Projection type methods in Banach space with application in topology optimization

DISSERTATION ZUR ERLANGUNG DES DOKTORGRADES
DER NATURWISSENSCHAFTEN (DR. RER. NAT.)
DER FAKULTÄT FÜR MATHEMATIK
DER UNIVERSITÄT REGENSBURG

vorgelegt von
Christoph Rupprecht aus
Weiden i. d. Opf.
im Jahr 2016

Promotionsgesuch eingereicht am: 14.1.2016

Die Arbeit wurde angeleitet von: Prof. Dr. Luise Blank

Prüfungsausschuss:	Vorsitzender:	Prof. Dr. Bernd Ammann
	1. Gutachter:	Prof. Dr. Luise Blank
	2. Gutachter:	Prof. Dr. Michael Hinze
	weiterer Prüfer:	Prof. Dr. Harald Garcke
	Ersatzprüfer:	Prof. Dr. Helmut Abels

Abstract

This thesis proposes a generalization of the projected gradient method with variable metric to an abstract Banach space setting. The motivation is the increasing interest in optimization problems posed in a Banach space and the current lack of general, globally convergent optimization methods therefor. Global convergence of the new variable metric projection type (VMPT) method is shown by adapting the assumptions of the finite dimensional setting appropriately using two different norms. Many aspects of the method are discussed, in particular different globalization strategies, the incorporation of second order information leading to Newton and quasi-Newton type methods and an application to proximal gradient methods. Similarities to existing numerical methods are pointed out and the application to a model problem is presented. The application of the new method to a topology optimization problem using a phase field model is discussed in detail. It is shown that the weak conditions for global convergence are satisfied. A semismooth Newton method for the solution of the arising subproblem is presented and local convergence is shown on the discrete level. The various numerical results are compared to the literature and to other state-of-the-art solvers, showing the superior performance of the new method. An existing modern time stepping scheme is enhanced by the introduction of adaptively chosen time step sizes, based on the theoretical results of the thesis. Multiple choices for the variable metric are discussed analytically and numerically and a problem specific scaling is derived. Moreover, reasonable choices for the problem parameters such as the stiffness tensor interpolation are analyzed. Numerical experiments show that the sensible choice of the mentioned parameters of the topology optimization problem and the numerical method lead to a huge boost in performance. The numerical experiments for the compliant mechanism problem disclose new difficulties for the used phase field model which have to be considered in future modeling.

Acknowledgements

I want to thank my supervisor Luise Blank for her continuous support and for providing a lot of helpful ideas and contributing with her great experience. Thank you for giving me so much freedom in the realization of this thesis. I am grateful for the interesting discussions with Harald Garcke and Helmut Abels. My thanks go to Michael Ulbrich, who had the idea to apply my results to the minimization of nonsmooth cost functionals. Moreover I want to thank Lars Müller for proof reading. Finally my thanks go to my wife and my family for their constant support.

Contents

1 Introduction

Over the last six decades much progress has been made in the theory, numerics and applications of optimization problems. Until now, many classes of optimization problems are well understood and efficient numerical solvers therefor have been developed. However, the majority of the results is concerned with finite dimensional problems. There is also much research done in Hilbert spaces. However, when it comes to Banach spaces, especially for constrained optimization problems, there are much less numerical methods available. In this thesis we will propose a new numerical method for convexly constrained optimization problems in Banach spaces.

A simple example of an optimization problem in a Banach space is the following. Find a measurable function $\varphi : [0,1] \to \mathbb{R}$ such that $j(\varphi) := \int_0^1 \cos(\varphi(x))\,\mathrm{d}x$ is minimized. It can be shown that $j : L^p((0,1)) \to \mathbb{R}$ is two times continuously differentiable if and only if $p > 2$ [Trö09]. As a consequence, the Hilbert space $L^2((0,1))$ cannot be chosen if second order derivatives of j are needed in the analysis or in the numerics.

In the following we give some important examples for infinite dimensional Banach spaces used in optimization and their applications. In most cases function spaces are considered, respectively spaces of generalized functions such as measures or distributions, defined on a domain Ω, which is an open subset of $\mathbb{R}^n$ or a lower dimensional object like a part of the boundary of another open subset. This also includes the lateral boundary of a space-time cylinder or its bottom appearing in parabolic optimal control problems. An example is the space $BV(\Omega)$ of functions with bounded variation, which are $L^1(\Omega)$-functions whose distributional derivative is a measure. The BV-seminorm favors piecewise constant functions and preserves edges. Thus the space appears for instance in topology optimization, image denoising or image segmentation, see [BDH12] and contained references. On the other hand the $L^1(\Omega)$-norm is known to promote sparsity and is used if sparse optimal solutions are desired, see [CK11, CRT13] and the references therein. The space $\mathcal{M}(\Omega)$ of regular Borel measures can be used instead of $L^1(\Omega)$ due to the lack of compactness of bounded sets in $L^1(\Omega)$ [CK11]. Often optimal control problems are only posed in the Banach space $L^p(\Omega)$, $p > 2$, instead of the Hilbert space $L^2(\Omega)$, since the reduced cost functional can only be shown to be differentiable in the former space, which can be due to nonlinearities in the unreduced cost functional itself or in the state equation [HUU99, ART02, Trö09]. Higher integrability of the control is also needed if a certain regularity, e.g. continuity, of the state is wanted. Another example for a problem in $L^p(\Omega)$-space is the seeking of controls with minimal $L^p(\Omega)$-norm [GLS05] and stochastic convex feasibility problems [BI12]. The space $L^\infty(\Omega)$ plays a special role in the analysis since it is not reflexive. If in addition derivatives or traces are needed then the Sobolev spaces $W_p^k(\Omega)$, $k \in \mathbb{N}$, $1 \le p \le \infty$, can be used (e.g. [KL94]), which consist of $L^p(\Omega)$ functions, whose partial derivatives up to order k are again $L^p(\Omega)$-functions. Finally, the space $C(\overline{\Omega})$ of continuous functions appears often in state constrained optimal control problems as space for the state variable [HPUU08, Trö09], since certain constraint qualifications are only fulfilled in this space.

We emphasize that it is crucial to understand the *infinite* dimensional setting. In practise it seems to be sufficient to consider only optimization methods for finite dimensional problems, since infinite dimensional problems have to be discretized in some way before they can be solved on a computer, which can only store finite dimensional vectors. However, in this case the method is only understood for fixed discretization parameter h. It can happen that the convergence of the method depends on h, e.g. that the numerical

method performs worse the smaller h is, or that the area of local convergence shrinks with h [UU00]. This is known as mesh-dependent behavior. We will study an example of such a method in Section 6.13.11. Thus it is important to analyze the numerical method in the infinite dimensional setting. Another aspect of infinite dimensional analysis is the following. Assume that the infinite dimensional problem is approximated by a sequence of finite dimensional problems, each having a global minimum. However, if the infinite dimensional problem does not have minimizers, it may happen that the sequence of global optima does not converge. For instance in compliance minimization this manifests itself in mesh dependent solutions and in the development of microstructures as $h \to 0$, see e.g. [BS03]. In optimization it is thus essential to understand the problem itself as well as the applied numerical method in the infinite dimensional setting, which is one of the goals of this thesis.

In this thesis we develop a generalization of the projected gradient method to a Banach space setting. In the following we will call this generalization 'variable metric projection type' (VMPT) method. The new method is designed to solve convexly constrained nonlinear optimization problems in Banach spaces of the form

$$\min j(\varphi), \quad \varphi \in \Phi_{ad} \subset X, \tag{1}$$

where X is a Banach space, Φ_{ad} is a convex subset and $j : X \to \mathbb{R}$ is the cost functional. If X is a Hilbert space then the projected gradient method is a well known solver for such problems. The projected gradient method is an iterative solver which determines for a given iterate $\varphi_k \in \Phi_{ad}$ the next iterate by the recursion $\varphi_{k+1} = P_\perp(\varphi_k - \nabla j(\varphi_k))$, where $P_\perp$ denotes the orthogonal projection onto Φ_{ad}. To ensure the existence of a projection, Φ_{ad} is assumed to be closed and convex. Thus, as in the method of steepest descent, the method moves along the direction of the negative gradient, whereupon a projection is applied to account for the constraints. As a consequence the iterates of the methods are feasible, i.e. we have $\varphi_k \in \Phi_{ad}$ for all k. To ensure that the constructed sequence converges to a solution of (1) one has to include a globalization, for which two possibilities are available. The first one is a curved search, where the new iterate φ_{k+1} is chosen along the projection arc $(0, \infty) \ni \lambda \mapsto \gamma(\lambda) := P_\perp(\varphi_k - \lambda \nabla j(\varphi_k))$, which is a curve in Φ_{ad}, see the left hand side of Figure 1. Thus the next iterate is $\varphi_{k+1} = \gamma(\lambda_k)$, where $\lambda_k > 0$ has to be chosen appropriately. The second possibility is a line search along the obtained descent direction, where φ_{k+1} is chosen on the line segment connecting φ_k and $P_\perp(\varphi_k - \nabla j(\varphi_k))$. In this case it holds $\varphi_{k+1} = \varphi_k + \alpha_k(P_\perp(\varphi_k - \nabla j(\varphi_k)) - \varphi_k)$, where $\alpha_k \in (0, 1]$ has to be chosen appropriately, see the right hand side of Figure 1.
We want to give a short survey of the projected gradient method. The method was first introduced in Hilbert space by Goldstein [Gol64] and Levitin and Polyak [LP66]. They use the curved search along the projection arc, where λ_k is chosen in an interval which depends on a priori unknown data like the Lipschitz constant of j' or some uniform upper bound of j''. McCormick and Tapia [MT72] choose λ_k as minimizer of j along the projection arc $\gamma(\lambda)$, $\lambda \geq 0$. On the one hand this is independent of the unknown Lipschitz constant, on the other hand it may be expensive to compute due to the nonsmoothness of the resulting one-dimensional problem. A more practicable choice for λ_k was introduced by Bertsekas [Ber76]. He adapted the well known Armijo-backtracking from unconstrained optimization, such that it can be used in the projected gradient method. In this backtracking, λ_k is chosen as $\lambda_k = \beta^{m_k} s$ with $s > 0$, $\beta \in (0, 1)$ and $m_k \in \mathbb{N}_0$ being the smallest integer such

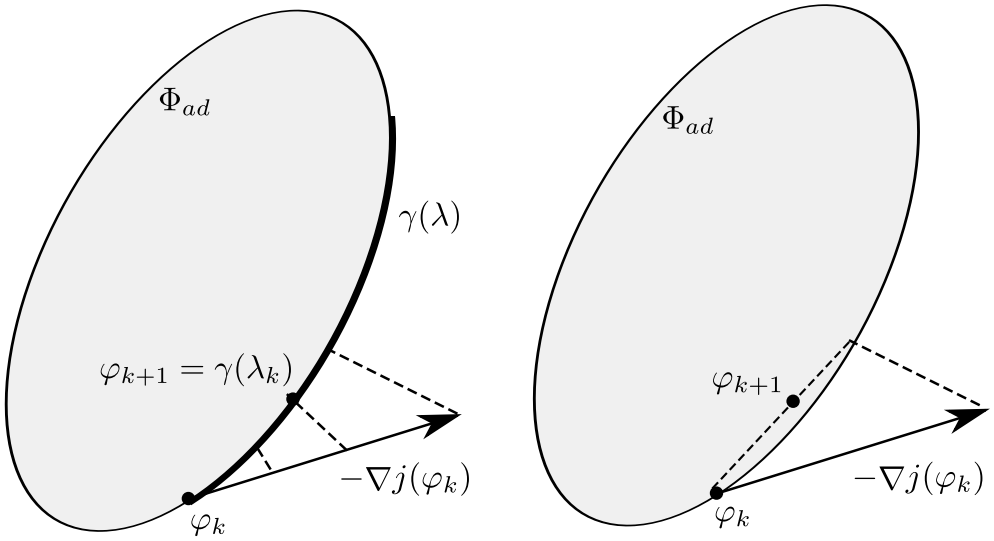


Figure 1: Projected gradient methods. Curved search (left) and line search (right).

that it holds

$$j(\gamma(\beta^{m_k} s)) \leq j(\varphi_k) - \sigma \frac{\|\varphi_k - \gamma(\beta^{m_k} s)\|^2}{\beta^{m_k} s} \tag{2}$$

for some $\sigma \in (0,1)$ independent of k. The minimal power m_k can be easily obtained by trial and error, i.e. one starts with $m_k = 0$ and increases m_k until the Armijo condition (2) is satisfied.

The alternative globalization, namely the line search along the feasible direction, was first proposed by Rosen [Ros60, Ros61] in finite dimension. However, he does not utilize the orthogonal projection onto Φ_{ad}, but rather projects the gradient onto the affine linear space which is determined by the linearization of the active constraints. A backtransport afterwards ensures the feasibility of the new iterate if the constraints are nonlinear. Rosen's projected gradient method was the first method capable of solving nonlinear constrained optimization problems. The line search method in combination with orthogonal projections in Hilbert space is treated in [DR70]. However, the step size α_k again depends on unknown data such as the Lipschitz constant of j'. The following more practical Armijo backtracking along the feasible direction is treated in [All80]. The feasible direction is given by $v_k := P_\perp(\varphi_k - \nabla j(\varphi_k)) - \varphi_k$ and the step size is chosen as $\alpha_k = \beta^{m_k}$ with $\beta \in (0,1)$ and $m_k \in \mathbb{N}_0$ being the smallest integer such that it holds

$$j(\varphi_k + \beta^{m_k} v_k) \leq j(\varphi_k) + \beta^{m_k} \sigma \left\langle j'(\varphi_k), v_k \right\rangle .$$

for some $\sigma \in (0,1)$ independent of k.

For more flexibility of the numerical method a variable metric can be introduced. In a Hilbert space setting this amounts to a change of the underlying inner product in each iteration. Thereby second order information can be included, giving rise to Newton and Quasi-Newton type methods. Let a_k denote the inner product used in the kth step of the method. The variable metric affects the orthogonal projection $P_\perp^{a_k}$ as well as the gradient $\nabla_{a_k} j(\varphi_k)$, which both depend on the inner product a_k. Recall that the gradient of j is

the Riesz representative of the Fréchet derivative. The resulting scaled projected gradient method is well-known in finite dimension, see [Ber99, Rus84], where global convergence is shown. There are also results available in Hilbert space, see [Dun87, GD88]. However, these results are only of local nature. To our knowledge there is no global convergence proof for the scaled projected gradient method in Hilbert spaces available in the literature. However, a closer look at the proofs in [Ber99] for the finite dimensional setting reveals that the transition from finite dimension to Hilbert space is straight forward. Only minor adaptions of the proof are necessary, like the replacement of convergence by weak convergence in some places. In [GB84] it is nonetheless claimed that the scaled projected gradient method is also well known in Hilbert spaces.
The typical assumption on the variable metric is of the type

$$\exists C, c > 0 : \quad c\|v\|^2 \leq a_k(v,v) \leq C\|v\|^2 \quad \forall v \in X, k \in \mathbb{N}_0 \tag{3}$$

where $\|.\|$ denotes the Hilbert space norm. Under this assumption each inner product a_k is equivalent to the inner product of the Hilbert space and thus (X, a_k) itself is a Hilbert space.
Up to now many contributions concerning the projected gradient method are available in finite dimension and Hilbert space. These include convergence rate estimates [LP66, DR70, All80, Dun81, Dun87, GD88], convergence of the whole sequence [Ius03, XWK07], local analysis including conditions for attractors [Dun81, Dun87, GD88], identification of active sets [Ber76, GD88, Dun88, KS92], two-metric extensions [Ber82, GB84, Kel99], mesh independency [KS92], projected Newton and quasi-Newton methods [LP66, Dun80, Ber82, Rus84, Dun88, Che96, Kel99], nonmonotone line search strategies and Barzilai-Borwein step length selection [BB88, BMR00], leading to the so called spectral projected gradient method. Applications to modern optimization problems can be found e.g. in [SKS11, PL14, ZXL15, NST15, MXW15]. Moreover, the review article [BMR14] provides many references for real world applications in optics, compressive sensing, geophysics, statistics, image restoration, atmospheric sciences, chemistry, dental radiology and others.

The analysis in this thesis is performed in an abstract Banach space setting, which includes from the examples above the spaces $W_p^k(\Omega)$, $k \in \mathbb{N}_0$, $1 < p < \infty$, and in particular $L^p(\Omega)$, as well as $L^\infty(\Omega)$. We will focus on the line search along the feasible direction, for which we are able to show global convergence. As it is pointed out in [Ber76], the line search variant may lead to poor convergence if many constraints are active at the solution. This can be seen in Figure 1, where on the left hand side much progress can be made by choosing $\lambda > 1$, whereas on the right hand side the largest possible step is given by the point $P_\perp(\varphi_k - \nabla j(\varphi_k))$. To circumvent this restriction, we therefore introduce an additional scaling of the gradient to be able to take larger steps, i.e. we consider the search direction $v_k := P_\perp^{a_k}(\varphi_k - \lambda_k \nabla_{a_k} j(\varphi_k)) - \varphi_k$ for some scaling parameter $\lambda_k > 0$ similar to [DR70, Ber76]. We note that the analysis of the curved search globalization cannot be carried over to our Banach space setting since the projection arc $\gamma(\lambda)$ may be discontinuous and therefore the existence of a positive λ_k fulfilling the Armijo condition (2) cannot be guaranteed. However, since we allow a variable scaling λ_k we at least propose a hybrid method, which is a mixture of both globalization strategies.
Although the transition from a finite dimensional setting to Hilbert space is straight forward, the transition from Hilbert to Banach space involves certain difficulties. The projected gradient method relies on the notion of an orthogonal projection and of a gradient. Both objects do not exist in general Banach spaces. We solve this issue by showing that at least a generalization of the mapping $\varphi \mapsto P_\perp^{a_k}(\varphi - \lambda_k \nabla_{a_k} j(\varphi))$ exists under suitable

assumptions. This can be achieved by viewing the orthogonal projection as a solution of a distance-minimization problem or equivalently of a variational inequality. In addition it is not obvious how to generalize the assumption (3) to Banach spaces. If we just assume (3) for the family of inner products a_k with $\|.\|$ being the Banach space norm then the following problem arises. From (3) it follows that the norm induced by a_k is equivalent to the Banach space norm for any k. Thus (X, a_k) is complete and hence a Hilbert space. We conclude that (3) can only hold if the Banach space X is isomorphic to some Hilbert space in the sense of linear homeomorphisms. Thus the assumption (3) is too strong for our purpose. We therefore use two different norms in our analysis and measure the lower bound in (3) with respect to a different norm than the upper bound. This gives the numerical method another flexibility and can account for two-norm discrepancies arising in optimal control problems [Trö09].
After coping with the mentioned difficulties we show global convergence of the method by adapting the techniques in [Ber99] for the finite dimensional setting. Moreover, we generalize our analysis also to apply to nonsmooth cost functionals. More precisely, this leads to a generalization of the proximal gradient method in Banach space using a variable metric. This method is known in Hilbert space using a variable metric [CV14] and also in Banach space using a constant metric [Bre09] (for more details see Section 4.11.2). However, the combination of a Banach space setting together with a variable metric is new. In addition we do not assume the convexity of the cost functional in contrast to other authors.

Most of the available numerical methods for problems in Banach space either converge only locally or have to assume strong assumptions like the convexity of the cost functional, or rely on a special structure of the admissible set, such as pointwise constraints in some L^p-space (see Section 3). In contrast, the generalization of the projected gradient method considered here converges *globally*, i.e. independent of the initial guess, it can handle *general nonlinear* cost functionals without assuming convexity, it does not rely on a special structure of the admissible set apart from convexity, some kind of boundedness and closedness, and the method can handle *two different norms*.

The usage of two different norms in our analysis has important consequences. For instance, in order to apply the classical projected gradient method with respect to the L^2-norm it is necessary that the cost functional is also differentiable with respect to this norm. Due to our generalization the projected gradient method can be applied with respect to the L^2-norm even if the cost functional is only differentiable in the much stronger L^∞-norm. We can show this under mild assumptions, e.g. that $j'(\varphi)$ is in $L^1 \subset (L^\infty)^*$.

