulation of Constraints

In this section we consider a reformulation of the constraints on subsets of locally bounded stresses, i.e.

$$\beta\big|\sigma_{ij}(\mathbf{x})\big| \leq 1, \qquad \text{a.e. in } \Omega,\ i,j = 1,\dots,d, \tag{5.6}$$

for some $\beta > 0$ so that $\beta \max_{ij} \sigma_{ij}^{\max} < 1$ and $\beta \min_{ij} \sigma_{ij}^{\min} > -1$. Condition (5.6) is an additional restriction on the design set, which will be needed for the reformulations below. Since the original stress constraint is stronger than (5.6) in regions with material, the additional constraint just states that the stresses should continued in a reasonable (in particular finite) way in regions without material. We recall that the constraints for the displacements $\mathbf{u}$ and

the density ρ are given by

$$\begin{aligned}
\operatorname{div}\boldsymbol{\sigma} &= 0, && \text{in } \Omega_{\mathrm{mat}},\\
\boldsymbol{\sigma} - \mathbf{C}\boldsymbol{\varepsilon}(\mathbf{u}) &= \mathbf{0}, && \text{in } \Omega,\\
\mathbf{u} &= \mathbf{0}, && \text{on } \Gamma_u,\\
\boldsymbol{\sigma}\cdot\mathbf{n} &= \mathbf{t}, && \text{on } \Gamma_t,\\
\boldsymbol{\sigma}\cdot\mathbf{n} &= \mathbf{0}, && \text{on } \big(\partial\Omega_{\mathrm{mat}}\setminus\Gamma_t\big)\cup\Gamma_{t_0},\\
\rho(\mathbf{x}) &\in \{0,1\}, && \text{a.e. in } \Omega,\\
\Phi^{\min} \le \Phi\big(\boldsymbol{\sigma}(\mathbf{x})\big) &\le \Phi^{\max}, && \text{a.e. in } \Omega_{\mathrm{mat}},\\
\mathbf{u}^{\min} \le \mathbf{u}(\mathbf{x}) &\le \mathbf{u}^{\max}, && \text{a.e. in } \Omega.
\end{aligned} \tag{5.7}$$

We will now reformulate the constraints as linear inequality constraints without the unknown set Ω_{mat} for the case of total stress and von Mises stress constraints (cf. Stolpe and Svanberg [140]). For this sake we introduce the set of feasible designs, displacements and stresses as

$$\mathcal{F}_\beta = \Big\{(\rho,\mathbf{u},\boldsymbol{\sigma}) \in BV\big(\Omega;[0,1]\big)\times H^1\big(\Omega;\mathbb{R}^d\big)\times L_\infty\big(\Omega;\mathbb{R}^{d\times d}\big)\ \Big|\ (\rho,\mathbf{u},\boldsymbol{\sigma}) \text{ satisfies (5.6) and (5.7)}\Big\}$$

and an additional artificial stress variable $\mathbf{s}\in L_\infty(\Omega;\mathbb{R}^{d\times d})$.

5.2.1 Reformulation of Total Stress Constraints

We begin with the reformulation in the case of total stress constraints, i.e. $\Phi(\boldsymbol{\sigma})=\boldsymbol{\sigma}$. Let $(\rho,\mathbf{u},\boldsymbol{\sigma})\in\mathcal{F}_\beta$, $\mathbf{x}\in\Omega$ and let $\mathbf{s}(\mathbf{x})=\boldsymbol{\sigma}(\mathbf{x})$ if $\rho(\mathbf{x})=1$ and $\mathbf{s}(\mathbf{x})=\mathbf{0}$ if $\rho(\mathbf{x})=0$, i.e. $\mathbf{s}=\rho\boldsymbol{\sigma}$. Then the constraints

$$-(1-\rho)\mathbf{1} \le \beta(\boldsymbol{\sigma}-\mathbf{s}) \le (1-\rho)\mathbf{1}, \qquad \text{in } \Omega, \tag{5.8}$$

with the matrix $\mathbf{1}=(1)_{ij}$ and

$$\rho(\mathbf{x})\boldsymbol{\sigma}^{\min} \le \mathbf{s}(\mathbf{x}) \le \rho(\mathbf{x})\boldsymbol{\sigma}^{\max}, \qquad \text{a.e. in } \Omega, \tag{5.9}$$

are fulfilled.

Vice versa, assume, that $(\rho,\mathbf{u},\boldsymbol{\sigma},\mathbf{s})\in BV(\Omega;[0,1])\times H^1(\Omega;\mathbb{R}^d)\times L_\infty(\Omega;\mathbb{R}^{d\times d})^2$ fulfills (5.8), (5.9) and $\rho(\mathbf{x})\in\{0,1\}$ almost everywhere in Ω. Let $\mathbf{x}\in\Omega$, then, for $\rho(\mathbf{x})=0$ the inequalities (5.9) imply $\mathbf{s}(\mathbf{x})=\mathbf{0}$ and the inequalities (5.8) yield $-\mathbf{1}\le\beta\boldsymbol{\sigma}(\mathbf{x})\le\mathbf{1}$. For the case $\rho(\mathbf{x})=1$ (5.8) implies $\mathbf{s}(\mathbf{x})=\boldsymbol{\sigma}(\mathbf{x})$ and (5.9) becomes $\boldsymbol{\sigma}^{\min}\le\boldsymbol{\sigma}(\mathbf{x})\le\boldsymbol{\sigma}^{\max}$.

Moreover, since either $\mathbf{s}(\mathbf{x})=\boldsymbol{\sigma}(\mathbf{x})$ or $\mathbf{s}(\mathbf{x})=\mathbf{0}$ almost everywhere in Ω, we obtain that $\operatorname{div}\mathbf{s}(\mathbf{x})=0$ almost everywhere in Ω.

The above arguments yield that $(\rho, \mathbf{u}, \boldsymbol{\sigma}) \in \mathcal{F}_\beta$ if and only if there exists $\mathbf{s}$ so that

$$\begin{aligned}
\operatorname{div} \mathbf{s} &= 0, && \text{in } \Omega, && (5.10a)\\
\boldsymbol{\sigma} - \mathbf{C}\boldsymbol{\varepsilon}(\mathbf{u}) &= \mathbf{0}, && \text{in } \Omega, && (5.10b)\\
\mathbf{u} &= \mathbf{0}, && \text{on } \Gamma_u, && (5.10c)\\
\mathbf{s} \cdot \mathbf{n} &= \mathbf{t}, && \text{on } \Gamma_t, && (5.10d)\\
\mathbf{s} \cdot \mathbf{n} &= \mathbf{0}, && \text{on } \Gamma_{t_0}, && (5.10e)\\
-(1-\rho)\mathbf{1} \le \beta(\boldsymbol{\sigma} - \mathbf{s}) &\le (1-\rho)\mathbf{1}, && \text{in } \Omega, && (5.10f)\\
\rho(\mathbf{x}) &\in \{0,1\}, && \text{a.e. in } \Omega, && (5.10g)\\
\rho(\mathbf{x})\boldsymbol{\sigma}^{\min} \le \mathbf{s}(\mathbf{x}) &\le \rho(\mathbf{x})\boldsymbol{\sigma}^{\max}, && \text{a.e. in } \Omega, && (5.10h)\\
\mathbf{u}^{\min} \le \mathbf{u}(\mathbf{x}) &\le \mathbf{u}^{\max}, && \text{a.e. in } \Omega. && (5.10i)
\end{aligned}$$

We mention that the conditions (5.10a), (5.10d), and (5.10e) have to be understood in a weak sense, namely as

$$\int_\Omega \mathbf{s} : \boldsymbol{\varepsilon}(\mathbf{v}) \, d\mathbf{x} = \int_{\Gamma_t} \mathbf{v} \cdot \mathbf{t} \, d\mathbf{a}, \qquad \forall \, \mathbf{v} \in H^1_{\Gamma_u}.$$

Note that, except of $\rho(\mathbf{x}) \in \{0,1\}$ almost everywhere in Ω, the constraints (5.10) are linear with respect to the new vector of unknowns $(\rho, \mathbf{u}, \boldsymbol{\sigma}, \mathbf{s})$. In particular, all constraints are formulated on Ω and not on the a-priori unknown set Ω_{mat}.

5.2.2 Reformulation of Von Mises Stress Constraints

In the following we discuss the reformulation of the inequalities in the case of von Mises stress constraints for spatial dimension $d = 2$ (similar arguing is possible for $d = 3$). Since both sides of (5.2) are positive, we can square them and since the constraint must only hold for $\rho(\mathbf{x}) = 1$, it can be written equivalently as

$$\rho(\mathbf{x})(\sigma_{11} - \sigma_{22})^2 + 4\rho(\mathbf{x})\sigma_{12}^2 \le \rho(\mathbf{x})\Phi^{\max^2}, \qquad \text{a.e. in } \Omega_{\text{mat}}.$$

As in the case of total stress we introduce an artificial stress variable $\mathbf{s}(\mathbf{x}) = \rho(\mathbf{x})\boldsymbol{\sigma}(\mathbf{x})$, and since $\rho(\mathbf{x})^2 = \rho(\mathbf{x})$ due to $\rho(\mathbf{x}) \in \{0,1\}$, we can reformulate the inequality above as

$$\big(\mathbf{s}_{11}(\mathbf{x}) - \mathbf{s}_{22}(\mathbf{x})\big)^2 + 4\big(\mathbf{s}_{12}(\mathbf{x})\big)^2 \le \rho(\mathbf{x})\Phi^{\max^2}, \qquad \text{a.e. in } \Omega. \tag{5.11}$$

This reformulation of the von Mises stress constraints results in the following set of constraints:

$$\begin{aligned}
\operatorname{div} \mathbf{s} &= 0, && \text{in } \Omega,\\
\boldsymbol{\sigma} - \mathbf{C}\boldsymbol{\varepsilon}(\mathbf{u}) &= \mathbf{0}, && \text{in } \Omega,\\
\mathbf{u} &= \mathbf{0}, && \text{on } \Gamma_u,\\
\mathbf{s} \cdot \mathbf{n} &= \mathbf{t}, && \text{on } \Gamma_t,\\
\mathbf{s} \cdot \mathbf{n} &= \mathbf{0}, && \text{on } \Gamma_{t_0},\\
-(1-\rho)\mathbf{1} \le \beta(\boldsymbol{\sigma} - \mathbf{s}) &\le (1-\rho)\mathbf{1}, && \text{in } \Omega,\\
\big(\mathbf{s}_{11}(\mathbf{x}) - \mathbf{s}_{22}(\mathbf{x})\big)^2 + 4\big(\mathbf{s}_{12}(\mathbf{x})\big)^2 &\le \rho(\mathbf{x})\Phi^{\max^2}, && \text{a.e. in } \Omega,\\
\rho(\mathbf{x}) &\in \{0,1\}, && \text{a.e. in } \Omega,\\
\mathbf{u}^{\min} \le \mathbf{u}(\mathbf{x}) &\le \mathbf{u}^{\max}, && \text{a.e. in } \Omega.
\end{aligned} \tag{5.12}$$

Note that for the reformulation (5.11) the von Mises stress constraint does not imply directly $\mathbf{s}(\mathbf{x}) = \mathbf{0}$ for the case $\rho(\mathbf{x}) = 0$, but only $s_{11}(\mathbf{x}) - s_{22}(\mathbf{x}) = 0$ and $s_{12}(\mathbf{x}) = 0$. But additionally, $\mathbf{s}$ is divergence free, and we may conclude for those spatial stresses that $\nabla s_{11}(\mathbf{x}) = \mathbf{0}$. Hence, $\mathbf{s}(\mathbf{x})$ is of the form

$$\mathbf{s}(\mathbf{x}) = \rho(\mathbf{x})\boldsymbol{\sigma}(\mathbf{x}) + \big(1 - \rho(\mathbf{x})\big) \begin{pmatrix} c & 0 \\ 0 & c \end{pmatrix}, \qquad \text{a.e. in } \Omega,$$

for any constant $c \in \mathbb{R}$. Since such artificial stresses will not change the von Mises stress, and since $\mathbf{s}(\mathbf{x})$ does not have a physical meaning for $\mathbf{x} \in \Omega$ with $\rho(\mathbf{x}) = 0$, the additional terms will not play a major role (we are basically free to define $\mathbf{s}(\mathbf{x})$ for $\mathbf{x} \in \Omega \setminus \Omega_{\text{mat}}$). The reformulation (5.11) involves convex quadratic constraints, but only with respect to the stress variable $\mathbf{s}$, which still yields constraint qualification after relaxation and discretization.

But for the approach in this chapter we rather use a conservative version of the von Mises criterion (cf. Stolpe and Svanberg [140]) that leads to linear constraints after reformulation and is given by

$$\rho(\mathbf{x})\big|\sigma_{11}(\mathbf{x}) - \sigma_{22}(\mathbf{x})\big| + 2\rho(\mathbf{x})\big|\sigma_{12}(\mathbf{x})\big| \leq \rho(\mathbf{x})\Phi^{\max}, \qquad \text{a.e. in } \Omega_{\text{mat}},$$

and can actually be reformulated into linear inequalities. Again use $\mathbf{s} = \rho\boldsymbol{\sigma}$ to obtain

$$\big|s_{11}(\mathbf{x}) - s_{22}(\mathbf{x})\big| + 2\big|s_{12}(\mathbf{x})\big| \leq \rho(\mathbf{x})\Phi^{\max}, \qquad \text{a.e. in } \Omega. \tag{5.13}$$

Moreover, we introduce functions p and q such that

$$0 \leq p(\mathbf{x}), \qquad 0 \leq q(\mathbf{x}), \qquad \text{a.e. in } \Omega, \tag{5.14}$$

and that

$$-p(\mathbf{x}) \leq s_{11}(\mathbf{x}) - s_{22}(\mathbf{x}) \leq p(\mathbf{x}), \qquad -q(\mathbf{x}) \leq s_{12}(\mathbf{x}) \leq q(\mathbf{x}), \qquad \text{a.e. in } \Omega. \tag{5.15}$$

These conditions yield that

$$\big|s_{11}(\mathbf{x}) - s_{22}(\mathbf{x})\big| \leq p(\mathbf{x}), \qquad \text{and} \qquad \big|s_{12}(\mathbf{x})\big| \leq q(\mathbf{x}), \qquad \text{a.e. in } \Omega.$$

Finally we impose the constraints

$$p(\mathbf{x}) + 2q(\mathbf{x}) \leq \rho(\mathbf{x})\Phi^{\max}, \qquad \text{a.e. in } \Omega, \tag{5.16}$$

and obtain as a consequence of the conservative von Mises stress (5.13) for $\mathbf{s}$. On the other hand, if the constraint (5.13) is satisfied by $\mathbf{s}$, the functions

$$\begin{aligned} p(\mathbf{x}) \;&=\; \max\big\{ -\big(s_{11}(\mathbf{x}) - s_{22}(\mathbf{x})\big),\, s_{11}(\mathbf{x}) - s_{22}(\mathbf{x})\big\}, && \text{a.e. in } \Omega, \\ q(\mathbf{x}) \;&=\; \max\big\{ -s_{12}(\mathbf{x}),\, s_{12}(\mathbf{x})\big\}, && \text{a.e. in } \Omega, \end{aligned}$$

satisfy the new linear constraints (5.14) - (5.16). Thus, we conclude the equivalence to the original conservative von Mises stress constraints. The reformulation of the conservative von

Mises stress constraints results in the following set of constraints:

$$\begin{aligned}
\operatorname{div}\mathbf{s} &= 0, && \text{in } \Omega,\\
\boldsymbol{\sigma} - \mathbf{C}\boldsymbol{\varepsilon}(\mathbf{u}) &= \mathbf{0}, && \text{in } \Omega,\\
\mathbf{u} &= \mathbf{0}, && \text{on } \Gamma_u,\\
\mathbf{s}\cdot\mathbf{n} &= \mathbf{t}, && \text{on } \Gamma_t,\\
\mathbf{s}\cdot\mathbf{n} &= \mathbf{0}, && \text{on } \Gamma_{t_0},\\
-(1-\rho)\mathbf{1} \;\le\; \beta(\boldsymbol{\sigma}-\mathbf{s}) &\le (1-\rho)\mathbf{1}, && \text{in } \Omega,\\
-p(\mathbf{x}) \;\le\; s_{11}(\mathbf{x}) - s_{22}(\mathbf{x}) &\le p(\mathbf{x}), && \text{a.e. in } \Omega,\\
-q(\mathbf{x}) \;\le\; s_{12}(\mathbf{x}) &\le q(\mathbf{x}), && \text{a.e. in } \Omega,\\
p(\mathbf{x}) + 2q(\mathbf{x}) &\le \rho(\mathbf{x})\Phi^{\max}, && \text{a.e. in } \Omega,\\
p(\mathbf{x}) \ge 0, \qquad q(\mathbf{x}) &\ge 0, && \text{a.e. in } \Omega,\\
\rho(\mathbf{x}) &\in \{0,1\}, && \text{a.e. in } \Omega,\\
\mathbf{u}^{\min} \;\le\; \mathbf{u}(\mathbf{x}) &\le \mathbf{u}^{\max}, && \text{a.e. in } \Omega.
\end{aligned} \tag{5.17}$$

A drawback of the reformulation is an increase in the number of unknowns and a high number of inequality constraints. On the other hand, this higher number of unknowns and constraints seems to be a reasonable price for the linear reformulation on Ω of the complicated original constraints (5.7). Note, that the reformulation is only possible due to the simultaneous formulation in ρ, $\mathbf{u}$ and $\boldsymbol{\sigma}$. As a consequence, the elasticity equations are formulated like in the Hellinger-Reissner principle (cf. Subsection 2.4.2).

5.3 Phase–Field Relaxation

All the constraints in (5.10) and in (5.17) are linear with respect to the vector of unknowns $(\rho, \mathbf{u}, \boldsymbol{\sigma}, \mathbf{s})$, except for $\rho(\mathbf{x}) \in \{0,1\}$ almost everywhere in Ω. In this section we will turn our attention to the relaxation of the minimal mass problems with the constraints (5.10) and (5.17). For this sake we replace the 0-1 constraint $\rho(\mathbf{x}) \in \{0,1\}$ by the following continuous version $\rho(\mathbf{x}) \in [0,1]$. Moreover, we approximate the perimeter term in the regularized objective functional (5.3) by the Cahn-Hilliard term P_ϵ (5.4):

$$J_\epsilon(\rho) = \gamma \int_\Omega \rho(\mathbf{x})\; d\mathbf{x} + \frac{\epsilon}{2}\int_\Omega \big|\nabla\rho(\mathbf{x})\big|^2\; d\mathbf{x} + \frac{1}{\epsilon}\int_\Omega W\big(\rho(\mathbf{x})\big)\; d\mathbf{x}. \tag{5.18}$$

The term $\int_\Omega W(\rho(\mathbf{x})\; d\mathbf{x}$ favorites those designs which take values close to 0 or 1 (*phase separation*), while the term $\int_\Omega |\nabla\rho(\mathbf{x})|^2\; d\mathbf{x}$ penalizes the spatial inhomogeneity of ρ. The transition between the phases occurs in a thin layer, in fact in thickness of order ϵ. When ϵ is small, the last term in (5.18) prevails, and the minimum of P_ϵ is attained by a function which takes mainly values close to 0 or 1. The theorem of Modica and Mortola tells that the minimizers of (5.18) converge to the minimizers of $\int_\Omega \rho(\mathbf{x})\; d\mathbf{x}$ in the sense of Γ-convergence.

Definition 5.1 (Γ-Convergence)**.** *Let X be a metric space, and for $\epsilon > 0$ let $F_\epsilon : X \to [0, +\infty]$ be given. F_ϵ Γ-converges to F on X as $\epsilon \to 0$, written as $F_\epsilon \xrightarrow{\Gamma} F$, if the following two conditions hold: Lower bound inequality: For every $u \in X$ and every sequence $\{u_\epsilon\}$ so that $u_\epsilon \to u$ in X there holds*

$$\liminf_{\epsilon\to 0} F_\epsilon\big(u_\epsilon\big) \ge F(u). \tag{5.19}$$

Upper bound inequality: For every $u \in X$ there exists $\{u_\epsilon\}$ so that $u_\epsilon \to u$ in X and

$$\lim_{\epsilon \to 0} F_\epsilon\big(u_\epsilon\big) = F(u). \tag{5.20}$$

When (5.19) *holds, then equality* (5.20) *can be replaced by* $\limsup_{\epsilon \to 0} F_\epsilon\big(u_\epsilon\big) \leq F(u)$.

The notion of Γ-convergence has the following properties (cf. Alberti [4] or Dal Maso [90]):

Remark 5.1 (Properties of Γ-convergence)**.**

a) The Γ-limit F is always lower semicontinuous on X.

b) Stability under continuous perturbations: If $F_\epsilon \xrightarrow{\Gamma} F$ and G is continuous, then $F_\epsilon + G \xrightarrow{\Gamma} F + G$.

c) Stability of minimizing sequences: If $F_\epsilon \xrightarrow{\Gamma} F$ and u_ϵ minimizes F_ϵ over X, then every cluster point of $\{u_\epsilon\}$ minimizes F over X.

Let X denote the space of all measurable functions $u : \Omega \to [0,1]$, so that $\int_\Omega u(\mathbf{x})\, d\mathbf{x} < |\Omega|$, endowed with the L_1-norm. Furthermore, let Su denote the measure theoretic boundary of $\{u = 1\}$ in Ω and $\mathcal{H}^{d-1}$ the $(d-1)$-dimensional Hausdorff measure. Then we can state the following theorem:

Theorem 5.1 (Modica and Mortola)**.** *Set $\alpha := 2\int_0^1 \sqrt{W(y)}\, dy$, and for every $\epsilon > 0$ let*

$$P_\epsilon(u) := \begin{cases} \epsilon \displaystyle\int_\Omega \big|\nabla u(\mathbf{x})\big|^2\, d\mathbf{x} + \frac{\epsilon}{2} \int_\Omega W\big(u(\mathbf{x})\big)\, d\mathbf{x} & \text{if } u \in H^1(\Omega) \cap X, \\ +\infty & \text{elsewhere in } X, \end{cases}$$

and

$$P(u) := \begin{cases} \alpha \mathcal{H}^{d-1}(Su) & \text{if } u \in BV\big(\Omega; \{0,1\}\big) \cap X, \\ +\infty & \text{elsewhere in } X, \end{cases}$$

where the parameter α is called the surface tension *between the two phases. Then the functionals P_ϵ Γ-converge to P in X.*

Moreover, let the sequences $\{\epsilon^{(k)}\}$ and $\{u^{(k)}\}$ be given so that $\epsilon^{(k)} \to 0$ and $P_{\epsilon^{(k)}}(u^{(k)})$ is bounded. Then $\{u^{(k)}\}$ is pre-compact in X.