In the second part of the thesis we present an application of the new VMPT method to a structural topology optimization problem. We consider a relaxed phase field formulation of the original perimeter penalized topology optimization problem which is a generalization of the models considered in [BGS+12, BFGS14]. The resulting nonlinear optimization problem is an elliptic optimal control problem posed in the Banach space $H^1 \cap L^\infty$ under linear constraints on the control, including also a nonlocal integral constraint. Since the analysis is performed in the space L^∞ we are able to allow any space dimension. The control enters the state equation in the second order coefficients, on the right hand side as distributed control and on the Neumann boundary in the sense of traces. Moreover, the control is a vector-valued function since we consider topology optimization of multiple materials. The objective we consider is a general smooth enough functional. The phase field model addresses typical issues in topology optimization: ill-posedness, non-smoothness of

the optimization problem, handling of topological changes and intermediate densities.

Global convergence of the VMPT method will be shown in $H^1 \cap L^\infty$, which facilitates together with the Γ-convergence result in [BGHR15] (see also Section 6.4) a rigorous tool for solving topology optimization problems. On the other hand, it turns out that for existing numerical methods for topology optimization problems, including pseudo time stepping methods, no convergence analysis is provided in the infinite dimensional setting and that mostly no rigorous stopping criterion is used (see Section 5 for details).

We generalize the results in [BFGS14] by showing well-posedness and C^2-regularity of the reduced cost functional. Therefor we utilize the direct method in the calculus or variations for the former and the implicit function theorem for the latter. Moreover we show existence and uniqueness of Lagrange multipliers for an arbitrary cost functional using the constraint qualification of Zowe and Kurcyusz [ZK79]. In the general setting these multipliers have only low regularity. Therefore no pointwise arguments can be used to show uniqueness of the Lagrange multipliers and we have to develop a new proof based on variational techniques. We also provide a numerical example where the Lagrange multiplier includes a measure concentrated on the boundary. Our results are a generalization of those in [BGSS13a], where only a concretely given objective is considered and where the Lagrange multipliers are L^2-functions, which can be treated pointwise.

We check the abstract assumptions for global convergence of the VMPT for various choices for the variable metric. These include the H^1-metric, a point-based choice including second order information and inner products coming from a time discrete Allen-Cahn and Cahn-Hilliard scheme. Moreover we consider a BFGS update based on the H^1-metric. We show that a sensible value for the scaling λ_k of the gradient has to depend linearly on the interfacial thickness and the phase field parameter ε, respectively.

Up to now we used pseudo time stepping methods to solve the considered topology optimization problem. However, the lack of convergence analysis and a stopping criterion therefor is the motivation to develop a new numerical scheme based on optimization techniques, from which the VMPT method originates. It turns out that in certain cases the used pseudo time stepping methods are a special instance of the VMPT method, which in particular proves global convergence of these methods. Moreover, due to this result we are able to propose a novel, easy-to-implement adaptive time stepping scheme for the Allen-Cahn and Cahn-Hilliard pseudo time stepping based on the Armijo condition. We show that changing the time step size is equivalent to changing the metric in the corresponding VMPT method. In addition the time step size is allowed to tend to infinity, in which case the projected H^1-gradient method is recovered. We provide numerical experiments to show that also in practice the time step size tends to infinity, which matches the results obtained in [DBH12] for a different adaptive scheme. By introducing adaptive time step sizes the presented Cahn-Hilliard method gets 80 times faster compared to a constant time step size as used in [BGS+12].

For the solution of the arising quadratic subproblem in the VMPT method we propose a primal dual active set (PDAS) method. This method is also used in [BGSS13a] for the solution of the quadratic optimization problem arising in a time step of the Allen-Cahn system. As in [BGSS13a] we show local convergence of the discrete method under the assumption that the graph of inactive sets is connected. This is proved using the equivalence

to a semismooth Newton method. However, our result generalizes that in [BGSS13a], since we do not prescribe a specific cost functional.
It is known that the PDAS method can be mesh dependent for such problems since it cannot be shown that the method converges in the infinite dimensional setting due to the lack of semismoothness. However, for our subproblem this is only a minor issue, since we use the active set of the last VMPT step as an initial guess, giving rise to a reasonable warm start. Moreover, we show numerically that the mesh dependency can be controlled well by introducing a nested iteration in the mesh parameter.

We obtain several results concerning the choices for the model parameters of the topology optimization problem. We show numerically that the usage of an obstacle potential in the Ginzburg-Landau energy is advantageous to a smooth potential. Moreover we show that a quadratic interpolation of the stiffness tensors leads to far better performance of the VMPT method as well as better separation of the phases in the optimal design compared to the linear interpolation used in [BGS+12, BFGS14]. Numerical examples demonstrate that the calculation time of the VMPT method can be reduced drastically by an elaborate choice of the variable metric, the scaling parameter λ, the stiffness interpolation scheme and the nesting strategy. In the experiment the computation time is decreased from 23 days to 9 minutes. We compare the different choices for the variable metric and discuss advantages and disadvantages. In particular it turns out that the objective value for local minima obtained by the variable metric including second order information is lower compared to e.g. the minima obtained by the projected H^1-gradient method.

Since the performance of the new VMPT method is much better than the previously used pseudo time stepping methods we are able to investigate the topology optimization problem in detail, even with data which make the problem difficult to solve. These include small values for the diffuse interface thickness, low penalization of the Ginzburg-Landau energy, a low volume fraction for the material and problems with multiple materials.
Moreover we present optimal designs for the compliant mechanism problem, which is known to be much harder to solve than the mean compliance problem. It turns out that some compliant mechanism solutions are not reasonable as the interface thickness approaches zero, since no full phase transition is formed. We argue that this is due to the poorly modeled cost functional which does not take a reaction force into account as in [Sig97]. We show numerically that this issue can be solved by introducing a workpiece which exerts a reaction force on the mechanism. However, this reaction force should be part of the model, for which further research is necessary.

Finally a comparison of the obtained optimal designs to the existing literature shows that on the one hand many optimal designs can be recovered by the VMPT method, but on the other hand the VMPT method converges to local minima which cannot be found in the existing literature. Moreover we show that the VMPT method is superior to state-of-the art numerical methods including pseudo time stepping, a reduced SQP method, and a semismooth Newton method, due to its fast and global convergence.

The goals of the thesis can be summarized as follows:

1. Analysis of a new abstract class of numerical methods in Banach space.

2. Discussion of the new method, i.e. special instances of the method, examples and comparison to other methods.

3. Analysis of a topology optimization problem using a phase field model.
4. Full numerical study of the VMPT method applied to the concrete optimization problem.
5. Numerical study of the topology optimization problem, including parameter choice and aspects of the phase field model.

The thesis is organized as follows.
Section 3 gives an overview of existing numerical methods for constrained optimization in Banach space.
Section 4 contains analysis and discussion of the VMPT method. The new method is derived on the basis of the projected gradient method in Section 4.1, and the algorithm for line search globalization is provided. Section 4.2 facilitates the precise assumptions together with the proof for the well-posedness of the subproblem and for global convergence. In Section 4.3 sufficient conditions for global convergence are discussed, where details for a point based choice of the variable metric are given in Section 4.4. Globalization along the projection arc is considered in Section 4.6. Only a weak continuity of the projection arc can be shown which does not imply the existence of positive step lengths. Thus a hybrid method is proposed combining both globalization strategies. In a Hilbert space setting no hybrid method is needed, for which a proof is given. A special instance of the VMPT method, namely the projected Newton's method, is included in Section 4.7. Global convergence is shown under a weak coercivity assumption on the Hessian, which takes two-norm discrepancies into account. Moreover, q-superlinear convergence rates are established under locally stronger conditions. These rates hold also if j' is only semismooth. In Section 4.8 a BFGS update of the variable metric is considered. Global convergence is shown under standard assumptions using an additional small shift of the metric. Section 4.9 contains remarks about the projection type subproblem. In particular a scaling of Barzilai-Borwein type is derived and the inexact solution of the subproblem is discussed. In Section 4.10 the VMPT method is generalized for the minimization of the sum of a differentiable and a convex nonsmooth functional, for which global convergence is established as in the smooth case. Section 4.11 comments on similarities of the VMPT method to other methods in the literature such as pseudo time stepping methods and operator splitting methods. Finally, a concrete application to a semilinear elliptic optimal control problem is presented in Section 4.12, where the problem is posed in the Banach space $L^\infty(\Omega)$ and the VMPT method is carried out with respect to the $L^2(\Omega)$-inner product.
Section 5 gives an introduction to general topology optimization. In particular, typical difficulties arising in topology optimization are pointed out and state-of-the-art numerical solvers and models are summarized.
Section 6 contains the analysis, discussion and numerics for an abstract class of structural topology optimization problems in linear elasticity using a vector-valued phase field model posed in the Banach space $H^1(\Omega)^N \cap L^\infty(\Omega)^N$. The abstract problem together with the used assumptions is stated in Section 6.1. Many examples for the abstract data are given. The special case of two phases is discussed separately, since it can be reduced to a problem using a scalar-valued phase field. Throughout the thesis the analysis is performed for the general vector-valued phase field problem and the results for the scalar-valued phase field are presented as corollaries. Section 6.2 is devoted to the analysis of the control-to-state

operator $S : H^1(\Omega)^N \cap L^\infty(\Omega)^N \to H^1(\Omega)^d$. It is shown that the state equation is well-posed and that the state depends locally Lipschitz continuously on the control. Moreover, C^2- regularity of S is shown and the PDEs for the linearized state and the second order derivatives are derived. It is discussed why the analysis is much easier in the Banach space $H^1(\Omega)^N \cap L^\infty(\Omega)^N$ than in the Hilbert space $H^1(\Omega)^N$. The existence of minimizers for the topology optimization problem in shown in Section 6.3. In Section 6.4 the Γ-convergence result of [BGHR15] is applied to our problem for the scalar-valued case if the boundary traction does not depend on the phase field. Comments on the vector-valued case are given. We establish first order optimality conditions in Section 6.5. The first order derivative of the reduced cost functional is represented by an adjoint approach. Existence and uniqueness of Lagrange multipliers with low regularity is proved. The resulting KKT system is also presented in a strong formulation. The second order derivatives of the reduced cost functional are computed in Section 6.6, again using an adjoint approach. In Section 6.7 we show that the topology optimization problem fulfills the assumptions for global convergence of the VMPT method. Moreover we introduce six choices for the variable metric, for which we also check the abstract assumptions. The theoretical results for the VMPT method are used in Section 6.8 to propose an adaptive choice for the time step size of pseudo time stepping methods of Allen-Cahn and Cahn-Hilliard type using different time discretizations. Moreover, a rigorous stopping criterion is proposed and global convergence is shown. A reduced SQP method is presented in Section 6.9, which is compared to the VMPT method later on. Section 6.10 contains details about the semismooth Newton (SSN) method and the primal dual active set method, respectively, which are applied to discretized problems with general objective functionals, including the functional of the topology optimization problem and the functional of the projection type subproblem. Local superlinear convergence is shown in case of the VMPT subproblem. This is also shown for a general objective functional under the assumption of a second order sufficiency condition and strict complementarity. Implementation details are given. It is shown that applying the SSN method to the unreduced or the reduced problem in case of two phases is equivalent under a certain assumption. Section 6.11 incorporates details about the used discretization of the VMPT method and the adaptive mesh. In Section 6.12 we derive a reasonable scaling parameter λ_k based on the interface width. For the used potential the corresponding surface tensions in the sharp interface problem are computed as well as the optimal phase transitions and the angle condition at the triple junctions.

Section 6.13 contains the various numerical results concerning the VMPT method applied to the mean compliance problem. These include results about the used potentials, the interpolation scheme for the stiffness tensors, mesh independency, nesting in the mesh parameter h, the performance of the PDAS method in the inner problem, the behavior of the different choices for the variable metric with respect to computation time and obtained local minimizers, as well as many numerical examples. Moreover, the dependency of the optimal designs and the VMPT method on the model parameters is studied. The obtained optimal designs are compared to existing results in the literature. The adaptivity for the time steps in the pseudo time stepping methods is evaluated numerically and the VMPT method is compared to the resulting method, as well as to the SQP method and the SSN method. Finally, numerical examples for Lagrange multipliers are given and counterexamples for the choice of the variable metric are discussed, which give rise to a mesh dependent VMPT method.

Section 6.14 contains the numerical results about the challenging compliant mechanism problem. As in the preceding section multiple inner products are used to compute the

optimal designs and a comparison is given. Difficulties for the phase field model in the limit $\varepsilon \to 0$ are pointed out on the basis of numerical experiments. It is demonstrated that the obtained local minimizers are undesired and thus a change in the used model is necessary. First numerical examples show that this issue can be circumvented by taking reaction forces into account.

Section 7 finally summarizes the new results of this thesis, its implications, and addresses open problems.

2 Notation and conventions

Fréchet derivatives are denoted by Df or f'. For functions between finite dimensional spaces this coincides with the Jacobian matrix. By $\nabla f := (Df)^T$ we denote the transposed Jacobian (the gradient). In particular for real valued functions Df is a row vector, whereas ∇f is a column vector. We emphasize that we strictly distinguish between the derivative and the gradient. For a real-valued function j in a Hilbert space the gradient ∇j is the Riesz representative of the derivative j', i.e. $\langle j'(\varphi), v\rangle = (\nabla j(\varphi), v)$ for all v, where $\langle .,. \rangle$ denotes the dual pairing and $(.,.)$ the inner product. Sometimes we annotate the space of the dual pairing as index like in $\langle .,. \rangle_{L^\infty,(L^\infty)^*}$. However, in most cases the space is clear from the context. Usually this will be the abstract space $\mathbb{X}\cap\mathbb{D}$ in Section 4 and the concrete space $H^1(\Omega)^N \cap L^\infty(\Omega)^N$ in Section 6. By $\nabla \cdot f$ we denote the divergence of a vector field. The divergence of a matrix-valued function $\nabla \cdot A$ is defined row-wise and gives a column vector-valued function as result. For partial (Fréchet-)derivatives we use the notation $f_x(x,y)$ or $D_x f(x,y)$ or $\partial_1 f(x_1,x_2)$. In finite dimension we write $\nabla_x f := (D_x f)^T$. We denote the space of linear and bounded operators from X to Y by $\mathcal{L}(X,Y)$ and identify $\mathcal{L}(X,\mathcal{L}(X,Y))$ with the space of bilinear continuous mappings $X\times X \to Y$. For the second order Fréchet derivative we write f''. Partial second order derivatives are denoted by $f_{x,y}$, where the differentiation with respect to x is applied first. Moreover we use the notation $D_x^2 f := f_{x,x}$, $D_x^1 f := f_x$ and $D_x^0 f := f$. Differentiation directions are placed in square brackets if more than one direction is present, i.e. $f_{x,y}(x,y)[a,b] := (D_y(D_x f))(x,y)ab$, where the direction a corresponds to y and the direction b to x. Directions are often named $\delta\varphi$ or $\tau\varphi$, which has to be read as single variable. Vector valued functions are typed in boldface. Set-valued mappings are denoted by $F : X \rightrightarrows Y$, which is a function from X to the power set of Y. For an introduction in differential calculus we refer to [Zei85].

For $1 \le p < \infty$ we denote by $L^p(\Omega)$ the space of p-integrable functions on Ω (where we identify functions which are equal almost everywhere). The special case $L^\infty(\Omega)$ is the space of all essentially bounded measurable functions (together with the mentioned identification). We denote by $W^{k,p}(\Omega)$ $(= W_p^k(\Omega))$ the Sobolev space of p-integrable real-valued functions on Ω with p-integrable weak derivatives up to order k, see also [AF03] for an introduction to Sobolev spaces. We abbreviate $H^1(\Omega) := W^{1,2}(\Omega)$. Sobolev spaces containing vector valued functions are denoted by $H^1(\Omega,\mathbb{R}^n)$ or in short $H^1(\Omega)^n$. We simply write $\|.\|_{L^2}$ instead of $\|.\|_{L^2(\Omega)^n}$, if the domain and the number of components is clear from the context. If no special qualifier for convergence is given (e.g. weak or weak-*), then strong convergence is meant. Often the $\mathrm{d}x$ is omitted in integral expressions, i.e. $\int_\Omega f := \int_\Omega f(x)\,\mathrm{d}x$. We also leave out the space variable x in superposition operators, i.e. $f(\varphi) := f(\varphi(x)) := f(x,\varphi(x))$ if the meaning is clear from the context.

In the presented estimates, $C > 0$ is always a generic constant, which can be different from estimate to estimate. We sometimes use Hölder's inequality $\|uv\|_{L^r} \le \|u\|_{L^p}\|v\|_{L^q}$ for $1/p + 1/q = 1/r$, the trace theorem and obvious inequalities like $\|\mathcal{E}(\boldsymbol{u})\|_{L^2} \le \|\boldsymbol{u}\|_{H^1}$, where $\mathcal{E}(u) = \frac{1}{2}(D\boldsymbol{u} + D\boldsymbol{u}^T)$, without reference.

The standard inner product for matrices is denoted by $A : B := \sum_{ij} a_{ij}b_{ij}$, the Euclidean inner product by $x\cdot y := \sum_i x_i y_i$ and the Euclidean norm by $|x|$. Also for matrix and tensor norms and the Lebesgue measure we write $|A|$, $|\boldsymbol{C}|$ and $|\Omega|$. Inequalities like $\boldsymbol{\varphi} \ge 0$ for vector valued functions are to be understood component-wise and almost everywhere.

We say that a sequence $(x_k)_k$ converges q-linearly to zero, if $|x_{k+1}|/|x_k| \to c$ for some $c \in (0,1)$, q-superlinearly, if $|x_{k+1}|/|x_k| \to 0$ and q-quadratically, if $|x_{k+1}|/|x_k|^2 \to c$ for some $c > 0$.

Since we consider a phase field model in Section 6, the term 'interface' always refers to the diffuse interface.

3 Existing numerical methods for constrained optimization in Banach space

We want to give an overview of state-of-the-art numerical methods for the solution of constrained optimization problems

$$\min j(\varphi), \quad \varphi \in \Phi_{ad} \subset X \tag{4}$$

posed in a Banach space X. For the Newton-type methods below we will also give the exact assumptions for convergence, since we will use these later in the thesis. We emphasize that we cite only results for a Banach space setting. In the case that X is a Hilbert space or a finite dimensional space much more results are available, of course.

First of all there are Newton-type methods, which are often used due to the typically fast convergence. For constrained optimization the **Josephy-Newton method** and the **semismooth Newton method** are appropriate, for which we cite some results which can be found e.g. in [HPUU08]. Both methods don't solve the optimization problem (4) directly, but are rather based on some optimality condition thereof. Suppose that Φ_{ad} is convex. Then a first order necessary condition for a minimizer φ is the variational inequality (VI)

$$\varphi \in \Phi_{ad}, \quad \left\langle j'(\varphi), \eta - \varphi \right\rangle \geq 0 \quad \forall \eta \in \Phi_{ad}. \tag{5}$$

On the other hand, if the admissible set is given by $\Phi_{ad} = \{\varphi \in X \mid e(\varphi) = 0,\ c(\varphi) \in K\}$ for some nonempty convex closed cone K and operators e, c, and if some constraint qualification is satisfied, then a first order necessary condition is given by the KKT system, which involves the primal variable φ and additional dual variables μ, λ, being Lagrange multipliers for the constraints $e(\varphi) = 0$ and $c(\varphi) \in K$, respectively. This KKT system can be written as a generalized equation of the form

$$0 \in G(x) + N(x), \tag{6}$$

where the unknown x contains the primal and dual variables, $G : Y \to Z$ is a continuously differentiable operator and $N : Y \rightrightarrows Z$ is a set-valued map with closed graph. If Φ_{ad} is nonempty, convex and closed then the variational inequality (5) can also be written as generalized equation with $Y = X$, $Z = X^*$, $x = \varphi$, $G = j' : X \to X^*$ and $N = N_{\Phi_{ad}} : X \rightrightarrows X^*$ the normal cone mapping of Φ_{ad}, i.e.

$$N_{\Phi_{ad}}(\varphi) = \begin{cases} \{\xi \in X^* \mid \langle \xi, \eta - \varphi \rangle \leq 0 \ \forall \eta \in \Phi_{ad}\} & \varphi \in \Phi_{ad} \\ \varnothing & \varphi \notin \Phi_{ad}. \end{cases}$$

The Josephy-Newton method can be used to solve the abstract inclusion (6). In particular the KKT system and the variational inequality (5) can be solved. As for the classical Newton method the algorithm involves successive linearization of the problem. However, since in general only G is differentiable, the linearization is only applied to G and not to N. This results in the recursion

$$0 \in G(x_k) + \left\langle G'(x_k), x_{k+1} - x_k \right\rangle + N(x_{k+1}) \tag{7}$$

for given initialization x_0. If the linearized inclusion possesses multiple solutions, then the solution nearest to x_k is taken. For the special case of the VI (5) the Josephy-Newton

method amounts to the linearized VI

$$\varphi_{k+1} \in \Phi_{ad}, \quad \left\langle j'(\varphi_k), \eta - \varphi_{k+1} \right\rangle + j''(\varphi_k)[\varphi_{k+1} - \varphi_k, \eta - \varphi_{k+1}] \geq 0 \quad \forall \eta \in \Phi_{ad}. \tag{8}$$

The Josephy-Newton method applied to the KKT system results in the SQP method, which calculates in each step a KKT triple $(d_k, \mu_{k+1}, \lambda_{k+1})$ of the quadratic program

$$\min_d \left\langle j'(\varphi_k), d \right\rangle + \frac{1}{2}\mathcal{L}_{\varphi\varphi}(\varphi_k, \mu_k, \lambda_k)[d, d] \tag{9}$$

$$e(\varphi_k) + e'(\varphi_k)d = 0 \tag{10}$$

$$c(\varphi_k) + c'(\varphi_k)d \in K \tag{11}$$

closest to $(0, \mu_k, \lambda_k)$, where $\mathcal{L}(\varphi, \mu, \lambda) = j(\varphi) + \langle \mu, e(\varphi) \rangle + \langle \lambda, c(\varphi) \rangle$ is the corresponding Lagrange functional. The primal variable is updated by $\varphi_{k+1} = \varphi_k + d_k$.
Local convergence of the Josephy-Newton method can be shown under the following regularity condition.

Definition 3.1. The generalized equation (6) is said to be strongly regular (in the sense of Robinson [Rob80]) at a solution x^* if the perturbed linearized generalized equation is locally uniquely solvable and the solution depends Lipschitz continuously on the perturbation, i.e. if there exist $\delta > 0$, $\varepsilon > 0$ and $L > 0$, such that for all $p \in Z$ with $\|p\| < \delta$ there exists a unique $x(p) \in Y$ with $\|x(p) - x^*\| < \varepsilon$ such that

$$p \in G(x^*) + \left\langle G'(x^*), x(p) - x^* \right\rangle + N(x(p))$$

and

$$\|x(p_1) - x(p_2)\| \leq L\|p_1 - p_2\| \quad \forall p_1, p_2 \in Z,\ \|p_1\| < \delta,\ \|p_2\| < \delta.$$

The following convergence result can be found verbatim in [HPUU08].

Theorem 3.2. *Let Y, Z be Banach spaces, $G : Y \to Z$ continuously differentiable and let $N : Y \rightrightarrows Z$ be set-valued with closed graph. If x^* is a strongly regular solution of the generalized equation* (6)*, then the Josephy-Newton method is locally q-superlinearly convergent in a neighborhood of x^*. If, in addition, G' is γ-Hölder continuous near x^*, then the order of convergence is $1 + \gamma$.*

A related method is the semismooth Newton method, which can be used to solve semismooth equations of the form

$$0 = G(x) \tag{12}$$

for some operator $G : Y \to Z$.

Definition 3.3. Let Y, Z be Banach spaces and $G : Y \to Z$ a continuous operator. Let $\partial G : Y \rightrightarrows \mathcal{L}(Y, Z)$ be given. Then G is called ∂G-semismooth at $x \in Y$ (in the sense of Ulbrich [Ulb01, Def. 3.1]) if

$$\sup_{M \in \partial G(x+h)} \frac{\|G(x+h) - G(x) - Mh\|}{\|h\|} \to 0 \quad \text{as } \|h\| \to 0.$$

G is called ∂G-semismooth of order $\gamma > 0$ at $x \in Y$ if

$$\sup_{M \in \partial G(x+h)} \|G(x+h) - G(x) - Mh\| = \mathcal{O}(\|h\|^{1+\gamma}) \quad \text{as } \|h\| \to 0.$$

For examples of semismooth operators we refer to [HPUU08].
The kth semismooth Newton step chooses some operator out of $\partial G(x_k)$ leading to the recursion

$$0 = G(x_k) + M_k(x_{k+1} - x_k) \quad \text{for some } M_k \in \partial G(x_k)$$

for given x_0. The following convergence result is available, see [HPUU08].

Theorem 3.4. *Let Y, Z be Banach spaces and $G : Y \to Z$ continuous and ∂G-semismooth at a solution x^* of (12). Let there exist $C, \delta > 0$, such that*

$$\|M^{-1}\| \le C \quad \forall M \in \partial G(x),\ x \in Y \text{ with } \|x - x^*\| < \delta.$$

Then the semismooth Newton method is locally q-superlinearly convergent in a neighborhood of x^*. If G is ∂G-semismooth of order $\gamma > 0$ at x^*, then the order of convergence is $1 + \gamma$.

In some cases the KKT system can be equivalently reformulated to a semismooth equation of the form (12). This can be done by replacing the complementarity condition in the KKT system by a semismooth projection equation, which is possible e.g. in finite dimension and for pointwise inequality constraints in L^2. We refer to Section 6.10.1 for an example. The semismooth Newton method can then be used to solve the resulting nonsmooth system.
There is also a Newton type method of Dunn [Dun80], which is not based on an optimality condition. It is similar to the Josephy-Newton method applied to the VI (5). However, an optimization problem corresponding to the linearized VI (8) is solved instead of the linearized VI itself in each Newton step, which is also well defined for non-convex Φ_{ad}.
A drawback of Newton type methods certainly is that convergence is only guaranteed if the initial guess is sufficiently close to the solution. To obtain global convergence additional effort has to be put into globalization strategies such as line search or trust region methods.

We have seen that optimality conditions can be written as an abstract generalized equation (6). Another class of numerical methods for solving generalized equations are **operator splitting methods**. They consider problems of the form

$$0 \in T(\varphi) \tag{13}$$

with either $T : X \rightrightarrows X$ or $T : X \rightrightarrows X^*$. For some splitting $T = T_1 + T_2$ the iterates of the method are given by the recursion

$$\frac{1}{\lambda_k}(\varphi_k - \varphi_{k+1}) \in T_1(\varphi_k) + T_2(\varphi_{k+1}),$$

where $\lambda_k > 0$ is a step size parameter. We refer to Section 4.11.2 for a detailed discussion. In the context of minimization problems, these methods include the proximal point method (the case $T_1 = 0$) and the proximal gradient method (the case $T_2 = \partial \chi_{\Phi_{ad}}$ and $T = j'$, see Section 4.10). Only few results are available if X is a Banach space compared to the Hilbert space or finite dimensional case. In [Bre09] convergence of a proximal gra-

dient method is shown if j is convex and the sum of a smooth and a nonsmooth functional (cf. also Section 4.10). Convergence results for the general problem (13) for accretive operators can be found in [LMMWX12, Cho15]. For the special case of the proximal point method the following results are available in Banach space: In [IB97] convergence is studied for convex optimization problems of the form (4). Convergence analysis for variational inequalities involving maximal monotone operators is covered in [BS00]. Finally a method for the general problem (13) with maximal hypomonotone T^{-1} is discussed in [OI07]. Except for the last reference all mentioned authors assume convexity of the optimization problem or some kind of monotonicity for the generalized equation (13).

Another method for solving constrained optimization problems in Banach spaces are **augmented Lagrangian methods**. These methods involve the primal and dual variables. Instead of the usual Lagrangian an augmented Lagrangian containing a penalty term is used. However, contrary to penalty methods the penalty parameter is not needed to tend to infinity, since an approximation of the Lagrange multiplier is maintained. As an example consider the case that $\Phi_{ad} = \{\varphi \in X \mid G(\varphi) \in K\}$, for some $G : X \to H$, a Hilbert space H and a convex closed cone $K \subset H$. The augmented Lagrangian then reads

$$\mathcal{L}_c(\varphi, \mu) = \inf_{y \in K - G(\varphi)} j(\varphi) - \langle \mu, y \rangle + \frac{c}{2} \|y\|^2$$

with penalty parameter $c > 0$, see [SS04]. For $c = 0$ the usual Lagrangian is recovered,

$$\mathcal{L}(\varphi, \mu) = j(\varphi) + \langle \mu, G(\varphi) \rangle$$

if $\mu \in K^- := \{\mu \in H^* \mid \langle \mu, \xi \rangle \leq 0 \ \forall \xi \in K\}$ (dual feasibility). The goal is to compute a solution of the augmented Lagrangian dual problem

$$\sup_{\mu \in H} \inf_{\varphi \in X} \mathcal{L}_c(\varphi, \mu), \tag{14}$$

while the primal problem

$$\inf_{\varphi \in X} \sup_{\mu \in H} \mathcal{L}_c(\varphi, \mu)$$

is equivalent to the optimization problem (4) for any $c \geq 0$. In every step of the method, the inner problem of (14) in the primal variable φ is solved and the dual variable μ is updated by some strategy, e.g. by a single step of the gradient method applied to the outer problem of (14) as in [IK08]. Moreover, an update rule for the penalty parameter c can be employed. In [BI12] global convergence is shown for an augmented Lagrangian method applied to a general convex optimization problem in a Banach space with pointwise inequality constraints in L^p (i.e. $H = L^p$ and $K = \{f \in L^p \mid f \leq 0 \text{ a.e.}\}$). Therefor it is shown that the augmented Lagrangian method coincides with the proximal point method applied to the unaugmented dual problem

$$0 \in \partial_\mu(-\inf_{\varphi \in X} \mathcal{L}(\varphi, \mu) + \chi_{K^-}(\mu)), \tag{15}$$

where χ_{K^-} is the indicator function of K^-,

$$\chi_{K^-}(\varphi) = \begin{cases} 0 & \varphi \in K^- \\ \infty & \varphi \notin K^- \end{cases},$$

and ∂_μ denotes the subdifferential with respect to μ (see [ET99] or Section 4.10). However, the inner primal subproblem has to be solved exactly to obtain convergence. Global convergence for an inexact version of the method (again only for convex problems) is studied in [BS00, IO01], where it is shown that the method is equivalent to a proximal point method applied to the saddle point problem

$$0 \in \begin{pmatrix} D_\varphi \mathcal{L}(\varphi, \mu) \\ \partial_\mu(-\mathcal{L}(\varphi, \mu) + \chi_{K^-}(\mu)) \end{pmatrix}.$$

An augmented Lagrangian SQP method for nonlinear optimal control problems in Banach spaces is analyzed in [ART02]. The augmentation is performed in the nonlinearity of the state constraint. It is shown that the method is a Josephy-Newton method applied to an augmented optimality system and thus local q-quadratic convergence is obtained under the assumption of strong regularity (Def. 3.1). Numerical results show that the augmented method performs better than the usual SQP method.
A weakness of augmented Lagrangian methods even in finite dimension is the possible convergence to infeasible points or to nonoptimal degenerate points [ISU12].

If Φ_{ad} is convex and weakly compact the **conditional gradient method** can be used to solve the optimization problem (4) [DR70]. In the kth step the subproblem

$$\min_{y \in \Phi_{ad}} \langle j'(\varphi_k), y - \varphi_k \rangle$$

is solved, where the original problem is replaced by a first order approximation. Let y_k be a minimizer of the subproblem. Then φ_{k+1} is determined by a line search along $\alpha \mapsto \varphi_k + \alpha(y_k - \varphi_k)$. In [DR70] global convergence is shown in case of an exact line search. However, typically the convergence rate is worse than for other methods such as projected gradient methods in finite dimension [Ber99].

If the optimization problem (4) cannot be solved directly by a numerical method then there is the possibility to **reformulate** the problem or to **approximate** the problem by a sequence of optimization problems which are easier to solve. For instance penalty and barrier methods approximate (4) by a sequence of unconstrained optimization problems. A Moreau-Yosida regularization can be used to approximate the indicator function $\chi_{\Phi_{ad}}$ for some convex Φ_{ad} by a sequence of Lipschitz continuously differentiable functions [IK08], which can be used to approximate (4) by a sequence of unconstrained problems. Also cone constraints of the form $G(\varphi) \in K$ can be handled by a Moreau-Yosida approximation by considering a regularization of $\chi_K(G(\varphi))$. This approach is often used to cope with state constraints in optimal control problems, see [HPUU08] and the references therein. Moreover, duality techniques can be employed to reformulate the problem. For instance in [CK11] the authors solve instead of the original optimization problem, which is posed in a measure space or in the space of functions with bounded variation (see Def. 6.21), its predual problem which is posed in a Hilbert space. Thus optimization methods in Hilbert spaces can be used to solve the predual problem, from which the solution of the original problem can be recovered. Another approximation approach is 'discretize-then-optimize', where the optimization problem is approximated by a sequence of finite dimensional problems, for which efficient solvers are available. However, mesh dependent behavior of the method can be an issue in this case. A further example is discussed in Section 6.4, where a topology optimization problem posed in the space of functions with bounded variation having discrete values is approximated by a sequence of regularized smooth optimization

problems posed in $H^1 \cap L^\infty$ in the sense of Γ-convergence.

There are some **specialized methods** available, in particular for optimal control problems and/or pointwise constraints in some L^p space. For instance trust-region methods for box-constrained optimal control problems can be found in [KS99, HUU99]. This method is also is used in [UU00] as a globalization of an affine-scaling interior-point Newton method for pointwise constraints in L^p. Interior-point methods are available for optimal control problems with pointwise box constraints in L^p, see [UU09] and the references therein. A special class of problems which are studied recently are optimal control problems with state constraints. These can be tackled by approximation methods such as Lavrentiev regularization, primal-dual path-following or barrier methods, see [Sch09] and the references therein. Of course there are more specialized methods available, which we do not mention here.

We can summarize that only few globally convergent methods for general constrained optimization problems in Banach spaces are available without assuming convexity of the problem or assuming a special structure such as pointwise constraints in L^p.

4 A new variable metric projection type (VMPT) method

4.1 Derivation as generalization of the scaled projected gradient method

The classical projected gradient method is a numerical method regarding the minimization of some nonlinear cost functional j within a convex admissible set Φ_{ad}, being a closed subset of some Hilbert space $\mathbb{H}$. The method makes use of the orthogonal projection $P_\perp(u)$ of some vector $u \in \mathbb{H}$ onto Φ_{ad}, as well as the gradient $\nabla_{\mathbb{H}} j(\varphi)$. Recall that the gradient $\nabla_{\mathbb{H}} j(\varphi) \in \mathbb{H}$ is defined as the Riesz representative of the derivative $j'(\varphi) \in \mathbb{H}^*$, i.e. it is characterized by the equation

$$(\nabla_{\mathbb{H}} j(\varphi), u)_{\mathbb{H}} = j'(\varphi)u \quad \forall u \in \mathbb{H}.$$

The fundamental iteration of the classical projected gradient method is given by the update

$$\varphi_{k+1} = P_\perp(\varphi_k - \nabla_{\mathbb{H}} j(\varphi_k)). \tag{16}$$

If the optimization problem is unconstrained, i.e. $\Phi_{ad} = \mathbb{H}$, then we have $P_\perp = \mathrm{id}$ and thus the projected gradient method becomes the usual gradient method for unconstrained optimization. As in the unconstrained case, care has to be taken to globalize the method. Here we consider the two possibilities as discussed in the introduction: The first is a line search along the descent direction given by

$$v_k := P_\perp(\varphi_k - \nabla_{\mathbb{H}} j(\varphi_k)) - \varphi_k, \tag{17}$$

i.e. find a step length $\alpha_k \in (0, 1]$ fulfilling some step length rule and set

$$\varphi_{k+1} = \varphi_k + \alpha_k v_k.$$

The second is a curved search along the projection arc, i.e. find $\lambda_k > 0$ fulfilling some step length rule and set

$$\varphi_{k+1} = P_\perp(\varphi_k - \lambda_k \nabla_{\mathbb{H}} j(\varphi_k)).$$

The curve $\lambda \mapsto P_\perp(\varphi_k - \lambda \nabla_{\mathbb{H}} j(\varphi_k))$ is called the projection arc.
For both globalization methods multiple step length rules are available. Amongst others there are the exact step length rule, Goldstein's step length rule, Powell's step length rule or Armijo backtracking [GS81, Ber99, HPUU08]. Each of them guarantees that $(j(\varphi_k))_k$ is a monotonically decreasing sequence. Also nonmonotone methods are available to possibly enhance convergence, see e.g. [BMR00].
In any case the orthogonal projection of some vector $\varphi_k - \lambda_k \nabla_{\mathbb{H}} j(\varphi_k)$ onto Φ_{ad} has to be calculated. The projection y is given as point with minimal distance, i.e. $y = P_\perp(\varphi_k - \lambda_k \nabla_{\mathbb{H}} j(\varphi_k))$ is the solution of the optimization problem

$$\min_{y \in \Phi_{ad}} \|y - (\varphi_k - \lambda_k \nabla_{\mathbb{H}} j(\varphi_k))\|_{\mathbb{H}}.$$

Equivalently, the following expression can be minimized:

$$\frac{1}{2}\|y - (\varphi_k - \lambda_k \nabla_{\mathbb{H}} j(\varphi_k))\|_{\mathbb{H}}^2 = \frac{1}{2}\|y - \varphi_k\|_{\mathbb{H}}^2 + \lambda_k (y - \varphi_k, \nabla_{\mathbb{H}} j(\varphi_k))_{\mathbb{H}} + \frac{1}{2}\|\lambda_k \nabla_{\mathbb{H}} j(\varphi_k))\|_{\mathbb{H}}^2.$$

The last term is a constant independent of y and can therefore be dropped. The second term can be reformulated using the characterization of the gradient, i.e. $(y - \varphi_k, \nabla_{\mathbb{H}} j(\varphi_k))_{\mathbb{H}} =$

$\langle j'(\varphi_k), y - \varphi_k \rangle$. For this calculation it is important that the gradient and the projection are taken with respect to the same inner product (here $(.,.)_{\mathbb{H}}$). We conclude that the projection is the solution of the optimization problem

$$\min_{y \in \Phi_{ad}} \frac{1}{2} \|y - \varphi_k\|_{\mathbb{H}}^2 + \lambda_k \left\langle j'(\varphi_k), y - \varphi_k \right\rangle.$$

In this formulation one realizes that only the directional derivative $\langle j'(\varphi_k), y - \varphi_k \rangle$ instead of the gradient appears. Thus, the computation of the gradient $\nabla_{\mathbb{H}} j(\varphi_k)$ can be circumvented in this way.
For a more general method we allow the inner product to change in every iteration, i.e. we use the inner product $a_k(.,.)$ in the kth iteration instead of $(.,.)_{\mathbb{H}}$ for the gradient and the projection. The resulting method is called scaled gradient projection in [Ber99]. The same calculation as above yields that the projection is given as solution of the problem

$$\min_{y \in \Phi_{ad}} \frac{1}{2} \|y - \varphi_k\|_{a_k}^2 + \lambda_k \left\langle j'(\varphi_k), y - \varphi_k \right\rangle, \tag{18}$$

where we define as usual $\|\varphi\|_{a_k} := \sqrt{a_k(\varphi, \varphi)}$. Note that $(\mathbb{H}, a_k)$ has to be a Hilbert space for any k in order to guarantee the existence of an orthogonal projection and a gradient (i.e. of the Riesz isomorphism). In this case it is trivial that the problem (18) is uniquely solvable.

The subproblem (18) is the starting point for the VMPT method. We will now leave the Hilbert space setting and move to a more general Banach space setting, i.e. the whole optimization problem is posed in a Banach space, the cost functional j is differentiable in a Banach space and $(a_k)_k$ is a sequence of inner products on this Banach space (for the precise assumptions we refer to the next section). Although inner products are still used, we do not demand that the Banach space is complete with respect to the a_k-norms, thus we will work in pre-Hilbert spaces rather than in Hilbert spaces. Note that in general pre-Hilbert spaces neither the orthogonal projection with respect to a_k, nor the gradient with respect to a_k has to exist. However, there is still a chance that the subproblem (18) is well posed. Since (18) is derived from a projection, we refer to it as 'projection type subproblem'. The subproblem (leaving away the index k) is parametrized by the metric a, the scaling factor λ and the current iterate φ. Therefore we denote its solution by $\mathcal{P}_{a,\lambda}(\varphi)$. To simplify notation in certain places we also write $\mathcal{P}_k(\varphi) := \mathcal{P}_{a_k,\lambda_k}(\varphi)$.