Proof. See Modica and Mortola [94]. □

Minimizing P over X means finding a set $\Omega_{\text{mat}} \subset \Omega$ among those with prescribed volume which minimizes the $(d-1)$-dimensional area of $\partial\Omega_{\text{mat}} \cap \Omega$. Note, that Theorem 5.1 reduces to the following three statements (see Alberti [4]):

Remark 5.2.

a) Compactness: Let the sequences $\{\epsilon^{(k)}\}$ and $\{u^{(k)}\}$ be given so that $\epsilon^{(k)} \to 0$ and $P_{\epsilon^{(k)}}(u^{(k)})$ is bounded. Then $\{u^{(k)}\}$ is pre-compact in $L_1(\Omega)$ and every limit point belongs to $BV(\Omega; \{0,1\})$.

b) Lower bound inequality: If $u \in BV(\Omega, \{0,1\})$, $\{u^{(k)}\} \subset H^1(\Omega)$ *and* $u^{(k)} \to u$ *in* $L_1(\Omega)$ *then*

$$\liminf_{\epsilon \to 0} P_\epsilon\big(u^{(k)}\big) \geq P(u).$$

c) Upper bound inequality: For every $u \in BV(\Omega, \{0,1\})$, *exists* $\{u^{(k)}\} \subset H^1(\Omega)$ *so that* $u^{(k)} \to u$ *in* $L_1(\Omega)$, $\int_\Omega u^{(k)}(\mathbf{x})\, d\mathbf{x} = \int_\Omega u(\mathbf{x})\, d\mathbf{x}$ *for every* ϵ *and*

$$\limsup_{\epsilon \to 0} P_\epsilon\big(u^{(k)}\big) \leq P(u).$$

The resulting relaxation in the case of total stress constraints is given by:

$$\begin{aligned}
J_\epsilon(\rho) \;&\to\; \min_{\rho,\mathbf{u},\mathbf{s}} && \\
\operatorname{div} \mathbf{s} \;&=\; 0, && \text{in } \Omega, \\
\boldsymbol{\sigma} - \mathbf{C}\boldsymbol{\varepsilon}(\mathbf{u}) \;&=\; \mathbf{0}, && \text{in } \Omega, \\
\mathbf{u} \;&=\; \mathbf{0}, && \text{on } \Gamma_u, \\
\mathbf{s} \cdot \mathbf{n} \;&=\; \mathbf{t}, && \text{on } \Gamma_t, \\
\mathbf{s} \cdot \mathbf{n} \;&=\; \mathbf{0}, && \text{on } \Gamma_{t_0}, \\
\rho \;&=\; 1, && \text{on } \Gamma_t, \\
-(1-\rho)\mathbf{1} \;\leq\; \beta(\boldsymbol{\sigma} - \mathbf{s}) \;&\leq\; (1-\rho)\mathbf{1}, && \text{in } \Omega, \\
0 \;\leq\; \rho(\mathbf{x}) \;&\leq\; 1, && \text{a.e. in } \Omega, \\
\rho(\mathbf{x})\boldsymbol{\sigma}^{\min} \;\leq\; \mathbf{s}(\mathbf{x}) \;&\leq\; \rho(\mathbf{x})\boldsymbol{\sigma}^{\max}, && \text{a.e. in } \Omega, \\
\mathbf{u}^{\min} \;\leq\; \mathbf{u}(\mathbf{x}) \;&\leq\; \mathbf{u}^{\max}, && \text{a.e. in } \Omega.
\end{aligned} \tag{5.21}$$

The function space setting for the relaxed problem (5.21) is given by

$$(\rho, \mathbf{u}, \boldsymbol{\sigma}, \mathbf{s}) \in \big(H^1(\Omega) \cap L_\infty(\Omega)\big) \times \big(H^1\big(\Omega; \mathbb{R}^d\big) \cap L_\infty\big(\Omega; \mathbb{R}^d\big)\big) \times L_\infty\big(\Omega; \mathbb{R}^{d \times d}\big)^2.$$

Note, that in addition to the reformulated set of constraints (5.10), there is a Dirichlet boundary condition for ρ a.e. on Γ_t, which is well-defined in the sense of traces of functions in $H^1(\Omega)$. The reasoning of adding this constraint is as follows: In the original set of constraints we have that $\mathbf{s} \cdot \mathbf{n} = \mathbf{t}$ on Γ_t. Hence, there exists a small open neighbourhood of Γ_t in Ω, where $\rho = 1$. Otherwise, $\rho = 0$ would imply $\mathbf{s} = \mathbf{0}$ and therefore $\mathbf{s} \cdot \mathbf{n} = \mathbf{0}$ on Γ_t. In the relaxed formulation, the trace of ρ is positive by analogous arguments, but not necessarily equal to one. Thus, the additional constraint will not change the limit of the constraint set. But on the other hand, it restricts the relaxation and simplifies the analysis of the relaxed problem.

In a similar way, the relaxed formulation of the problem with conservative von Mises stress

($d = 2$) is stated as

$$
\begin{aligned}
J_\epsilon(\rho) \;&\to\; \min_{\rho,\mathbf{u},\mathbf{s},p,q} \\
\operatorname{div}\mathbf{s} \;&=\; 0, && \text{in } \Omega, \\
\boldsymbol{\sigma} - \mathbf{C}\boldsymbol{\varepsilon}(\mathbf{u}) \;&=\; \mathbf{0}, && \text{in } \Omega, \\
\mathbf{u} \;&=\; \mathbf{0}, && \text{on } \Gamma_u, \\
\mathbf{s}\cdot\mathbf{n} \;&=\; \mathbf{t}, && \text{on } \Gamma_t, \\
\mathbf{s}\cdot\mathbf{n} \;&=\; \mathbf{0}, && \text{on } \Gamma_{t_0}, \\
\rho \;&=\; 1, && \text{on } \Gamma_t, \\
-(1-\rho)\mathbf{1} \;\le\; \beta(\boldsymbol{\sigma}-\mathbf{s}) \;&\le\; (1-\rho)\mathbf{1}, && \text{in } \Omega, \\
-p(\mathbf{x}) \;\le\; s_{11}(\mathbf{x}) - s_{22}(\mathbf{x}) \;&\le\; p(\mathbf{x}), && \text{a.e. in } \Omega, \\
-q(\mathbf{x}) \;\le\; s_{12}(\mathbf{x}) \;&\le\; q(\mathbf{x}), && \text{a.e. in } \Omega, \\
p(\mathbf{x}) + 2q(\mathbf{x}) \;&\le\; \rho(\mathbf{x})\Phi^{\max}, && \text{a.e. in } \Omega, \\
p(\mathbf{x}) \;\ge\; 0, \qquad q(\mathbf{x}) \;&\ge\; 0, && \text{a.e. in } \Omega, \\
0 \;\le\; \rho(\mathbf{x}) \;&\le\; 1 && \text{a.e. in } \Omega, \\
\mathbf{u}^{\min} \;\le\; \mathbf{u}(\mathbf{x}) \;&\le\; \mathbf{u}^{\max}, && \text{a.e. in } \Omega.
\end{aligned}
\tag{5.22}
$$

with the function space setting

$$
(\rho, \mathbf{u}, \boldsymbol{\sigma}, \mathbf{s}, p, q) \in \big(H^1(\Omega) \cap L_\infty(\Omega)\big) \times \big(H^1\big(\Omega;\mathbb{R}^d\big) \cap L_\infty\big(\Omega;\mathbb{R}^d\big)\big) \times L_\infty\big(\Omega;\mathbb{R}^{d\times d}\big)^2 \times L_\infty(\Omega)^2.
$$

The basic of the relaxation idea is the Γ-convergence of problem (5.21) to (5.10) as ϵ tends to zero (respectively Γ-convergence of (5.22) to (5.12)) for the von Mises stress constraints). This can be derived form the Γ-convergence of the perimeter term P_ϵ to which only a continuous function is added (compare Remark 5.1).

So far we have not discussed possible choices for the function W in (5.4). Commonly used in phase–field simulations of phase-transitions problems (e.g. in the Allen-Cahn and Cahn-Hilliard equation, cf. e.g. Barles, Soner and Souganidis [13] and Caginalp and Socolovsky [50]) is the *double-well potential*:

$$
W(y) = y^2(1-y)^2, \qquad y \in \mathbb{R}.
$$

Recently, the so-called *double-obstacle potential*

$$
W(y) = y(1-y), \qquad y \in [0,1], \tag{5.23}
$$

has received further attention (cf. Blowey and Elliott [25]). In the case of evolutions like the Allen-Cahn equation, the use of the double-obstacle potential is rather a computational complication, because $\rho(\mathbf{x}) \in [0,1]$ has to be enforced, in contrast to the evolution with the double-well potential. Moreover, the partial differential equation has to be reformulated as a variational inequality. In our case, the choice of the double-obstacle potential (5.23) seems more attractive, since we enforce the bound constraints on ρ in (5.21) and (5.22) anyway. Moreover, the double-obstacle potential causes a polynomial nonlinearity of lower degree than the double-well potential.

Obviously there is no restriction just to use two phases, 0 (void) and 1 (material), e.g. with

$$W(y) = \prod_{i=1}^{n} (y - \overline{y}_i)^2$$

separates the phases $\overline{y}_i$, $i = 1, \ldots, n$.

In the rest of the section we further examine the structure of the objective functional of the relaxed problems. Due to the second term in the Cahn-Hilliard functional P_ϵ, we have to expect the objective functional $J_\epsilon(\rho)$ to be nonconvex. In particular, for small ϵ, when minimizers are forced to take values close to 0 or 1. For large values of ϵ, the first term in P_ϵ dominates and thus, the objective functional, and as a consequence, the optimization problem is convex.

Theorem 5.2. *Let $W \in C^2([0,1])$. Then there exists $\epsilon_0 > 0$ dependent on the ground structure Ω only, so that the objective functional $J_\epsilon(\rho)$ is convex for all $\epsilon > \epsilon_0$.*

Proof. The objective functional $J_\epsilon(\rho)$ is twice continuously differentiable with the derivatives

$$J'_\epsilon(\rho)\psi = \gamma \int_\Omega \psi(\mathbf{x}) \, d\mathbf{x} + \epsilon \int_\Omega \nabla\rho(\mathbf{x})^T \nabla\psi(\mathbf{x}) \, d\mathbf{x} + \frac{1}{\epsilon} \int_\Omega W'\big(\rho(\mathbf{x})\big)\psi(\mathbf{x}) \, d\mathbf{x}$$

and

$$J''_\epsilon(\rho)(\psi_1, \psi_2) = \epsilon \int_\Omega \nabla\psi_1(\mathbf{x})^T \nabla\psi_2(\mathbf{x}) \, d\mathbf{x} + \frac{1}{\epsilon} \int_\Omega W''\big(\rho(\mathbf{x})\big)\psi_1(\mathbf{x})\psi_2(\mathbf{x}) \, d\mathbf{x}.$$

Because of the constraint $0 \le \rho(\mathbf{x}) \le 1$ almost everywhere in Ω, we obtain $|W''(\rho(\mathbf{x}))| \le W_0$, where $W_0 \in \mathbb{R}$ denotes the maximum of $W''(\rho(\mathbf{x}))$ for all $\mathbf{x} \in [0,1]$. Moreover, since $\rho(\mathbf{x}) = 1$ for all $\mathbf{x} \in \Gamma_t$, admissible variations satisfy $\psi(\mathbf{x}) = 0$ for all $\mathbf{x} \in \Gamma_t$. Due to a Friedrich-type inequality, there exists a constant $C_F > 0$ so that

$$\int_\Omega \psi(\mathbf{x})^2 \, d\mathbf{x} \le C_F \int_\Omega \big|\nabla\psi(\mathbf{x})\big|^2 \, d\mathbf{x}$$

for admissible variations $\psi \in H^1(\Omega)$ with $\psi(\mathbf{x}) = 0$ a.e. on Γ_t. Hence,

$$J''_\epsilon(\rho)(\psi, \psi) \ge \left(\epsilon - \frac{C_F W_0}{\epsilon}\right) \int_\Omega \big|\nabla\psi(\mathbf{x})\big| \, d\mathbf{x}.$$

Consequently, for all $\epsilon \ge \epsilon_0 := \sqrt{C_F W_0}$ the objective functional is convex. Since all constraints are linear, the relaxed problem is convex. □

Using the double-obstacle potential (5.23) in P_ϵ we observe that (5.21) and (5.22) are actually quadratic programming problems.

5.4 Existence of Solutions

In this section we will investigate the existence of solutions to the relaxed problem (5.21). Before we start we introduce the Banach space of functions with essentially bounded strain

$$BS_\infty(\Omega) := \Big\{ \mathbf{u} \in L_\infty\big(\Omega; \mathbb{R}^d\big) \;\Big|\; \boldsymbol{\varepsilon}(\mathbf{u}) \in L_\infty\big(\Omega; \mathbb{R}^{d\times d}\big) \Big\}$$

and the Hilbert space with square-integrable strain

$$BS_2(\Omega) := \left\{ \mathbf{u} \in L_2\big(\Omega;\mathbb{R}^d\big) \;\middle|\; \boldsymbol{\varepsilon}(\mathbf{u}) \in L_2\big(\Omega;\mathbb{R}^{d\times d}\big) \right\},$$

with norms

$$\|\mathbf{u}\|_{BS_\infty} := \max\left\{ \|\mathbf{u}\|_{L_\infty(\Omega)}, \|\boldsymbol{\varepsilon}(\mathbf{u})\|_{L_\infty(\Omega)} \right\}$$

and

$$\|\mathbf{u}\|_{BS_2} := \sqrt{\|\mathbf{u}\|_0^2 + \|\boldsymbol{\varepsilon}(\mathbf{u})\|_0^2},$$

respectively. One can verify by standard arguments that $BS_\infty(\Omega)$ is a Banach space, including all elements of the Sobolev space $W^1_\infty(\Omega;\mathbb{R}^d)$ and that $BS_2(\Omega)$ is a Hilbert space with inner product

$$\langle \mathbf{u}, \mathbf{v} \rangle_{BS_2} := \langle \mathbf{u}, \mathbf{v} \rangle_{L_2} + \big\langle \boldsymbol{\varepsilon}(\mathbf{u}), \boldsymbol{\varepsilon}(\mathbf{v}) \big\rangle_{L_2}.$$

As usual for weak solutions of partial differential equations, we understand the equality constraints on $\mathbf{s}$ in a standard weak sense:

$$\int_\Omega \mathbf{s}(\mathbf{x}) : \boldsymbol{\varepsilon}\big(\boldsymbol{\psi}(\mathbf{x})\big)\; d\mathbf{x} = \int_{\Gamma_t} \mathbf{t}(\mathbf{a})^T \boldsymbol{\psi}(\mathbf{a})\; d\mathbf{a}, \qquad \forall\; \boldsymbol{\psi} \in H^1_{\Gamma_u}\big(\Omega;\mathbb{R}^d\big).$$

Similarly, we interpret the stress-strain relation in an L_2-sense, i.e.

$$\int_\Omega \boldsymbol{\sigma}(\mathbf{x}) : \boldsymbol{\Psi}(\mathbf{x}) - \mathbf{C}\boldsymbol{\varepsilon}\big(\mathbf{u}(\mathbf{x})\big) : \boldsymbol{\Psi}(\mathbf{x})\; d\mathbf{x} = 0, \qquad \forall\; \boldsymbol{\Psi} \in L_2\big(\Omega;\mathbb{R}^{d\times d}\big).$$

We start the analysis with the lower semicontinuity property of the objective funtional:

Lemma 5.1. *Let W be defined by (5.23). Then the objective functional $J_\epsilon : H^1(\Omega) \to \mathbb{R}$ is sequentially weakly lower semicontinuous.*

Proof. Due to the compact embedding $H^1(\Omega) \hookrightarrow L_2(\Omega)$, the linear functional $\rho \mapsto \gamma \int_\Omega \rho(\mathbf{x})\; d\mathbf{x}$ and the quadratic functional $\rho \mapsto \frac{1}{\epsilon}\int_\Omega W(\rho(\mathbf{x}))\; d\mathbf{x}$ are weakly continuous. Together with the sequential weak lower semicontinuity of the square of the norm in Hilbert spaces applied to the second term in J_ϵ, we obtain the assertion. □

Besides lower semicontinuity, a fundamental ingredient for the existence of solutions to optimization problems is the compactness of the feasible set in appropriate topologies (cf. Section 4.2). In order to obtain some weak compactness, we examine the boundedness of the constraint set:

Lemma 5.2. *Let $\epsilon > 0$ and let $(\rho, \mathbf{u}, \boldsymbol{\sigma}, \mathbf{s}) \in (H^1(\Omega) \cap L_\infty(\Omega)) \times BS_\infty(\Omega) \times L_\infty(\Omega;\mathbb{R}^{d\times d})^2$ satisfy the constraints in (5.21). Then, $(\rho, \mathbf{u}, \boldsymbol{\sigma}, \mathbf{s})$ lies in a bounded set with respect to the corresponding norms.*

Proof. From the bound constraints $0 \le \rho(\mathbf{x}) \le 1$ a.e. in Ω we immediately conclude that ρ lies in the unit ball of $L_\infty(\Omega)$. Consequently, we deduce that

$$\min\left\{\mathbf{0}, \boldsymbol{\sigma}^{\min}\right\} \le \mathbf{s} \le \max\left\{\boldsymbol{\sigma}^{\max}, \mathbf{0}\right\}$$

and hence, $\mathbf{s}$ is bounded in the norm of $L_\infty(\Omega;\mathbb{R}^{d\times d})$. Due to

$$\mathbf{s} - \frac{1-\rho}{\beta}\mathbf{1} \le \boldsymbol{\sigma} \le \mathbf{s} + \frac{1-\rho}{\beta}\mathbf{1}$$

we further conclude the boundedness of $\boldsymbol{\sigma}$ in the norm of $L_\infty(\Omega;\mathbb{R}^{d\times d})$. Finally, the bound constraints on $\mathbf{u}$ imply its boundedness in the norm of $L_\infty(\Omega;\mathbb{R}^d)$ and together with the stress-strain relation and the positive definiteness of $\mathbf{C}$ we may conclude the boundedness of $\mathbf{u}$ in the norm of $BS_\infty(\Omega)$. □

With these preliminary results we can provide an existence result for the relaxed topology optimization problem for arbitrary positive ϵ:

Theorem 5.3. *Let $\epsilon > 0$, $\beta > 0$, and let W be defined by* (5.23)*. Moreover, let the feasible set defined by the constraints in* (5.21) *be nonempty. Then there exists a solution*

$$(\overline{\rho}, \overline{\mathbf{u}}, \overline{\boldsymbol{\sigma}}, \overline{\mathbf{s}}) \in \big(H^1(\Omega) \cap L_\infty(\Omega)\big) \times BS_\infty(\Omega) \times L_\infty\big(\Omega;\mathbb{R}^{d\times d}\big)^2$$

of the optimization problem (5.21)*.*

Proof. For admissible densities $\rho \geq 0$, the objective functional J_ϵ is bounded below by zero and thus, the infimum j_0 of J_ϵ on the admissible set is finite. Hence, we can find a minimizing sequence

$$\big\{\big(\rho^{(k)}, \mathbf{u}^{(k)}, \boldsymbol{\sigma}^{(k)}, \mathbf{s}^{(k)}\big) \in \big(H^1(\Omega) \cap L_\infty(\Omega)\big) \times BS_\infty(\Omega) \times L_\infty\big(\Omega;\mathbb{R}^{d\times d}\big)^2$$

so that $J_\epsilon(\rho^{(k)}) \to j_0$. Since $J_\epsilon(\rho^{(k)})$ converges, the sequence is bounded in particular and since

$$\frac{2}{\epsilon} J_\epsilon\big(\rho^{(k)}\big) \geq \int_\Omega \big|\nabla \rho^{(k)}(\mathbf{x})\big|^2 \, d\mathbf{x},$$

we obtain boundedness of $\rho^{(k)}$ in $H^1(\Omega)$. Due to Lemma 5.2 and standard precompactness results for bounded sets in weak or weak* topologies, we can extract a subsequence (again denoted by the superscript k) so that

$$\begin{aligned}
\rho^{(k)} &\to \overline{\rho} && \text{weak in } H^1(\Omega), \text{ and weak}^* \text{ in } L_\infty(\Omega),\\
\mathbf{u}^{(k)} &\to \overline{\mathbf{u}} && \text{weak in } BS_2(\Omega), \text{ and weak}^* \text{ in } L_\infty\big(\Omega;\mathbb{R}^d\big),\\
\boldsymbol{\sigma}^{(k)} &\to \overline{\boldsymbol{\sigma}} && \text{weak}^* \text{ in } L_\infty\big(\Omega;\mathbb{R}^{d\times d}\big),\\
\mathbf{s}^{(k)} &\to \overline{\mathbf{s}} && \text{weak}^* \text{ in } L_\infty\big(\Omega;\mathbb{R}^{d\times d}\big).
\end{aligned}$$

Because of the closedness of simple bounds with respect to weak* convergence in L_∞, we can conclude that the limit $(\overline{\rho}, \overline{\mathbf{u}}, \overline{\boldsymbol{\sigma}}, \overline{\mathbf{s}})$ satisfies all the inequalities in (5.21). Moreover, since for $\boldsymbol{\Psi} \in H^1_{\Gamma_u}(\Omega;\mathbb{R}^d)$, we have in particular $\nabla\boldsymbol{\Psi} \in L_1(\Omega;\mathbb{R}^{d\times d})$. Thus we may conclude that

$$\int_{\Gamma_t} \mathbf{t}(\mathbf{a})^T \boldsymbol{\psi}(\mathbf{a}) \, d\mathbf{a} = \int_\Omega \mathbf{s}^{(k)}(\mathbf{x}) : \nabla\boldsymbol{\psi}(\mathbf{x}) \, d\mathbf{x} \quad \to \quad \int_\Omega \overline{\mathbf{s}}(\mathbf{x}) : \nabla\boldsymbol{\psi}(\mathbf{x}) \, d\mathbf{x}$$

due to weak* convergence in L_∞. Hence, $\overline{\mathbf{s}}$ satisfies the associated equality constraints. From the weak convergence of $\mathbf{u}^{(k)}$ in $BS_2(\Omega)$ we conclude that $\overline{\mathbf{u}}$ satisfies the stress-strain relation and hence, $(\overline{\rho}, \overline{\mathbf{u}}, \overline{\boldsymbol{\sigma}}, \overline{\mathbf{s}})$ is in the feasible set. With the sequential lower semicontinuity from Lemma 5.1 we finally obtain that $(\overline{\rho}, \overline{\mathbf{u}}, \overline{\boldsymbol{\sigma}}, \overline{\mathbf{s}})$ is a solution of the optimization problem (5.21). □

A similar analysis can be accomplished for the relaxed problem (5.22) with local von Mises stress constraints.

5.5 Discretization

In the following we consider the discretization of the relaxed problems for $\Omega \subset \mathbb{R}^d$, with $d = 2$, detailing again the analysis for the case of total stress constraints (5.21). For sake of simplicity and motivated by the typical choices of ground structures, we assume that Ω is of polygonal shape.

In order to construct a finite element approximation for problem (5.21), we decompose the ground structure $\Omega = \bigcup_{i=1}^{n} \overline{\tau}_i$ into a suitable triangulation $\mathcal{T}_h = \{\tau_i \mid i = 1, \ldots, n\}$ with n elements and m nodes (cf. Subsection 2.2.3). We shall use two different finite elements for the density ρ, the displacements $\mathbf{u}$ and the stresses $\mathbf{s}$. For the density ρ and the displacement components u_1 and u_2 we use the discrete H^1-subspace of linear elements

$$V^h := \big\{\tilde{v} \in C(\Omega) \;\big|\; \tilde{v}|_{\tau_i} \in \mathcal{P}_1(\tau_i),\ i = 1, \ldots, n\big\}.$$

The stress components of $\mathbf{s}$ are approximated by the L_∞-subspace of constant elements

$$Q^h := \big\{\tilde{q} \in L_\infty(\Omega) \;\big|\; \tilde{q}|_{\tau_i} \in \mathcal{P}_0(\tau_i),\ i = 1, \ldots, n\big\}.$$

Again $\mathcal{P}_k(\tau_i)$ represents the space of polynomials of maximal degree k over the triangle τ_i. Note that for $\mathbf{u} \in V^h \times V^h \subset BS_\infty(\Omega)$, we obtain $\frac{\partial u_i}{\partial x_j} \in Q^h$ for $i, j = 1, 2$.

The equality constraints are discretized by a standard finite element approach, i.e. we look for $\tilde{\mathbf{s}} \in (Q^h)^{2\times 2}$ and $\tilde{\mathbf{u}} \in V^h \times V^h$ satisfying

$$\int_\Omega \tilde{\mathbf{s}}(\mathbf{x}) : \nabla \tilde{\mathbf{v}}(\mathbf{x})\ d\mathbf{x} = \int_{\Gamma_t} \mathbf{t}(\mathbf{a})^T \tilde{\mathbf{v}}(\mathbf{a})\ d\mathbf{a}, \qquad \forall\ \tilde{\mathbf{v}} \in V^h \times V^h,\ \tilde{\mathbf{v}}|_{\Gamma_u} = \mathbf{0},$$

and

$$\int_\Omega \tilde{\boldsymbol{\sigma}}(\mathbf{x}) : \tilde{\mathbf{q}}(\mathbf{x}) - \mathbf{C}\boldsymbol{\varepsilon}\big(\tilde{\mathbf{u}}(\mathbf{x})\big) : \tilde{\mathbf{q}}(\mathbf{x})\ d\mathbf{x} = 0, \qquad \forall\ \tilde{\mathbf{q}} \in (Q^h)^{2\times 2}.$$

The bound constraints on the displacements and density can be enforced directly, since for piecewise linear functions, the constraints hold if and only if they hold in all nodes of the grid.