Although orthogonal projections, gradients and the Riesz isomorphism do not exist in a general normed space X, there are similar concepts available, which we review briefly. First of all there exists a generalized projection $\pi : X^* \to X$ in uniformly convex and uniformly smooth Banach spaces [Alb96]. Formally, the projection type subproblem can be seen as generalized projection of some functional in X^*, as we will discuss in Section 4.9. However, we use much weaker assumptions. In particular, the normed space we use for the generalized projection is not even assumed to be complete.
A generalization of the (negative) gradient in Hilbert spaces is e.g. the anti-gradient defined in [KA64] as a vector v minimizing $\langle j'(\varphi), v \rangle / \|v\|$. Existence or uniqueness of such an anti-gradient is not given in general normed spaces. A similar concept is given in [Dun09], where a solution v of the equations $\|v\|_X = \|f'(x)\|_{X^*}$, $f'(x)v = \|f'(x)\|_{X^*} \|v\|_X$ in a normed space X is called gradient vector at x. Again existence or uniqueness is not given in general. However, if the space X is reflexive, at least existence can be shown. A relaxed notion is the ν-approximate gradient v for $\nu \in (0,1)$, defined as a solution of

$\|v\|_X = \|f'(x)\|_{X^*}$, $f'(x)v \geq (1-\nu)\|f'(x)\|_{X^*}\|v\|_X$, which is guaranteed to exist in any normed space. The definition is closely connected to gradient related search directions in finite dimension in the sense that the angle between $\nabla j(x)$ and v is bounded away from 90° [NW06]. In this thesis we will not assume the existence of a gradient. However, we will show the existence and uniqueness of a *projected* gradient in the sense of the projection type subproblem (18).
A generalization of the Riesz isomorphism to normed spaces is the p-duality map $J_p := \partial\|.\|_X^p/p : X \rightrightarrows X^*$, where ∂ denotes the subdifferential [Sho97]. In general J_p is set-valued. If X is a Hilbert space then J_2 is the usual Riesz isomorphism. It turns out that v is a gradient vector in the sense of [Dun09] if and only if $j'(x) \in J_2(v)$. The p-duality map is used in certain numerical methods (e.g. [SLS06, OI07, Bre09]), but we don't need it here.

The VMPT method formally coincides with the scaled projected gradient method (using the variable metric a_k), where the basic iteration (16) is replaced by

$$\varphi_{k+1} = y_k,$$

with y_k being the solution of the subproblem (18). Still a globalization technique is needed to ensure convergence of the method. We will consider here only Armijo backtracking, because it is easy to implement and very popular in today's literature. For the theoretical results we use Armijo backtracking along the search direction v_k. We emphasize that Armijo backtracking along the projection arc is not possible under the assumptions stated in the next section, since the cost functional is in general not continuous along the projection arc. However, we will present a workaround in terms of a hybrid method in Section 4.6.

The VMPT method for solving the optimization problem

$$\min_{\varphi \in \Phi_{ad}} j(\varphi) \tag{19}$$

is summarized in Algorithm 4.1. The search direction v_k defined in line 5 is analogous to (17) in the projected gradient method. The Armijo rule we use in (20) is the same as in [Ber99] in finite dimension. The $\|.\|_{\mathbb{X}}$-norm appearing in the stopping criterion in line 6 will be introduced in the next section. Also the motivation of the stopping criterion is given in the next section, see Remark 4.15.

We note that the scaling parameter λ_k can be eliminated by dividing the cost functional of the projection type subproblem (18) by λ_k and defining the new inner product $\tilde{a}_k := a_k/\lambda_k$, i.e. we have $\mathcal{P}_{a,\lambda} \equiv \mathcal{P}_{a/\lambda,1}$. However, it is convenient to have λ_k as an independent parameter, amongst others to define the curved search without changing the inner product a_k.

4.2 Global convergence

In this section we formulate the precise assumptions used for the global convergence proof. Recall that global convergence means convergence independent of the choice of the initial guess. As typical for such type of methods, convergence has to be understood in the sense that each accumulation point of the iterates generated by the method is a stationary point of the optimization problem, and that a first order necessary optimality condition is fulfilled in the limit $k \to \infty$, respectively. As usual for the analysis of iterative methods, we

Algorithm 4.1 VMPT method with line search

1: Choose $\varphi_0 \in \Phi_{ad}$, $0 < \beta < 1$ and $0 < \sigma < 1$
2: $k := 0$
3: **while** $k \leq k_{\max}$ **do**
4: Calculate the minimum $y_k := \mathcal{P}_k(\varphi_k)$ of the subproblem (18).
5: Set $v_k := y_k - \varphi_k$
6: **if** $\|v_k\|_{\mathbb{X}} < \text{tol}$ **then**
7: **return**
8: **end if**
9: Calculate the step length $0 < \alpha_k \leq 1$ by Armijo backtracking direction in v_k, i.e. find the minimal power $m_k \in \mathbb{N}_0$ such that $\alpha_k := \beta^{m_k}$ fulfills

$$j(\varphi_k + \alpha_k v_k) \leq j(\varphi_k) + \alpha_k \sigma \left\langle j'(\varphi_k), v_k \right\rangle. \tag{20}$$

10: Update $\varphi_{k+1} := \varphi_k + \alpha_k v_k$
11: $k := k + 1$
12: **end while**

ignore the stopping criterion in line 6 of Algorithm 4.1, i.e. we assume that the VMPT method generates an infinite sequence. Moreover, we assume that no iterate of the VMPT method is a stationary point of j. Otherwise, the iterates stay constant after finitely many steps and the main statements of this section become trivial.

We assume that the admissible set Φ_{ad} is a subset of the intersection of two normed spaces $\mathbb{X}$ and $\mathbb{D}$. Since we work with two different norms, $\|.\|_{\mathbb{X}}$ and $\|.\|_{\mathbb{D}}$, we are able to formulate the assumptions in a weak way. For instance we assume the differentiability of the cost functional with respect to the strong norm $\|.\|_{\mathbb{X}\cap\mathbb{D}} := \|.\|_{\mathbb{X}} + \|.\|_{\mathbb{D}}$, whereas the coercivity of the inner products is claimed only with respect to the weaker norm $\|.\|_{\mathbb{X}}$, cf. **(A5)** and **(A9)** below. The precise assumptions on the spaces $\mathbb{X}$ and $\mathbb{D}$ are as follows.

(A1) $\mathbb{X}$ is a real reflexive Banach space. $\mathbb{B}$ is a separable real Banach space and $\mathbb{D}$ is a real Banach space which is isometrically isomorphic to $\mathbb{B}^*$. Moreover, for each sequence $(\varphi_i)_i \subset \mathbb{X} \cap \mathbb{D}$ with $\varphi_i \to \varphi$ weakly in $\mathbb{X}$ and $\varphi_i \to \overline{\varphi}$ weakly-* in $\mathbb{D}$ for some $\varphi \in \mathbb{X}$, $\overline{\varphi} \in \mathbb{D}$, it holds $\varphi = \overline{\varphi}$.

In order for the intersection $\mathbb{X} \cap \mathbb{D}$ to make sense we assume that there exists a common superspace of $\mathbb{X}$ and $\mathbb{D}$. In the case of function spaces this can be e.g. the space of measurable functions, or the space of distributions.
We identify $\mathbb{D}$ with $\mathbb{B}^*$ and therefore we say a sequence converges weakly-* in $\mathbb{D}$ if it converges weakly-* in $\mathbb{B}^*$. The separability of $\mathbb{B}$ is needed to get (sequential) weak-* compactness in $\mathbb{D}$ by the Banach-Alaoglu theorem. We note that it is also possible that $\mathbb{D}$ is a reflexive Banach space instead of some dual space. In this case weak-* convergence in $\mathbb{D}$ has to be replaced by weak convergence everywhere, see Theorem 4.18. However, we have the space $\mathbb{D} = L^\infty(\Omega) \cong (L^1(\Omega))^*$ in mind for applications, where Ω is some σ-finite measure space. The benefit of the space L^∞ is that e.g. for optimal control problems it is often much easier to show the differentiability of the (reduced) cost functional with respect to L^∞ than with respect to some other norm like L^2, see e.g. [KS92]. Also the analysis of Nemytskii operators (superposition operators) appearing often in nonlinear PDEs is much easier in L^∞ than in other L^p spaces, see [Trö09]. In some cases regularity theory for the state equation can be skipped if the analysis is performed for controls in $L^\infty(\Omega)$.

Therefore, no assumptions on the space dimension and the smoothness of $\partial\Omega$ are needed, cf. Remark 6.14.
An example for spaces fulfilling the assumption **(A1)** is $\mathbb{X} = W^{k,p}(\Omega)$ for $k \in \mathbb{N}_0$, $1 < p < \infty$ and $\mathbb{D} = L^q(\Omega)$ for $1 < q \leq \infty$, where $\Omega \subset \mathbb{R}^d$ is a bounded domain.

In addition let the following assumptions on the problem hold:

(A2) $\Phi_{ad} \subset \mathbb{X} \cap \mathbb{D}$ is convex and non-empty.

(A3) Φ_{ad} is closed in $\mathbb{X}$.

(A4) Φ_{ad} is bounded in $\mathbb{D}$.

(A5) j is continuously differentiable in a neighborhood of $\Phi_{ad} \subset \mathbb{X} \cap \mathbb{D}$.

(A6) $j(\varphi) \geq -C$ for for some $C > 0$ and all $\varphi \in \Phi_{ad}$.

(A7) For each $\varphi \in \Phi_{ad}$ the derivative $j'(\varphi)$ has the following continuity: For each sequence $(\varphi_i)_i \subset \mathbb{X} \cap \mathbb{D}$ with $\varphi_i \to 0$ weakly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$ it holds $\langle j'(\varphi), \varphi_i \rangle \to 0$.

For the parameters of the VMPT method we assume the following properties:

(A8) $(a_k)_k$ is a sequence of inner products on $\mathbb{X} \cap \mathbb{D}$.

(A9) There exists $C > 0$, s.t. $C\|u\|^2_{\mathbb{X}} \leq a_k(u,u)$ for all $u \in \mathbb{X} \cap \mathbb{D}$ and $k \in \mathbb{N}_0$.

(A10) a_k is (not necessarily uniformly) bounded in $\mathbb{X} \cap \mathbb{D}$, i.e. for all $k \in \mathbb{N}_0$ there exists $C_k > 0$ such that $a_k(p,v) \leq C_k \|p\|_{\mathbb{X}\cap\mathbb{D}} \|v\|_{\mathbb{X}\cap\mathbb{D}}$ for all $p, v \in \mathbb{X} \cap \mathbb{D}$.

(A11) For each $k \in \mathbb{N}_0$, $v \in \Phi_{ad}$ and for each sequence $(p_i)_i \subset \mathbb{X} \cap \mathbb{D}$ with $p_i \to 0$ weakly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$ it holds $a_k(v, p_i) \to 0$ as $i \to \infty$.

(A12) Let $(\varphi_k)_k$ be the sequence of iterates generated by Algorithm 4.1. For any subsequence with $\varphi_{k_i} \to \varphi$ in $\mathbb{X} \cap \mathbb{D}$ for some $\varphi \in \mathbb{X} \cap \mathbb{D}$, and for any sequences $(v_i)_i, (p_i)_i \subset \mathbb{X} \cap \mathbb{D}$ with $v_i \to 0$ strongly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$ and $p_i \to p$ in $\mathbb{X} \cap \mathbb{D}$ for some $p \in \mathbb{X} \cap \mathbb{D}$ it holds for the corresponding subsequence of inner products that $a_{k_i}(p_i, v_i) \to 0$ as $i \to \infty$.

(A13) There exist $\lambda_{min}, \lambda_{max}$, s.t. $0 < \lambda_{min} \leq \lambda_k \leq \lambda_{max}$ for all $k \in \mathbb{N}_0$.

We call **(A1)**-**(A13)** standard assumptions, which are assumed for the rest of Section 4 if not stated otherwise.

We emphasize that the differentiability of j in **(A5)** is assumed with respect to the $\mathbb{X} \cap \mathbb{D}$-norm. We also note that **(A5)** implies Fréchet differentiability of j with respect to that norm.
Note that the above assumptions do not imply the existence of a minimizer for the optimization problem (19).

Remark 4.1. We say that some inner product a (resp. $\lambda \in \mathbb{R}$) fulfills the assumptions, if the constant sequence $a_k = a$ (resp. $\lambda_k = \lambda$) for all $k \geq 0$ fulfills the assumptions. This will be helpful for statements with fixed k. Note that the assumptions on λ reduces to the requirement $\lambda > 0$.

Remark 4.2. The assumptions are stated as weak as possible. In Section 4.3 and Section 4.4 we give sufficient conditions for the assumptions to be fulfilled, which may be checked easier than the abstract assumptions for the concrete problems. For instance **(A12)** is fulfilled if a_k is uniformly bounded or if it depends continuously on φ_k. However, in Section 6 the given problem and the used variable metrics only fulfill the weak assumptions stated above.

Remark 4.3. Under assumption **(A8)**, the assumption **(A10)** is equivalent to

$$\forall k \in \mathbb{N}_0 \ \exists C_k > 0 : \ |a_k(p,p)| \leq C_k \|p\|^2_{\mathbb{X}\cap\mathbb{D}} \quad \forall p \in \mathbb{X}\cap\mathbb{D},$$

since the Cauchy Schwarz inequality holds for a_k, thus $|a_k(p,v)| \leq \sqrt{a_k(p,p)}\sqrt{a_k(v,v)} \leq C_k\|p\|_{\mathbb{X}\cap\mathbb{D}}\|v\|_{\mathbb{X}\cap\mathbb{D}}$.

In the literature concerning variable metric methods in a Hilbert space $\mathbb{H}$ or in $\mathbb{R}^n$ it is typically assumed that j is continuously differentiable in $\mathbb{H}$ and that it holds

$$\exists C, c > 0 : \quad c\|p\|^2_{\mathbb{H}} \leq a_k(p,p) \leq C\|p\|^2_{\mathbb{H}} \quad \forall p \in \mathbb{H}, \tag{21}$$

see e.g. [Gol65, Han77, GS81, GB84, Dun87, GD88, Ber99, Kel99]. We weaken these conditions in the following way. In our setting the space $\mathbb{X}$ plays the role of $\mathbb{H}$, thus we replace the assumption of a Hilbert space by a more general reflexive Banach space. Moreover, the differentiability of j in $\mathbb{H}$ is relaxed to the differentiability with respect to the stronger $\mathbb{X}\cap\mathbb{D}$-norm. Finally, we also relax the boundedness of a_k in $\mathbb{H}$ to the boundedness of a_k in the stronger $\mathbb{X}\cap\mathbb{D}$-norm. This is natural, since one should be able to use $a_k = j''(\varphi_k)$, see Section 4.7, and $j''(\varphi_k)$ is only bounded in the $\mathbb{X}\cap\mathbb{D}$-norm. We also weaken the *uniform* boundedness of a_k by the assumptions **(A10)**-**(A12)**, cf. Lemma 4.19.
Although each a_k defines an inner product on $\mathbb{X}\cap\mathbb{D}$, the norm induced by a_k is in general not equivalent to the $\mathbb{X}\cap\mathbb{D}$-norm, nor equivalent to the $\mathbb{X}$-norm. If we assume (21), then the a_k-norm is equivalent to the $\mathbb{H}$-norm, and thus $(\mathbb{H}, a_k)$ is a Hilbert space. This is not the case in our generalization. Also the different inner products a_k, $k = 0, 1, \ldots$, don't have to be equivalent in our formulation.
In the existing literature the boundedness condition **(A4)** is usually not required. However, it is crucial in our setting and cannot be dropped, not even to show the existence of a solution of the projection type subproblem. This is a main difference in the assumptions for the VMPT method. Because of the boundedness condition **(A4)** the VMPT method cannot be applied to unconstrained problems, i.e. to the case $\Phi_{ad} = \mathbb{X}\cap\mathbb{D}$. Note that Φ_{ad} is assumed to be bounded only in $\mathbb{D}$ and can therefore be unbounded in $\mathbb{X}$.

In the following we show well posedness of the projection type subproblem, as well as the global convergence result. We start by some auxiliary lemmas.

Lemma 4.4. *Let $(\varphi_k)_k \subset \Phi_{ad}$ be a sequence with $\varphi_k \to \varphi$ in $\mathbb{X}\cap\mathbb{D}$ for some $\varphi \in \mathbb{X}\cap\mathbb{D}$ and $(p_k)_k \subset \mathbb{X}\cap\mathbb{D}$ with $p_k \to p$ weakly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$ for some $p \in \mathbb{X}\cap\mathbb{D}$. Then $\langle j'(\varphi_k), p_k\rangle \to \langle j'(\varphi), p\rangle$.*

Proof. We observe that $\varphi \in \Phi_{ad}$ because of **(A3)**. We use **(A5)** and **(A7)** to obtain

$$\begin{aligned} &|\left\langle j'(\varphi_k), p_k\right\rangle - \left\langle j'(\varphi), p\right\rangle| \leq |\left\langle j'(\varphi_k), p_k\right\rangle - \left\langle j'(\varphi), p_k\right\rangle| + |\left\langle j'(\varphi), p_k\right\rangle - \left\langle j'(\varphi), p\right\rangle| \leq \\ &\leq \underbrace{\|j'(\varphi_k) - j'(\varphi)\|_{(\mathbb{X}\cap\mathbb{D})^*}}_{\to 0} \underbrace{\|p_k\|_{\mathbb{X}\cap\mathbb{D}}}_{\leq C} + \underbrace{|\left\langle j'(\varphi), p_k - p\right\rangle|}_{\to 0} \to 0. \end{aligned}$$

□

Lemma 4.5. *Let $(p_k)_k \subset \Phi_{ad}$ with $p_k \to p$ weakly in $\mathbb{X}$ for some $p \in \Phi_{ad}$. Then $p_k \to p$ weakly-* in $\mathbb{D}$.*

Proof. We show that given an arbitrary subsequence of p_k we can extract another subsequence, which converges to p weakly-* in $\mathbb{D}$. The claim then follows from Lemma 7.3.

Let us denote the arbitrary subsequence of p_k again by p_k. Due to **(A4)**, p_k is uniformly bounded in $\mathbb{D}$ and by virtue of the Banach-Alaoglu theorem and the separability of $\mathbb{B}$ we can extract a subsequence (denoted by p_k) with $p_k \to \tilde{p}$ weakly-* in $\mathbb{D}$ for some $\tilde{p} \in \mathbb{D}$. Due to **(A1)** we have $\tilde{p} = p$. □

We are now able to show the well posedness of the projection type subproblem.