Finally, we need to discretize the inequality constraints involving both, the stresses and the density. Since the components of $\mathbf{s}$ are in a different subspace than ρ, the discretization is not straightforward. In particular, we cannot pose local constraints in the grid nodes or on edges, because functions in Q^h are discontinuous over edges. Consequently, the more promising approach is to interpret the inequality constraints as constraints in Q^h. For this sake we introduce the discrete projection operator (with respect to the L_2-norm) $\mathcal{P}^h : V^h \to Q^h$,

$$\big(\mathcal{P}^h \tilde{v}\big)\big|_\tau := \frac{1}{|\tau|} \int_\tau \tilde{v}(\mathbf{x})\ d\mathbf{x}, \qquad \forall\ \tau \in \mathcal{T}^h,\ \forall\ \tilde{v} \in V^h. \tag{5.24}$$

Let the vectors $\boldsymbol{\rho}^h \in \mathbb{R}^m$, $\mathbf{u}^h \in \mathbb{R}^{2m}$, and $\mathbf{s}^h \in \mathbb{R}^{3n}$ (using the symmetry $s_{ij} = s_{ji}$) contain the coefficients of the finite element functions $\tilde{\rho} \in V^h$, $\tilde{\mathbf{u}} \in V^h \times V^h$, and $\tilde{\mathbf{s}} \in (Q^h)^{2\times 2}$, respectively. The discretized problem can now be written equivalently as a quadratic

programming problem:

$$\gamma \mathbf{e}^{h^T}\boldsymbol{\rho}^h + \frac{\epsilon}{2}\boldsymbol{\rho}^{h^T}\mathbf{K}\boldsymbol{\rho}^h + \frac{1}{\epsilon}\boldsymbol{\rho}^{h^T}\big(\mathbf{e}^h - \mathbf{M}\boldsymbol{\rho}^h\big) \;\to \min_{(\boldsymbol{\rho}^h,\mathbf{u}^h,\mathbf{s}^h)\in\mathbb{R}^{3m+3n}} \tag{5.25a}$$

$$\text{subject to} \quad \big(\mathbf{I} - \mathbf{B}_u^T\mathbf{B}_u\big)\mathbf{D}^T\mathbf{s}^h = \mathbf{t}^h, \tag{5.25b}$$

$$\mathbf{B}_u\mathbf{u}^h = \mathbf{0}, \tag{5.25c}$$

$$\mathbf{B}_t\boldsymbol{\rho}^h = \mathbf{1}, \tag{5.25d}$$

$$-\mathbf{Q}\big(\mathbf{1} - \mathbf{P}\boldsymbol{\rho}^h\big) \le \beta\big(\mathbf{C}^h\mathbf{D}\mathbf{u}^h - \mathbf{s}^h\big) \le \mathbf{Q}\big(\mathbf{1} - \mathbf{P}\boldsymbol{\rho}^h\big), \tag{5.25e}$$

$$\boldsymbol{\Sigma}^{\min}\mathbf{Q}\mathbf{P}\boldsymbol{\rho}^h \le \mathbf{s}^h \le \boldsymbol{\Sigma}^{\max}\mathbf{Q}\mathbf{P}\boldsymbol{\rho}^h, \tag{5.25f}$$

$$\mathbf{0} \le \boldsymbol{\rho}^h \le \mathbf{1}, \tag{5.25g}$$

$$\mathbf{u}^{\min} \le \mathbf{u}^h \le \mathbf{u}^{\max}. \tag{5.25h}$$

In the objective functional (5.25a), $\mathbf{e}^h \in \mathbb{R}^n$ is a vector representing the coefficients of the constant function 1 with respect to the linear basis function of the finite dimensional subspace V^h. $\mathbf{K} \in \mathbb{R}^{n\times n}$ is the stiffness matrix arising from the finite element discretization of the negative Laplacian in V^h and $\mathbf{M} \in \mathbb{R}^{n\times n}$ is the mass matrix for the identity in V^h. Note that in the above formulation (5.25) of the discretized problem, the stresses $\boldsymbol{\sigma}$ are eliminated through the stress-strain relationship (5.1c). In the discretized formulation of the boundary conditions (5.25c) - (5.25d) the matrices $\mathbf{B}_u \in \mathbb{R}^{2n_u\times 2n}$ and $\mathbf{B}_t \in \mathbb{R}^{n_\rho\times n}$ with entries 0 or 1 realize the boundary conditions, where n_u is the number of nodes on Γ_u and n_ρ is the number of nodes on Γ_t. In the discretized partial differential equation (5.25b) $\mathbf{t}^h$ is a discrete representation of the traction forces. Moreover, the matrix $\mathbf{D}^T \in \mathbb{R}^{3m\times 2n}$ in (5.25e) is the discretization of the divergence operator (restricted to symmetric stress tensors) and $\mathbf{C}^h$ is the discrete analogon of the elastic tensor $\mathbf{C}$. In (5.25f) $\boldsymbol{\Sigma}^{\min}, \boldsymbol{\Sigma}^{\max} \in \mathbb{R}^{3m\times 3m}$ and are diagonal matrices, representing the corresponding entries of $\boldsymbol{\sigma}^{\min}$ and $\boldsymbol{\sigma}^{\max}$, respectively. Finally, $\mathbf{Q} \in \mathbb{R}^{3m\times m}$ is an extension matrix and $\mathbf{P} \in \mathbb{R}^{m\times n}$ is the matrix representation of the projection operator $\mathcal{P}^h$ (5.24).

The above reasoning shows that after discretization we end up with a linearly constrained quadratic programming problem for the variables $(\boldsymbol{\rho}^h, \mathbf{u}^h, \mathbf{s}^h) \in \mathbb{R}^{3m+3n}$ with $2n + 2n_u + n_\rho$ equality, $12m$ inequality and $6n$ bound constraints. Note that $2n_u$ equalities corresponding to the divergence constraints (5.25b) for the nodal points in Γ_u are actually redundant and can be eliminated. Additionally, the components of $\mathbf{u}^h$ corresponding to the nodal points on Γ_u and the corresponding bound constraints can be eliminated, since we have to assume $\mathbf{u}^{\min} \le \mathbf{0} \le \mathbf{u}^{\max}$ anyway in order to obtain feasible points. Furthermore, the components of $\boldsymbol{\rho}^h$ related to the nodal values of Γ_t and the related bound constraints can be eliminated. So, we consequently end up with a smaller programming problem with $3n + 3m - 2n_u - n_\rho$ unknowns, $2n - 2n_u$ equality constraints, $12m$ inequality constraints, and $6n - 4n_u - 2n_\rho$ bound constraints.

The existence of solutions for the reduced discretized programming problem can be verified in an analogous way to the infinite-dimensional situation under the assumption that there exists a feasible point.

5.5.1 Constraint Qualification

For linear constraints, the common notions of constraint qualifications such as the LICQ (see Definition 2.1), the MFCQ (see Definition 2.2), and the Slater qualification (see Definition 2.3)

are equivalent. In order to show that constraint qualification hold for the linear constraints of (5.25), the equality constraints have to be linearly independent and there has to exist a feasible point, that fulfills all inequality constraints strictly. Since the feasible set does not depend on the relaxation parameter ϵ, the constraint qualification is always uniform with respect to the relaxation.

We obtain the linear independence of the equality constraints by standard reasoning for finite element discretizations. Therefore, let us turn our attention to the inequality constraints. Here we shall use a natural assumption, namely that stresses and displacements obtained from a design domain completely occupied with material ($\boldsymbol{\rho}_1^h = \mathbf{1}$) satisfy the displacement and stress constraints strictly. Thus, if the constraints are active at maximal mass already, it is quite unlikely to find an optimal design with lower mass anyway. In other words, the constraints would be then too severe to allow a different optimal design. In the following let us formulate this assumption in mathematical terms. Let $\overline{\mathbf{u}}^h$ and $\overline{\mathbf{s}}^h$ ($\overline{\mathbf{s}}^h = \overline{\boldsymbol{\sigma}}^h$) be the solution of the corresponding discrete elasticity problem:

$$\big(\mathbf{I} - \mathbf{B}_u^T\mathbf{B}_u\big)\mathbf{D}^T\mathbf{s}^h = \mathbf{t}^h, \qquad \mathbf{s}^h - \mathbf{C}^h\mathbf{D}\mathbf{u}^h = \mathbf{0}, \qquad \mathbf{B}_u\mathbf{u}^h = \mathbf{0},$$

which can be shown to be uniquely defined from standard finite element theory. Following the above assumption we then assume that

$$\boldsymbol{\Sigma}^{\min}\mathbf{Q}\mathbf{P}\boldsymbol{\rho}_1^h < \overline{\mathbf{s}}^h < \boldsymbol{\Sigma}^{\max}\mathbf{Q}\mathbf{P}\boldsymbol{\rho}_1^h \quad \text{and} \quad \mathbf{u}^{\min} < \overline{\mathbf{u}}^h < \mathbf{u}^{\max}, \tag{5.26}$$

where the $<$ means strict inequality for each component. Then we obtain the following result:

Theorem 5.4. *Let* (5.26) *be satisfied and let* $\beta > 0$. *Then, the inequality constraints of problem* (5.25)*, with the above mentioned elimination of variables, satisfy constraint qualifications.*

Proof. As noticed above, it suffices to find a feasible point satisfying the inequality constraints strictly. For this sake we choose a design $\boldsymbol{\rho}_\delta^h = \delta\mathbf{1}$ with $0 < \delta < 1$. Then $\|\boldsymbol{\rho}_1^h - \boldsymbol{\rho}_\delta^h\| = (n - n_\rho)(1 - \delta)$, i.e., the distance to $\boldsymbol{\rho}_1^h$ becomes arbitrarily small as $\delta \to 1$. Because of continuity we can find $0 < \delta < 1$ and a solution $\mathbf{u}_\delta^h$, $\mathbf{s}_\delta^h$ ($\mathbf{s}_\delta^h = \delta\boldsymbol{\sigma}_\delta^h$) of

$$\big(\mathbf{I} - \mathbf{B}_u^T\mathbf{B}_u\big)\mathbf{D}^T\mathbf{s}^h = \mathbf{t}^h, \qquad \delta^{-1}\mathbf{s}^h - \mathbf{C}^h\mathbf{D}\mathbf{u}^h = \mathbf{0}, \qquad \mathbf{B}_u\mathbf{u}^h = \mathbf{0},$$

so that

$$\boldsymbol{\Sigma}^{\min}\mathbf{Q}\mathbf{P}\boldsymbol{\rho}_\delta^h < \mathbf{s}_\delta^h < \boldsymbol{\Sigma}^{\max}\mathbf{Q}\mathbf{P}\boldsymbol{\rho}_\delta^h \quad \text{and} \quad \mathbf{u}^{\min} < \mathbf{u}_\delta^h < \mathbf{u}^{\max},$$

is fulfilled. Moreover, we have $0 < \boldsymbol{\rho}_\delta^h = \delta\mathbf{1} < \mathbf{1}$ and that

$$\begin{aligned} -\mathbf{Q}(1-\delta)\mathbf{1} = -\mathbf{Q}\big(\mathbf{1} - \mathbf{P}\boldsymbol{\rho}_\delta^h\big) &< \beta\big(\mathbf{C}^h\mathbf{D}\mathbf{u}^h - \mathbf{s}^h\big) \\ &< \mathbf{Q}\big(\mathbf{1} - \mathbf{P}^h\boldsymbol{\rho}_\delta^h\big) = \mathbf{Q}(1-\delta)\mathbf{1}, \end{aligned}$$

provided β is sufficiently small. Hence, all (reduced) inequality constraints are satisfied strictly by $(\boldsymbol{\rho}_\delta^h, \mathbf{u}_\delta^h, \mathbf{s}_\delta^h)$, which implies the assertion. □

5.5.2 First-Order Optimality

If the constraints satisfy constraint qualifications, which indeed holds under suitable assumptions as verified above, one can formulate the first-order necessary optimality conditions of problem (5.25). So we introduce the Lagrangian $\mathcal{L}$ and the Lagrangian multipliers $\boldsymbol{\lambda}_i^h$, for $i = 1, \dots, 11$, whose dimension will be clear from their appearance in the Lagrangian $\mathcal{L}$, via

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\rho}^h, \mathbf{u}^h, \mathbf{s}^h, \boldsymbol{\lambda}_1^h, \dots, \boldsymbol{\lambda}_{11}^h) \;=\; & \gamma \mathbf{e}^{h^T} \boldsymbol{\rho}^h + \frac{\epsilon}{2} \boldsymbol{\rho}^{h^T} \mathbf{K} \boldsymbol{\rho}^h + \frac{1}{\epsilon} \boldsymbol{\rho}^{h^T} \big(\mathbf{e}^h - \mathbf{M}\boldsymbol{\rho}^h\big) + \\
& \boldsymbol{\lambda}_1^{h^T} \Big(\big(\mathbf{I} - \mathbf{B}_u^T \mathbf{B}_u\big) \mathbf{D}^T \mathbf{s}^h - \mathbf{t}^h \Big) + \boldsymbol{\lambda}_2^{h^T} \big(\mathbf{B}_u \mathbf{u}^h\big) + \\
& \boldsymbol{\lambda}_3^{h^T} \big(\mathbf{B}_\rho \boldsymbol{\rho}^h - \mathbf{1}\big) - \boldsymbol{\lambda}_4^{h^T} \Big(\mathbf{Q}\big(\mathbf{1} - \mathbf{P}\boldsymbol{\rho}^h\big) + \beta\big(\mathbf{C}\mathbf{D}\mathbf{u}^h - \mathbf{s}^h\big) \Big) + \\
& \boldsymbol{\lambda}_5^{h^T} \Big(\beta\big(\mathbf{C}^h \mathbf{D}\mathbf{u}^h - \mathbf{s}^h\big) - \mathbf{Q}\big(\mathbf{1} - \mathbf{P}\boldsymbol{\rho}^h\big) \Big) + \\
& \boldsymbol{\lambda}_6^{h^T} \big(\boldsymbol{\Sigma}^{\min} \mathbf{Q}\mathbf{P}\boldsymbol{\rho}^h - \mathbf{s}^h\big) + \boldsymbol{\lambda}_7^{h^T} \big(\mathbf{s}^h - \boldsymbol{\Sigma}^{\max} \mathbf{Q}\mathbf{P}^h \boldsymbol{\rho}^h\big) - \\
& \boldsymbol{\lambda}_8^{h^T} \boldsymbol{\rho}^h + \boldsymbol{\lambda}_9^{h^T} \big(\boldsymbol{\rho}^h - \mathbf{1}\big) + \boldsymbol{\lambda}_{10}^{h\,T} \big(\mathbf{u}^{\min} - \mathbf{u}^h\big) + \boldsymbol{\lambda}_{11}^{h\,T} \big(\mathbf{u}^h - \mathbf{u}^{\max}\big),
\end{aligned}
$$

with $\boldsymbol{\lambda}_i \geq \mathbf{0}$ for $i = 4, \dots, 11$.

Using the Lagrangian we can finally state the first-first optimality conditions for the optimization problem (5.25):

$$
\begin{aligned}
\nabla_{\boldsymbol{\rho}^h} \mathcal{L} \;=\; & \gamma \mathbf{e}^h + \epsilon \mathbf{K}\boldsymbol{\rho}^h + \frac{1}{\epsilon}\big(\mathbf{e}^h - 2\mathbf{M}\boldsymbol{\rho}^h\big) + \mathbf{B}_\rho^T \boldsymbol{\lambda}_3^h - \boldsymbol{\lambda}_8^h + \boldsymbol{\lambda}_9^h + \\
& + \mathbf{P}^T \mathbf{Q}^T \big(\boldsymbol{\lambda}_4^h + \boldsymbol{\lambda}_5^h + \boldsymbol{\Sigma}^{\max T} \boldsymbol{\lambda}_6^h - \boldsymbol{\Sigma}^{\min^T} \boldsymbol{\lambda}_7^h\big) && = \mathbf{0}, \\
\nabla_{\mathbf{u}^h} \mathcal{L} \;=\; & \mathbf{B}_\rho^T \boldsymbol{\lambda}_3^h + \beta \mathbf{C}^{h^T} \mathbf{D}^T \big(\boldsymbol{\lambda}_5^h - \boldsymbol{\lambda}_4^h\big) - \boldsymbol{\lambda}_{10}^h + \boldsymbol{\lambda}_{11}^h && = \mathbf{0}, \\
\nabla_{\mathbf{s}^h} \mathcal{L} \;=\; & \mathbf{D}\big(\mathbf{I} - \mathbf{B}_u^T \mathbf{B}_u\big) \boldsymbol{\lambda}_1^h + \beta\big(\boldsymbol{\lambda}_4^h - \boldsymbol{\lambda}_5^h\big) - \boldsymbol{\lambda}_6^h + \boldsymbol{\lambda}_7^h && = \mathbf{0}, \\
\nabla_{\boldsymbol{\lambda}_1^h} \mathcal{L} \;=\; & \big(\mathbf{I} - \mathbf{B}_u^T \mathbf{B}_u\big) \mathbf{D}^T \mathbf{s}^h - \mathbf{t}^h && = \mathbf{0}, \\
\nabla_{\boldsymbol{\lambda}_2^h} \mathcal{L} \;=\; & \mathbf{B}_u \mathbf{u}^h \;=\; \mathbf{0}, \qquad \nabla_{\boldsymbol{\lambda}_3^h} = \mathbf{B}_\rho \boldsymbol{\rho}^h - \mathbf{1} && = \mathbf{0}, \\
\nabla_{\boldsymbol{\lambda}_4^h} \mathcal{L} \;=\; & \mathbf{Q}\big(\mathbf{1} - \mathbf{P}\boldsymbol{\rho}^h\big) + \beta\big(\mathbf{C}^h \mathbf{D}\mathbf{u}^h - \mathbf{s}^h\big) \;\leq\; \mathbf{0}, \\
& \boldsymbol{\lambda}_4^{h^T} \Big(\mathbf{Q}\big(\mathbf{1} - \mathbf{P}\boldsymbol{\rho}^h\big) + \beta\big(\mathbf{C}^h \mathbf{D}\mathbf{u}^h - \mathbf{s}^h\big) \Big) \;=\; \mathbf{0}, \;\text{ and } \boldsymbol{\lambda}_4^h \;\geq\; \mathbf{0}, \\
\nabla_{\boldsymbol{\lambda}_5^h} \mathcal{L} \;=\; & \beta\big(\mathbf{C}^h \mathbf{D}\mathbf{u}^h - \mathbf{s}^h\big) - \mathbf{Q}\big(\mathbf{1} - \mathbf{P}\boldsymbol{\rho}^h\big) \;\leq\; \mathbf{0}, \\
& \boldsymbol{\lambda}_5^{h^T} \Big(\beta\big(\mathbf{C}\mathbf{D}\mathbf{u}^h - \mathbf{s}^h\big) - \mathbf{Q}\big(\mathbf{1} - \mathbf{P}\boldsymbol{\rho}^h\big) \Big) \;=\; \mathbf{0}, \quad \text{and } \boldsymbol{\lambda}_5^h \;\geq\; \mathbf{0}, \\
\nabla_{\boldsymbol{\lambda}_6^h} \mathcal{L} \;=\; & \boldsymbol{\Sigma}^{\min} \mathbf{Q}\mathbf{P}\boldsymbol{\rho}^h - \mathbf{s}^h \;\leq\; \mathbf{0}, \quad \boldsymbol{\lambda}_6^{h^T} \big(\boldsymbol{\Sigma}^{\min} \mathbf{Q}\mathbf{P}\boldsymbol{\rho}^h - \mathbf{s}^h\big) \;=\; \mathbf{0}, && \text{and } \boldsymbol{\lambda}_6^h \;\geq\; \mathbf{0}, \\
\nabla_{\boldsymbol{\lambda}_7^h} \mathcal{L} \;=\; & \mathbf{s}^h - \boldsymbol{\Sigma}^{\max} \mathbf{Q}\mathbf{P}\boldsymbol{\rho}^h \;\leq\; \mathbf{0}, \quad \boldsymbol{\lambda}_7^{h^T} \big(\mathbf{s}^h - \boldsymbol{\Sigma}^{\max} \mathbf{Q}\mathbf{P}\boldsymbol{\rho}^h\big) \;=\; \mathbf{0}, && \text{and } \boldsymbol{\lambda}_7^h \;\geq\; \mathbf{0}, \\
\nabla_{\boldsymbol{\lambda}_8^h} \mathcal{L} \;=\; & -\boldsymbol{\rho}^h \;\leq\; \mathbf{0}, \qquad \boldsymbol{\lambda}_8^{h^T} \boldsymbol{\rho}^h \;=\; \mathbf{0}, && \text{and } \boldsymbol{\lambda}_8^h \;\geq\; \mathbf{0}, \\
\nabla_{\boldsymbol{\lambda}_9^h}^T \mathcal{L} \;=\; & \boldsymbol{\rho}^h - \mathbf{1} \;\leq\; \mathbf{0}, \qquad \boldsymbol{\lambda}_9^{h^T} \big(\boldsymbol{\rho}^h - \mathbf{1}\big) \;=\; \mathbf{0}, && \text{and } \boldsymbol{\lambda}_9^h \;\geq\; \mathbf{0}, \\
\nabla_{\boldsymbol{\lambda}_{10}^h} \mathcal{L} \;=\; & \mathbf{u}^{\min} - \mathbf{u}^h \;\leq\; \mathbf{0}, \qquad \boldsymbol{\lambda}_{10}^{h\,T} \big(\mathbf{u}^{\min} - \mathbf{u}^h\big) \;=\; \mathbf{0}, && \text{and } \boldsymbol{\lambda}_{10}^h \;\geq\; \mathbf{0}, \\
\nabla_{\boldsymbol{\lambda}_{11}^h} \mathcal{L} \;=\; & \mathbf{u}^h - \mathbf{u}^{\max} \;\leq\; \mathbf{0}, \qquad \boldsymbol{\lambda}_{11}^{h\,T} \big(\mathbf{u}^h - \mathbf{u}^{\max}\big) \;=\; \mathbf{0}, && \text{and } \boldsymbol{\lambda}_{11}^h \;\geq\; \mathbf{0}.
\end{aligned}
$$

Of course the discretization of the problem (5.22) with local von Mises constraints can be accomplished in the same way as for the problem (5.21). By similar reasoning it is also possible to show the validity of constraint qualifications and to deduce the first-order optimality conditions.

5.6 Numerical Experiments

5.6.1 Continuation in ϵ

As motivated above the optimization problem (5.25) will be solved for a decreasing sequence of the relaxation parameter ϵ. As $\epsilon \to 0$, in analogy to Γ-convergence of the perimeter functional, we expect convergence of the sequence of minimum solutions to a final solution. Moreover, since the double obstacle term is of leading order in ϵ, such a final optimal design will have a sharp interface between material Ω_{mat} and void $(\Omega \setminus \Omega_{\text{mat}})$.

To achieve this we will use a continuation method such that we choose a decreasing sequence $\{\epsilon_i\}$ with $\epsilon_i \to 0$ for $i = 0, \ldots, l$, where l describes the total number of continuation levels. The corresponding optimization problems are then solved by an interior-point method, as described in the next Subsection 5.6.2. Between the levels ϵ can be reduced, e.g. like $\epsilon_{i+1} = \delta\epsilon_i$ with $0 < \delta < 1$ or like $\epsilon_{i+1} = \epsilon_0^i$ if $0 < \epsilon_0 < 1$. If we decrease ϵ too slow, we may expect from theory and observations from numerical tests, that the final solution is not changed, but we end up with a possible unnecessary high number of levels l. On the other hand, if ϵ is decreased too fast, the optimization process might get stuck in some undesired local minimum, since the objective functional J_ϵ is turned from convex to non-convex too quickly.

5.6.2 Adaption of the Problem to an Interior-Point Method

We solve the problem (5.25) with Ipopt (see Wächter et al [146]), which is a free available optimization code realizing a primal-dual interior-point optimization method (cf., Subsection 2.1.3). Ipopt, implemented by A. Wächter and L. T. Biegler, is able to solve nonlinear programming problems of the following form:

$$\begin{aligned} f(\mathbf{x}) &\to \min_{\mathbf{x}\in\mathbb{R}^n} \\ \text{subject to} \quad \mathbf{c}_{\mathcal{E}}(\mathbf{x}) &= \mathbf{0}, \\ \mathbf{x}^{\min} \le \mathbf{x} &\le \mathbf{x}^{\max}, \end{aligned}$$

where $\mathcal{E}$ denotes the index set of the equality constraints. More information about the implementation and further issues of Ipopt are given in Wächter and Biegler [147].

General nonlinear programming problems with inequality constraints can be written in the above framework using slack variables. So we reformulate (5.25) by introducing some

vector $\mathbf{z}^h = (\mathbf{z}_1^h, \mathbf{z}_2^h, \mathbf{z}_3^h, \mathbf{z}_4^h) \in \mathbb{R}^{12m}$ of slack variables, leading to:

$$\begin{aligned}
\gamma \mathbf{e}^{h^T} \boldsymbol{\rho}^h + \frac{\epsilon}{2} \boldsymbol{\rho}^{h^T} \mathbf{K} \boldsymbol{\rho}^h + \frac{1}{\epsilon} \boldsymbol{\rho}^{h^T} \big(\mathbf{e}^h - \mathbf{M}\boldsymbol{\rho}^h\big) \;&\rightarrow \min_{(\boldsymbol{\rho}^h, \mathbf{u}^h, \mathbf{s}^h, \mathbf{z}^h) \in \mathbb{R}^{15m+3n}} \\
\text{subject to} \qquad \big(\mathbf{I} - \mathbf{B}_u^T \mathbf{B}_u\big) \mathbf{D}^T \mathbf{s}^h &= \mathbf{t}^h, \\
\mathbf{B}_u \mathbf{u}^h &= \mathbf{0}, \\
\mathbf{B}_t \boldsymbol{\rho}^h &= \mathbf{1}, \\
-\mathbf{Q}\big(\mathbf{1} - \mathbf{P}\boldsymbol{\rho}^h\big) - \beta\big(\mathbf{C}^h \mathbf{D}\mathbf{u}^h - \mathbf{s}^h\big) + \mathbf{z}_1^h &= \mathbf{0}, \\
\beta\big(\mathbf{C}^h \mathbf{D}\mathbf{u}^h - \mathbf{s}^h\big) - \mathbf{Q}\big(\mathbf{1} - \mathbf{P}\boldsymbol{\rho}^h\big) + \mathbf{z}_2^h &= \mathbf{0}, \\
\boldsymbol{\Sigma}^{\min} \mathbf{Q}\mathbf{P}\boldsymbol{\rho}^h - \mathbf{s}^h + \mathbf{z}_3^h &= \mathbf{0}, \\
\mathbf{s}^h - \boldsymbol{\Sigma}^{\max} \mathbf{Q}\mathbf{P}\boldsymbol{\rho}^h + \mathbf{z}_4^h &= \mathbf{0}, \\
\mathbf{0} \;\leq\; \boldsymbol{\rho}^h \;&\leq\; \mathbf{1}, \\
\mathbf{u}^{\min} \;\leq\; \mathbf{u}^h \;&\leq\; \mathbf{u}^{\max}, \\
\mathbf{0} \;\leq\; \mathbf{z}^h &.
\end{aligned}$$

Finally we solve a programming problem with $3n + 15m - 2n_u - n_\rho$ unknowns, $2n + 12m - 2n_u$ equality constraints and $6n + 12m - 4n_u - 2n_\rho$ bound constraints. A similar discrete programming problem for the problem with conservative von Mises stress constraints (5.22) can be deduced in a analogous way.

5.6.3 Numerical Examples

For numerical examples we have chosen two simple examples where the global optimal designs are known on very coarse grids (cf. Stope and Svanberg [140]), which provides some reference for our solutions. For sake of simplicity the Young's modulus and the Poisson's ratio of the given material are $E = 1N/m^2$ and $\nu = 0.3$. We used the plain strain model for the computations and also, for simplicity, all structures have a unit thickness of $1m$ and are loaded with the half of the unit load. Reasonable bounds for the displacements $\mathbf{u}$ and the stresses $\boldsymbol{\sigma}$ are provided by the unique solutions $\overline{\mathbf{u}}$ and $\overline{\boldsymbol{\sigma}}$ of the corresponding elasticity problem when the whole design domain Ω is filled with material. Then the displacement bounds are determined by

$$u_i^{\max} = -u_i^{\min} = 2 \max\big\{ |\overline{u}_i(\mathbf{x})| \;\big|\; \forall\, \mathbf{x} \in \Omega \big\}, \qquad i = 1, 2,$$

and the stress bounds are set to

$$\begin{aligned}
\sigma_{11}^{\max} = \sigma_{22}^{\max} &= \max\big\{ |\overline{\sigma}_{11}(\mathbf{x})|, |\overline{\sigma}_{22}(\mathbf{x})| \;\big|\; \mathbf{x} \in \Omega \big\}, \\
\sigma_{12}^{\max} = \sigma_{21}^{\max} &= \max\big\{ |\overline{\sigma}_{12}(\mathbf{x})|, |\overline{\sigma}_{21}(\mathbf{x})| \;\big|\; \mathbf{x} \in \Omega \big\},
\end{aligned}$$

with $\boldsymbol{\sigma}^{\min} = -\boldsymbol{\sigma}^{\max}$. The von Mises stress bound is given by

$$\Phi^{\max} = \max\big\{ \Phi\big(\overline{\boldsymbol{\sigma}}\big) \;\big|\; \mathbf{x} \in \Omega \big\}.$$

All numerical examples are performed on a PC using a 2.4 GHz Intel CPU and 2 GB memory. For the mesh generation the software package NETGEN/NGSolve was used (see Schöberl

ET AL. [117]). The optimization part was done using the interior-point code Ipopt. As Ipopt is used as a 'black-box', we did not adjust its linear solver and its stopping criterion to our needs. So we stop the optimization process per continuation level if the stopping criterion of Ipopt is fulfilled with a tolerance of 10^{-5} or a maximum number of iterations is reached. For an approximation of the Hessian of the Lagrange functional a BFGS routine is used. We know that tailoring the problem to a black-box optimizer is not the most efficient approach for the numerical solution. However, the aim of this chapter is rather to develop and to test the general relaxation approach than to construct an efficient optimization method. In both examples the scaling parameter γ of the mass term in the objective (5.18) is set to $\gamma = 2$ and we start the ϵ-continuation for $l = 4$ levels with $\epsilon_0 = 0.1$. Between the levels ϵ is reduced like $\epsilon^{i+1} = 0.5\epsilon^i$.

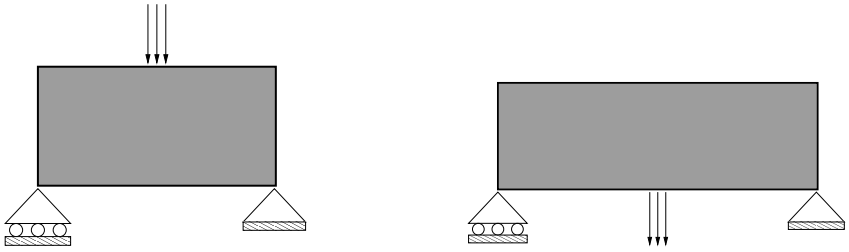

Figure 5.2: Sketches of simple beam examples.

As a first example we treat the problem shown left in Figure 5.2. There, the load condition, bearings and geometry are illustrated with a design domain of dimension $2m \times 1m$. Here we consider stress constraints with respect to local stresses. The corresponding bound constraints are:

$$\mathbf{u}^{\max} = -\mathbf{u}^{\min} = \begin{pmatrix} 0.525 \\ 0.525 \end{pmatrix} m \qquad \text{and} \qquad \boldsymbol{\sigma}^{\max} = -\boldsymbol{\sigma}^{\min} = \begin{pmatrix} \sigma_{11}^{\max} \\ \sigma_{22}^{\max} \\ \sigma_{12}^{\max} \end{pmatrix} = \begin{pmatrix} 0.65 \\ 0.65 \\ 0.262 \end{pmatrix} \frac{N}{m^2}.$$

A mesh with 14182 elements is used for the optimization process, so we finally end up with 234532 unknowns. In more detail we have 7260, 14182, 42546, and 170184 degrees of freedom for the density, displacements, stresses, and slacks, respectively. The total number of equality constraints is 184725 and 191986 for the bound constraints, respectively. The overall computational time for the 4 levels is about 16 hours and the volume of the final optimal design is 0.344 m^3 (17.2% of $|\Omega|$). In Figure 5.3 we see the final optimal design and the corresponding final values of the σ_{11} stress component. Moreover, the optimal designs concerning the levels of the ϵ-continuation are shown in Figure 5.4.

For the second example we consider the right problem in Figure 5.2, where the load condition, bearings and geometry is shown. The dimensions of the design domain are $3m \times 1m$. Here we choose to calculate an optimal design with respect to bounded conservative von Mises stress and again we list the corresponding bound constraints:

$$\mathbf{u}^{\max} = -\mathbf{u}^{\min} = \begin{pmatrix} 0.99 \\ 0.99 \end{pmatrix} m \qquad \text{and} \qquad \Phi^{\max} = 0.78 \frac{N}{m^2}.$$

For the discretization and the optimization process we use a mesh with 17291 elements, which results in 355098 unknowns, 8849, 17698, 51873, 51871, and 224784 degrees of freedom

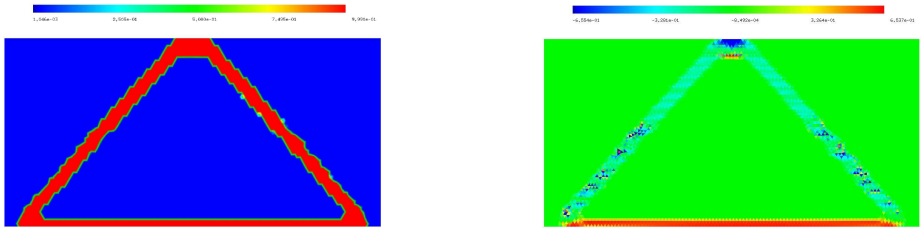

Figure 5.3: Solution of the short beam example. Left: Optimal material distribution. Right: Final values of the σ_{11} stress component.

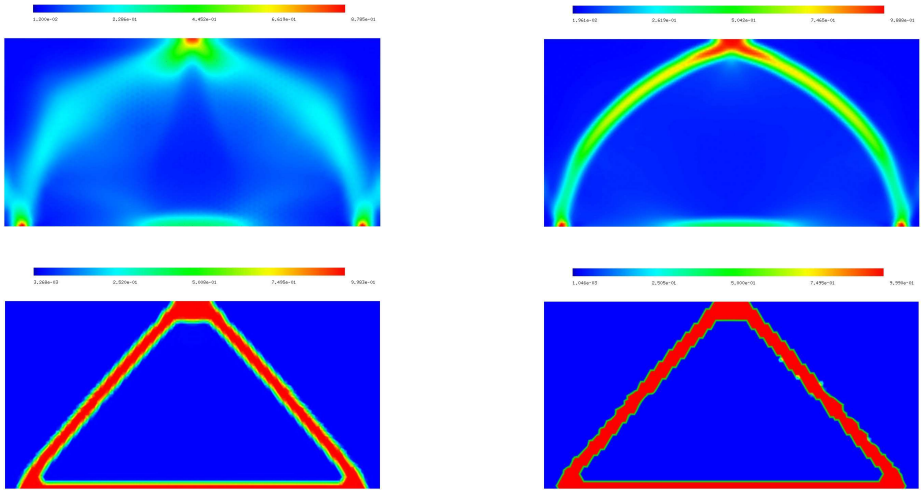

Figure 5.4: Optimal designs of the 4 level ϵ-continuation with $\epsilon_0 = 0.1$, $\epsilon_1 = 0.05$, $\epsilon_2 = 0.025$, and $\epsilon_3 = 0.0125$, respectively.

for the density, displacements, stresses, conservative von Mises stress approximation and slacks, respectively. Here we end up with a total number of 242503 equality constraints and 251352 bound constraints. In Figure 5.5 we see the optimal design of the problem and the

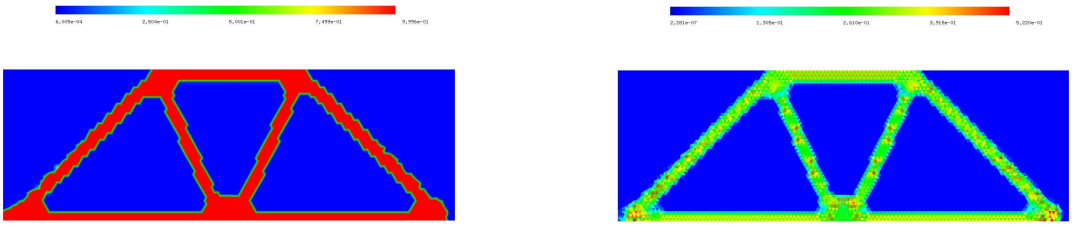

Figure 5.5: Solution of the long beam example. Left: Optimal material distribution. Right: Final values of the conservative von Mises stress.

corresponding final values of the conservative von Mises stress. The solution time for 4 levels of ϵ-continuation was about 19 hours and total volume is reduced to $1.134m^3$ (37.8% of $|\Omega|$). For comparison, we show in Figure 5.6 an optimal design with respcect to minimal compliance with the same volume as the minimal mass solution. Finally, the Figures 5.7 and 5.8 show

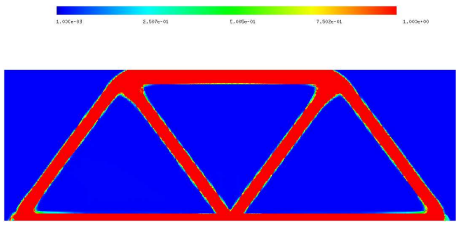

Figure 5.6: An optimal design of the long beam example with respect to minimal compliance.

the final values of the displacement and stress components.

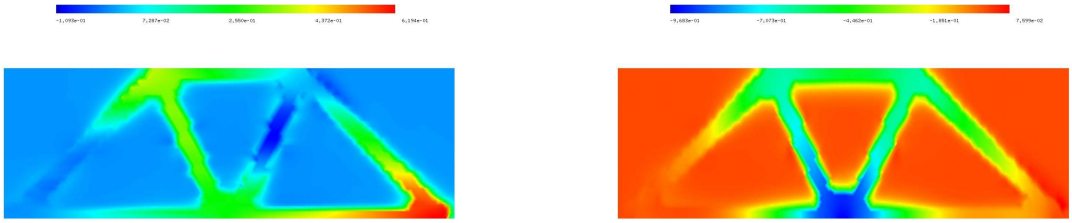

Figure 5.7: Final values of the displacement components u_1 and u_2, respectively.

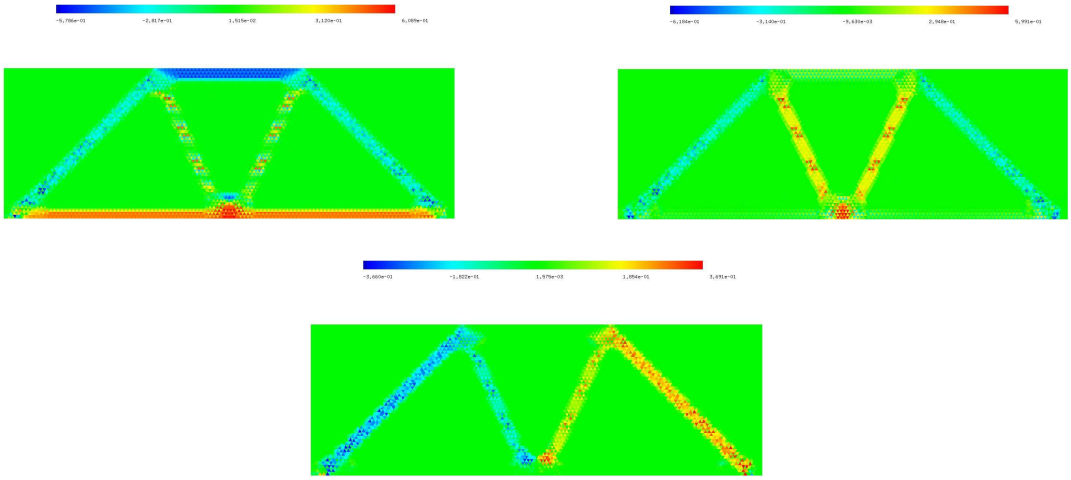

Figure 5.8: Final values of the stress components σ_{11}, σ_{22} and σ_{12}, respectively.

Chapter 6

An Optimal Solver to a KKT-System

Over the last two decades interior-point methods turned out to be efficient optimization methods for solving large-scale nonlinear optimization problems (cf. Subsection 2.1.3). Most of the computing time is actually spent in the solution of linear systems like (2.8) arising from the linearization of the primal-dual equations (2.7). Instead of solving the nonsymmetric system (2.8), a symmetric system like

$$\begin{pmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \triangle \mathbf{x} \\ \triangle \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix} \tag{6.1}$$

can be achieved by some elimination steps. For this system we have to ensure that $\mathbf{A}$ is positive definite when projected onto the null space of the matrix $\mathbf{B}$ (cf. the V_0-ellipticity (2.16)). It might also happen that $\mathbf{B}^T$ does not have full rank, as a consequence the system matrix in (6.1) is singular. Thus, it might be necessary to modify the matrix in the following way to sustain regularity:

$$\begin{pmatrix} \mathbf{A} + \delta_1 \mathbf{I} & \mathbf{B}^T \\ \mathbf{B} & -\delta_2 \mathbf{I} \end{pmatrix} \begin{pmatrix} \triangle \mathbf{x} \\ \triangle \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix},$$

with some small $\delta_1, \delta_2 \geq 0$ (see also Section 6.3).

Multigrid methods certainly belong to the most efficient methods for solving large-scale systems, arising from discretized partial differential equations. While the construction of such methods for symmetric and positive definite systems, like resulting from a discretization from (2.11), is quite standard, this is not the case for saddle point problems. A successful construction of a solver with optimal complexity for linear systems like (6.1) would yield a significant speedup for an interior-point method. One of the most important ingredients of an efficient multigrid method is an appropriate smoother, i.e. a simple iterative smoothing procedure (cf. Subsection 2.3.3). In this chapter we consider a multiplicative Schwarz-type iteration method as a smoother in a multigrid method. Each iteration step of such a multiplicative Schwarz-type smoother consists of the solution of several small local saddle point problems, i.e. small local version of the problem (6.1).

More information about this kind of smoother will be given in Section 6.2. In the next section, as a starting point, we will deduce a saddle point problem from the primal-dual optimality conditions for the optimization problem (5.21) presented in the previous chapter.

Finally, in Section 6.3 we will present some numerical experiments from the application of a multigrid method with the mentioned smoother to the derived saddle point problem.

6.1 The Optimality System

In this section we will derive an optimality system for the optimization systems (5.21) and (5.22) from the previous Chapter 5. The derivation can be performed for both stress criteria, total stress and conservative von Mises stress, but we restrict ourselves to the case of total stress for sake of simplicity. As a starting point we reconsider the following optimization problem:

$$
\begin{aligned}
J_\epsilon(\rho) &\rightarrow \min_{\rho,\mathbf{u},\mathbf{s}} && && \text{(6.2a)}\\
\operatorname{div}\mathbf{s} &= 0, && \text{in } \Omega, && \text{(6.2b)}\\
\boldsymbol{\sigma} - \mathbf{C}\boldsymbol{\varepsilon}(\mathbf{u}) &= \mathbf{0}, && \text{in } \Omega, && \text{(6.2c)}\\
\mathbf{u} &= \mathbf{0}, && \text{on } \Gamma_u, && \text{(6.2d)}\\
\mathbf{s}\cdot\mathbf{n} &= \mathbf{t}, && \text{on } \Gamma_t, && \text{(6.2e)}\\
\mathbf{s}\cdot\mathbf{n} &= \mathbf{0}, && \text{on } \Gamma_{t_0}, && \text{(6.2f)}\\
\rho &= 1, && \text{on } \Gamma_t, && \text{(6.2g)}\\
-(1-\rho)\mathbf{1} \;\le\; \beta(\boldsymbol{\sigma}-\mathbf{s}) &\le\; (1-\rho)\mathbf{1}, && \text{in } \Omega, && \text{(6.2h)}\\
0 \;\le\; \rho(\mathbf{x}) &\le\; 1, && \text{a.e. in } \Omega, && \text{(6.2i)}\\
\rho(\mathbf{x})\boldsymbol{\sigma}^{\min} \;\le\; \mathbf{s}(\mathbf{x}) &\le\; \rho(\mathbf{x})\boldsymbol{\sigma}^{\max}, && \text{a.e. in } \Omega, && \text{(6.2j)}\\
\mathbf{u}^{\min} \;\le\; \mathbf{u}(\mathbf{x}) &\le\; \mathbf{u}^{\max}, && \text{a.e. in } \Omega. && \text{(6.2k)}
\end{aligned}
$$

with

$$
J_\epsilon(\rho) = \gamma \int_\Omega \rho(\mathbf{x})\; d\mathbf{x} + \frac{\epsilon}{2}\int_\Omega \left|\nabla\rho(\mathbf{x})\right|^2\; d\mathbf{x} + \frac{1}{\epsilon}\int_\Omega \rho(\mathbf{x})\big(1-\rho(\mathbf{x})\big)\; d\mathbf{x},
$$

and the function space setting

$$
(\rho,\mathbf{u},\mathbf{s}) \in \big(H^1(\Omega)\cap L_\infty(\Omega)\big)\times\big(H^1\big(\Omega;\mathbb{R}^2\big)\cap L_\infty\big(\Omega;\mathbb{R}^2\big)\big)\times L_\infty\big(\Omega;\mathbb{R}^{2\times 2}\big).
$$

Again, as in Section 5.2, we understand the equality constraints (6.2b), (6.2e), and (6.2f) in a weak sense. Because we aim at an interior-point optimization method to solve this problems, we will perform the derivation in a primal-dual interior-point framework. Especially, we consider a primal-dual barrier method to solve nonlinear optimization problems of the form

$$
\begin{aligned}
f(\mathbf{x}) &\rightarrow \min_{\mathbf{x}\in\mathbb{R}^n}\\
\text{subject to}\qquad c(\mathbf{x}) &= \mathbf{0},\\
\mathbf{x}^{\min} \;\le\; \mathbf{x} &\le\; \mathbf{x}^{\max},
\end{aligned}
$$

see, e.g. Wächter and Biegler [147]. Problems with inequality constraints, like (6.2), can be reformulated in the above form by introducing slack variables. Thus, we rewrite the inequality constraints (6.2h) and (6.2i) as equalities with the additional functions $\mathbf{z}_i \in L_2(\Omega;\mathbb{R}^{2\times 2})$, $i = 1,\dots,4$. Furthermore, to ease the representation of the problem, we **omit** the boundary conditions (6.2d) - (6.2g), and rely on their proper treatment throughout the

reformulations. As an additional simplification we eliminate $\boldsymbol{\sigma}$ using the identity (6.2c). Consequently, we end up with the following optimization problem:

$$
\begin{aligned}
J_\epsilon(\rho) \;&\to\; \min_{\rho,\mathbf{u},\mathbf{s},\mathbf{z}} && && (6.3a)\\
\operatorname{div}\mathbf{s}(\mathbf{x}) \;&=\; \mathbf{0}, && \text{in } \Omega, && (6.3b)\\
-\big(1-\rho(\mathbf{x})\big)\mathbf{1} - \beta\big(\mathbf{C}\boldsymbol{\varepsilon}\big(\mathbf{u}(\mathbf{x})\big) - \mathbf{s}(\mathbf{x})\big) + \mathbf{z}_1(\mathbf{x}) \;&=\; \mathbf{0}, && \text{in } \Omega, && (6.3c)\\
-\big(1-\rho(\mathbf{x})\big)\mathbf{1} + \beta\big(\mathbf{C}\boldsymbol{\varepsilon}\big(\mathbf{u}(\mathbf{x})\big) - \mathbf{s}(\mathbf{x})\big) + \mathbf{z}_2(\mathbf{x}) \;&=\; \mathbf{0}, && \text{in } \Omega, && (6.3d)\\
\rho(\mathbf{x})\boldsymbol{\sigma}^{\min} - \mathbf{s}(\mathbf{x}) + \mathbf{z}_3(\mathbf{x}) \;&=\; \mathbf{0}, && \text{in } \Omega, && (6.3e)\\
\mathbf{s}(\mathbf{x}) - \rho(\mathbf{x})\boldsymbol{\sigma}^{\max} + \mathbf{z}_4(\mathbf{x}) \;&=\; \mathbf{0}, && \text{in } \Omega, && (6.3f)\\
0 \;\le\; \rho(\mathbf{x}) \;&\le\; 1, && \text{a.e. in } \Omega, && (6.3g)\\
\mathbf{u}^{\min} \;\le\; \mathbf{u}(\mathbf{x}) \;&\le\; \mathbf{u}^{\max}, && \text{a.e. in } \Omega, && (6.3h)\\
\mathbf{z}_i(\mathbf{x}) \;&\ge\; \mathbf{0}, && \text{a.e. in } \Omega,\ i=1,\dots,4. && (6.3i)
\end{aligned}
$$

Interior-point methods propose to add the bound constraints to the objective functional and treat them implicitly by using a barrier function. With a barrier parameter $\mu > 0$ this leads to the barrier objective functional

$$
\begin{aligned}
J_{\epsilon,\mu}(\rho) = J_\epsilon(\rho) - \mu \Bigg(&\int_\Omega \ln\big(\rho(\mathbf{x})\big) + \ln\big(1-\rho(\mathbf{x})\big)\, d\mathbf{x} + \\
&+ \int_\Omega \ln\big(\mathbf{u}(\mathbf{x}) - \mathbf{u}^{\min}\big) + \ln\big(\mathbf{u}^{\max} - \mathbf{u}(\mathbf{x})\big)\, d\mathbf{x} + \int_\Omega \ln\big(\mathbf{z}(\mathbf{x})\big)\, d\mathbf{x} \Bigg),
\end{aligned}
$$

where we write $\ln\mathbf{u}$ instead of $\ln u_1 + \ln u_2$ for $\mathbf{u} \in \mathbb{R}^2$ and $\ln\mathbf{z}$ instead of $\sum_{i,j=1}^2 \ln z_{ij}$ for $\mathbf{z} \in \mathbb{R}^{2\times 2}$ for simplicity. The corresponding barrier problem then looks like

$$
\begin{aligned}
J_{\epsilon,\mu}(\rho) \;&\to\; \min_{\rho,\mathbf{u},\mathbf{s},\mathbf{z}} && \\
\operatorname{div}\mathbf{s}(\mathbf{x}) \;&=\; \mathbf{0}, && \text{in } \Omega,\\
-\big(1-\rho(\mathbf{x})\big)\mathbf{1} - \beta\big(\mathbf{C}\boldsymbol{\varepsilon}\big(\mathbf{u}(\mathbf{x})\big) - \mathbf{s}(\mathbf{x})\big) + \mathbf{z}_1(\mathbf{x}) \;&=\; \mathbf{0}, && \text{in } \Omega,\\
-\big(1-\rho(\mathbf{x})\big)\mathbf{1} + \beta\big(\mathbf{C}\boldsymbol{\varepsilon}\big(\mathbf{u}(\mathbf{x})\big) - \mathbf{s}(\mathbf{x})\big) + \mathbf{z}_2(\mathbf{x}) \;&=\; \mathbf{0}, && \text{in } \Omega,\\
\rho(\mathbf{x})\boldsymbol{\sigma}^{\min} - \mathbf{s}(\mathbf{x}) + \mathbf{z}_3(\mathbf{x}) \;&=\; \mathbf{0}, && \text{in } \Omega,\\
\mathbf{s}(\mathbf{x}) - \rho(\mathbf{x})\boldsymbol{\sigma}^{\min} + \mathbf{z}_4(\mathbf{x}) \;&=\; \mathbf{0}, && \text{in } \Omega,
\end{aligned}
\tag{6.4}
$$