Lemma 4.6. *Let a and λ fulfill the assumptions (in the sense of Remark 4.1). Then the operator $\mathcal{P}_{a,\lambda} : \Phi_{ad} \to \Phi_{ad}$ is well defined, i.e. the corresponding projection type subproblem (see (18)) with $\varphi \in \Phi_{ad}$ is uniquely solvable. Moreover, $y = \mathcal{P}_{a,\lambda}(\varphi)$ is given as unique solution of the variational inequality*

$$y \in \Phi_{ad}, \quad a(y-\varphi, \eta - y) + \lambda \left\langle j'(\varphi), \eta - y \right\rangle \geq 0 \quad \forall \eta \in \Phi_{ad}. \tag{22}$$

Proof. Let $\varphi \in \Phi_{ad}$ be arbitrary. We show the existence and uniqueness of $\mathcal{P}_{a,\lambda}(\varphi)$. Problem (18), leaving away the index k, is equivalent to the problem

$$\min_{y \in \Phi_{ad}} \quad g(y) = \frac{1}{2} a(y,y) + \langle b, y \rangle \tag{23}$$

with $\langle b, y \rangle := \lambda \langle j'(\varphi), y \rangle - a(\varphi, y)$. From **(A5)** and **(A10)** we get $b \in (\mathbb{X} \cap \mathbb{D})^*$.
We show the existence of minimizers by the direct method in the calculus of variations. First we show that g is bounded from below. Let $y \in \Phi_{ad}$. By **(A9)** and **(A4)** we get

$$\begin{aligned} g(y) &\geq C\|y\|_{\mathbb{X}}^2 - \|b\|_{(\mathbb{X}\cap\mathbb{D})^*}\|y\|_{\mathbb{X}\cap\mathbb{D}} = C\|y\|_{\mathbb{X}}^2 - \tilde{C}(\|y\|_{\mathbb{X}} + \underbrace{\|y\|_{\mathbb{D}}}_{\leq C}) \geq \\ &\geq C\|y\|_{\mathbb{X}}^2 - \tilde{C}\|y\|_{\mathbb{X}} - \tilde{C} > -C \end{aligned} \tag{24}$$

where $C > 0$ and $\tilde{C} > 0$ are generic constants.
From this we conclude that $\inf_{y \in \Phi_{ad}} g(y) > -\infty$ and we can choose an infimizing sequence $y_i \in \Phi_{ad}$, such that $g(y_i) \to \inf_{y \in \Phi_{ad}} g(y)$. From the estimate (24) we conclude that y_i is bounded in $\mathbb{X}$. Therefore we can extract a subsequence (still denoted by y_i) which converges weakly in $\mathbb{X}$ to some $y^* \in \mathbb{X}$. Since Φ_{ad} is convex and closed in $\mathbb{X}$, it is also weakly sequentially closed in $\mathbb{X}$ and thus $y^* \in \Phi_{ad}$. By Lemma 4.5 we get $y_i \to y^*$ weakly-* in $\mathbb{D}$. Until now we showed:

$$\begin{aligned} &g(y_i) \to \inf_{y \in \Phi_{ad}} g(y), \\ &\quad y_i \to y^* \text{ weakly in } \mathbb{X} \text{ and weakly-* in } \mathbb{D}, \\ &\quad y^* \in \Phi_{ad}. \end{aligned}$$

It remains to show $g(y^*) = \inf_{y \in \Phi_{ad}} g(y)$. From **(A7)** and **(A11)** we get $\langle b, y_i \rangle \to \langle b, y^* \rangle$. On the other hand, we get by **(A9)** and **(A11)** that

$$\liminf_i a(y_i, y_i) = \liminf_i (\underbrace{a(y_i - y^*, y_i - y^*)}_{\geq 0} + \underbrace{a(y^*, y_i - y^*)}_{\to 0} + \underbrace{a(y_i, y^*)}_{\to a(y^*, y^*)}) \geq a(y^*, y^*), \tag{25}$$

thus $\liminf_i g(y_i) \geq g(y^*)$. This yields

$$\inf_{y \in \Phi_{ad}} g(y) \leq g(y^*) \leq \liminf_i g(y_i) = \inf_{y \in \Phi_{ad}} g(y)$$

and we conclude $g(y^*) = \inf_{y \in \Phi_{ad}} g(y)$. To show the uniqueness of the minimizer, we note

that g is strictly convex, which can be seen by the estimate

$$g(tu+(1-t)v) = tg(u)+(1-t)g(v) - \underbrace{\frac{1}{2}t(1-t)a(u-v,u-v)}_{>0} < tg(u)+(1-t)g(v),$$

which holds for all $u, v \in \mathbb{X} \cap \mathbb{D}$, $u \neq v$ and $t \in (0,1)$, where we again used **(A9)**. Since the subproblem is convex, it is equivalent to the variational inequality

$$y \in \Phi_{ad}, \quad \langle g'(y), \eta - y \rangle \geq 0 \text{ for all } \eta \in \Phi_{ad},$$

which in turn corresponds to (22). □

For ease of reference we note the following special case.

Corollary 4.7. *Let $(\varphi_k)_k$ be the sequence generated by Algorithm 4.1. Then $y_k := \mathcal{P}_k(\varphi_k)$ is given as the unique solution of the variational inequality*

$$y_k \in \Phi_{ad}, \quad a_k(y_k - \varphi_k, \eta - y_k) + \lambda_k \langle j'(\varphi_k), \eta - y_k \rangle \geq 0 \quad \forall \eta \in \Phi_{ad}. \tag{26}$$

A vector $\varphi \in \Phi_{ad}$ is called a stationary point of j if

$$\langle j'(\varphi), \eta - \varphi \rangle \geq 0 \text{ for all } \eta \in \Phi_{ad}, \tag{27}$$

see e.g. [Trö09]. This is a first order necessary condition for a minimizer and it is sufficient if j is convex.

Lemma 4.8. *Let a and λ fulfill the assumptions (in the sense of Remark 4.1). Then $\varphi \in \Phi_{ad}$ is a stationary point of j if and only if $\mathcal{P}_{a,\lambda}(\varphi) = \varphi$.*

Proof. Let $\varphi \in \Phi_{ad}$. For arbitrary $\eta \in \Phi_{ad}$ the cost functional g defined in (23) fulfills

$$\begin{aligned}\langle g'(\varphi), \eta - \varphi \rangle &= a(\varphi, \eta - \varphi) + \lambda \langle j'(\varphi), \eta - \varphi \rangle - a(\varphi, \eta - \varphi) = \\ &= \lambda \langle j'(\varphi), \eta - \varphi \rangle,\end{aligned}$$

which proves that φ is a stationary point of g if and only if it is a stationary point of j. From the convexity of g (see the proof of Lemma 4.6) On the other hand we get that φ is a stationary point of g if and only if it is the minimum of g (see [Trö09]), i.e. if and only if $\mathcal{P}_{a,\lambda}(\varphi) = \varphi$. □

In particular we get that an iterate φ_k of the algorithm is a stationary point of j if and only if $v_k = \mathcal{P}_k(\varphi_k) - \varphi_k = 0$. In the classical Hilbert space case the lemma above corresponds to

$$\varphi \in \Phi_{ad} \text{ stationary} \iff \varphi = P_{\perp}(\varphi - \lambda \nabla_{\mathbb{H}} j(\varphi)).$$

In the unconstrained case we thus recover that φ is stationary if and only if $\nabla_{\mathbb{H}} j(\varphi) = 0$.

The following lemma yields a crucial inequality, which is typically obtained for projected gradient type methods [Ber99, GS81]. In particular the inequality guarantees that v_k is a descent direction.

Lemma 4.9. *Let $(\varphi_k)_k$ be the sequence generated by Algorithm 4.1 and let $v_k := \mathcal{P}_k(\varphi_k) - \varphi_k$ as in the algorithm. Then it holds*

$$\left\langle j'(\varphi_k), v_k \right\rangle \leq -\frac{C}{\lambda_{max}} \|v_k\|_{\mathbb{X}}^2 \leq 0 \tag{28}$$

for some $C > 0$.

Proof. Let $y_k := \mathcal{P}_k(\varphi_k)$. We test the corresponding variational inequality (26) by $\eta = \varphi_k \in \Phi_{ad}$ and use **(A9)** to obtain

$$\begin{aligned} 0 &\leq -a_k(y_k - \varphi_k, y_k - \varphi_k) + \lambda_k \left\langle j'(\varphi_k), \varphi_k - y_k \right\rangle \\ &\leq -C\|v_k\|_{\mathbb{X}}^2 - \lambda_k \left\langle j'(\varphi_k), v_k \right\rangle. \end{aligned}$$

Finally, **(A13)** yields

$$\left\langle j'(\varphi_k), v_k \right\rangle \leq -\frac{C}{\lambda_k} \|v_k\|_{\mathbb{X}}^2 \leq -\frac{C}{\lambda_{max}} \|v_k\|_{\mathbb{X}}^2.$$

□

Remark 4.10. From the previous lemma we get $\langle j'(\varphi_k), v_k \rangle < 0$ as long as $v_k \neq 0$, i.e. as long as φ_k is not stationary. Thus, the Armijo backtracking is well defined, i.e. one can always find a positive step length $\alpha_k = \beta^{m_k}$, such that the Armijo condition (20) is fulfilled. This can be shown as in the finite dimensional case, see e.g. [Ber99].

Lemma 4.11. *Let for a sequence $(\varphi_i)_i \subset \Phi_{ad}$ hold $\varphi_i \to \varphi$ in $\mathbb{X} \cap \mathbb{D}$ for some $\varphi \in \mathbb{X} \cap \mathbb{D}$. Then there exists $C > 0$, such that $\|\mathcal{P}_k(\varphi_i)\|_{\mathbb{X}\cap\mathbb{D}} \leq C$ for all $k, i \in \mathbb{N}_0$.*

Proof. Let $i, k \in \mathbb{N}_0$ be arbitrary. We convince ourselves that Lemma 4.9 still holds if φ_k is replaced by φ_i. From (28) we then get using **(A4)**

$$\begin{aligned} C\|\mathcal{P}_k(\varphi_i) - \varphi_i\|_{\mathbb{X}}^2 \leq \left\langle j'(\varphi_i), \varphi_i - \mathcal{P}_k(\varphi_i) \right\rangle &\leq \underbrace{\|j'(\varphi_i)\|_{(\mathbb{X}\cap\mathbb{D})^*}}_{\leq C} \|\mathcal{P}_k(\varphi_i) - \varphi_i\|_{\mathbb{X}\cap\mathbb{D}} \\ &\leq C(\|\mathcal{P}_k(\varphi_i) - \varphi_i\|_{\mathbb{X}} + 1). \end{aligned}$$

Thus, $\|\mathcal{P}_k(\varphi_i) - \varphi_i\|_{\mathbb{X}} \leq C$. Since φ_i is uniformly bounded in $\mathbb{X}$ we get $\|\mathcal{P}_k(\varphi_i)\|_{\mathbb{X}} \leq C$. Finally, the statement follows form **(A4)**. □

In the following we prove that the search directions v_k are gradient related, where we adapt the notion of gradient relation from the finite dimensional case covered in [Ber99].

Lemma 4.12. *Let $(\varphi_k)_k$ be the sequence generated by Algorithm 4.1. Then the corresponding search directions v_k are gradient related in the following sense: Let $\varphi_{k_i} \to \varphi$ in $\mathbb{X} \cap \mathbb{D}$ for a subsequence, where $\varphi \in \Phi_{ad}$ is not a stationary point of j. Then $(v_{k_i})_i$ is bounded in $\mathbb{X} \cap \mathbb{D}$ and $\limsup_i \langle j'(\varphi_{k_i}), v_{k_i} \rangle < 0$.*

Proof. Let $\varphi \in \Phi_{ad}$ be non-stationary and $\varphi_{k_i} \to \varphi$ in $\mathbb{X} \cap \mathbb{D}$ for a subsequence. Then from Lemma 4.11 we get that $v_{k_i} = \mathcal{P}_{k_i}(\varphi_{k_i}) - \varphi_{k_i}$ is bounded in $\mathbb{X} \cap \mathbb{D}$.
From (28) we get

$$\limsup \left\langle j'(\varphi_{k_i}), v_{k_i} \right\rangle \leq \limsup -\frac{C}{\lambda_{max}} \|v_{k_i}\|_{\mathbb{X}}^2 = -\frac{C}{\lambda_{max}} \liminf \|v_{k_i}\|_{\mathbb{X}}^2.$$

We prove $\liminf \|v_{k_i}\|_{\mathbb{X}} \neq 0$ by contradiction. Assume $\liminf \|v_{k_i}\|_{\mathbb{X}} = 0$. Then $v_{k_i} \to 0$ in $\mathbb{X}$ for a subsequence (we denote it again by v_{k_i}). Moreover, it holds $y_{k_i} := \mathcal{P}_{k_i}(\varphi_{k_i}) = v_{k_i} + \varphi_{k_i} \to \varphi$ in $\mathbb{X}$. By Lemma 4.5 we also get $y_{k_i} \to \varphi$ weakly-* in $\mathbb{D}$. The corresponding variational inequality (26) reads

$$a_{k_i}(y_{k_i} - \varphi_{k_i}, \eta - y_{k_i}) + \lambda_{k_i} \left\langle j'(\varphi_{k_i}), \eta - y_{k_i} \right\rangle \geq 0 \quad \forall \eta \in \Phi_{ad}. \tag{29}$$

We use the estimate **(A9)** to get

$$\begin{aligned} a_{k_i}(y_{k_i} - \varphi_{k_i}, \eta - y_{k_i}) &= -a_{k_i}(y_{k_i} - \varphi_{k_i}, y_{k_i} - \varphi_{k_i}) + a_{k_i}(y_{k_i} - \varphi_{k_i}, \eta - \varphi_{k_i}) \\ &\leq a_{k_i}(y_{k_i} - \varphi_{k_i}, \eta - \varphi_{k_i}). \end{aligned}$$

We plug this into (29) and divide by λ_{k_i}, which leads to

$$\frac{1}{\lambda_{k_i}} a_{k_i}(y_{k_i} - \varphi_{k_i}, \eta - \varphi_{k_i}) + \left\langle j'(\varphi_{k_i}), \eta - y_{k_i} \right\rangle \geq 0 \quad \forall \eta \in \Phi_{ad}.$$

Now we can pass to the limit. From Lemma 4.4 we get $\langle j'(\varphi_{k_i}), \eta - y_{k_i} \rangle \to \langle j'(\varphi), \eta - \varphi \rangle$. From **(A12)** we get $a_{k_i}(y_{k_i} - \varphi_{k_i}, \eta - \varphi_{k_i}) \to 0$. Since, by **(A13)**, $\frac{1}{\lambda_{k_i}} \leq \frac{1}{\lambda_{min}}$ is bounded we end up with

$$\left\langle j'(\varphi), \eta - \varphi \right\rangle \geq 0 \quad \forall \eta \in \Phi_{ad},$$

which shows that φ is stationary and which is a contradiction. □

Lemma 4.13. *Let the standard assumptions hold and let in addition j be convex. Then the following lower-semicontinuity holds.*
Let $(\varphi_k)_k \subset \Phi_{ad}$ be a sequence with $\varphi_k \to \varphi$ weakly in $\mathbb{X}$ for some $\varphi \in \mathbb{X} \cap \mathbb{D}$. Then

$$\liminf_{k \to \infty} j(\varphi_k) \geq j(\varphi).$$

Proof. Since Φ_{ad} is convex and closed in $\mathbb{X}$ we have $\varphi \in \Phi_{ad}$ and from Lemma 4.5 we get $\varphi_k \to \varphi$ weakly-* in $\mathbb{D}$. From the convexity of j we get

$$j(\varphi_k) \geq j(\varphi) + \left\langle j'(\varphi), \varphi_k - \varphi \right\rangle$$

and **(A7)** yields $\langle j'(\varphi), \varphi_k - \varphi \rangle \to 0$. Taking the lim inf of both sides proves the statement. □

Note that it is well known that continuous convex functionals on a Banach space are weakly lower-semicontinuous [ET99]. However, this is not applicable since j is neither continuous in $\mathbb{X}$, nor does it hold that $\varphi_k \to \varphi$ weakly in $\mathbb{X} \cap \mathbb{D}$, thus Lemma 4.13 is not trivial.

Now we are able to prove global convergence of the VMPT method.

Theorem 4.14. *Let $(\varphi_k)_k$ be the sequence generated by Algorithm 4.1. Then*

1. *$\lim_{k \to \infty} j(\varphi_k)$ exists,*
2. *every accumulation point of φ_k in $\mathbb{X} \cap \mathbb{D}$ is a stationary point of j,*
3. *for each subsequence with $\varphi_{k_i} \to \varphi$ in $\mathbb{X} \cap \mathbb{D}$ for some $\varphi \in \Phi_{ad}$, it holds $v_{k_i} \to 0$ in $\mathbb{X}$.*

4. *Let additionally* $j \in C^{1,\gamma}(\Phi_{ad})$ *for some* $0 < \gamma \leq 1$. *Then*

$$\left\langle j'(\varphi_k), v_k \right\rangle \to 0 \quad \textit{and} \quad v_k \to 0 \textit{ in } \mathbb{X}$$

(for the whole sequence).

5. *Let additionally* $j \in C^{1,\gamma}(\Phi_{ad})$ *for some* $0 < \gamma \leq 1$ *and let* j *be convex. Moreover, let* a_k *be uniformly bounded in the sense that there exists* $C > 0$ *such that*

$$|a_k(p,v)| \leq C\|p\|_{\mathbb{X}}\|v\|_{\mathbb{X}} \quad \forall k \in \mathbb{N}_0,\ p, v \in \mathbb{X} \cap \mathbb{D}. \tag{30}$$

Then not only strong accumulation points but even weak accumulation points are stationary in the following sense: Let there exist $\varphi \in \mathbb{X} \cap \mathbb{D}$ *such that* $\varphi_{k_i} \to \varphi$ *weakly in* $\mathbb{X}$ *for a subsequence. Then* φ *is a global minimizer of* j *in* Φ_{ad}.

Proof. We note that the idea of the proof for statement 1 and 2 is the same as in [Ber99] for the finite dimensional case.
1. From the Armijo condition (20) and since v_k is a descent direction (see (28)), we get

$$j(\varphi_{k+1}) - j(\varphi_k) \leq \alpha_k \sigma \left\langle j'(\varphi_k), v_k \right\rangle \leq 0, \tag{31}$$

thus $(j(\varphi_k))_k$ is monotonically decreasing. Since j is bounded from below, see **(A6)**, we get convergence $j(\varphi_k) \to j^*$ for some $j^* \in \mathbb{R}$, which proves 1.

2. The proof is by contradiction. Let φ_k be as in the theorem. We assume that there is a subsequence for which $\varphi_{k_i} \to \varphi$ holds in $\mathbb{X} \cap \mathbb{D}$ for some $\varphi \in \Phi_{ad}$. We assume that φ is non-stationary. Because the left hand side of (31) goes to zero, we get $\alpha_k \left\langle j'(\varphi_k), v_k \right\rangle \to 0$. By Lemma 4.12 we conclude $|\langle j'(\varphi_{k_i}), v_{k_i}\rangle| \geq C > 0$ and thus $\alpha_{k_i} \to 0$. Thus there exists some $\bar{i} \in \mathbb{N}$ such that $\alpha_{k_i}/\beta \leq 1$ for all $i \geq \bar{i}$, hence $\alpha_{k_i}/\beta = \beta^{m_k - 1}$ does not fulfill the Armijo condition (20) due to the minimality of m_k, i.e.

$$j\left(\varphi_{k_i} + \frac{\alpha_{k_i}}{\beta} v_{k_i}\right) - j(\varphi_{k_i}) > \frac{\alpha_{k_i}}{\beta} \sigma \left\langle j'(\varphi_{k_i}), v_{k_i} \right\rangle \text{ for all } i \geq \bar{i}. \tag{32}$$

We apply the mean value theorem to find some $0 \leq \overline{\alpha}_{k_i} \leq \alpha_{k_i}/\beta$ such that

$$\left\langle j'\left(\varphi_{k_i} + \overline{\alpha}_{k_i} v_{k_i}\right), v_{k_i} \right\rangle \frac{\alpha_{k_i}}{\beta} = j\left(\varphi_{k_i} + \frac{\alpha_{k_i}}{\beta} v_{k_i}\right) - j(\varphi_{k_i}).$$

Together with (32) and $\alpha_{k_i} > 0$ (see Remark 4.10), this yields

$$\left\langle j'\left(\varphi_{k_i} + \overline{\alpha}_{k_i} v_{k_i}\right), v_{k_i} \right\rangle > \sigma \left\langle j'(\varphi_{k_i}), v_{k_i} \right\rangle \text{ for all } i \geq \bar{i}. \tag{33}$$

We pass to the limit in the inequality. Therefor, we note that $\varphi_{k_i} + \overline{\alpha}_{k_i} v_{k_i} \to \varphi$ in $\mathbb{X} \cap \mathbb{D}$, since $\alpha_{k_i} \to 0$ and v_{k_i} is uniformly bounded in $\mathbb{X} \cap \mathbb{D}$ (Lemma 4.12). Moreover, we get that $y_{k_i} := \varphi_{k_i} + v_{k_i} \in \Phi_{ad}$ is uniformly bounded in $\mathbb{X}$ and thus we can extract a subsequence (denoted by y_{k_i}) with $y_{k_i} \to y$ weakly in $\mathbb{X}$ for some $y \in \Phi_{ad}$. Due to Lemma 4.5 we also get $y_{k_i} \to y$ weakly-* in $\mathbb{D}$. We conclude $v_{k_i} \to y - \varphi$ weakly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$. Due to Lemma 4.4 we can pass to the limit in (33) to obtain

$$(1 - \sigma) \left\langle j'(\varphi), y - \varphi \right\rangle \geq 0,$$

thus $\langle j'(\varphi), y-\varphi\rangle \geq 0$ (recall $\sigma \in (0,1)$). On the other hand, we get by Lemma 4.12

$$\left\langle j'(\varphi), y-\varphi\right\rangle = \limsup_i \left\langle j'(\varphi_{k_i}), v_{k_i}\right\rangle < 0,$$

which is a contradiction.