In order to formulate the first order necessary conditions for (6.4), we consider the Lagrangian for the above problem. For this sake we introduce Lagrange multipliers $\boldsymbol{\lambda}_0 \in H_0^1(\Omega;\mathbb{R}^2)$, $\boldsymbol{\lambda}_1,\dots,\boldsymbol{\lambda}_4 \in L_2(\Omega;\mathbb{R}^{2\times 2})$ and state

$$
\begin{aligned}
\mathcal{L}(\rho,\mathbf{u},\mathbf{s},\mathbf{z},\boldsymbol{\lambda}) \;= J_{\epsilon,\mu}(\rho) - \langle \mathbf{s}, \varepsilon(\boldsymbol{\lambda}_0)\rangle \;+\; &\big\langle -(1-\rho)\mathbf{1} - \beta\big(\mathbf{C}\boldsymbol{\varepsilon}(\mathbf{u}) - \mathbf{s}\big) + \mathbf{z}_1, \boldsymbol{\lambda}_1\big\rangle + \\
+\; &\big\langle -(1-\rho)\mathbf{1} + \beta\big(\mathbf{C}\boldsymbol{\varepsilon}(\mathbf{u}) - \mathbf{s}\big) + \mathbf{z}_2, \boldsymbol{\lambda}_2\big\rangle + \\
+\; &\big\langle \rho\boldsymbol{\sigma}^{\min} - \mathbf{s} + \mathbf{z}_3, \boldsymbol{\lambda}_3\big\rangle \;+\; \big\langle \mathbf{s} - \rho\boldsymbol{\sigma}^{\max} + \mathbf{z}_4, \boldsymbol{\lambda}_4\big\rangle,
\end{aligned}
\tag{6.5}
$$

with the notation $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_0, \dots, \boldsymbol{\lambda}_4)$ and $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_4)$. The optimality conditions then look as:

$$
\begin{aligned}
\nabla_\rho \mathcal{L} &= \gamma + \epsilon\Delta\rho + \frac{1}{\epsilon}(1-2\rho) - \frac{\mu}{\rho} + \frac{\mu}{1-\rho} + \mathbf{1} : \boldsymbol{\lambda}_1 + \\
&\qquad + \mathbf{1} : \boldsymbol{\lambda}_2 + \boldsymbol{\sigma}^{\min} : \boldsymbol{\lambda}_3 - \boldsymbol{\sigma}^{\max} : \boldsymbol{\lambda}_4 = 0, && \text{(6.6a)}\\
\nabla_{\mathbf{u}} \mathcal{L} &= -\frac{\mu}{\mathbf{u} - \mathbf{u}^{\min}} + \frac{\mu}{\mathbf{u}^{\max} - \mathbf{u}} + \beta\mathbf{C}\,\mathrm{div}\,\boldsymbol{\lambda}_1 - \beta\mathbf{C}\,\mathrm{div}\,\boldsymbol{\lambda}_2 = \mathbf{0}, && \text{(6.6b)}\\
\nabla_{\mathbf{s}} \mathcal{L} &= -\boldsymbol{\varepsilon}(\boldsymbol{\lambda}_0) + \beta\boldsymbol{\lambda}_1 - \beta\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_3 + \boldsymbol{\lambda}_4 = \mathbf{0}, && \text{(6.6c)}\\
\nabla_{\mathbf{z}_i} \mathcal{L} &= -\frac{\mu}{\mathbf{z}_i} + \boldsymbol{\lambda}_i = \mathbf{0}, && \text{(6.6d)}\\
\nabla_{\boldsymbol{\lambda}_0} \mathcal{L} &= \mathrm{div}\,\mathbf{s} = \mathbf{0}, && \text{(6.6e)}\\
\nabla_{\boldsymbol{\lambda}_1} \mathcal{L} &= -(1-\rho)\mathbf{1} - \beta\big(\mathbf{C}\boldsymbol{\varepsilon}(\mathbf{u}) - \mathbf{s}\big) + \mathbf{z}_1 = \mathbf{0}, && \text{(6.6f)}\\
\nabla_{\boldsymbol{\lambda}_2} \mathcal{L} &= -(1-\rho)\mathbf{1} + \beta\big(\mathbf{C}\boldsymbol{\varepsilon}(\mathbf{u}) - \mathbf{s}\big) + \mathbf{z}_2 = \mathbf{0}, && \text{(6.6g)}\\
\nabla_{\boldsymbol{\lambda}_3} \mathcal{L} &= \rho\boldsymbol{\sigma}^{\min} - \mathbf{s} + \mathbf{z}_3 = \mathbf{0}, && \text{(6.6h)}\\
\nabla_{\boldsymbol{\lambda}_4} \mathcal{L} &= \mathbf{s} - \rho\boldsymbol{\sigma}^{\max} + \mathbf{z}_4 = \mathbf{0}. && \text{(6.6i)}
\end{aligned}
$$

In (6.6b) and (6.6d) (and further on) the fractions are meant by components. Moreover, we use the identity

$$\int_\Omega \mathrm{div}\,\boldsymbol{\sigma} \cdot \mathbf{v}\; d\mathbf{x} = -\int_\Omega \boldsymbol{\sigma} : \boldsymbol{\varepsilon}(\mathbf{v})\; d\mathbf{x}$$

for $\mathbf{v} \in H_0^1(\Omega)$ for the derivation of (6.6) and for the statement of the Lagrangian (6.5). Moreover, the equality (6.6b) again has to be understood in a weak sense. In the spirit of primal-dual interior point methods we now introduce new independent variables ν_i that act as multipliers for the bound constraints (6.3g) - (6.3i). In particular we choose $\nu_1, \nu_2 \in H^1(\Omega)$, $\boldsymbol{\nu}_3, \boldsymbol{\nu}_4 \in H^1(\Omega; \mathbb{R}^2)$, and $\boldsymbol{\nu}_i \in L_2(\Omega; \mathbb{R}^{2\times 2})$ for $i = 5, \dots, 8$, such that

$$
\begin{gathered}
\nu_1 = \frac{\mu}{\rho}, \quad \nu_2 = \frac{\mu}{1-\rho}, \quad \boldsymbol{\nu}_3 = \frac{\mu}{\mathbf{u} - \mathbf{u}^{\min}}, \quad \boldsymbol{\nu}_4 = \frac{\mu}{\mathbf{u}^{\max} - \mathbf{u}},\\
\boldsymbol{\nu}_5 = \frac{\mu}{\mathbf{z}_1}, \quad \boldsymbol{\nu}_6 = \frac{\mu}{\mathbf{z}_2}, \quad \boldsymbol{\nu}_7 = \frac{\mu}{\mathbf{z}_3}, \quad \boldsymbol{\nu}_8 = \frac{\mu}{\mathbf{z}_4}.
\end{gathered}
\tag{6.7}
$$

Using the definition (6.7) of the dual variables, the optimality conditions (6.6) turn into the following system in the primal variables ρ, $\mathbf{u}$, $\mathbf{s}$, $\mathbf{z}$, and the dual variables $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$:

$$\gamma + \epsilon\Delta\rho + \frac{1}{\epsilon}(1-2\rho) - \nu_1 + \nu_2 + \mathbf{1} : \boldsymbol{\lambda}_1 + \mathbf{1} : \boldsymbol{\lambda}_2 + \boldsymbol{\sigma}^{\min} : \boldsymbol{\lambda}_3 - \boldsymbol{\sigma}^{\max} : \boldsymbol{\lambda}_4 = 0, \tag{6.8a}$$

$$-\boldsymbol{\nu}_3 + \boldsymbol{\nu}_4 + \beta\mathbf{C}\,\mathrm{div}\,\boldsymbol{\lambda}_1 - \beta\mathbf{C}\,\mathrm{div}\,\boldsymbol{\lambda}_2 = \mathbf{0}, \tag{6.8b}$$

$$-\boldsymbol{\varepsilon}(\boldsymbol{\lambda}_0) + \beta\boldsymbol{\lambda}_1 - \beta\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_3 + \boldsymbol{\lambda}_4 = \mathbf{0}, \tag{6.8c}$$

$$-\boldsymbol{\nu}_5 + \boldsymbol{\lambda}_1 = -\boldsymbol{\nu}_6 + \boldsymbol{\lambda}_2 = -\boldsymbol{\nu}_7 + \boldsymbol{\lambda}_3 = -\boldsymbol{\nu}_8 + \boldsymbol{\lambda}_4 = \mathbf{0}, \tag{6.8d}$$

$$\mathrm{div}\,\mathbf{s} = \mathbf{0}, \tag{6.8e}$$

$$-\big(1-\rho\big)\mathbf{1} - \beta\big(\mathbf{C}\boldsymbol{\varepsilon}(\mathbf{u}) - \mathbf{s}\big) + \mathbf{z}_1 = \mathbf{0}, \tag{6.8f}$$

$$-\big(1-\rho\big)\mathbf{1} + \beta\big(\mathbf{C}\boldsymbol{\varepsilon}(\mathbf{u}) - \mathbf{s}\big) + \mathbf{z}_2 = \mathbf{0}, \tag{6.8g}$$

$$\rho\boldsymbol{\sigma}^{\min} - \mathbf{s} + \mathbf{z}_3 = \mathbf{0}, \tag{6.8h}$$

$$\mathbf{s} - \rho\boldsymbol{\sigma}^{\max} + \mathbf{z}_4 = \mathbf{0}, \tag{6.8i}$$

$$-\rho + \frac{\mu}{\nu_1} = 0, \tag{6.8j}$$

$$\rho - 1 + \frac{\mu}{\nu_2} = 0, \tag{6.8k}$$

$$-\big(\mathbf{u} - \mathbf{u}^{\min}\big) + \frac{\mu}{\boldsymbol{\nu}_3} = \mathbf{0}, \tag{6.8l}$$

$$-\big(\mathbf{u}^{\max} - \mathbf{u}\big) + \frac{\mu}{\boldsymbol{\nu}_4} = \mathbf{0}, \tag{6.8m}$$

$$-\mathbf{z}_1 + \frac{\mu}{\boldsymbol{\nu}_5} = -\mathbf{z}_2 + \frac{\mu}{\boldsymbol{\nu}_6} = -\mathbf{z}_3 + \frac{\mu}{\boldsymbol{\nu}_7} = -\mathbf{z}_4 + \frac{\mu}{\boldsymbol{\nu}_8} = \mathbf{0}, \tag{6.8n}$$

where the form of the equalities (6.8j) - (6.8n) is motivated to get a symmetric system matrix after discretization.

In the following we consider the discretization of the primal-dual equations (6.8). In order to construct a finite element approximation we assume that $\overline{\Omega} = \bigcup_{i=1}^{n} \overline{\tau}_i$ is partitioned into a proper triangulation $\mathcal{T} = \{\tau_i \mid i = 1, \ldots, n\}$ with n triangles τ_i. We shall use two different finite elements for the primal and dual variables. For the density ρ, the components of the displacements $\mathbf{u}$, the dual variables ν_1, ν_2, and the components of the dual variables $\boldsymbol{\nu}_3$, $\boldsymbol{\nu}_4$, we use the discrete H^1-subspace of linear elements

$$V^h := \big\{\tilde{v} \in C(\Omega);\ \big|\ \tilde{v}|_{\tau_i} \in \mathcal{P}_1(\tau_i),\ i = 1, \ldots, n\big\}.$$

For the components of the Lagrangian multiplier $\boldsymbol{\lambda}_0$ we use the discrete H_0^1-subspace V_0^h of linear elements with zero boundary conditions. The components of the stress $\mathbf{s}$, the slack variables $\mathbf{z}$, the dual variables $\boldsymbol{\nu}_5, \ldots \boldsymbol{\nu}_8$, and of the Lagrange multipliers $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_4$ are approximated by the L_∞-subspace of constant elements

$$Q^h := \big\{\tilde{q} \in L_\infty(\Omega)\ \big|\ \tilde{q}|_{\tau_i} \in \mathcal{P}_0(\tau_i),\ i = 1, \ldots, n\big\}.$$

As in the previous chapters, $\mathcal{P}_k(\tau_i)$ represents the space of polynomials of maximal degree k over the triangle τ_i. Using these finite element approximations we discretize the system (6.8) by a standard finite element approach, i.e. we consider the weak formulations of the equations (6.8a) - (6.8n) and perform a partial integration for the divergence terms in the equations (6.8b) and (6.8e). Taking into account the Hilbert spaces $V = H^1(\Omega)$, $V_0 = H_0^1(\Omega)$, $V_{\Gamma_u} = H^1_{\Gamma_u}(\Omega)$, and $Q = L_2(\Omega)$ and having the application of a Newton type method in mind,

we can write the weak linearized formulation of (6.8) in the following way: Find updates $\rho, \nu_1, \nu_2 \in V$, $\mathbf{u}, \boldsymbol{\nu}_3, \boldsymbol{\nu}_4 \in V_{\Gamma_u}^2$, $\boldsymbol{\lambda}_0 \in V_0^2$, and $\mathbf{s}, \mathbf{z}_1, \ldots, \mathbf{z}_4, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_4, \boldsymbol{\nu}_5, \ldots, \boldsymbol{\nu}_8 \in Q^{2\times 2}$ for the current values of $\overline{\rho}, \overline{\mathbf{u}}, \overline{\mathbf{s}}, \overline{\mathbf{z}}, \overline{\boldsymbol{\lambda}}, \overline{\nu}_1, \overline{\nu}_2, \overline{\boldsymbol{\nu}}_3, \ldots, \overline{\boldsymbol{\nu}}_8$, respectively, such that

$$
\begin{aligned}
-\epsilon(\nabla\rho, \nabla v)_0 - \frac{2}{\epsilon}(\rho, v)_0 - (\nu_1, v)_0 + (\nu_2, v)_0 + (\mathbf{1} : \boldsymbol{\lambda}_1, v)_0 + \\
+ (\mathbf{1} : \boldsymbol{\lambda}_2, v)_0 + (\boldsymbol{\sigma}^{\min} : \boldsymbol{\lambda}_3, v)_0 + (\boldsymbol{\sigma}^{\max} : \boldsymbol{\lambda}_4, v)_0 &= -\big(\gamma + \frac{1}{\epsilon}\big)(1, v)_0, \\
-(\boldsymbol{\nu}_3, \boldsymbol{\phi})_0 + (\boldsymbol{\nu}_4, \boldsymbol{\phi})_0 - \beta\big(\mathbf{C}\boldsymbol{\lambda}_1, \boldsymbol{\varepsilon}(\boldsymbol{\phi})\big)_0 + \beta\big(\mathbf{C}\boldsymbol{\lambda}_2, \boldsymbol{\varepsilon}(\boldsymbol{\phi})\big)_0 &= 0, \\
-\big(\boldsymbol{\varepsilon}(\boldsymbol{\lambda}_0), \mathbf{q}\big)_0 + \beta(\boldsymbol{\lambda}_1, \mathbf{q})_0 - \beta(\boldsymbol{\lambda}_2, \mathbf{q})_0 - (\boldsymbol{\lambda}_3, \mathbf{q})_0 + (\boldsymbol{\lambda}_4, \mathbf{q})_0 &= 0, \\
-(\boldsymbol{\nu}_5, \mathbf{q})_0 + (\boldsymbol{\lambda}_1, \mathbf{q})_0 = -(\boldsymbol{\nu}_6, \mathbf{q})_0 + (\boldsymbol{\lambda}_2, \mathbf{q})_0 &= \\
-(\boldsymbol{\nu}_7, \mathbf{q})_0 + (\boldsymbol{\lambda}_3, \mathbf{q})_0 = -(\boldsymbol{\nu}_8, \mathbf{q})_0 + (\boldsymbol{\lambda}_4, \mathbf{q})_0 &= 0, \\
-\big(\mathbf{s}, \boldsymbol{\varepsilon}(\boldsymbol{\psi})\big)_0 &= \int_{\Gamma_t} \mathbf{t} \cdot \boldsymbol{\psi}\, d\mathbf{x}, \\
(\rho\mathbf{1}, \mathbf{q})_0 - \beta\big(\mathbf{C}\boldsymbol{\varepsilon}(\mathbf{u}), \mathbf{q}\big)_0 + \beta(\mathbf{s}, \mathbf{q})_0 + (\mathbf{z}_1, \mathbf{q})_0 &= (\mathbf{1}, \mathbf{q})_0, \\
(\rho\mathbf{1}, \mathbf{q})_0 + \beta\big(\mathbf{C}\boldsymbol{\varepsilon}(\mathbf{u}), \mathbf{q}\big)_0 - \beta(\mathbf{s}, \mathbf{q})_0 + (\mathbf{z}_2, \mathbf{q})_0 &= (\mathbf{1}, \mathbf{q})_0, \\
(\rho\boldsymbol{\sigma}^{\min}, \mathbf{q})_0 - (\mathbf{s}, \mathbf{q})_0 + (\mathbf{z}_3, \mathbf{q})_0 &= 0, \\
(\mathbf{s}, \mathbf{q})_0 - (\rho\boldsymbol{\sigma}^{\max}, \mathbf{q})_0 + (\mathbf{z}_4, \mathbf{q})_0 &= 0, \\
-(\rho, v)_0 + \frac{\mu}{\overline{\nu}_1}(\nu_1, v)_0 &= 0, \\
(\rho, v)_0 + \frac{\mu}{\overline{\nu}_2}(\nu_2, v)_0 &= (1, v)_0, \\
-(\mathbf{u}, \boldsymbol{\phi})_0 + \frac{\mu}{\overline{\boldsymbol{\nu}}_3}(\boldsymbol{\nu}_3, \boldsymbol{\phi})_0 &= -(\mathbf{u}^{\min}, \boldsymbol{\phi})_0, \\
(\mathbf{u}, \boldsymbol{\phi})_0 + \frac{\mu}{\overline{\boldsymbol{\nu}}_4}(\boldsymbol{\nu}_4, \boldsymbol{\phi})_0 &= (\mathbf{u}^{\max}, \boldsymbol{\phi})_0, \\
-(\mathbf{z}_1, \mathbf{q})_0 + \frac{\mu}{\overline{\boldsymbol{\nu}}_5}(\boldsymbol{\nu}_5, \mathbf{q})_0 = -(\mathbf{z}_2, \mathbf{q})_0 + \frac{\mu}{\overline{\boldsymbol{\nu}}_6}(\boldsymbol{\nu}_6, \mathbf{q})_0 &= \\
= -(\mathbf{z}_3, \mathbf{q})_0 + \frac{\mu}{\overline{\boldsymbol{\nu}}_7}(\boldsymbol{\nu}_7, \mathbf{q})_0 = -(\mathbf{z}_4, \mathbf{q})_0 + \frac{\mu}{\overline{\boldsymbol{\nu}}_8}(\boldsymbol{\nu}_8, \mathbf{q})_0 &= 0,
\end{aligned}
$$

where the above equalities shall hold for all test functions $v \in V$, $\boldsymbol{\phi} \in V_0^2$, $\boldsymbol{\psi} \in V_{\Gamma_u}^2$, and $\mathbf{q} \in Q^{2\times 2}$. Let the vectors $\triangle\boldsymbol{\rho}^h$, $\triangle\mathbf{u}^h$, and so on, contain the coefficients of the finite element functions $\tilde{\rho} \in V^h$ and $\tilde{\mathbf{u}} \in (V_{\Gamma_u}^h)^2$, and so on, respectively. We add the symbol $\triangle$ to emphasis that we consider the updates of current iterates in a Newton type iteration method. Moreover, we use the symmetries in the occurring variables $\mathbf{s}$, $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_4$, $\mathbf{z}_1, \ldots, \mathbf{z}_4$, and $\boldsymbol{\nu}_5, \ldots, \boldsymbol{\nu}_8$ (e.g. $s_{ij} = s_{ji}$) to reduce the number of unknowns. The discretized problem can now be written

as:

$$
\begin{aligned}
-\epsilon\mathbf{K}\triangle\boldsymbol{\rho}^h - \frac{2}{\epsilon}\mathbf{M}\triangle\boldsymbol{\rho}^h - \mathbf{M}\triangle\boldsymbol{\nu}_1^h + \mathbf{M}\triangle\boldsymbol{\nu}_2^h + \tilde{\mathbf{N}}^T\triangle\boldsymbol{\lambda}_1^h + \\
+ \tilde{\mathbf{N}}^T\triangle\boldsymbol{\lambda}_2^h + \tilde{\mathbf{N}}^T\boldsymbol{\Sigma}^{\min}\triangle\boldsymbol{\lambda}_3^h - \tilde{\mathbf{N}}^T\boldsymbol{\Sigma}^{\max}\triangle\boldsymbol{\lambda}_4^h &= -\big(\gamma + \frac{1}{\epsilon}\big)\mathbf{e}_{V^h}^h, && \text{(6.9a)}\\
-\mathbf{M}_2\triangle\boldsymbol{\nu}_3^h + \mathbf{M}_2\triangle\boldsymbol{\nu}_4^h - \beta\mathbf{D}^T\mathbf{C}^h\triangle\boldsymbol{\lambda}_1^h + \beta\mathbf{D}^T\mathbf{C}^h\triangle\boldsymbol{\lambda}_2^h &= \mathbf{0}, && \text{(6.9b)}\\
-\mathbf{D}\triangle\boldsymbol{\lambda}_0^h + \beta\mathbf{N}\triangle\boldsymbol{\lambda}_1^h - \beta\mathbf{N}\triangle\boldsymbol{\lambda}_2^h - \mathbf{N}\triangle\boldsymbol{\lambda}_3^h + \mathbf{N}\triangle\boldsymbol{\lambda}_4^h &= \mathbf{0}, && \text{(6.9c)}\\
-\mathbf{N}\triangle\boldsymbol{\nu}_5^h + \mathbf{N}\triangle\boldsymbol{\lambda}_1^h = -\mathbf{N}\triangle\boldsymbol{\nu}_6^h + \mathbf{N}\triangle\boldsymbol{\lambda}_2^h &= \\
= -\mathbf{N}\triangle\boldsymbol{\nu}_7^h + \mathbf{N}\triangle\boldsymbol{\lambda}_3^h = -\mathbf{N}\triangle\boldsymbol{\nu}_8^h + \mathbf{N}\triangle\boldsymbol{\lambda}_4^h &= \mathbf{0}, && \text{(6.9d)}\\
-\mathbf{D}^T\triangle\mathbf{s}^h &= \mathbf{t}^h, && \text{(6.9e)}\\
\tilde{\mathbf{N}}\triangle\boldsymbol{\rho}^h - \beta\mathbf{C}^h\mathbf{D}\triangle\mathbf{u}^h + \beta\mathbf{N}\triangle\mathbf{s}^h + \mathbf{N}\triangle\mathbf{z}_1^h &= \mathbf{e}_{(Q^h)^3}^h, && \text{(6.9f)}\\
\tilde{\mathbf{N}}\triangle\boldsymbol{\rho}^h + \beta\mathbf{C}^h\mathbf{D}\triangle\mathbf{u}^h - \beta\mathbf{N}\triangle\mathbf{s}^h + \mathbf{N}\triangle\mathbf{z}_2^h &= \mathbf{e}_{(Q^h)^3}^h, && \text{(6.9g)}\\
\boldsymbol{\Sigma}^{\min}\tilde{\mathbf{N}}\triangle\boldsymbol{\rho}^h - \mathbf{N}\triangle s^h + \mathbf{N}\triangle\mathbf{z}_3^h &= \mathbf{0}, && \text{(6.9h)}\\
-\boldsymbol{\Sigma}^{\max}\tilde{\mathbf{N}}\triangle\boldsymbol{\rho}^h + \mathbf{N}\triangle\mathbf{s}^h + \mathbf{N}\triangle\mathbf{z}_4^h &= \mathbf{0}, && \text{(6.9i)}\\
-\mathbf{M}\triangle\boldsymbol{\rho}^h + \mu\mathbf{M}_{\nu_1}\triangle\boldsymbol{\nu}_1^h &= \mathbf{0}, && \text{(6.9j)}\\
\mathbf{M}\triangle\boldsymbol{\rho}^h + \mu\mathbf{M}_{\nu_2}\triangle\boldsymbol{\nu}_2^h &= \mathbf{e}_{V^h}^h, && \text{(6.9k)}\\
-\mathbf{M}_2\triangle\mathbf{u}^h + \mu\mathbf{M}_{\nu_3}\triangle\boldsymbol{\nu}_3 &= -\mathbf{M}_2\mathbf{u}^{\min h}, && \text{(6.9l)}\\
\mathbf{M}_2\triangle\mathbf{u}^h + \mu\mathbf{M}_{\nu_4}\triangle\boldsymbol{\nu}_4 &= \mathbf{M}_2\mathbf{u}^{\max h}, && \text{(6.9m)}\\
-\mathbf{N}\triangle\mathbf{z}_1^h + \mu\mathbf{N}_{\nu_5}\triangle\boldsymbol{\nu}_5^h = -\mathbf{N}\triangle\mathbf{z}_2^h + \mu\mathbf{N}_{\nu_6}\triangle\boldsymbol{\nu}_6^h &= \\
-\mathbf{N}\triangle\mathbf{z}_3^h + \mu\mathbf{N}_{\nu_7}\triangle\boldsymbol{\nu}_7^h = -\mathbf{N}\triangle\mathbf{z}_4^h + \mu\mathbf{N}_{\nu_8}\triangle\boldsymbol{\nu}_8^h &= \mathbf{0}. && \text{(6.9n)}
\end{aligned}
$$

In (6.9a), $\mathbf{K}$ is a stiffness matrix arising from the finite element discretization of the Laplacian in V^h and $\mathbf{M}$ is a mass matrix for the identity in V^h. Furthermore, $\tilde{\mathbf{N}}$ is mixed mass matrix between the spaces V^h and $(Q^h)^3$. $\mathbf{e}_{V^h}^h$ and $\mathbf{e}_{(Q^h)^3}^h$ are vectors representing the coefficients of the constant function 1 with respect to the spaces V^h and $(Q^h)^3$, respectively. $\boldsymbol{\Sigma}^{\min}$ and $\boldsymbol{\Sigma}^{\max}$ are diagonal matrices representing the corresponding entries of $\boldsymbol{\sigma}^{\min}$ and $\boldsymbol{\sigma}^{\max}$, respectively. In (6.9b), $\mathbf{M}_2$ is a mass matrix for the identity in $V_{\Gamma_u}^h \times V_{\Gamma_u}^h$ and $\mathbf{C}^h$ is the discrete analogon of elasticity tensor $\mathbf{C}$. Moreover, $\mathbf{D}^T$ is the representation of the divergence operator (restricted to symmetric stress tensors). The mass matrix $\mathbf{N}$ in (6.9c) represents the mass matrix for the identity in $(Q^h)^3$. In the discretized partial differential equation (6.9e) $\mathbf{t}^h$ is a discrete representation of the traction forces. Moreover, in the equations (6.9j) - (6.9m) the matrices $\mathbf{M}_{\nu_1}, \ldots, \mathbf{M}_{\nu_4}$, and the matrices $\mathbf{N}_{\nu_5}, \ldots, \mathbf{N}_{\nu_8}$, in (6.9n) are weighted mass matrices with the weights $\nu_1, \ldots, \boldsymbol{\nu}_8$, respectively.