3. Let φ_{k_i} be a subsequence as in the assumption. We show $\langle j'(\varphi_{k_i}), v_{k_i}\rangle \to 0$, which proves together with (28) the statement.
We choose an arbitrary subsequence of $(\langle j'(\varphi_{k_i}), v_{k_i}\rangle)_i$, which we denote the same. From Lemma 4.11 we get that $y_{k_i} := \mathcal{P}_{k_i}(\varphi_{k_i})$ is bounded in $\mathbb{X}$. Thus we can extract a subsequence (denoted again by y_{k_i}), for which $y_{k_i} \to y$ weakly in $\mathbb{X}$ for some $y \in \Phi_{ad}$. From Lemma 4.5 we get also $y_{k_i} \to y$ weakly-* in $\mathbb{D}$. Hence we have that $v_{k_i} = y_{k_i} - \varphi_{k_i} \to y - \varphi$ weakly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$. We apply Lemma 4.4 to get $\langle j'(\varphi_{k_i}), v_{k_i}\rangle \to \langle j'(\varphi), y-\varphi\rangle$. Since φ is stationary we get $\langle j'(\varphi), y-\varphi\rangle \geq 0$. On the other hand we know from (28) that $\langle j'(\varphi_{k_i}), v_{k_i}\rangle \leq 0$ for all i and thus also $\langle j'(\varphi), y-\varphi\rangle \leq 0$. We conclude $\langle j'(\varphi), y-\varphi\rangle = 0$. Hence, from any subsequence of $\langle j'(\varphi_{k_i}), v_{k_i}\rangle$ we can choose another subsequence, which converges to 0 and thus $\langle j'(\varphi_{k_i}), v_{k_i}\rangle \to 0$, see Lemma 7.3.

4. We show $\langle j'(\varphi_k), v_k\rangle \to 0$. The second statement follows from (28). We choose an arbitrary subsequence of $(\langle j'(\varphi_k), v_k\rangle)_k$, which we denote the same. From (31) we get

$$\alpha_k \left\langle j'(\varphi_k), v_k\right\rangle \to 0. \tag{34}$$

If there exists some constant $C > 0$ such that $\alpha_k > C$ for all k we get $\langle j'(\varphi_k), v_k\rangle \to 0$ and we are finished. If this is not the case there exists a subsequence (again denoted by index k) such that $\alpha_k \to 0$ and $0 < \alpha_k < \beta$ for all k. As above we conclude that the step length α_k/β does not fulfill the Armijo condition (20), i.e.

$$j\left(\varphi_k + \frac{\alpha_k}{\beta} v_k\right) - j(\varphi_k) > \frac{\alpha_k}{\beta}\sigma \left\langle j'(\varphi_k), v_k\right\rangle. \tag{35}$$

Since $j \in C^{1,\gamma}(\Phi_{ad})$ we get by Lemma 7.2

$$j\left(\varphi_k + \frac{\alpha_k}{\beta} v_k\right) - j(\varphi_k) \leq \frac{\alpha_k}{\beta}\left\langle j'(\varphi_k), v_k\right\rangle + \frac{1}{1+\gamma} L \left(\frac{\alpha_k}{\beta}\right)^{1+\gamma} \|v_k\|^{1+\gamma}_{\mathbb{X}\cap\mathbb{D}}. \tag{36}$$

Because of the boundedness of Φ_{ad} in $\mathbb{D}$, see **(A4)**, we get

$$\|v_k\|^{1+\gamma}_{\mathbb{X}\cap\mathbb{D}} \leq C(\|v_k\|^{1+\gamma}_{\mathbb{X}} + \|v_k\|^{1+\gamma}_{\mathbb{D}}) \leq C(\|v_k\|^{1+\gamma}_{\mathbb{X}} + 1). \tag{37}$$

Putting (35), (36) and (37) together gives

$$0 < (\sigma - 1)\left\langle j'(\varphi_k), v_k\right\rangle < \frac{1}{1+\gamma} C \frac{\alpha_k^\gamma}{\beta^\gamma}(\|v_k\|^{1+\gamma}_{\mathbb{X}} + 1).$$

Estimate (28) yields

$$|\left\langle j'(\varphi_k), v_k\right\rangle| < C\alpha_k^\gamma (|\left\langle j'(\varphi_k), v_k\right\rangle|^{\frac{1+\gamma}{2}} + 1). \tag{38}$$

We get $x_k := |\langle j'(\varphi_k), v_k\rangle| \to 0$. Otherwise there exists a subsequence still denoted by x_k with $x_k \to c$ for some $c > 0$. Rearranging (38) gives $1 < C\alpha_k^\gamma (x_k^{\frac{-1+\gamma}{2}} + x_k^{-1}) \to 0$, which is a contradiction. By the same argument as in the proof of 3. we get $\langle j'(\varphi_k), v_k\rangle \to 0$ for the whole sequence.

5. Without loss of generality we assume $\varphi_k \to \varphi$ weakly in $\mathbb{X}$ for the whole sequence. We note that it holds $\varphi \in \Phi_{ad}$ because of **(A3)** and **(A2)**. Let $y_k := \mathcal{P}_k(\varphi_k)$. From the variational inequality (26) and (30) we get

$$\left\langle j'(\varphi_k), \eta - y_k \right\rangle \geq \frac{1}{\lambda_k} a_k(y_k - \varphi_k, y_k - \eta) \geq -C\|v_k\|_{\mathbb{X}} \|y_k - \eta\|_{\mathbb{X}}, \quad \forall \eta \in \Phi_{ad}$$

thus

$$\left\langle j'(\varphi_k), \eta - \varphi_k \right\rangle \geq \left\langle j'(\varphi_k), \eta - y_k \right\rangle + \left\langle j'(\varphi_k), y_k - \varphi_k \right\rangle \geq -C\|v_k\|_{\mathbb{X}} \|y_k - \eta\|_{\mathbb{X}} + \left\langle j'(\varphi_k), v_k \right\rangle$$

for all $\eta \in \Phi_{ad}$. Statement 4. of the theorem yields $\langle j'(\varphi_k), v_k \rangle \to 0$ and $\|v_k\|_{\mathbb{X}} \to 0$. Moreover, $\|y_k - \eta\|_{\mathbb{X}} = \|\varphi_k + v_k - \eta\|_{\mathbb{X}}$ is uniformly bounded since v_k and φ_k are uniformly bounded in $\mathbb{X}$. Thus,

$$\liminf_{k \to \infty} \left\langle j'(\varphi_k), \eta - \varphi_k \right\rangle \geq 0 \quad \forall \eta \in \Phi_{ad}. \tag{39}$$

From the convexity of j we get

$$j(\eta) \geq j(\varphi_k) + \left\langle j'(\varphi_k), \eta - \varphi_k \right\rangle \quad \forall \eta \in \Phi_{ad}. \tag{40}$$

From statement 1. of the theorem we get $j(\varphi_k) \searrow j^*$ for some $j^* \geq \inf_{\varphi \in \Phi_{ad}} j(\varphi)$. We take the lim inf of both sides of (40) and use (39) to obtain

$$j(\eta) \geq j^* \quad \forall \eta \in \Phi_{ad},$$

hence $j^* = \inf_{\varphi \in \Phi_{ad}} j(\varphi)$. From Lemma 4.13 and $\varphi \in \Phi_{ad}$ we finally get

$$j^* = \liminf_{k \to \infty} j(\varphi_k) \geq j(\varphi) \geq \inf_{\varphi \in \Phi_{ad}} j(\varphi) = j^*,$$

thus $j(\varphi) = \inf_{\varphi \in \Phi_{ad}} j(\varphi)$.

□

Remark 4.15. The result $v_k \to 0$ in statement 4. of Theorem 4.14 is similar to $\nabla j(\varphi_k) \to 0$ for unconstrained optimization methods, since we have $v_k = 0$ if and only if φ_k is stationary. In this sense the result states that a stationarity condition is satisfied in the limit. However, it is insightful that $v_k \to 0$ in the $\mathbb{X}$-norm and not in the $\mathbb{X} \cap \mathbb{D}$-norm. These arguments also motivate the stopping criterion $\|v_k\|_{\mathbb{X}} < tol$ in line 6 of Algorithm 4.1. The expression $\|v_k\|$ is called stationarity measure in [HPUU08, Kel99].

Remark 4.16. In the classical case $j \in C^1(\mathbb{H})$ for some Hilbert space $\mathbb{H}$ and $a_k = (.,.)_{\mathbb{H}}$ (i.e. no variable metric), the statement 4. of Theorem 4.14 is shown in [HPUU08] (for curved search instead of line search) under the same assumption $j \in C^{1,\gamma}$ and in [Ber76] assuming $j \in C^{1,1}$. Again in the classical case, statement 5. of Theorem 4.14 is shown in [GS81] under the same assumption that j is convex and $j \in C^{1,\gamma}$, using different step size rules.

Remark 4.17. We already discussed that the VMPT method relaxes the assumptions used by the classical projected gradient method. On the other hand, the global convergence result is tightened in the following sense. Assume that there is a Hilbert space $\mathbb{H} \hookrightarrow \mathbb{X} \cap \mathbb{D}$, for which the assumptions of the classical projected gradient method are fulfilled. Then the classical theory yields that every accumulation point in $\mathbb{H}$ is stationary. According to Theorem 4.14 even every $\mathbb{X} \cap \mathbb{D}$ accumulation point is stationary. Note that there can be

accumulation points in $\mathbb{X} \cap \mathbb{D}$ which are not accumulation points in $\mathbb{H}$. As an example, in [KU14] the embedding $H^2 \hookrightarrow W^{2-1/4,1/4} \hookrightarrow W^{1,\infty}$ in 1D is used to perform a gradient method in H^2. However, the regularized problem in [KU14] is unconstrained, thus the VMPT method does not apply.

According to **(A1)**, $\mathbb{D}$ is isometrically isomorphic to some dual space. We now discuss an alternative assumption, namely that $\mathbb{D}$ is a reflexive Banach space. More precisely, we assume

(A1') $\mathbb{X}$ and $\mathbb{D}$ are real reflexive Banach spaces. Moreover, for each sequence $(\varphi_i)_i \subset \mathbb{X} \cap \mathbb{D}$ with $\varphi_i \to \varphi$ weakly in $\mathbb{X}$ and $\varphi_i \to \overline{\varphi}$ weakly in $\mathbb{D}$ for some $\varphi \in \mathbb{X}$, $\overline{\varphi} \in \mathbb{D}$, it holds $\varphi = \overline{\varphi}$.

Theorem 4.18. *Let the standard assumptions hold with **(A1)** replaced by **(A1')** and 'weak-* convergence in $\mathbb{D}$' replaced by 'weak convergence in $\mathbb{D}$'. Then all statements of Section 4.2 remain valid.*

Proof. The only lemma which explicitly uses that $\mathbb{D}$ is isometrically isomorphic to some dual space is Lemma 4.5, where sequential weak-* compactness in $\mathbb{D}$ is established by the Banach-Alaoglu theorem. This argument has to be replaced by the Eberlein-Šmulian theorem, which establishes sequential weak compactness instead. In the remaining proofs one has to replace 'weak-* convergence in $\mathbb{D}$' by 'weak convergence in $\mathbb{D}$'. □

4.3 Sufficient conditions for the abstract assumptions

We give sufficient conditions for the inner product a_k.

Lemma 4.19. *Let $a_k : (\mathbb{X} \cap \mathbb{D}) \times (\mathbb{X} \cap \mathbb{D}) \to \mathbb{R}$ be an inner product for all $k \in \mathbb{N}_0$. Let there exist constants $c > 0$, $C > 0$ such that*

$$c\|u\|_{\mathbb{X}}^2 \le a_k(u,u) \le C\|u\|_{\mathbb{X}}^2 \quad \forall u \in \mathbb{X} \cap \mathbb{D},\ k \in \mathbb{N}_0.$$

*Then a_k fulfills the assumptions **(A8)-(A12)**.*

Proof. By assumption, **(A8)** and **(A9)** are fulfilled. Since a_k defines an inner product, we can apply the Cauchy-Schwarz inequality,

$$a_k(p,v) \le \|p\|_{a_k}\|v\|_{a_k} \le C\|p\|_{\mathbb{X}}\|v\|_{\mathbb{X}} \le C\|p\|_{\mathbb{X}\cap\mathbb{D}}\|v\|_{\mathbb{X}\cap\mathbb{D}} \quad \forall p,v \in \mathbb{X} \cap \mathbb{D},\ k \in \mathbb{N}_0,$$

thus **(A10)** is fulfilled. Since the map $\mathbb{X} \cap \mathbb{D} \ni p \mapsto a_k(\varphi, p)$ is continuous with respect to the $\mathbb{X}$ norm, we can extend it to $\mathbb{X}$ by the Hahn-Banach theorem. We get $a_k(\varphi, .) \in \mathbb{X}^*$ and hence $a_k(\varphi, p_i) \to 0$ as $i \to \infty$, for all sequences $(p_i)_i \subset \mathbb{X}$ with $p_i \to 0$ weakly in $\mathbb{X}$. Thus, **(A11)** is fulfilled. To prove assumption **(A12)**, let $(v_i)_i, (p_i)_i \subset \mathbb{X} \cap \mathbb{D}$ be given with $v_i \to 0$ in $\mathbb{X}$ and $p_i \to p$ in $\mathbb{X}$ for some $p \in \mathbb{X}$. Then $|a_{k_i}(p_i, v_i)| \le C\|p_i\|_{\mathbb{X}}\|v_i\|_{\mathbb{X}} \to 0$ as $i \to \infty$ for any subsequence a_{k_i}. □

In the case that $\mathbb{X}$ is a Hilbert space, one can thus take a_k to be the Hilbert space inner product $(.,.)_{\mathbb{X}}$. This is still a new result since $\mathbb{D}$ is a Banach space.

In a pure Hilbert space setting we get the following sufficient conditions for the assumptions of Theorem 4.18.

Lemma 4.20. *Let $\mathbb{H}$ be a real Hilbert space. Let $\Phi_{ad} \subset \mathbb{H}$ be a convex, closed, bounded and nonempty subset and let $j \in C^1(\mathbb{H})$ be bounded from below in Φ_{ad}. Assume that $(a_k)_k$ is a sequence of inner products on $\mathbb{H}$. Moreover assume*

$$\begin{aligned} c\|u\|_{\mathbb{H}}^2 \le a_k(u,u) \le C\|u\|_{\mathbb{H}}^2 \quad & \forall u \in \mathbb{H},\ k \in \mathbb{N}_0, \\ 0 < \lambda_{min} \le \lambda_k \le \lambda_{max} \quad & \forall k \in \mathbb{N}_0 \end{aligned} \tag{41}$$

*for some λ_{min}, $\lambda_{max}, C, c > 0$. Then the standard assumptions hold for the choice $\mathbb{D} = \mathbb{X} = \mathbb{H}$ with **(A1)** replaced by **(A1')** and 'weak-* convergence in $\mathbb{D}$' replaced by 'weak convergence in $\mathbb{D}$'.*

Proof. Assumptions **(A1')**,**(A2)**-**(A6)** and **(A13)** are obvious. **(A8)**-**(A12)** can be shown as in Lemma 4.19. It remains **(A7)**, which is fulfilled since $j'(\varphi) \in \mathbb{H}^*$ is weakly continuous for all $\varphi \in \Phi_{ad}$. □

Remark 4.21. We note that the boundedness of Φ_{ad} (assumption **(A4)**) is not needed in the Hilbert space setting of Lemma 4.20 to show the statements of Section 4.2. In the proofs, the boundedness is only needed to control the $\mathbb{D}$-norm. But since this norm now coincides with the $\mathbb{X}$-norm, which can be controlled otherwise, the assumption **(A4)** is not needed anymore.

It turns out that the assumptions in Lemma 4.20 are used throughout the literature concerning variable metric methods. For instance Bertsekas [Ber99] needs the same assumptions as in Lemma 4.20 (except for the boundedness of Φ_{ad}) to show global convergence of the scaled projected gradient method with $\lambda_k = 1$ in finite dimension. In [GB84] a Hilbert space setting is considered, where also the assumptions of Lemma 4.20 are used to show global convergence. However, their algorithm is quite different, since different inner products are used for the gradient and the projection. In [GD88, Dun87], assumption (41) is used to show local properties of a similar variable metric method in Hilbert space. For the unconstrained case in a Hilbert space setting we refer to [GS81], where it is assumed that the condition numbers of the inner products a_k (resp. of the corresponding linear operators) are uniformly bounded, which is slightly weaker than the assumption (41).

Sufficient conditions for **(A5)** and **(A7)** are e.g. that j is continuously differentiable in $\mathbb{D}$ with $j'(\varphi) \in \mathbb{B} \subset (\mathbb{D})^*$ for all $\varphi \in \Phi_{ad}$, where the space $\mathbb{B}$ is as in **(A1)**. In case of $\mathbb{D} = L^\infty(\Omega)$ this amounts to the continuous differentiability of j in $L^\infty(\Omega)$ with $j'(\varphi) \in L^1(\Omega) \subset (L^\infty(\Omega))^*$. This condition is often fulfilled for optimal control problems of ODEs [MQ80, KS89, KS92, KZ13]. Also for optimal control problems of PDEs the control-to-state operator is often only differentiable in L^∞, if e.g. the control appears in the highest order coefficient [BFGS14, DES15] or if certain nonlinearities are present [Trö09]. See also the semilinear elliptic optimal control problem discussed in Section 4.12. Note that differentiability in $L^\infty(\Omega)$ is weaker than differentiability in $L^2(\Omega)$ if the measure of Ω is finite.

In the case that $\mathbb{D}$ is not a dual space, we get a sufficient criterion for assumption **(A7)**:

Lemma 4.22. *Let $\mathbb{X}$ and $\mathbb{D}$ be Banach spaces with a common normed superspace $\mathbb{A}$ and continuous embeddings $\mathbb{X} \hookrightarrow \mathbb{A}$ and $\mathbb{D} \hookrightarrow \mathbb{A}$. Additionally, let $\mathbb{X} \cap \mathbb{D}$ be a dense subset of $\mathbb{X}$ and of $\mathbb{D}$. Then **(A5)** implies **(A7)** (with weak-* convergence replaced by weak convergence in $\mathbb{D}$).*

Proof. Under the assumptions we get from the duality theorem in [BL76] (see Theorem 7.6) that $(\mathbb{X}\cap\mathbb{D})^* = \mathbb{X}^* + \mathbb{D}^*$. It follows that $j'(\varphi) \in (\mathbb{X}\cap\mathbb{D})^*$ can be written as $j'(\varphi) = l_1 + l_2$

with some $l_1 \in \mathbb{X}^*$ and $l_2 \in \mathbb{D}^*$. Let now $(\varphi_i)_i \subset \mathbb{X} \cap \mathbb{D}$ with $\varphi_i \to 0$ weakly in $\mathbb{X}$ and weakly in $\mathbb{D}$. Then $\langle j'(\varphi), \varphi_i \rangle = \langle l_1, \varphi_i \rangle + \langle l_2, \varphi_i \rangle \to \langle j'(\varphi), \varphi^* \rangle$. □

In case that the variable metric is chosen point based, i.e. a_k depends on φ_k rather than on k, and the metric depends continuously on φ_k, the assumptions are also fulfilled. This will be covered in the next section.

4.4 Point based choice of the variable metric

In the standard assumptions the variable inner product a_k depends on the iteration number k. A special case thereof is particularly of interest, namely when the inner product depends on the current iterate φ_k rather than on k itself. Thus we assume in this section that we have given a family $(a_\varphi)_{\varphi \in \Phi_{ad}}$ of inner products, out of which we choose $a_k := a_{\varphi_k}$ in the kth step of the VMPT method. Note that the index φ in a_φ does not denote a differentiation with respect to φ here, as e.g. in f_x. Moreover, for simplicity we assume in this section that λ_k does not depend on k, which is no restriction, since a variable scaling λ_k can be put into the inner product a_k as discussed before.

The most important application for a point based choice of a_k is the projected Newton's method (discussed in Section 4.7), which uses $a_\varphi = j''(\varphi)$. Often a_φ is chosen to be only an approximation of $j''(\varphi)$, leading to a quasi-Newton type method.

Since the inner product depends now on the current iterate φ_k rather than on k and λ_k is independent of k, the solution operator of the projection type subproblem is also independent of k, thus we introduce the notation $\mathcal{P}(\varphi) := \mathcal{P}_{a_\varphi,\lambda}(\varphi)$.
We assume the following properties of a_φ and λ_k:

(A8') $(a_\varphi)_{\varphi \in \Phi_{ad}}$ is a family of inner products on $\mathbb{X} \cap \mathbb{D}$.