The linear system (6.9) can be written in a compact representation as

$$\boldsymbol{\mathcal{K}}\triangle\mathbf{x} = \mathbf{f}^h, \tag{6.10}$$

with

$$\triangle\mathbf{x} = \big(\triangle\boldsymbol{\rho}^h, \triangle\mathbf{u}^h, \triangle\mathbf{s}^h, \triangle\mathbf{z}_1^h, \ldots, \triangle\mathbf{z}_4^h, \triangle\boldsymbol{\lambda}_0^h, \ldots, \triangle\boldsymbol{\lambda}_4^h, \triangle\boldsymbol{\nu}_1^h, \ldots, \triangle\boldsymbol{\nu}_8^h\big)$$

and

$$\mathbf{f}^h = \Big(-(\gamma+\frac{1}{\epsilon})\mathbf{e}^h_{V^h}, \mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{t}^h, \mathbf{e}^h_{(Q^h)^3}, \mathbf{e}^h_{(Q^h)^3}, \mathbf{0},\mathbf{0},\mathbf{0},$$
$$\mathbf{e}^h_{V^h}, -\mathbf{M}_2\mathbf{u}^{\min^h}, \mathbf{M}_2\mathbf{u}^{\max h}, \mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0}\Big).$$

The coefficient matrix $\boldsymbol{\mathcal{K}}$ in (6.10) contains the matrices in (6.9) as block matrices and turns out to be large (even too large to be printed on one page). In order to reduce the size of the system to a more reasonable one, we reduce the system (but we keep the notation $\boldsymbol{\mathcal{K}}$ for the system matrix and $\mathbf{f}^h$ for the right-hand side after each of the following eliminations) using the following eliminations of the dual variables:

$$\begin{aligned}
\triangle\boldsymbol{\nu}_1^h &= \frac{1}{\mu}\mathbf{M}_{\nu_1}^{-1}\mathbf{M}\triangle\boldsymbol{\rho}^h, \qquad \triangle\boldsymbol{\nu}_2^h = -\frac{1}{\mu}\mathbf{M}_{\nu_2}^{-1}\mathbf{M}\triangle\boldsymbol{\rho}^h + \frac{1}{\mu}\mathbf{M}_{\nu_2}^{-1}\mathbf{e}^h_{V^h},\\
\triangle\boldsymbol{\nu}_3^h &= \frac{1}{\mu}\mathbf{M}_{\nu_3}^{-1}\mathbf{M}_2\triangle\mathbf{u}^h - \frac{1}{\mu}\mathbf{M}_{\nu_3}^{-1}\mathbf{M}_2\mathbf{u}^{\min^h},\\
\triangle\boldsymbol{\nu}_4^h &= -\frac{1}{\mu}\mathbf{M}_{\nu_4}^{-1}\mathbf{M}_2\triangle\mathbf{u}^h + \frac{1}{\mu}\mathbf{M}_{\nu_4}^{-1}\mathbf{M}_2\mathbf{u}^{\max h},\\
\triangle\boldsymbol{\nu}_5^h &= \frac{1}{\mu}\mathbf{N}_{\nu_5}^{-1}\mathbf{N}\triangle\mathbf{z}_1^h, \qquad \triangle\boldsymbol{\nu}_6^h = \frac{1}{\mu}\mathbf{N}_{\nu_6}^{-1}\mathbf{N}\triangle\mathbf{z}_2^h,\\
\triangle\boldsymbol{\nu}_7^h &= \frac{1}{\mu}\mathbf{N}_{\nu_7}^{-1}\mathbf{N}\triangle\mathbf{z}_3^h, \qquad \triangle\boldsymbol{\nu}_8^h = \frac{1}{\mu}\mathbf{N}_{\nu_8}^{-1}\mathbf{N}\triangle\mathbf{z}_4^h.
\end{aligned} \tag{6.11}$$

This first elimination leads to a smaller linear system like

$$\boldsymbol{\mathcal{K}}\triangle\mathbf{x} = \mathbf{f}^h, \tag{6.12}$$

with

$$\triangle\mathbf{x} = \big(\triangle\boldsymbol{\rho}^h, \triangle\mathbf{u}^h, \triangle\mathbf{s}^h, \triangle\mathbf{z}_1^h, \ldots, \triangle\mathbf{z}_4^h, \triangle\boldsymbol{\lambda}_0^h, \ldots, \triangle\boldsymbol{\lambda}_4^h\big)$$

and

$$\mathbf{f}^h = \Big(-(\gamma+\frac{1}{\epsilon})\mathbf{e}^h_{V^h} - \frac{1}{\mu}\mathbf{M}\mathbf{M}_{\nu_2}^{-1}\mathbf{e}^h_{V^h}, -\frac{1}{\mu}\mathbf{M}_2\big(\mathbf{M}_{\nu_3}^{-1}\mathbf{u}^{\min^h} + \mathbf{M}_{\nu_4}^{-1}\mathbf{u}^{\max h}\big),$$
$$\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{t}^h, \mathbf{e}^h_{(Q^h)^3}, \mathbf{e}^h_{(Q^h)^3}, \mathbf{0},\mathbf{0}\Big).$$

The system matrix $\boldsymbol{\mathcal{K}}$ of (6.12) is given by

$$\boldsymbol{\mathcal{K}} = \begin{pmatrix}
\boldsymbol{\mathcal{K}}_{\rho\rho} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \tilde{\mathbf{N}}^T & \tilde{\mathbf{N}}^T & \tilde{\mathbf{N}}^T\boldsymbol{\Sigma}^{\min} & -\tilde{\mathbf{N}}^T\boldsymbol{\Sigma}^{\max}\\
\mathbf{0} & \boldsymbol{\mathcal{K}}_{uu} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\beta\mathbf{D}^T\mathbf{C}^h & \beta\mathbf{D}^T\mathbf{C}^h & \mathbf{0} & \mathbf{0}\\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{D} & \beta\mathbf{N} & -\beta\mathbf{N} & -\mathbf{N} & \mathbf{N}\\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\mathcal{K}}_{z_1z_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{N} & \mathbf{0} & \mathbf{0} & \mathbf{0}\\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\mathcal{K}}_{z_2z_2} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{N} & \mathbf{0} & \mathbf{0}\\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\mathcal{K}}_{z_3z_3} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{N} & \mathbf{0}\\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\mathcal{K}}_{z_4z_4} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{N}\\
\mathbf{0} & \mathbf{0} & -\mathbf{D}^T & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}\\
\tilde{\mathbf{N}} & -\beta\mathbf{C}^h\mathbf{D} & \beta\mathbf{N} & \mathbf{N} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}\\
\tilde{\mathbf{N}} & \beta\mathbf{C}^h\mathbf{D} & -\beta\mathbf{N} & \mathbf{0} & \mathbf{N} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}\\
\boldsymbol{\Sigma}^{\min}\tilde{\mathbf{N}} & \mathbf{0} & -\mathbf{N} & \mathbf{0} & \mathbf{0} & \mathbf{N} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}\\
-\boldsymbol{\Sigma}^{\max}\tilde{\mathbf{N}} & \mathbf{0} & \mathbf{N} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{N} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}
\end{pmatrix},$$

with

$$\begin{aligned}
\boldsymbol{\mathcal{K}}_{\rho\rho} &= -\epsilon\mathbf{K} - \frac{2}{\epsilon}\mathbf{M} - \frac{1}{\mu}\mathbf{M}\big(\mathbf{M}_{\nu_1}^{-1} + \mathbf{M}_{\nu_2}^{-1}\big)\mathbf{M},\\
\boldsymbol{\mathcal{K}}_{uu} &= -\frac{1}{\mu}\mathbf{M}_2\big(\mathbf{M}_{\nu_3}^{-1} + \mathbf{M}_{\nu_4}^{-1}\big)\mathbf{M}_2,\\
\boldsymbol{\mathcal{K}}_{z_iz_i} &= -\frac{1}{\mu}\mathbf{N}\mathbf{N}_{\nu_i}^{-1}\mathbf{N}, \qquad i = 1,\ldots,4.
\end{aligned}$$

As further eliminations we eliminate the slack variables $\triangle \mathbf{z}_i^h$, for $i = 1, \ldots, 4$, from the system (6.12):

$$\triangle \mathbf{z}_i^h = \mu \mathbf{N}^{-1} \mathbf{N}_{\nu_i} \triangle \boldsymbol{\lambda}_i^h, \qquad i = 1, \ldots, 4. \tag{6.13}$$

Hence, we come up with the following system

$$\boldsymbol{\mathcal{K}} \begin{pmatrix} \triangle \boldsymbol{\rho}^h \\ \triangle \mathbf{u}^h \\ \triangle \mathbf{s}^h \\ \triangle \boldsymbol{\lambda}_0^h \\ \triangle \boldsymbol{\lambda}_1^h \\ \triangle \boldsymbol{\lambda}_2^h \\ \triangle \boldsymbol{\lambda}_3^h \\ \triangle \boldsymbol{\lambda}_4^h \end{pmatrix} = \begin{pmatrix} -\big(\gamma + \frac{1}{\epsilon}\big)\mathbf{e}_{V^h}^h - \frac{1}{\mu}\mathbf{M}\mathbf{M}_{\nu_2}^{-1}\mathbf{e}_{V^h}^h \\ -\frac{1}{\mu}\mathbf{M}_2\big(\mathbf{M}_{\nu_3}^{-1}\mathbf{u}^{\min^h} + \mathbf{M}_{\nu_4}^{-1}\mathbf{u}^{\max h}\big) \\ \mathbf{0} \\ \mathbf{t}^h \\ \mathbf{e}_{(Q^h)^3}^h \\ \mathbf{e}_{(Q^h)^3}^h \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} \tag{6.14}$$

with the coefficient matrix

$$\boldsymbol{\mathcal{K}} = \begin{pmatrix} \boldsymbol{\mathcal{K}}_{\rho\rho} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \tilde{\mathbf{N}}^T & \tilde{\mathbf{N}}^T & \tilde{\mathbf{N}}^T\boldsymbol{\Sigma}^{\min} & -\tilde{\mathbf{N}}^T\boldsymbol{\Sigma}^{\max} \\ \mathbf{0} & \boldsymbol{\mathcal{K}}_{uu} & \mathbf{0} & \mathbf{0} & -\beta\mathbf{D}^T\mathbf{C}^h & \beta\mathbf{D}^T\mathbf{C}^h & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{D} & \beta\mathbf{N} & -\beta\mathbf{N} & -\mathbf{N} & \mathbf{N} \\ \mathbf{0} & \mathbf{0} & -\mathbf{D}^T & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \tilde{\mathbf{N}} & -\beta\mathbf{C}^h\mathbf{D} & \beta\mathbf{N} & \mathbf{0} & \mu\mathbf{N}_{\nu_5} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \tilde{\mathbf{N}} & \beta\mathbf{C}^h\mathbf{D} & -\beta\mathbf{N} & \mathbf{0} & \mathbf{0} & \mu\mathbf{N}_{\nu_6} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{\Sigma}^{\min}\tilde{\mathbf{N}} & \mathbf{0} & -\mathbf{N} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mu\mathbf{N}_{\nu_7} & \mathbf{0} \\ -\boldsymbol{\Sigma}^{\max}\tilde{\mathbf{N}} & \mathbf{0} & \mathbf{N} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mu\mathbf{N}_{\nu_8} \end{pmatrix}.$$

As a last reduction we eliminate the updates concerning the Lagrange multipliers $\boldsymbol{\lambda}_1, \ldots \boldsymbol{\lambda}_4$, from the linear system (6.14) using

$$\begin{aligned} \triangle \boldsymbol{\lambda}_1^h &= \frac{1}{\mu}\mathbf{N}_{\nu_5}^{-1}\mathbf{e}_{(Q^h)^3}^h - \frac{1}{\mu}\mathbf{N}_{\nu_5}^{-1}\tilde{\mathbf{N}}\triangle\boldsymbol{\rho}^h + \frac{\beta}{\mu}\mathbf{N}_{\nu_5}^{-1}\mathbf{C}^h\mathbf{D}\triangle\mathbf{u}^h - \frac{\beta}{\mu}\mathbf{N}_{\nu_5}^{-1}\mathbf{N}\triangle\mathbf{s}^h, \\ \triangle \boldsymbol{\lambda}_2^h &= \frac{1}{\mu}\mathbf{N}_{\nu_6}^{-1}\mathbf{e}_{(Q^h)^3}^h - \frac{1}{\mu}\mathbf{N}_{\nu_6}^{-1}\tilde{\mathbf{N}}\triangle\boldsymbol{\rho}^h - \frac{\beta}{\mu}\mathbf{N}_{\nu_6}^{-1}\mathbf{C}^h\mathbf{D}\triangle\mathbf{u}^h + \frac{\beta}{\mu}\mathbf{N}_{\nu_6}^{-1}\mathbf{N}\triangle\mathbf{s}^h, \\ \triangle \boldsymbol{\lambda}_3^h &= -\frac{1}{\mu}\boldsymbol{\Sigma}^{\min}\mathbf{N}_{\nu_7}^{-1}\tilde{\mathbf{N}}\triangle\boldsymbol{\rho}^h + \frac{1}{\mu}\mathbf{N}_{\nu_7}^{-1}\mathbf{N}\triangle\mathbf{s}^h, \\ \triangle \boldsymbol{\lambda}_4^h &= \frac{1}{\mu}\boldsymbol{\Sigma}^{\max}\mathbf{N}_{\nu_8}^{-1}\tilde{\mathbf{N}}\triangle\boldsymbol{\rho}^h - \frac{1}{\mu}\mathbf{N}_{\nu_8}^{-1}\mathbf{N}\triangle\mathbf{s}^h. \end{aligned} \tag{6.15}$$

This elimination finally results in the symmetric saddle point problem

$$\begin{pmatrix} \boldsymbol{\mathcal{K}}_{\rho\rho} & \boldsymbol{\mathcal{K}}_{\rho u} & \boldsymbol{\mathcal{K}}_{\rho s} & \mathbf{0} \\ \boldsymbol{\mathcal{K}}_{\rho u}^T & \boldsymbol{\mathcal{K}}_{uu} & \boldsymbol{\mathcal{K}}_{us} & \mathbf{0} \\ \boldsymbol{\mathcal{K}}_{\rho s}^T & \boldsymbol{\mathcal{K}}_{us}^T & \boldsymbol{\mathcal{K}}_{ss} & \mathbf{D}^T \\ \mathbf{0} & \mathbf{0} & \mathbf{D} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \triangle \boldsymbol{\rho}^h \\ \triangle \mathbf{u}^h \\ \triangle \mathbf{s}^h \\ \triangle \boldsymbol{\lambda}_0^h \end{pmatrix} = \mathbf{f}^h, \tag{6.16}$$

with

$$\mathbf{f}^h = \begin{pmatrix} -\big(\gamma + \frac{1}{\epsilon}\big)\mathbf{e}_{V^h}^h - \frac{1}{\mu}\mathbf{M}\mathbf{M}_{\nu_2}^{-1}\mathbf{e}_{V^h}^h - \frac{1}{\mu}\tilde{\mathbf{N}}^T\big(\mathbf{N}_{\nu_5}^{-1} + \mathbf{N}_{\nu_6}^{-1}\big)\mathbf{e}_{(Q^h)^3}^h \\ -\frac{1}{\mu}\mathbf{M}_2\big(\mathbf{M}_{\nu_3}^{-1}\mathbf{u}^{\min^h} + \mathbf{M}_{\nu_4}^{-1}\mathbf{u}^{\max h}\big) - \frac{\beta}{\mu}\mathbf{C}^h\mathbf{D}\big(\mathbf{N}_{\nu_5}^{-1} - \mathbf{N}_{\nu_6}^{-1}\big)\mathbf{e}_{(Q^h)^3}^h \\ -\frac{\beta}{\mu}\mathbf{N}_3\big(\mathbf{N}_{\nu_5}^{-1} - \mathbf{N}_{\nu_6}^{-1}\big)\mathbf{e}_{(Q^h)^3}^h \\ \mathbf{0} \end{pmatrix}$$

and the final block matrices

$$
\begin{aligned}
\boldsymbol{\mathcal{K}}_{\rho\rho} &= -\epsilon\mathbf{K} - \frac{2}{\epsilon}\mathbf{M} - \frac{1}{\mu}\mathbf{M}\big(\mathbf{M}_{\nu_1}^{-1} + \mathbf{M}_{\nu_2}^{-1}\big)\mathbf{M} - \\
&\quad - \frac{1}{\mu}\tilde{\mathbf{N}}^T\big(\mathbf{N}_{\nu_5}^{-1} + \mathbf{N}_{\nu_6}^{-1} + \boldsymbol{\Sigma}^{\min 2}\mathbf{N}_{\nu_7}^{-1} + \boldsymbol{\Sigma}^{\max 2}\mathbf{N}_{\nu_8}^{-1}\big)\tilde{\mathbf{N}}, \\
\boldsymbol{\mathcal{K}}_{uu} &= -\frac{1}{\mu}\mathbf{M}_2\big(\mathbf{M}_{\nu_3}^{-1} + \mathbf{M}_{\nu_4}^{-1}\big)\mathbf{M}_2 - \frac{\beta^2}{\mu}\mathbf{D}^T\mathbf{C}^{h^2}\big(\mathbf{N}_{\nu_5}^{-1} + \mathbf{N}_{\nu_6}^{-1}\big)\mathbf{D}, \\
\boldsymbol{\mathcal{K}}_{ss} &= -\frac{1}{\mu}\mathbf{N}\big(\beta^2\mathbf{N}_{\nu_5}^{-1} + \beta^2\mathbf{N}_{\nu_6}^{-1} + \mathbf{N}_{\nu_7}^{-1} + \mathbf{N}_{\nu_8}^{-1}\big)\mathbf{N}, \\
\boldsymbol{\mathcal{K}}_{\rho u} &= \frac{\beta}{\mu}\tilde{\mathbf{N}}^T\big(\mathbf{N}_{\nu_5}^{-1} - \mathbf{N}_{\nu_6}^{-1}\big)\mathbf{C}^h\mathbf{D}^T, \\
\boldsymbol{\mathcal{K}}_{\rho s} &= \frac{1}{\mu}\tilde{\mathbf{N}}^T\big(\beta\mathbf{N}_{\nu_6}^{-1} - \beta\mathbf{N}_{\nu_5}^{-1} + \boldsymbol{\Sigma}^{\min}\mathbf{N}_{\nu_7}^{-1} + \boldsymbol{\Sigma}^{\max}\mathbf{N}_{\nu_8}^{-1}\big)\mathbf{N}, \\
\boldsymbol{\mathcal{K}}_{us} &= \frac{\beta^2}{\mu}\mathbf{C}^h\mathbf{D}^T\big(\mathbf{N}_{\nu_5}^{-1} + \mathbf{N}_{\nu_6}^{-1}\big)\mathbf{N}.
\end{aligned}
\tag{6.17}
$$

The linear system (6.16) yields a solution in the variables $\triangle\boldsymbol{\rho}^h, \triangle\mathbf{u}^h, \triangle\mathbf{s}^h, \triangle\boldsymbol{\lambda}_0^h$. Using this solution, the other variables are determined by the substitutions (6.15), (6.13), and finally (6.15).

6.2 A Multigrid KKT Solver

In this section we consider (additive and multiplicative) Schwarz-type iteration methods as smoothers in a multigrid method for saddle point problems. Each iteration step of such a Schwarz-type smoother consists of the solution of several small local saddle point problems in a Jacobi- or Gauss-Seidel-type manner. The computational domain is therefore divided into overlapping cells, also called *patches*. One iteration step of a Schwarz-type smoother consists now of solving a local saddle point problem for each patch. This is done in a Jacobi- or Gauß-Seidel-type manner and thus, called additive or multiplicative Schwarz-type smoother.

To begin with, we state the two most basic iterative methods for a linear system

$$\mathbf{K}\mathbf{u} = \mathbf{f},$$

which are used as smoothing methods, namely the Jacobi- and the Gauss-Seidel iterations, being the origins of the additive and multiplicative Schwarz methods, respectively. In Algorithm 6.1 we present the *Jacobi* iteration. The algorithm is simple, but with the disadvantage of slow convergence (note the analogy to the Richardson iteration in Algorithm 2.1). We state the Jacobi iteration without any consideration about convergence, but refer e.g. to Jung and Langer [80]. One criterion for the convergence of the damped Jacobi iteration is the symmetry and positive definiteness of the system matrix $\mathbf{K}$. A similar method to the Jacobi iteration is the *Gauss-Seidel* method. In difference to the Jacobi iteration we use for the computation of the i-th component u_i^k the already updated components u_j^k, for $j = 1, \ldots, i-1$, in iteration k. The Gauß-Seidel method is presented in Algorithm 6.2. Again, the Gauss-Seidel iteration converges if the system matrix $\mathbf{K}$ is symmetric and positive definite. For more information we refer again e.g. to Jung and Langer [80] or to Hackbusch [71]. Both iteration methods (Jacobi iterations after under-relaxation, damped Jacobi) have smoothing properties in the

Algorithm 6.1 Damped Jacobi iteration

Choose a damping parameter τ, $0 < \tau < \frac{2}{\lambda_{\max}(\mathrm{diag}(\mathbf{K})^{-1}\mathbf{K})}$.
Choose a relative error bound $\varepsilon > 0$.
Initialize start value $\mathbf{u}^0 \in \mathbb{R}^n$.
$k = 0$;

while not converged **do**
  for $i = 1, \ldots, n$ **do**

$$u_i^{k+1} = (1-\tau)u_i^k + \frac{\tau}{K_{ii}}\left(f_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} K_{ij}u_j^k\right);$$

  end for
  $k = k + 1$;
end while

Algorithm 6.2 Gauß-Seidel iteration

Choose a relative error bound $\varepsilon > 0$.
Initialize start value $\mathbf{u}^0 \in \mathbb{R}^n$.
$k = 0$;

while not converged **do**

$$u_1^{k+1} = \frac{1}{K_{11}}\left(f_1 - \sum_{j=2}^{n} K_{1j}u_j^k\right);$$

  for $i = 2, \ldots, n-1$ **do**

$$u_i^{k+1} = \frac{1}{K_{ii}}\left(f_i - \sum_{j=1}^{i-1} K_{ij}u_j^{k+1} - \sum_{j=i+1}^{n} K_{ij}u_j^k\right);$$

  end for

$$u_n^{k+1} = \frac{1}{K_{nn}}\left(f_n - \sum_{j=1}^{n-1} K_{nj}u_j^k\right);$$

  $k = k + 1$;
end while

sense that they reduce the high frequency part of the error components and are cheap to apply.

The above definitions of the iteration methods would lead to *pointwise* methods. Below we shall define smoothing operators in terms of subspace decompositions, which will lead to a *blockwise* iteration method. These procedures are related to overlapping domain decomposition algorithms and to the classical Schwarz method. They are generalizations of Jacobi and Gauß-Seidel iteration procedures.

We start to introduce the Schwarz-type smoothers in an abstract framework of mixed variational problems (cf. Subsection 2.2.2). For this sake let V and Q be Hilbert spaces and let $a(\cdot,\cdot) : V \times V \to \mathbb{R}$, $b(\cdot,\cdot) : V \times Q \to \mathbb{R}$, and $c(\cdot,\cdot) : Q \times Q \to \mathbb{R}$ be continuous bilinear forms. Furthermore, let $f(\cdot) : V \to \mathbb{R}$ and $g(\cdot) : Q \to \mathbb{R}$ be continuous linear forms. Then we

can formulate the following mixed variational problem: Find $u \in V$ and $p \in Q$ such that

$$\begin{aligned} a(u,v) \;+\; b(v,p) \;&=\; f(v), \qquad \forall\, v \in V, \\ b(u,q) \;-\; c(p,q) \;&=\; g(q), \qquad \forall\, q \in Q. \end{aligned} \tag{6.18}$$

Following on the framework of multigrid methods we would introduce now a hierarchy of finite element spaces $V_0 \subset \ldots \subset V_l \subset V$, $Q_0 \subset \ldots \subset Q_l \subset Q$ on a corresponding hierarchy of increasingly finer meshes and so on. But since the smoothing procedure involves only one level of the sequence of spaces, we will omit these notations and fix one level i, $0 < i < l$. For simplification of notation we will also drop the subindex k when denoting spaces, matrices and so on. Following a standard finite element discretization let the vectors $\mathbf{v} \in \mathbb{R}^n$ and $\mathbf{q} \in \mathbb{R}^n$ contain the coefficients of the corresponding finite element functions with respect to some bases of V and Q. Moreover, we introduce the matrix representation of the mixed variational problem (6.18):

$$\begin{pmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & -\mathbf{C} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix}. \tag{6.19}$$

In the sequel we will again abbreviate the system matrix with $\boldsymbol{\mathcal{K}}$, i.e.

$$\boldsymbol{\mathcal{K}} = \begin{pmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & -\mathbf{C} \end{pmatrix}.$$

As a consequence from the properties from the bilinear forms, we assume that $\mathbf{A}$ is a symmetric positive semi-definite $n \times n$ matrix, $\mathbf{C}$ is a symmetric positive semi-definite $m \times m$ matrix, that $\mathbf{B}$ is a $m \times n$ matrix, and that $\boldsymbol{\mathcal{K}}$ is regular.

We shall start with a decomposition of the spaces

$$V = \sum_{i=1}^{l} V^i \qquad \text{and} \qquad Q = \sum_{i=1}^{l} Q^i.$$

Before we define the additive and multiplicative smoother we have to introduce linear operators for each subspace to set up the local sub-problems:

$$\mathbf{P}_{V\,i} : \mathbb{R}^{n_i} \to \mathbb{R}^n \qquad \text{and} \qquad \mathbf{P}_{Q_i} : \mathbb{R}^{m_i} \to \mathbb{R}^m, \qquad \text{for } i = 1, \ldots, l, \tag{6.20}$$

with n_i, m_i denoting the dimensions of the local subspaces V_i and Q_i, respectively. The matrices $\mathbf{P}_{V\,i}$ and $\mathbf{P}_{Q_i}$ denote prolongation operators with the associated restriction operators $\mathbf{P}_{V\,i}^T$ and $\mathbf{P}_{Q_i}^T$, respectively. Furthermore, let the operators (6.20) satisfy the conditions

$$\sum_{i=1}^{l} \mathbf{P}_{V\,i} \mathbf{P}_{V\,i}^T = \mathbf{I} \qquad \text{and} \qquad \sum_{i=1}^{l} \mathbf{P}_{Q_i} \mathbf{P}_{Q_i}^T \text{ is regular.} \tag{6.21}$$

With these preliminaries we can now define two Schwarz-type smoothers assuming that $\mathbf{u}^k$ and $\mathbf{p}^k$ are some approximations for the exact solutions $\mathbf{u}$ and $\mathbf{p}$ of (6.19).

The first one will be called an *additive Schwarz smoother* and is defined by

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \sum_{i=1}^{l} \mathbf{P}_{V\,i} \mathbf{v}_i, \qquad \mathbf{p}^{k+1} = \mathbf{p}^k + \sum_{i=1}^{l} \mathbf{P}_{Q_i} \mathbf{q}_i,$$

with $\mathbf{v}_i$ and $\mathbf{q}_i$, $i = 1, \dots, l$, solving the local saddle point problem

$$\begin{pmatrix} \hat{\mathbf{A}}_i & \mathbf{B}_i^T \\ \mathbf{B}_i & \mathbf{B}_i \hat{\mathbf{A}}_i^{-1} \mathbf{B}_i^T - \hat{\mathbf{S}}_i \end{pmatrix} \begin{pmatrix} \mathbf{v}_i \\ \mathbf{q}_i \end{pmatrix} = \begin{pmatrix} \mathbf{P}_{V_i}^T (\mathbf{f} - \mathbf{A}\mathbf{u}^k - \mathbf{B}^T \mathbf{p}^k) \\ \mathbf{P}_{Q_i}^T (\mathbf{g} - \mathbf{B}\mathbf{u}^k + \mathbf{C}\mathbf{p}^k) \end{pmatrix},$$

where $\hat{\mathbf{S}}_i = \frac{1}{\tau}(\mathbf{C}_i + \mathbf{B}_i \hat{\mathbf{A}}_i^{-1} \mathbf{B}_i^T)$, with some damping parameter $\tau > 0$. Thus, the actual residuum is restricted to the smaller spaces. Then the local saddle point problems are solved for all patches, and the solutions are finally prolongated back onto the whole space. This Jacobi-type process can be seen as an additive Schwarz method and the corresponding smoothing operator $\boldsymbol{\mathcal{S}}_A$ can be written as

$$\boldsymbol{\mathcal{S}}_A \begin{pmatrix} \mathbf{u}^k \\ \mathbf{p}^k \end{pmatrix} = \begin{pmatrix} \mathbf{u}^k \\ \mathbf{p}^k \end{pmatrix} + \sum_{i=1}^{l} \mathbf{P}_i \hat{\boldsymbol{\mathcal{K}}}_i^{-1} \mathbf{P}_i^T \left(\begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix} - \boldsymbol{\mathcal{K}} \begin{pmatrix} \mathbf{u}^k \\ \mathbf{p}^k \end{pmatrix} \right),$$

where we used the abbreviations

$$\hat{\boldsymbol{\mathcal{K}}}_i = \begin{pmatrix} \hat{\mathbf{A}}_i & \mathbf{B}_i^T \\ \mathbf{B}_i & \mathbf{B}_i \hat{\mathbf{A}}_i^{-1} \mathbf{B}_i^T - \hat{\mathbf{S}}_i \end{pmatrix} \qquad \text{and} \qquad \mathbf{P}_i = \begin{pmatrix} \mathbf{P}_{Vi} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{Q_i} \end{pmatrix}.$$

Moreover, we define the *multiplicative Schwarz smoother* based on the above subspace decomposition as the following procedure: Set $\mathbf{w}^0 = \mathbf{0}$ and $\mathbf{r}^0 = \mathbf{0}$ and compute

$$\begin{pmatrix} \mathbf{w}^i \\ \mathbf{r}^i \end{pmatrix} = \begin{pmatrix} \mathbf{w}^{i-1} \\ \mathbf{r}^{i-1} \end{pmatrix} + \mathbf{P}_i \hat{\boldsymbol{\mathcal{K}}}_i^{-1} \mathbf{P}_i^T \left(\begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix} - \boldsymbol{\mathcal{K}} \begin{pmatrix} \mathbf{w}^{i-1} \\ \mathbf{r}^{i-1} \end{pmatrix} \right), \qquad \text{for } i = 1, \dots, l, \quad (6.22)$$

where we set $\tau = 1$, i.e. the local saddle point problems resemble the global saddle point problem in shape. Finally we define the multiplicative smoother as

$$\boldsymbol{\mathcal{S}}_M \begin{pmatrix} \mathbf{u}^k \\ \mathbf{p}^k \end{pmatrix} = \begin{pmatrix} \mathbf{u}^k \\ \mathbf{p}^k \end{pmatrix} + \begin{pmatrix} \mathbf{w}^l \\ \mathbf{r}^l \end{pmatrix}. \quad (6.23)$$

So far we did not pose any conditions on the local matrices $\hat{\mathbf{A}}_i$, $\mathbf{B}_i$, and $\mathbf{C}_i$. For the additive case we can state the following theorem, under which assumptions it is possible to interpret the additive Schwarz iteration as a symmetric inexact Uzawa method. Then, the smoothing property, an important part of a convergence proof for multigrid methods, can be shown (cf. SCHÖBERL AND ZULEHNER [116]).

Theorem 6.1. *Assume that* (6.21) *is satisfied, the matrices* $\hat{\mathbf{A}}_i$ *and* $\hat{\mathbf{S}}_i$ *are symmetric and positive definite, and there is a symmetric positive definite* $n \times n$ *matrix* $\hat{\mathbf{A}}$ *such that*

$$\mathbf{P}_{V_i} \hat{\mathbf{A}} = \hat{\mathbf{A}}_i \mathbf{P}_{V_i}^T, \qquad \textit{for } i = 1, \dots, l.$$

Furthermore, assume that the matrices $\mathbf{B}_i$ *satisfy the condition*

$$\mathbf{P}_{Q_i}^T \mathbf{B} = \mathbf{B}_i \mathbf{P}_{Q_i}^T, \qquad \textit{for } i = 1, \dots, l.$$

Then we have

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \mathbf{v}^k \qquad \textit{and} \qquad \mathbf{p}^{k+1} = \mathbf{p}^k + \mathbf{q}^k, \quad (6.24)$$

where $\mathbf{v}^k$ *and* $\mathbf{q}^k$ *satisfy the equation*

$$\begin{pmatrix} \hat{\mathbf{A}} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{B}\hat{\mathbf{A}}^{-1}\mathbf{B}^T - \hat{\mathbf{S}} \end{pmatrix} \begin{pmatrix} \mathbf{v}^k \\ \mathbf{q}^k \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix} - \boldsymbol{\mathcal{K}} \begin{pmatrix} \mathbf{u}^k \\ \mathbf{p}^k \end{pmatrix} \quad (6.25)$$

and

$$\hat{\mathbf{S}} = \left(\sum_{i=1}^{l} \mathbf{P}_{Q_i} \hat{\mathbf{S}}_i^{-1} \mathbf{P}_{Q_i}^T \right)^{-1}.$$

Proof. See SCHÖBERL AND ZULEHNER [116]. □

Up to the knowledge of the author, there is no theory available for the multiplicative Schwarz-type smoother. But in practice, the multiplicative version turns out to much more efficient than the additive iteration scheme. So, we realized our numerical test examples with the multiplicative Schwarz-type smoother, as presented in the next section. The verification of the assumptions of Theorem 6.1 for our particular case, as well as numerical experiments for the additive version, are still missing and will be part of future research.

Notes and remarks for Section 6.2

- Note, that the Gauß-Seidel iteration depends on the ordering of the unknowns and that the Jacobi iteration is independent of the ordering of the unknowns, see e.g. HACKBUSCH [73]. In order to get a symmetric multigrid operator (2.25) the post smoothing has to be arranged in a backward fashion for the Gauß-Seidel-type iteration. Moreover, the same number of pre- and post-smoothing steps has to be used.

- Usually numerical experiments show that the multiplicative Schwarz smoother leads to significantly better convergence rates as the additive variant. We refer to SCHÖBERL AND ZULEHNER [116] for a theoretical analysis for the convergence and smoothing properties of the additive smoother. A theoretical analysis for the multiplicative version, however, is still missing.

6.3 Numerical experiments

For the numerical results we choose $\Omega = (0,1) \times (0,1)$ and decompose it into a regular triangulation $\mathcal{T}_h^k = \{\tau_i \mid i = 1, \ldots, n_k\}$ for each level k of a hierarchy of l nested meshes with $3 \leq k \leq l$. That means that level $k = 3$ is the coarsest grid where the corresponding linear system is solved exactly. For each level k we assemble the block matrices that finally build up the saddle point system (6.16). For convenience we state the system matrix again:

$$\boldsymbol{\mathcal{K}}_k = \begin{pmatrix} \boldsymbol{\mathcal{K}}_{\rho\rho} & \boldsymbol{\mathcal{K}}_{\rho u} & \boldsymbol{\mathcal{K}}_{\rho s} & \mathbf{0} \\ \boldsymbol{\mathcal{K}}_{\rho u}^T & \boldsymbol{\mathcal{K}}_{uu} & \boldsymbol{\mathcal{K}}_{us} & \mathbf{0} \\ \boldsymbol{\mathcal{K}}_{\rho s}^T & \boldsymbol{\mathcal{K}}_{us}^T & \boldsymbol{\mathcal{K}}_{ss} & \mathbf{D}^T \\ \mathbf{0} & \mathbf{0} & \mathbf{D} & \mathbf{0} \end{pmatrix}, \tag{6.26}$$

with the block matrices (6.17). In order to test the multiplicative patch smoother (6.22) - (6.23) we solved the saddle point system (6.16) on a hierarchy with an increasing number of meshes. We set $\mathbf{f}_k = \mathbf{0}$ and used randomly chosen starting values for $\triangle \mathbf{x}_k^0$ for the exact solutions $\triangle \mathbf{x}_k$. For constructing the local subproblems we decomposed the grid $\mathcal{T}_h^k$ into m_k overlapping patches, where m_k denotes the number of nodes on level k. Each patch consists of the at most 6 surrounding triangles for each node. As mentioned in the previous section, we approximated the density ρ, the displacements $\mathbf{u}$, and the Lagrangian multiplier $\boldsymbol{\lambda}_0$ with linear elements and the stresses $\mathbf{s}$ with constant elements. The corresponding subspaces V_i, for $i =$

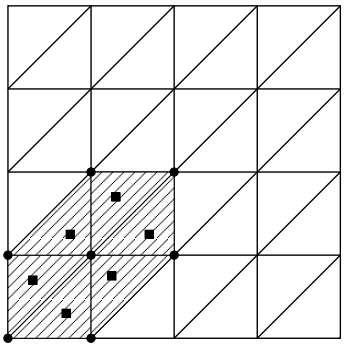

Figure 6.1: Patch of a local saddle point problem.

$1, \ldots, m_k$, consist now of the degrees of freedom of the node i, related to the approximations of the density and the displacement components, and the degrees of freedom in the surrounding elements, related to the stress components. The subspaces Q_i, for $i = 1, \ldots, m_k$, consist of the unknowns at node i with respect to the approximation of the Lagrangian multiplier $\boldsymbol{\lambda}_0$. Figure 6.1 shows an example of a patch, where the places marked with a '■' indicate the unknowns of the constant elements and the places marked with a '●' indicate the unknowns of the linear elements. For the actual numerical tests we chose the local block matrix $\hat{\mathbf{A}}_i = \mathbf{A}_i$ and

Level	Unknowns	Smoothing steps			
		2		4	
		Iterations	Conv. Factor	Iterations	Conv. Factor
4	725	25	0.478	14	0.255
5	2853	27	0.510	15	0.269
6	11333	26	0.489	14	0.255
7	45189	25	0.479	13	0.230
8	180485	23	0.445	12	0.209

Table 6.1: Convergence rates for a W-cycle and an error reduction by a factor of 10^{-8} ($\epsilon = 0.1$, $\mu = 0.1$, $\boldsymbol{\nu}_i^h = \mathbf{1}$ for $i = 1, \ldots, 8$).

used a W-cycle with 2 smoothing steps (one pre- and one post-smoothing step. We stopped the iteration process when the initial defect was reduced by a factor of 10^{-8}, measured by the Euclidean norm. In Table 6.1 we list the convergence data for the following choice of parameters: $\epsilon = 0.1$, $\mu = 0.1$, and $\boldsymbol{\nu}_i^h = \mathbf{1}$ for $i = 1, \ldots, 8$. The table shows the typical multigrid convergence behavior, i.e., convergence rates that are asymptotic independent of the grid level and an asymptotic constant number of iterations. For the next test example we set μ and ϵ so smaller values, as these parameters are supposed to tend to zero in actual computations. We chose $\mu = 10^{-6}$ and $\epsilon = 10^{-4}$, smaller values as have actually been used in the computations in Subsection 5.6.3. All in all, Table 6.2 shows again the expected behavior.

Also arbitrary values of the dual variables $\boldsymbol{\nu}_1^h, \ldots, \boldsymbol{\nu}_4^h$, in $[10^{-6}, 10^1]$, possibly after suitable scaling, do not change this behaviour. However, the dual variables $\boldsymbol{\nu}_5^h, \ldots, \boldsymbol{\nu}_8^h$, that act as Lagrange multipliers for the slack variables $\mathbf{z}_1^h, \ldots, \mathbf{z}_4^h$ (see (6.7) in Subsection 6.1) may cause troubles. In the case that $\mathbf{v}_5^h \leq \mathbf{v}_6^h$ and $\mathbf{v}_7^h \leq \mathbf{v}_8^h$ the condition number of the system

Level	Unknowns	Smoothing steps			
		2		4	
		Iterations	Conv. Factor	Iterations	Conv. Factor
4	725	39	0.621	19	0.376
5	2853	25	0.478	14	0.258
6	11333	24	0.460	13	0.226
7	45189	22	0.427	12	0.210
8	180485	22	0.425	12	0.211

Table 6.2: Convergence rates for a W-cycle and an error reduction by a factor of 10^{-8} ($\epsilon = 10^{-4}$, $\mu = 10^{-6}$, $\boldsymbol{\nu}_i^h = \mathbf{1}$ for $i = 1, \ldots, 8$).

matrix is raising, but the multigrid iteration still achieves convergence with more then 4 smoothing steps. See Table 6.3 for the convergence data for extreme values $\boldsymbol{\nu}_5^h = \boldsymbol{\nu}_7^h = \mathbf{10}$ and $\boldsymbol{\nu}_6^h = \boldsymbol{\nu}_8^h = \mathbf{10^{-6}}$ with 5 and 9 pre- and post-smoothing steps, respectively. Unfortunately,

Level	Unknowns	Smoothing steps			
		10		18	
		Iterations	Conv. Factor	Iterations	Conv. Factor
4	725	52	0.701	26	0.490
5	2853	45	0.664	23	0.446
6	11333	34	0.582	18	0.358
7	45189	29	0.529	17	0.321
8	180485	27	0.500	16	0.298

Table 6.3: Convergence rates for a W-cycle and an error reduction by a factor of 10^{-8}. ($\epsilon = 10^{-4}$, $\mu = 10^{-6}$, $\boldsymbol{\nu}_5^h = \boldsymbol{\nu}_7^h = \mathbf{10}$, and $\boldsymbol{\nu}_6^h = \boldsymbol{\nu}_8^h = \mathbf{10^{-6}}$).

for some choices $\mathbf{v}_5^h > \mathbf{v}_6^h$ and $\mathbf{v}_7^h > \mathbf{v}_8^h$ the upper left block of the system matrix (6.26) becomes almost indefinite, e.g., $\lambda_{\min} \in [-9 \cdot 10^{-6}, 9 \cdot 10^{-6}]$. Similar behaviour is reported, e.g. in Maar and Schulz [89] and Wächter and Biegler [147]. If the upper left block looses its property to be positive definite, the smoother fails and the multigrid iteration diverges. For instance for the choices $\mathbf{v}_5^h = \mathbf{v}_7^h = 10^2$ and $\mathbf{v}_6^h = \mathbf{v}_8^h = 10^{-6}$ we get $\lambda_{\max} = 336$ and $\lambda_{\min} = -8 \cdot 10^6$.

A known remedy for the above situation is to add a small multiple of the identity matrix to the upper left block, which is called *inertia correction* in literature and is, e.g. used in the software packages Ipopt and LOQO. In our test examples a addition of $\delta\mathbf{I}$, with $\delta = 10^{-3}$, to the upper left part removed the mentioned difficulties.

Chapter 7

Conclusions and Outlook

In this thesis we presented two new approaches, one to the minimal compliance problem with respect to limited mass and one to the minimal mass problem with respect to limited local stresses. Moreover, we gave a brief discussion of the two conceptual approaches how to formulate an optimal design problem and developed an optimal solver for a KKT-system.

The discussion in Chapter 3 highlight the pros and cons of the nested and simultaneous approach. Here, the construction of optimal solvers for the KKT-systems for various design problems and a comparison to the nested approach in terms of computational effort is an interesting and hot topic.

In Chapter 4 we presented an adaptive multilevel approach to the minimal compliance problem, which turned out to work successful for several examples in 2D and 3D. Furthermore, due to a multigrid method, we solve the systems of linear equations, resulting from the discretized state equations, with optimal complexity. The bottleneck of the high computational time for the applying the discrete filter operator on very fine unstructured grids can be overcome by using, e.g. $\mathcal{H}$-matrices or other data-sparse approximating techniques. This could serve as a starting point for future research.

A new method for solving topology optimization problems with local stress constraints is presented in Chapter 5. The reformulation of the set of constraints and the phase-field relaxation lead to a parameter-dependent family of large-scale optimization problems, satisfying uniform constraint qualifications. So far no particular emphasis has been laid on the efficient solution of the discretized problems. Tailoring a suitable optimization routine, e.g. a QP solver or a tuned interior-point method for nonlinear constraints, will reduce the computational times tremendously in comparison to the use of a black-box solver. Moreover, the possibility of a better tuning of the optimization method will raise the possibility to solve more advanced examples. This will be treated in future, due to the high importance of local stress constraints in the field of structural optimization and due to high interest of the author.

The optimal solver of the KKT-system presented in Chapter 6 shows the potential of Schwarz-type smoothers in the multigrid framework. The linear complexity solver should be embedded in a dual-primal interior-point optimization method to show its true potential. This is a task that can be accomplished within the construction of a suitable optimization routine for the discretized optimization problems resulting from the approach in Chapter 5.

Bibliography

[1] W. Achtziger. Local stability of trusses in the context of topology optimization, part I: Exact modelling. *Structural and Multidisciplinary Optimization*, 17(4):235–246, 1999.

[2] W. Achtziger. Local stability of trusses in the context of topology optimization, part II: A numerical approach. *Structural and Multidisciplinary Optimization*, 17(4):247–258, 1999.

[3] R.A. Adams. *Sobolev Spaces*. Academic Press, New York, 1976.

[4] G. Alberti. Variational models for phase transitions. An approach via γ-convergence. In L. Ambrosio and N. Dancer, editors, *Calculus of variations and partial differential equations. Topics on geometrical evolution problems and degree theory*, pages 95–114. Springer, Berlin-Heidelberg, 2000.

[5] G. Allaire. *Shape Optimization by the Homogenization Method*. Springer, New York - Berlin - Heidelberg, 2002.

[6] G. Allaire, E. Bonnetier, G. Francfort, and F. Jouve. Shape optimization by the homogenization method. *Numerische Mathematik*, 76:27–68, 1997.

[7] G. Allaire, F. De Gournay, F. Jouve, and A.-M. Toader. Structural optimization using topological and shape sensitivity via a level set method. *Control and Cybernetics*, 34:59–80, 2005.

[8] G. Allaire, F. Jouve, and A.-M. Toader. Structural optimization using sensitivity analysis and a level set method. *Journal of Computational Physics*, 194:363–393, 2004.

[9] G. Allaire and R.V. Kohn. Optimal design for minimum weight and compliance in plane stress using extremal microstructures. *European Journal of Mechanics. A. Solids*, 12(6):839–878, 1993.

[10] L. Ambrosio and G. Buttazzo. An optimal design problem with perimeter penalization. *Calculus of Variations and Partial Differential Equations*, 1(1):55–69, 1993.

[11] E. Arian, A. Battermann, and E.W. Sachs. Approximation of the newton step by a defect correction process. Technical Report 99-12, Department of Mathematics/Informatics, University of Trier, 1999.

[12] O. Axelsson. *Iterative solution methods*. Cambridge University Press, second edition, 1996.

[13] G. Barles, H.M. Soner, and P.E. Souganidis. Front propagation and phase–field theory. *SIAM Journal of Control and Optimization*, 31(2):439–469, 1993.

[14] A. Battermann and M. Heinkenschloss. Preconditioners for Karush-Kuhn-Tucker matrices in the optimal control of distributed systems. In W. Desch, F. Kappel, and K. Kunisch, editors, *Control and estimation of distributed parameter systems. International conference in Vorau, Austria, July 14–20, 1996*, volume 126 of *ISNM, International Series Numerical Mathematics*, pages 15–32, Basel, 1998. Birkhäuser.

[15] A. Battermann and E.W. Sachs. Block preconditioners for KKT systems in PDE-governed optimal control problems. In K.-H. Hoffmann et al., editor, *Fast solution of discretized optimization problems. Workshop held at the Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany, May 8-12, 2000*, volume 138 of *ISNM, International Series Numerical Mathematics*, pages 1–18, Basel, 2001. Birkhäuser.