(A9') There exists $C > 0$, s.t. $a_\varphi(u,u) \geq C\|u\|_{\mathbb{X}}^2$ for all $u \in \mathbb{X} \cap \mathbb{D}$ and $\varphi \in \Phi_{ad}$.

(A10') For each $\varphi \in \Phi_{ad}$ there exists $C(\varphi) > 0$, s.t. $|a_\varphi(u,v)| \leq C(\varphi)\|u\|_{\mathbb{X}\cap\mathbb{D}}\|v\|_{\mathbb{X}\cap\mathbb{D}}$ for all $u, v \in \mathbb{X} \cap \mathbb{D}$.

(A11') For each $\varphi \in \Phi_{ad}$, $v \in \mathbb{X} \cap \mathbb{D}$ and for each sequence $(p_i)_i \subset \mathbb{X} \cap \mathbb{D}$ with $p_i \to 0$ weakly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$ it holds $a_\varphi(v, p_i) \to 0$ as $i \to \infty$.

(A12') For each sequence $(\varphi_i)_i \subset \Phi_{ad}$ with $\varphi_i \to \varphi$ in $\mathbb{X} \cap \mathbb{D}$ for some $\varphi \in \Phi_{ad}$ it holds $a_{\varphi_i} \to a_\varphi$ with respect to the bilinear operator norm.

(A13') It holds $\lambda_k = \lambda$ for some $\lambda > 0$ and all $k \in \mathbb{N}_0$.

For the special case $a_\varphi = j''(\varphi)$ sufficient conditions for **(A8')**-**(A12')** are given in Theorem 4.38. In particular **(A11')** is trivial in this case, see Lemma 4.37.

Lemma 4.23. *Let $(a_\varphi)_{\varphi \in \Phi_{ad}}$ fulfill the assumptions **(A8')-(A12')** and let φ_k be the iterates of the VMPT method using $a_k := a_{\varphi_k}$. Then $(a_k)_k$ fulfills **(A8)-(A12)**.*

Proof. **(A8)**-**(A11)** are obvious. To show **(A12)**, let φ_{k_i} be a subsequence with $\varphi_{k_i} \to \varphi$ in $\mathbb{X} \cap \mathbb{D}$ for some $\varphi \in \mathbb{X} \cap \mathbb{D}$ and let sequences $(v_i)_i, (p_i)_i \subset \mathbb{X} \cap \mathbb{D}$ be given with $v_i \to 0$ in $\mathbb{X}$ and weakly-* in $\mathbb{D}$ and $p_i \to p$ in $\mathbb{X} \cap \mathbb{D}$ for some $p \in \mathbb{X} \cap \mathbb{D}$. Note that $\varphi \in \Phi_{ad}$ due to the closedness of Φ_{ad}. We estimate

$$|a_{\varphi_{k_i}}(p_i, v_i)| \leq \underbrace{|(a_{\varphi_{k_i}} - a_\varphi)(p_i, v_i)|}_{\leq \|a_{\varphi_{k_i}} - a_\varphi\| \|p_i\|_{\mathbb{X}\cap\mathbb{D}} \|v_i\|_{\mathbb{X}\cap\mathbb{D}} \to 0} + \underbrace{|a_\varphi(p_i - p, v_i)|}_{\leq \|a_\varphi\| \|p_i - p\|_{\mathbb{X}\cap\mathbb{D}} \|v_i\|_{\mathbb{X}\cap\mathbb{D}} \to 0} + \underbrace{|a_\varphi(p, v_i)|}_{\to 0} \to 0,$$

where we use **(A11')**. □

In the following we show some properties of the solution operator $\mathcal{P}$.
In the Hilbert space case (i.e. if $\mathbb{D} = \mathbb{X} = \mathbb{H}$ for some Hilbert space $\mathbb{H}$), it follows from the continuity of the projection and of j', that $\varphi \mapsto P_\perp(\varphi - \lambda \nabla_{\mathbb{H}} j(\varphi))$ is continuous in $\mathbb{H}$. Recall that this map coincides with $\mathcal{P}$ in the Hilbert space setting. We now prove a similar result for the Banach space setting. However, since j is only differentiable in $\mathbb{X} \cap \mathbb{D}$, we show continuity from $\mathbb{X} \cap \mathbb{D}$ into $\mathbb{X}$. Additionally, we allow for a variable point-based metric a_φ. Since this metric depends continuously on φ, it does not influence the continuity of the operator $\mathcal{P}$.

Lemma 4.24. *Under the assumptions **(A1)-(A7),(A8')-(A13')** the mapping $\mathcal{P}$ is continuous from $\Phi_{ad} \subset \mathbb{X} \cap \mathbb{D}$ into $\mathbb{X}$, i.e. for $(\varphi_k)_k \subset \Phi_{ad}$ and $\varphi_k \to \varphi$ in $\mathbb{X} \cap \mathbb{D}$ for some $\varphi \in \Phi_{ad}$ we get $\mathcal{P}(\varphi_k) \to \mathcal{P}(\varphi)$ in $\mathbb{X}$.*

Proof. Let $(\varphi_k)_k$ and φ as in the assumption. Let $y_k := \mathcal{P}(\varphi_k)$ and $y := \mathcal{P}(\varphi)$. The corresponding variational inequality (22) for y_k reads

$$a_{\varphi_k}(y_k - \varphi_k, \eta - y_k) + \lambda \left\langle j'(\varphi_k), \eta - y_k \right\rangle \geq 0 \quad \forall \eta \in \Phi_{ad},$$

as well as the variational inequality for y

$$a_{\varphi}(y - \varphi, \eta - y) + \lambda \left\langle j'(\varphi), \eta - y \right\rangle \geq 0 \quad \forall \eta \in \Phi_{ad}.$$

We test the first inequality by $\eta = y$ and the second by $\eta = y_k$ and add up the two inequalities, which yields

$$\begin{aligned} 0 &\leq a_{\varphi_k}(y_k - \varphi_k, y - y_k) + a_{\varphi}(y - \varphi, y_k - y) + \lambda \left\langle j'(\varphi) - j'(\varphi_k), y_k - y \right\rangle \\ &= a_{\varphi}(y_k - y, y - y_k) + a_{\varphi_k}(y_k - \varphi_k, y - y_k) + a_{\varphi}(\varphi - y_k, y - y_k) + \lambda \left\langle j'(\varphi) - j'(\varphi_k), y_k - y \right\rangle \end{aligned}$$

Utilizing **(A9')** for a_φ leads to

$$C\|y - y_k\|_{\mathbb{X}}^2 \leq a_{\varphi_k}(y_k - \varphi_k, y - y_k) + a_{\varphi}(\varphi - y_k, y - y_k) + \lambda \left\langle j'(\varphi) - j'(\varphi_k), y_k - y \right\rangle. \quad (42)$$

We prove that the right hand side goes to zero. Recall that y_k is uniformly bounded in $\mathbb{X} \cap \mathbb{D}$, which can be shown as in Lemma 4.11. Thus we get

$$|\left\langle j'(\varphi) - j'(\varphi_k), y_k - y \right\rangle| \leq \|j'(\varphi) - j'(\varphi_k)\|_{(\mathbb{X}\cap\mathbb{D})^*} \|y_k - y\|_{\mathbb{X}\cap\mathbb{D}} \to 0.$$

For the remaining terms we get by **(A12')**

$$\begin{aligned} |a_{\varphi_k}(y_k - \varphi_k, y - y_k) + a_{\varphi}(\varphi - y_k, y - y_k)| &\leq |(a_{\varphi_k} - a_\varphi)(y_k - \varphi_k, y - y_k)| + |a_\varphi(\varphi - \varphi_k, y - y_k)| \\ &\leq \underbrace{\|a_{\varphi_k} - a_\varphi\|}_{\to 0} \underbrace{\|y_k - \varphi_k\|_{\mathbb{X}\cap\mathbb{D}} \|y - y_k\|_{\mathbb{X}\cap\mathbb{D}}}_{\leq C} + C \underbrace{\|\varphi - \varphi_k\|_{\mathbb{X}\cap\mathbb{D}}}_{\to 0} \underbrace{\|y - y_k\|_{\mathbb{X}\cap\mathbb{D}}}_{\leq C} \to 0. \end{aligned}$$

□

The previous result is in particular true if $a_\varphi = a$ is taken independently of φ, i.e. the inner product is the same for each step of the VMPT method.

It is well known that orthogonal projections in Hilbert spaces are Lipschitz continuous. In the case that this also holds for j', the map $\varphi \mapsto P_\perp(\varphi - \lambda \nabla_{\mathbb{H}} j(\varphi))$ is Lipschitz continuous. The following generalization of this result can be shown for the operator $\mathcal{P}$.

Corollary 4.25. *Let **(A1)-(A7),(A8')-(A13')** hold, let $a_k = a$ be independent of k and let $j \in C^{1,1}(\Phi_{ad})$, then it holds the Lipschitz-type estimate*

$$\|\mathcal{P}(\varphi_1) - \mathcal{P}(\varphi_2)\|_{\mathbb{X}}^2 \leq C\|\varphi_1 - \varphi_2\|_{\mathbb{X}\cap\mathbb{D}}\|\mathcal{P}(\varphi_1) - \mathcal{P}(\varphi_2)\|_{\mathbb{X}\cap\mathbb{D}}$$

for some $C > 0$ and all $\varphi_1, \varphi_2 \in \Phi_{ad}$.

Proof. The inequality (42) becomes in this case

$$\begin{aligned} C\|\mathcal{P}(\varphi_1) - \mathcal{P}(\varphi_2)\|_{\mathbb{X}}^2 &\leq a(\varphi_1 - \varphi_2, \mathcal{P}(\varphi_1) - \mathcal{P}(\varphi_2)) + \lambda\left\langle j'(\varphi_1) - j'(\varphi_2), \mathcal{P}(\varphi_2) - \mathcal{P}(\varphi_1)\right\rangle \\ &\leq C\|\varphi_1 - \varphi_2\|_{\mathbb{X}\cap\mathbb{D}}\|\mathcal{P}(\varphi_1) - \mathcal{P}(\varphi_2)\|_{\mathbb{X}\cap\mathbb{D}}. \end{aligned}$$

□

In the case $\mathbb{X} = \mathbb{D}$ the original Lipschitz continuity can be recovered. In the other case we can at least show local Hölder continuity with exponent $\frac{1}{2}$.

Corollary 4.26. *Let **(A1)-(A7),(A8')-(A13')** hold, let $a_k = a$ be independent of k and let $j \in C^{1,1}(\Phi_{ad})$, then $\mathcal{P} : \mathbb{X} \cap \mathbb{D} \to \mathbb{X}$ is locally $\frac{1}{2}$-Hölder continuous.*

Proof. Let $M > 0$, $\varphi_i \in \Phi_{ad}$ be arbitrary with $\|\varphi_i\|_{\mathbb{X}\cap\mathbb{D}} \leq M$, $i = 1, 2$. From **(A4)**, we get with Corollary 4.25

$$\begin{aligned} \|\mathcal{P}(\varphi_1) - \mathcal{P}(\varphi_2)\|_{\mathbb{X}}^2 &\leq C\|\varphi_1 - \varphi_2\|_{\mathbb{X}\cap\mathbb{D}}\|\mathcal{P}(\varphi_1) - \mathcal{P}(\varphi_2)\|_{\mathbb{X}\cap\mathbb{D}} \\ &\leq C(M)(\|\mathcal{P}(\varphi_1) - \mathcal{P}(\varphi_2)\|_{\mathbb{X}} + 1). \end{aligned}$$

We conclude $\|\mathcal{P}(\varphi_1) - \mathcal{P}(\varphi_2)\|_{\mathbb{X}} \leq C(M)$. Applying Corollary 4.25 and **(A4)** again yields

$$\|\mathcal{P}(\varphi_1) - \mathcal{P}(\varphi_2)\|_{\mathbb{X}} \leq C(M)\|\varphi_1 - \varphi_2\|_{\mathbb{X}\cap\mathbb{D}}^{\frac{1}{2}}.$$

□

4.5 Translation invariance

Like the projected gradient method the VMPT method is translation invariant. This property seems to be trivial, but it can be very helpful, for instance if the abstract assumptions can only be shown for some translated optimization problem, as it is the case for the problem in Section 6.1.1.

Theorem 4.27. *The VMPT method is translation invariant, i.e.:*
Let $t \in \mathbb{X} \cap \mathbb{D}$ arbitrary. Then it holds for the iterates $(\tilde{\varphi}_k)_k$ of the method applied to the translated problem

$$\begin{aligned} &\min j(\varphi - t) \\ &\varphi \in \Phi_{ad} + t \end{aligned}$$

with initial guess $\tilde{\varphi}_0 = \varphi_0 + t$, that $\tilde{\varphi}_k = \varphi_k + t$, where $(\varphi_k)_k$ are the iterates of the method applied to the untranslated problem

$$\begin{aligned} &\min j(\varphi) \\ &\varphi \in \Phi_{ad}. \end{aligned}$$

Proof. We prove the statement by induction in the iteration number k. For $k = 0$ the statement holds by assumption. Let the statement hold for k. We add a tilde to all variables appearing in the method for the translated problem. The projection type subproblem for the translated problem reads

$$\begin{aligned} \min \quad & \frac{1}{2}\|\tilde{y} - \tilde{\varphi}_k\|_{a_k}^2 + \lambda_k \left\langle j'(\tilde{\varphi}_k - t), \tilde{y} - \tilde{\varphi}_k \right\rangle \\ & \tilde{y} \in \Phi_{ad} + t, \end{aligned}$$

which is equivalent to

$$\begin{aligned} \min \quad & \frac{1}{2}\|(\tilde{y} - t) - (\tilde{\varphi}_k - t)\|_{a_k}^2 + \lambda_k \left\langle j'(\tilde{\varphi}_k - t), (\tilde{y} - t) - (\tilde{\varphi}_k - t) \right\rangle \\ & \tilde{y} - t \in \Phi_{ad}. \end{aligned}$$

Since $\tilde{\varphi}_k - t = \varphi_k$ we get that the projection type subproblems are equivalent. The unique solvability yields

$$\tilde{y}_k - t = y_k.$$

Thus for the search direction it holds

$$\tilde{v}_k = \tilde{y}_k - \tilde{\varphi}_k = y_k + t - (\varphi_k + t) = v_k.$$

Now consider the line search. For the translated problem the Armijo condition (20) is

$$j(\tilde{\varphi}_k - t + \tilde{\alpha}_k \tilde{v}_k) \leq j(\tilde{\varphi}_k - t) + \tilde{\alpha}_k \sigma \left\langle j'(\tilde{\varphi}_k - t), \tilde{v}_k \right\rangle,$$

which is equivalent to

$$j(\varphi_k + \tilde{\alpha}_k v_k) \leq j(\varphi_k) + \tilde{\alpha}_k \sigma \left\langle j'(\varphi_k), v_k \right\rangle.$$

Since also the step length resulting from the Armijo backtracking is unique we get $\tilde{\alpha}_k = \alpha_k$ and

$$\tilde{\varphi}_{k+1} = \tilde{\varphi}_k + \tilde{\alpha}_k \tilde{v}_k = \varphi_k + t + \alpha_k v_k = \varphi_{k+1} + t.$$

□

In the case that $a_k = a$ is chosen independent of k one can show by the same way that the method is invariant under a-orthogonal transformations, which are of the form $\tilde{\varphi}_k = A\varphi_k + t$, where $A : \mathbb{X} \cap \mathbb{D} \to \mathbb{X} \cap \mathbb{D}$ is linear, bijective and it holds $\|A\varphi\|_a = \|\varphi\|_a$ for all $\varphi \in \mathbb{X} \cap \mathbb{D}$. If A is not a-orthogonal, it still holds $\tilde{\varphi}_k = A\varphi_k + t$, but φ_k are then the iterates of the method using the transformed metric $(x, y) \mapsto a(Ax, Ay)$.

4.6 Curved search along the projection arc

The VMPT method in Algorithm 4.1 is globalized using Armijo backtracking along the straight line given by the search direction v_k. However, in some cases it can be preferable to perform a curved search along the projection arc, i.e. $\alpha_k = 1$ is fixed and λ_k is determined, such that it fulfills some step length rule. Note that the projection arc is in our case given by $\lambda \mapsto \mathcal{P}_{a,\lambda}(\varphi)$ for fixed inner product a and $\varphi \in \Phi_{ad}$. This is analog to the map $\lambda \mapsto P_\perp(\varphi - \lambda \nabla_{\mathbb{H}} j(\varphi))$ for the classical projected gradient method in Hilbert space.

We restrict ourselves again to Armijo backtracking along the projection arc, which can be defined as follows (see e.g. [KS92, Ber99]):

Definition 4.28. Find the minimal power $m_k \in \mathbb{N}_0$, such that $\lambda_k := \beta^{m_k}\bar{\lambda}_k$ fulfills

$$j(\mathcal{P}_{a_k,\lambda_k}(\varphi_k)) \le j(\varphi_k) + \sigma \left\langle j'(\varphi_k), \mathcal{P}_{a_k,\lambda_k}(\varphi_k) - \varphi_k \right\rangle, \tag{43}$$

where $\beta \in (0,1)$ and $\sigma \in (0,1)$ are fixed numbers and $\bar{\lambda}_k > 0$ is some initial guess of the step length.

It is proposed in [Ber76] that the initial step length $\bar{\lambda}_k$ can depend on k. This is very useful in practice, because an initial guess of the step length can be used to reduce the number of backtracking steps and thus the number of projections.

The benefits of the curved search are as follows: In finite dimension it can be shown that the active constraints of a nondegenerate minimum are determined after finitely many steps, see e.g. [Kel99, Ber76]. This can also be shown in Hilbert space if *finitely* many inequality constraints are present [GD88]. As soon as the active constraints are known, an unconstrained method can be used to compute the minimum (resp. a method dealing only with equality constraints). However, it is not trivial to decide at which iterate the active constraints are known. Moreover, if the number of (real-valued) inequalities is infinite, the active constraints cannot be identified in a finite number of iterations, see [KS92]. An example for infinitely many constraints are the box constraints $u_a \le u \le u_b$ in L^∞, which often occur in optimal control problems.
The drawback of a curved search is that in every backtracking step the operator $\mathcal{P}_{a_k,\lambda_k}(\varphi_k)$ has to evaluated, which can be expensive. Thus for each trial value for λ_k the projection type subproblem has to be solved. On the other hand, for a line search along the direction v_k as in Algorithm 4.1, only a single subproblem has to be solved in each step of the VMPT method. Moreover, if j is quadratic, then the exact step length (the minimum of j along the search path) can be computed analytically for the line search, but not for the curved search. In general, smoothness and convexity of j is preserved when restricting j to the line search path, whereas this is not the case for the restriction to the projection arc.

4.6.1 Properties of the projection arc and difficulties arising in the Banach space setting

We begin by showing some properties of the projection arc which are analog to the Hilbert space case.

The following lemma corresponds to the monotonicity of the orthogonal projection in Hilbert space $(P_\perp(x) - P_\perp(y), x - y)_{\mathbb{H}} \ge 0$:

Lemma 4.29. *Let $\lambda_1, \lambda_2 > 0$, $\varphi, \eta \in \Phi_{ad}$, and the inner product a fulfill the standard assumptions. Then it holds*

$$a(\varphi - \eta, \mathcal{P}_{a,\lambda_1}(\varphi) - \mathcal{P}_{a,\lambda_2}(\eta)) - \left\langle \lambda_1 j'(\varphi) - \lambda_2 j'(\eta), \mathcal{P}_{a,\lambda_1}(\varphi) - \mathcal{P}_{a,\lambda_2}(\eta) \right\rangle \ge 0$$

with equality if and only if $\mathcal{P}_{a,\lambda_1}(\varphi) = \mathcal{P}_{a,\lambda_2}(\eta)$.

In particular, by choosing $\eta = \varphi$ it holds

$$(\lambda_1 - \lambda_2)\left\langle j'(\varphi), \mathcal{P}_{a,\lambda_1}(\varphi) - \mathcal{P}_{a,\lambda_2}(\varphi)\right\rangle \leq 0, \tag{44}$$

and thus the map

$$\lambda \mapsto \left\langle j'(\varphi), \mathcal{P}_{a,\lambda}(\varphi)\right\rangle$$

is nonincreasing.

Proof. We skip the proof, since it is analog to the Hilbert space case, see [HPUU08, Lemma 1.10 (d)]. □

Lemma 4.30. *For all $\varphi \in \Phi_{ad}$ and inner product a fulfilling the standard assumptions, the function*

$$\Phi(\lambda) := \frac{1}{\lambda}\|P_{a,\lambda}(\varphi) - \varphi\|_a, \quad \lambda > 0$$

is nonincreasing.