[16] M.P. Bendsøe. Optimal shape design as a material distribution problem. *Structural Optimization*, 1:193–202, 1989.

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

[18] M.P. Bendsøe, A.R. Diaz, R. Lipton, and J.E. Taylor. Optimal design of material properties and material distribution for multiple load cases. *International Journal for Numerical Methods in Engineering*, 38:1149–1170, 1995.

[19] M.P. Bendsøe, J.M. Guedes, R.B. Haber, P. Pederson, and J.E. Taylor. An analytical model to predict optimal material properties in the context of optimal structural design. *Transactions of the ASME, Journal of Applied Mechanics*, 61(4):930–937, 1994.

[20] M.P. Bendsøe and N. Kikuchi. Generating optimal topologies in structural design using a homogenization method. *Computer Methods in Applied Mechanics and Engineering*, 71(2):197–224, 1988.

[21] M.P. Bendsøe and O. Sigmund. Material interpolation schemes in topology optimization. *Archive of Applied Mechanics*, 69:635–654, 1999.

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

[23] M. Benzi, G.H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14:1–137, 2005.

[24] G. Biros and O. Ghattas. Inexactness issues in the Lagrange-Krylov-Schur method for PDE-constrained optimization. In L.T. Biegler et al., editor, *Large-scale PDE-constrained optimization*, volume 30 of *Lecture Notes Comput. Sci. Eng.*, pages 91–114. Springer, Berlin, 2003.

[25] J.F. Blowey and C.M. Elliot. A phase–field model with a double obstacle potential. In G. Buttazzo and A. Visintin, editors, *Motion by Mean Curvature and Related Topics*, pages 1–22. de Gruyter, Berlin, 1994.

[26] P.T. Boggs and J.W. Tole. Sequential quadratic programming. *Acta Numerica*, 4:1–51, 1995.

[27] T. Borrvall. Topology optimization of elastic continua using restriction. *Archives of Computational Methods in Engineering*, 8(4):351–385, 2001.

[28] T. Borrvall and J. Petersson. Topology optimization using regularized intermediate density control. *Computer Methods in Applied Mechanics and Engineering*, 190:4911–4928, 2001.

[29] T. Borrvall and J. Petersson. Topology optimization of fluids in stokes flow. *International Journal for Numerical Methods in Fluids*, 41:77–107, 2003.

[30] B. Bourdin. Filters in topology optimization. *International Journal for Numerical Methods in Engineering*, 50(9):2143–2158, 2001.

[31] B. Bourdin and A. Chambolle. Design-dependent loads in topology optimization. *ESIAM: Control, Optimisation and Calculus of Variations*, 9:19–48, 2003.

[32] D. Braess. *Finite Elements: Theory, Fast Solvers and Applications in Solid Mechanics*. Cambridge University Press, second edition, 2001.

[33] D. Braess and R. Sarazin. An efficient smoother for the stokes problem. *Applied Numerical Mathematics*, 23(1):3–19, 1997.

[34] A. Braides. *"Gamma"-Convergence for Beginners*, volume 22 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2002.

[35] J.H. Bramble. *Multigrid methods*, volume 294 of *Pitman Research Notes in Mathematics Series*. Longman Scientific & Technical, Harlow, 1993.

[36] J.H. Bramble and J.E. Pasciak. A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems. *Mathematics of Computation*, 50:1–17, 1988.

[37] J.H. Bramble, J.E. Pasciak, and A.T. Vassilev. Analysis of the inexact Uzawa algorithm for saddle point problems. *SIAM Journal on Numerical Analysis*, 34:1072–1092, 1997.

[38] S.C. Brenner and L.R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer, New York, 2002.

[39] F. Brezzi. On the existance, uniquness and approximation of saddle point problems arising from lagrangian multipliers. *RAIRO Modél Math. Anal. Numér*, 2:129–151, 1974.

[40] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. Springer, New York, 1991.

[41] M. Bruyneel, P. Duysinx, and C. Fleury. A family of mma approximations for structural optimization. *Structural and Multidisciplinary Optimization*, 24:263–276, 2002.

[42] T. Buhl, C.B.W. Pedersen, and O. Sigmund. Stiffness design of geometrically non-linear structures using topology optimization. *Structural and Multidisciplinary Optimization*, 19(2):93–104, 2000.

[43] M. Burger. A framework for the construction of level set methods for shape optimization and reconstruction. *Interfaces and Free Boundaries*, 5:301–329, 2003.

[44] M. Burger, B. Hackl, and W. Ring. Incorporating topological derivatives into level set methods. *Journal of Computational Physics*, 194(1):344–362, 2004.

[45] M. Burger and W. Mühlhuber. Iterative regularization of parameter identification problems by sequential quadratic programming methods. *Inverse Problems*, 18(4):943–969, 2002.

[46] M. Burger and W. Mühlhuber. Numerical approximation of an SQP-type method for parameter identification. *SIAM Journal on Numercial Analysis*, 40(5):1775–1797, 2002.

[47] M. Burger and S. Osher. A survey on level set methods for inverse problems and optimal design. *European Journal of Applied Mathematics*, 16(2):263–301, 2005.

[48] M. Burger and R. Stainko. Phase–field relaxation of topology optimization with local stress constraints. To appear in *SIAM Journal on Control and Optimization*, 2006.

[49] R.H. Byrd, M.E. Hribar, and J. Nocedal. An interior point algorithm for large–scale nonlinear programming. *SIAM Journal of Optimization*, 9(4):877–900, 1999.

[50] G. Caginalp and E. Socolovsky. Phase field computations of single-needle crystals, crystal growth, and motion by mean curvature. *SIAM Journal on Scientific Computing*, 15(1):106–126, 1994.

[51] J.W. Cahn and J.E. Hilliard. Free energy of a nonuniform system I - Interfacial free energy. *Journal of Chemical Physics*, 28:258–267, 1958.

[52] G. Chen and J. Zhou. *Boundary Element Methods*. Computational Mathematics and Applications. Academic Press, Harcourt Brace Jovanovich, London - San Diego - New York, 1992.

[53] G.D. Cheng and X. Gou. ϵ-relaxed approach in topology optimization. *Structural and Multidisciplinary Optimization*, 13:258–267, 1997.

[54] G.D. Cheng and Z. Jiang. Study on topology optimization with stess constraints. *Engineering Optimization*, 20:129–148, 1992.

[55] G.D. Cheng and N. Olhoff. An investigation concerning optimal design of solid elastic plates. *International Journal of Solids and Structures*, 17:305–323, 1981.

[56] P.G. Ciarlet. *Mathematical Elasticity. Volume I: Three-dimensional elasticity*, volume 20 of *Studies in Mathematics and its Applications*. North-Holland, 1988.

[57] P.G. Ciarlet. *The Finite Element Method for Elliptic Problems*, volume 40 of *Classics in Applied Mathematics*. SIAM Society for Industrial and Applied Mathematics, 2002. Reprint, unabridged republication of the original 1987.

[58] G. Duvant and J.L. Lions. *Inequalities in mechanics an physics*, volume 219 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1976.

[59] P. Duysinx and M.P. Bendsøe. Topology optimization of continuum structures with local stress constraints. *International Journal for Numerical Methods in Engineering*, 43:1453–1478, 1998.

[60] P. Duysinx and O. Sigmund. New developements in handling stress constraints in optimal material distributions. In *7th Symposium on Multidisciplinary Analysis and Optimization*, pages 1501–1509. AIAA/USAF/NASA/ISSMO, AIAA-98-4906, 1998.

[61] H.A. Eschenauer, V.V. Kobolev, and A. Schuhmacher. Bubble method for topology and shape optimization of structures. *Structural Optimization*, 8:42–51, 1994.

[62] H.A. Eschenauer and N. Olhoff. Topology optimization of continuum structures: A review. *Applied Mechanics Rev*, 54(4), 2001.

[63] P. Fernandes, J.M. Guedes, and H. Rodrigues. Topology optimization of three-dimensional linear elastic structures with a constraint on 'perimeter'. *Computers and Structures*, 73(6), 1999.

[64] R. Fletcher. *Practical Methods of Optimization*, volume 2: Constrained Optimization. John Wiley & Sons, New York, 1981.

[65] C. Fleury and V. Braibant. Structural optimization: A new dual method using mixed variables. *International Journal for Numerical Methods in Engineering*, 23:409–428, 1986.

[66] A. Forsgren, P.E. Gill, and M.H. Wright. Interior methods for nonlinear optimization. *SIAM Review*, 55(4):525–597, 2002.

[67] R.W. Freund and N.M. Nachtigal. QMR: A quasi-minimal residual method for non-Hermitian linear systems. *Numerische Mathematik*, 60(3):315–339.

[68] C. Geiger and Chr. Kanzow. *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer, Berlin - Heidelberg, 2002.

[69] A. Gersborg-Hansen, O. Sigmund, and R.B. Haber. Topology optimization of channel flow problems. *Structural and Multidisciplinary Optimization*, 30(3):181–192, 2005.

[70] Ch. Großmann and H.G. Roos. *Numerik partieller Differentialgleichungen*. B.G. Teubner, Stuttgart, 1994.

[71] W. Hackbusch. *Iterative Lösung großer schwachbesetzter Gleichungssysteme*. B.G. Teubner, Stuttgart, 1991.

[72] W. Hackbusch. A sparse matrix arithmetic based on $\mathcal{H}$-matrices. I: Introduction to $\mathcal{H}$-matrices. *Computing*, 62(2):89–108, 1999.

[73] W. Hackbusch. *Multi-Grid Methods and Applications*. Springer, Berlin, 2003.

[74] W. Han and B.D. Reddy. *Plasticity: Mathematical theory and numerical analysis*, volume 9 of *Interdisciplinary Applied Mathematics*. Springer, 1999.

[75] B. Heinrich. *Finite Difference Methods on Irregular Networks. A Generalized Approach to Second Order Elliptic Problems*, volume 82 of *International Series of Numerical Mathematics*. Birkhäuser, Basel, 1987.

[76] R.H.W. Hoppe and S.I. Petrova. Primal-dual newton interior point methods in shape and topology optimization. *Numerical Linear Algebra with Applications*, 11:413–429, 2004.

[77] R.H.W. Hoppe, S.I. Petrova, and V.H. Schulz. Numerical analysis and its applications - topology optimization of conductive media described by maxwell's equations. *Lecture Notes in Computer Science*, pages 414–422, 2001.

[78] J. Jonsmann, O. Sigmund, and S. Bouwstra. Multi degrees of freedom electro-thermal microactuators. *TRANSDUCERS'99*, pages 1372–1375, 1999.

[79] M. Jung and U. Langer. Applications of multilevel methods to practical problems. *Surveys on Mathematics for Industry*, 1:217–257, 1991.

[80] M. Jung and U. Langer. *Methode der finiten Elemente für Ingenieure*. B.G. Teubner, Stuttgart - Leipzig - Wiesbaden, 2001.

[81] A. Kawamoto. *Generation of Articulated Mechanisms by Optimization Techniques*. PhD thesis, Technical University of Denmark, Department of Mathematics, 2004.

[82] U. Langer and W. Queck. On the convergence factor of Uzawa's algorithm. *Journal of Computational and Applied Mathematics*, 15:191–202, 1986.

[83] U. Langer and W. Queck. Preconditioned Uzawa-type iterative methods for solving mixed finite element equations. Theory - application - software. *Wissenschaftliche Schriftenreihe Technische Universitaet Karl-Marx-Stadt 3*, page 90p., 1987.

[84] U.D. Larsen, O. Sigmund, and S. Bouwstra. Design and fabrication of compliant micromechanisms and structures with negative poisson's ratio. *IEEE Journal of Microelectromechanical Systems*, 6(2):99–106, 1997.

[85] R. Lipton. Homogenization theory and the assessment of extreme field values in composites with random microstructure. *SIAM Journal on Applied Mathematics*, 65:475–493, 2004.

[86] R. Lipton and M. Stuebner. Optimal design of graded microstructure through inverse homogenization of control of pointwise stress. To appear in *Quarterly Journal of Mechanics and Applied Mathematics*, 2005.

[87] D. Lukáš. On solution to an optimal shape design problem in 3-dimensional linear magnetostatics. *Applications of Mathematics*, 49(5):441–464.

[88] D. Lukáš, D. Ciprian, J. Pištora, K. Postava, and M. Foldyna. Multilevel solvers for 3-dimensional optimal shape design with an application to magneto-optics. In *Proceedings of the 9th International Symposium on Microwave and Optical Technology (ISMOT)*, pages 5445–5451, 2003.

[89] B. Maar and V. Schulz. Interior point multigrid methods for topology optimization. *Structural and Multidisciplinary Optimization*, 19:214–224, 2000.

[90] G. Dal Maso. *An introduction to* Γ*-convergence*. Birkhäuser, Basel, 1993.

[91] G. Maurant. *Computer solution of large linear systems*, volume 28 of *Studies in Mathematics and its Applications*. Elsevier, 1999.

[92] K. Maute, S. Schwarz, and E. Ramm. Adaptive topology optimization of elastoplastic structures. *Structural and Multidisciplinary Optimization*, 15(2):81–91, 1998.

[93] A.G.M. Michell. The limit of economy of material in frame structures. *Philosophical Magazine*, 8(6):589–597, 1904.

[94] L. Modica and S. Mortola. Un esempio di Γ-convergenza. *Boll. Unione Mat. Ital.*, 14(B):285–299, 1977.

[95] M.G. Mullender, R. Huiskes, and H. Weinans. A physiological approach to the simulation of bone remodelling as a self-organizational control process. *Journal of Biomechanics*, 11:1389–1394, 1994.

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

[97] G. Of and O. Steinbach. A fast multipol boundary element method for a modified hypersingular boundary integral equation. In W. Wendland et al., editor, *Analysis and simulation of multifield problems. Selected papers of the international conference on multifield problems, Stuttgard, Germany, April 8-10, 2002*, volume 12 of *Lecture Notes in Applied Computational Mechanics*, pages 163–169. Springer, 2003.

[98] S. Osher and R.P. Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*. Springer, New York, 2002.

[99] S. Osher and J.A. Sethian. Fronts propagating with curvature dependent speed: Algorithms based on hamilton-jacobi formulations. *Journal of Computational Physics*, 79(1):12–49, 1988.

[100] Chr.C. Paige and M.A. Saunders. Solution of sparse indefinite linear systems of linear equations. *SIAM Journal on Numerical Analysis*, 12:617–629, 1975.

[101] J.T. Pereira, E.A. Fancello, and C.S. Barcellos. Topology optimization of continuum structures with material failure constraints. *Structural and Multidisciplinary Optimization*, 26:50–66, 2004.

[102] J. Peterson and O. Sigmund. Slope constrained topology optimization. *International Journal for Numerical Methods in Engineering*, 41:1417–1434, 1998.

[103] J. Petersson. A finite element analysis of optimal variable thickness sheets. *SIAM Journal on Numerical Analysis*, 36(6):1759–1778, 1999.

[104] J. Petersson. Some convergence results in perimeter-controlled topology optimization. *Computer Methods in Applied Mechanics and Engineering*, 171, 1999.

[105] St. Reitzinger. *Algebraic Multigrid Methods for Large Scale Finite Element Equations*. Schriften der Johannes Kepler Universität Linz. Universitätsverlag Rudolf Trauner, Linz, Austria, 2001.

[106] A. Rietz. Sufficiency of a finite exponent in the simp (power law) method. *Structural and Multidisciplinary Optimization*, 12:159–163, 2001.

[107] M.P. Rossow and J.E. Taylor. A finite element method for the optimal design of variable thickness sheets. *AIAA J.*, (11):1566–1569, 1973.

[108] G.I.N. Rozvany. Aims, scope, methods, history and unified terminology of computer-aided topology optimization in structural mechanics. *Structural and Multidisciplinary Optimization*, 21:90–108, 2001.

[109] G.I.N. Rozvany. On design-dependent constraints and singular topologies. *Structural and Multidisciplinary Optimization*, 21(2):164–173, 2001.

[110] G.I.N. Rozvany and M. Zhou. The coc algorithm, part I: Cross-section optimization or sizing. *Computer Methods in Applied Mechanics and Engineering*, 89:281–308, 1991.

[111] Y. Saad. *Iterative methods for sparse linear systems*. SIAM Society for Industrial and Applied Mathematics, Philadelphia, second edition, 2003.

[112] Y. Saad and M.H. Schulz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linera systems. *SIAM Journal on Scientific Computing*, 7(3):856–869, 1986.

[113] E.W. Sachs. Control applications of reduced SQP-methods. In R. Bulirsch and D. Kraft, editors, *Proceedings of the 9th IFAC Workshop on Control Applications of Optimization*, pages 89–104, 1994.

[114] S. Sauter and C. Schwab. *Randelementmethoden*. Teubner, Wiesbaden, 2004.

[115] A.H. Schatz, V. Thomée, and W.L. Wendland. *Mathematical Theorie of Finite and Boundary Element Methods*. Birkhäuser, 1990.

[116] J. Schöberl and W. Zulehner. On Schwarz-type smoothers for saddle point problems. *Numerische Mathematik*, 95(2):377–399, 2003.

[117] J. Schöberl et al. NETGEN/NGSolve software package, 2006. Home page: http:\\www.hpfem.jku.at [04 January 2006].

[118] V.H. Schulz and H.G. Bock. Partially reduced sqp methods for large-scale nonlinear optimization problems. *Nonlinear Analysis, Theory, Methods & Applications*, 30(8):4723–4734, 1997.

[119] Ch. Schwab. *p- and hp- Finite Element Methods*. Oxford University Press, New York, 1998.

[120] J. Sethian and A. Wiegmann. Structural boundary design via level set and immersed interface methods. *Journal of Computational Physics*, 163(2):489–528, 2000.

[121] O. Sigmund. *Design of material structures using topology optimization*. PhD thesis, Technical University of Denmark, Department of Solid Mechanics, 1994.

[122] O. Sigmund. On the design of compliant mechanisms using topology optimization. *Mechanics of Structures and Machines*, 25(4):495–526, 1997.

[123] O. Sigmund. On the optimality of bone microstructure. In P. Pederson and M.P. Bendsøe, editors, *Synthesis in Bio Solid Mechanics*, IUTAM, pages 221–234. Kluwer, 1999.

[124] O. Sigmund. Design of multiphysics actuators using topology optimization - Part I: One-material structures. *Computer Methods in Applied Mechanics and Engineering*, 190(49-50):6577–6604, 2001.

[125] O. Sigmund. Design of multiphysics actuators using topology optimization - Part II: Two-material structures. *Computer Methods in Applied Mechanics and Engineering*, 190(49-50):6605–6627, 2001.

[126] O. Sigmund and P.M. Clausen. Topology optimization using a mixed-formulation: An alternative way to solve pressure load problems. Submitted, 2005.

[127] O. Sigmund and J.S. Jensen. Topology optimization of elastic band gap structures and waveguides. In H.A. Mang, F.G. Rammerstorfer, and J. Eberhardsteiner, editors, *Proceedings of the Fifth World Congress on Computational Mechanics*. Vienna University of Technology, Austria, 2002.

[128] O. Sigmund and J.S. Jensen. Topology optimization of photonic crystal structures: A high-bandwidth low-loss T-junction waveguide. *J. Opt. Soc. Am. B*, pages 1191–1198, 2005.

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

[130] O. Sigmund and S. Torquato. Design of materials with extreme thermal expansion using a three-phase topology optimization method. *Journal of the Mechanics and Physics of Solids*, 45(6):1037–1067, 1997.

[131] O. Sigmund, S. Torquato, and I.A. Aksay. On the design of 1-3 piezocomposites using topology optimization. *Journal of Material Research*, 13(4):1038–1048.

[132] J. Sokolowski and A. Zochowski. On the topological derivative in shape optimization. *SIAM Journal on Control and Optimization*, 37(4):1251–1272, 1999.

[133] R. Stainko. An adaptive multilevel approach to minimal compliance topology optimization. *Communications in Numerical Methods in Engineering*, 22:109–118, 2006.

[134] O. Steinbach. *Numerische Näherungsverfahren für elliptische Randwertprobleme: Finite Elemente und Randelemente*. Teubner, Stuttgart, Leipzig, Wiesbaden, 2003.

[135] M. Stolpe. Global optimization of minimum weight truss topology problems with stress, displacement, and local buckling constraints using branch-and-bound. *International Journal for Numerical Methods in Engineering*, 61:1270–1309, 2004.

[136] M. Stolpe. On the reformulation of topology optimization problems as linear or convex quadratic mixed 0–1 programs. Technical Report 2004-13, Department of Mathematics, Technical University of Denmark, September 2004.

[137] M. Stolpe and K. Svanberg. An alternative interpolation scheme for minimum compliance topology optimization. *Structural and Multidisciplinary Optimization*, 22:116–124, 2001.

[138] M. Stolpe and K. Svanberg. On the trajectories of penalization methods for topology optimization. *Structural and Multidisciplinary Optimization*, 21:128–139, 2001.

[139] M. Stolpe and K. Svanberg. On the trajectories of the epsilon-relaxation approach for stress-constrained truss topology optimization. *Structural and Multidisciplinary Optimization*, 21:140–151, 2001.

[140] M. Stolpe and K. Svanberg. Modeling topology optimization problems as linear mixed 0–1 programs. *International Journal for Numerical Methods in Engineering*, 57(5):723–739, 2003.

[141] K. Svanberg. The method of moving asymptotes - a new method for structural optimization. *International Journal for Numerical Methods in Engineering*, 24:359–373, 1987.

[142] K. Svanberg. On the convexity and concavity of compliances. *Structural Optimization*, 7:42–46, 1994.

[143] K. Svanberg. A class of globally convergent optimization methods based on conservative convex separable approximations. *SIAM Journal of Optimization*, 12(2):555–573, 2002.

[144] J.W. Thomas. *Numerical Partial Differential Equations: Finite Difference Methods*. Springer, New York, 1995.

[145] S. Turteltaub. Optimal control and optimization of functionally graded materials for thermomechanical processes. *International Journal of Solids and Structures*, 39(12):3175–3197, 2002.

[146] A. Wächter et al. Ipopt software package, 2006. Home page: http:\\projects.coin-or.org \Ipopt [01 February 2006].

[147] A. Waechter and L.T. Biegler. On the implementation of an interior–point filter line–search algorithm for large–scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.

[148] M.Y. Wang, X.M. Wang, and D.M. Guo. A level set method for structural topology optimization. *Computer Methods in Applied Mechanics and Engineering*, 192(1-2):227–246, 2003.

[149] M.Y. Wang and S. Zhou. Phase transition: A variational method for structural topology optimization. Technical report, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, June 2003.

[150] G. Wittum. On the convergence om multi-grid methods with transforming smoothers. *Numerische Mathematik*, 57(1):15–38, 1990.

[151] S.J. Wright. *Primal-Dual Interior-Point Methods*. SIAM, Society for Industrial and Applied Mathematics, Philadelphia, 1997.

[152] J. Yoo and H. Hong. A modified density approach for topology optimization in magnetic fields. *International Journal of Solids and Structures*, 41(9-10):2461–2477, 2004.

[153] O.C. Zienkiewics. *The Finite Element Method in Engineering Science*. McGraw-Hill, London, third edition, 1977.

[154] Chr. Zillober. A combined convex approximation - interior point approach for large scale nonlinear programming. *Optimization and Engineering*, 2:51–73, 2001.

[155] Chr. Zillober. Global convergence of a nonlinear programming method using convex approximations. *Numerical Algorithms*, 27:265–289, 2001.

[156] Chr. Zillober. Scpip - an efficient software tool for the solution of structural optimization problems. *Structural and Multidisciplinary Optimization*, 24:362–371, 2002.

[157] W. Zulehner. A class of smoothers for saddle point problems. *Computing*, 65(3):227–246.

[158] W. Zulehner. Analysis of iterative methods for saddle point problems: A unified approach. *Mathematics of Computation*, 71:479–505, 2002.

Eidestattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Linz, im Februar 2006 DI Roman Stainko

Curriculum Vitae

Name: Roman Stainko

Date of birth: June 1, 1976

Place of birth: Linz, Austria.

Nationality: Austria.

Affiliation: Special Research Program (SFB) F013
Numerical and Symbolic Scientific Computing
Johannes Kepler University Linz
Altenbergerstraße 69, A-4040 Linz, Austria
`http://www.sfb013.uni-linz.ac.at`

Education:

September 1982 - July 1986	Elementary school.
September 1986 - July 1994	Grammar school.
October 1994	Started studying *Technical Mathematics* at the Johannes Kepler University Linz.
February 1998 - June 1998	Stay at DTU (Danish Technical University) in Lyngby, Copenhagen.
November 2000	Diploma in Nonsmooth Optimization.
November 2001	Started as a PhD student at SFB F013, Project F1309: *Multilevel Solvers for Large Scale Discretized Optimization Problems.*
May 2003 - June 2003	Stay at the department of mathematics (MAT), at DTU, Lyngby, Copenhagen.
March 2006	Doctorate in Computational Mathematics.

Employment:

November 2001 - March 2006	Research Assistant at the SFB F013.