Proof. Again analog to the Hilbert space case, see [HPUU08, Lemma 1.10 (e)], where inequality (44) is used. □

Lemma 4.31. *For any $\varphi \in \Phi_{ad}$ and a, λ fulfilling the standard assumptions it holds*

$$\left\langle j'(\varphi), \mathcal{P}_{a,\lambda}(\varphi) - \varphi\right\rangle \leq -\frac{1}{\lambda}\|\mathcal{P}_{a,\lambda}(\varphi) - \varphi\|_a^2.$$

Proof. The proof of Lemma 4.9 can be repeated verbatim. □

Lemma 4.32. *For all non-stationary $\varphi \in \Phi_{ad}$ and inner product a fulfilling the standard assumptions, it holds*

$$\begin{aligned}
\|\mathcal{P}_{a,\lambda}(\varphi) - \varphi\|_a &\to 0 \quad \text{as } \lambda \searrow 0,\\
\frac{\|\mathcal{P}_{a,\lambda}(\varphi) - \varphi\|_a^2}{\lambda} &\to 0 \quad \text{as } \lambda \searrow 0,\\
\frac{\|\mathcal{P}_{a,\lambda}(\varphi) - \varphi\|_a}{\lambda} &\geq C > 0 \quad \text{for } 0 < \lambda \leq 1,
\end{aligned}$$

i.e. $\mathcal{P}_{a,\lambda}(\varphi)$ converges to φ faster than $\sqrt{\lambda}$ but not faster than λ.

Proof. As in Lemma 4.11 it can be shown that $(\|\mathcal{P}_{a,\lambda}(\varphi)\|_{\mathbb{X}})_{\tilde{\lambda}>\lambda>0}$ is bounded for any $\tilde{\lambda} > 0$. From Lemma 4.31 we get

$$-\lambda\|j'(\varphi)\|_{(\mathbb{X}\cap\mathbb{D})^*}\underbrace{\|\mathcal{P}_{a,\lambda}(\varphi) - \varphi\|_{\mathbb{X}\cap\mathbb{D}}}_{\leq C} \leq \lambda\left\langle j'(\varphi), \mathcal{P}_{a,\lambda}(\varphi) - \varphi\right\rangle \leq -\|\mathcal{P}_{a,\lambda}(\varphi) - \varphi\|_a^2 \leq 0,$$

from which the first statement follows.
From the first statement, **(A9)** and Lemma 4.5 we get $\mathcal{P}_{a,\lambda}(\varphi) \to \varphi$ weakly-* in $\mathbb{D}$ and thus $\left\langle j'(\varphi), \mathcal{P}_{a,\lambda}(\varphi) - \varphi\right\rangle \to 0$ due to **(A7)**. The second statement then follows from Lemma 4.31.
The third statement holds since $\Phi(\lambda)$ is nonincreasing (see Lemma 4.30) and thus

$$\frac{\|\mathcal{P}_{a,\lambda}(\varphi) - \varphi\|_a}{\lambda} = \Phi(\lambda) \geq \Phi(1) = \|\mathcal{P}_{a,1}(\varphi) - \varphi\|_a.$$

It holds $\Phi(1) > 0$ since φ is not stationary, see Lemma 4.8. □

From the previous lemma and the coercivity **(A9)** we get

Corollary 4.33. *For all $\varphi \in \Phi_{ad}$ and a fulfilling the standard assumptions it holds $\mathcal{P}_{a,\lambda}(\varphi) \to \varphi$ in $\mathbb{X}$ as $\lambda \searrow 0$.*

It turns out that the continuity of the projection arc in $\mathbb{X}$ (in contrast to $\mathbb{X} \cap \mathbb{D}$) is not sufficient to show global convergence of the curved search method: A standard way to prove global convergence of the projected gradient method with curved search is to show the existence of a positive lower bound for λ_k. However, it is not possible to transfer the proofs in [Kel99, Thm. 5.4.5] , [GS81, Thm. 8.4], [LP66, Thm. 5.1] or [DR70, Thm. 2.5] to the Banach space setting considered in this thesis, since two different norms appear in the respective estimates. From estimates connected to Taylor expansions of j we always get the $\mathbb{X} \cap \mathbb{D}$-norm because of the differentiability in $\mathbb{X} \cap \mathbb{D}$ **(A5)**. On the other hand, in the estimates coming from the variational inequality (22) the weaker $\mathbb{X}$-norm appears due to the coercivity of a_k in $\mathbb{X}$ **(A9)**. Because of the same reason, also the global convergence proofs in [HPUU08, Ber99, Gol64] cannot be adapted. The same holds if we use the alternative Armijo condition (see e.g. [Ber76])

$$j(\mathcal{P}_{a_k,\lambda_k}(\varphi_k)) \le j(\varphi_k) - \frac{\sigma}{\lambda_k}\|\mathcal{P}_{a_k,\lambda_k}(\varphi_k) - \varphi_k\|_{\mathbb{X}}^2$$

or, using the $\mathbb{X} \cap \mathbb{D}$-norm,

$$j(\mathcal{P}_{a_k,\lambda_k}(\varphi_k)) \le j(\varphi_k) - \frac{\sigma}{\lambda_k}\|\mathcal{P}_{a_k,\lambda_k}(\varphi_k) - \varphi_k\|_{\mathbb{X}\cap\mathbb{D}}^2.$$

Not even the existence of a positive step length $\lambda_k > 0$ fulfilling the Armijo condition can be shown. The proof in [Ber99, GB82] or [GK02] therefor could be adapted if we could prove that $\mathcal{P}_{a_k,\lambda}(\varphi_k) \to \varphi_k$ in $\mathbb{X} \cap \mathbb{D}$ as $\lambda \searrow 0$. However, we only have the weaker convergence $\mathcal{P}_{a_k,\lambda}(\varphi_k) \to \varphi_k$ in $\mathbb{X}$ as $\lambda \searrow 0$, see Corollary 4.33. Thus, we can only show that the projection arc is continuous in $\mathbb{X}$ (at $\lambda = 0$) and hence the cost functional j is in general not continuous along the projection arc. Note that in [GB82] it is exploited that j' is continuous along the projection arc for $\lambda = 0$, which we cannot show here. On the other hand, the straight line $\alpha \mapsto \varphi_k + \alpha v_k$ used in the line search globalization is smooth in the stronger $\mathbb{X} \cap \mathbb{D}$-norm, hence j is C^1 along this line. Recall that we used $\varphi_k + \alpha v_k \to \varphi_k$ in $\mathbb{X} \cap \mathbb{D}$ as $\alpha \to 0$ in the global convergence proof of Theorem 4.14.

To overcome these difficulties we propose two approaches. The first is a hybrid method, which first tries to perform a curved search in λ. In case the curved search fails, a line search in α is done as backup. The second approach is to claim stronger assumptions than the standard assumptions. Based on the insight that the coercivity assumption **(A9)** is too weak to be able to perform a curved search, we claim the stronger coercivity

$$\exists c > 0: \quad c\|u\|_{\mathbb{X}\cap\mathbb{D}}^2 \le a_k(u,u) \quad \forall u \in \mathbb{X} \cap \mathbb{D},\ k \in \mathbb{N}_0.$$

However, under this assumption we have $c\|u\|_{\mathbb{X}\cap\mathbb{D}}^2 \le a_k(u,u) \le C\|u\|_{\mathbb{X}\cap\mathbb{D}}^2$, thus the a_k-norm is equivalent to the $\mathbb{X} \cap \mathbb{D}$-norm and $(\mathbb{X} \cap \mathbb{D}, a_k)$ becomes a Hilbert space. Hence we restrict ourselves to the Hilbert space case for the second approach.

4.6.2 Compromise: A hybrid method

As already mentioned, the curved search in λ may fail, e.g. if the Armijo condition is not fulfilled for any $\lambda > 0$. The hybrid method described in Algorithm 4.2 performs a line search in α in this case to guarantee global convergence. It turns out that Algorithm 4.2 is a special case of the VMPT method in Algorithm 4.1 and thus the global convergence proof of Theorem 4.14 applies.

Algorithm 4.2 A hybrid method

1: Choose $\varphi_0 \in \Phi_{ad}$, $0 < \beta < 1$, $0 < \sigma < 1$, $\lambda_{min} > 0$.
2: $k := 0$.
3: **while** $k \leq k_{max}$ **do**
4: Find the minimal power $m_k \in \mathbb{N}_0$, such that $\lambda_k := \beta^{m_k} \bar{\lambda}_k$ fulfills

$$j(\mathcal{P}_{a_k,\lambda_k}(\varphi_k)) \leq j(\varphi_k) + \sigma \left\langle j'(\varphi_k), \mathcal{P}_{a_k,\lambda_k}(\varphi_k) - \varphi_k \right\rangle.$$

5: **if** such an m_k does not exist or $\lambda_k < \lambda_{min}$ **then**
6: $\lambda_k := \lambda_{min}$.
7: $v_k := \mathcal{P}_{a_k,\lambda_k}(\varphi_k) - \varphi_k$.
8: Calculate the step length $0 < \alpha_k \leq 1$ by Armijo backtracking in direction v_k, i.e. find the minimal power $m_k \in \mathbb{N}_0$ such that $\alpha_k := \beta^{m_k}$ fulfills

$$j(\varphi_k + \alpha_k v_k) \leq j(\varphi_k) + \alpha_k \sigma \left\langle j'(\varphi_k), v_k \right\rangle.$$

9: **else**
10: $\alpha_k := 1$
11: $v_k := \mathcal{P}_{a_k,\lambda_k}(\varphi_k) - \varphi_k$
12: **end if**
13: Update $\varphi_{k+1} := \varphi_k + \alpha_k v_k$
14: **if** $\|v_k\|_{\mathbb{X}} < tol$ **then**
15: **return**
16: **end if**
17: $k := k + 1$
18: **end while**

Theorem 4.34. *Let the standard assumptions hold except for **(A13)** and let there exist $\lambda_{max} > 0$, such that $\bar{\lambda}_k \leq \lambda_{max}$ for all $k \in \mathbb{N}_0$. Then the statements of the global convergence theorem 4.14 hold for Algorithm 4.2.*

Proof. By induction we show that the iterates of the hybrid algorithm 4.2 coincide with the iterates of the VMPT method with line search (Algorithm 4.1), using the sequence $(\lambda_k)_k$ from Algorithm 4.2, and apply Theorem 4.14. Assume that the same initial guess $\varphi_0 \in \Phi_{ad}$ is used for Algorithm 4.2 and Algorithm 4.1. Then the base case $k = 0$ is shown. For the inductive step, assume that the kth iterate of Algorithm 4.2 and Algorithm 4.1 coincide. Consider as first case that the curved search in Algorithm 4.2 fails. Then it holds $\lambda_k \geq \lambda_{min}$ and α_k is determined by Armijo backtracking in both algorithms. Thus the iterates in the $(k+1)$th step coincide. In the second case, the curved search succeeds, $\lambda_k \geq \lambda_{min}$ and $\alpha_k = 1$ is chosen by Algorithm 4.2. Since the curved search succeeds, the step length $\alpha_k = 1$ fulfills the Armijo condition (20) and thus Algorithm 4.1 also chooses $\alpha_k = 1$, which completes the induction. From the assumption $\bar{\lambda}_k \leq \lambda_{max}$ we conclude that

Algorithm 4.2 produces step lengths with $0 < \lambda_{min} \le \lambda_k \le \lambda_{max}$ for all $k \in \mathbb{N}_0$, thus **(A13)** is fulfilled. Finally we apply Theorem 4.14. □

4.6.3 The Hilbert space case

Under the standard assumptions, the existence of a positive step length λ_k fulfilling the Armijo condition cannot be shown. We now impose stricter assumptions to show that the Armijo backtracking along the projection arc is well defined and the resulting step lengths are bounded away from zero. The algorithm is summarized in Algorithm 4.3, which coincides with Algorithm 4.2 except for the line search backup.

Algorithm 4.3 VMPT method in Hilbert space with curved search

Choose $\varphi_0 \in \Phi_{ad}$, $0 < \beta < 1$, $0 < \sigma < 1$.
$k := 0$.
while $k \le k_{max}$ **do**
  Find the minimal power $m_k \in \mathbb{N}_0$, such that $\lambda_k := \beta^{m_k}\bar{\lambda}_k$ fulfills

$$j(\mathcal{P}_{a_k,\lambda_k}(\varphi_k)) \le j(\varphi_k) + \sigma \left\langle j'(\varphi_k), \mathcal{P}_{a_k,\lambda_k}(\varphi_k) - \varphi_k \right\rangle .$$

  Update $\varphi_{k+1} := \mathcal{P}_{a_k,\lambda_k}(\varphi_k)$
  if $\|\varphi_{k+1} - \varphi_k\|_{\mathbb{H}} < tol$ **then**
    return
  end if
  $k := k + 1$
end while

For simplicity we use the assumptions of Lemma 4.20 and assume $j \in C^{1,1}$. The following theorem is well known if a_k does not depend on k. However, we did not find the statements in the literature for variable metric a_k. Thus we include the proof here.

Theorem 4.35. *Let $\mathbb{H}$ be a real Hilbert space. Let $\Phi_{ad} \subset \mathbb{H}$ be a convex, closed, bounded and nonempty subset and let $j \in C^{1,1}(\mathbb{H})$ be bounded from below in Φ_{ad}. Assume that $(a_k)_k$ is a sequence of inner products on $\mathbb{H}$. Moreover assume*

$$\begin{aligned} c\|u\|_{\mathbb{H}}^2 \le a_k(u,u) \le C\|u\|_{\mathbb{H}}^2 \quad & \forall u \in \mathbb{H},\ k \in \mathbb{N}_0, \\ 0 < \bar{\lambda}_{min} \le \bar{\lambda}_k \le \bar{\lambda}_{max} \quad & \forall k \in \mathbb{N}_0 \end{aligned} \tag{45}$$

for some $\bar{\lambda}_{min}$, $\bar{\lambda}_{max}, C, c > 0$. Then the statements of the global convergence theorem 4.14 hold for Algorithm 4.3 with $\mathbb{X} \cap \mathbb{D}$ and $\mathbb{X}$ replaced by $\mathbb{H}$ and $v_k := \varphi_{k+1} - \varphi_k$.

Proof. By virtue of (45) we can adapt the proof in [Kel99] to get a lower bound on the step length λ_k. Let $k \in \mathbb{N}_0$, $\varphi \in \Phi_{ad}$ and $\lambda > 0$ be fixed. As in the proof of Lemma 4.9 it can be shown that

$$\left\langle j'(\varphi), \mathcal{P}_{a_k,\lambda}(\varphi) - \varphi \right\rangle \le -\frac{C}{\lambda}\|\mathcal{P}_{a_k,\lambda}(\varphi) - \varphi\|_{\mathbb{H}}^2.$$

Combining this with the Hölder estimate in Lemma 7.2 we get

$$\begin{aligned} j(\mathcal{P}_{a_k,\lambda}(\varphi)) - j(\varphi) &\leq \left\langle j'(\varphi), \mathcal{P}_{a_k,\lambda}(\varphi) - \varphi \right\rangle + \frac{L}{2} \|\mathcal{P}_{a_k,\lambda}(\varphi) - \varphi\|_{\mathbb{H}}^2 \\ &\leq \left\langle j'(\varphi), \mathcal{P}_{a_k,\lambda}(\varphi) - \varphi \right\rangle - \frac{\lambda L}{2C} \left\langle j'(\varphi), \mathcal{P}_{a_k,\lambda}(\varphi) - \varphi \right\rangle \\ &= \left(1 - \frac{\lambda L}{2C}\right) \underbrace{\left\langle j'(\varphi), \mathcal{P}_{a_k,\lambda}(\varphi) - \varphi \right\rangle}_{\leq 0} \end{aligned}$$

There exists a $\tilde{\lambda} > 0$ such that $1 - \frac{\lambda L}{2C} \geq \sigma$ for all $0 < \lambda \leq \tilde{\lambda}$ and thus the Armijo condition is fulfilled for these λ. We conclude that the Armijo backtracking in line 4 of Algorithm 4.3 produces step lengths λ_k for which it holds $\lambda_k \geq \lambda_{min} := \min\{\tilde{\lambda}\beta, \bar{\lambda}_{min}\}$ and $\lambda_k \leq \lambda_{\max} := \bar{\lambda}_{max}$. As in the proof of Theorem 4.34 we observe that the iterates of Algorithm 4.3 are the same as for Algorithm 4.1 using the same parameters λ_k. Note that Algorithm 4.1 always accepts $\alpha_k = 1$ for this choice of λ_k. Hence we can apply Lemma 4.20, Theorem 4.18 and Theorem 4.14 to show the statement. We also note that in this case it holds $v_k = \varphi_{k+1} - \varphi_k$, since $\alpha_k = 1$. □

We note that the boundedness of Φ_{ad} in $\mathbb{H}$ is not needed in the Hilbert space case as argued in Remark 4.21.

4.7 Projected Newton's method

The projected Newton's method corresponds to the point based choice of $a_\varphi = j''(\varphi)$ in the VMPT method. Typically one uses $\lambda_k = 1$ for all k, thus in every step of the projected Newton's method one has to solve the subproblem (cf. (18))

$$\min_{y \in \Phi_{ad}} \frac{1}{2} j''(\varphi_k)[y - \varphi_k, y - \varphi_k] + \left\langle j'(\varphi_k), y - \varphi_k \right\rangle. \tag{46}$$

In the case that $j''(\varphi_k)$ is potitive definite the subproblem is equivalent to the linear variational inequality (cf. (26))

$$y \in \Phi_{ad}, \quad j''(\varphi_k)[y - \varphi_k, \eta - y] + \left\langle j'(\varphi_k), \eta - y \right\rangle \geq 0 \quad \forall \eta \in \Phi_{ad}.$$

If $j''(\varphi_k)$ is not positive definite the VI is only a necessary condition for a minimizer of (46). It turns out that except for the line search, the projected Newton's method coincides with the Josephy-Newton method applied to the variational inequality

$$\varphi \in \Phi_{ad}, \quad \left\langle j'(\varphi), \eta - \varphi \right\rangle \geq 0 \quad \forall \eta \in \Phi_{ad},$$

see (8), which is a first order optimality condition for the considered optimization problem (19). Subproblem (46) corresponds to the minimization of the second order Taylor polynomial of j in Φ_{ad}. Thus, j is approximated by a quadratic functional, whereas Φ_{ad} stays unchanged. This is similar to the idea of the SQP method, where j is also replaced by a quadratic functional and in addition the constraints are linearized. One can establish that the projected Newton's method coincides with the SQP method if the constraints determining Φ_{ad} are already linear, see e.g. Section 6.9.

The projected Newton's method is well known and results are available in finite dimension, Hilbert space and Banach space. The results are mainly global and local convergence

statements. Regarding global convergence, one shows convergence independent of the initial guess φ_0 if some globalization technique, e.g. line search, is added. Usually, for line search methods the cost functional j has to be convex in order to get a descent direction. In fact, a typical assumption is of the form

$$\exists C, c > 0: \quad c\|u\|^2 \leq j''(\varphi)[u,u] \leq C\|u\|^2 \quad \forall \varphi, u, \tag{47}$$

see [Ber99] for the finite dimensional case and [Gol65] for the unconstrained Hilbert space case. In [Dun80], (47) is assumed for $c = 0$ to obtain global convergence. For local convergence statements, one usually considers the projected Newton's method without globalization. As typical for Newton type methods, local q-superlinear and q-quadratic rates can be shown [Ber99, Gol65, Dun80, Dun88]. For local convergence results in Hilbert space, (47) is assumed in [LP66] and at least the lower bound of (47) is assumed in [Ber99] in finite dimension. Since the projected Newton's method is also a Josephy-Newton method, we have Robinson's strong regularity as sufficient condition for superlinear convergence. This condition can also be fulfilled for nonconvex cost functionals in contrast to (47).

Using the global convergence proof for the VMPT method in Theorem 4.14 we will show global convergence of the projected Newton's method, where we weaken the assumption (47) to

$$\exists C(\varphi), c > 0: \quad c\|u\|_{\mathbb{X}}^2 \leq j''(\varphi)[u,u] \leq C(\varphi)\|u\|_{\mathbb{X}\cap\mathbb{D}}^2 \quad \forall \varphi, u,$$

i.e. we request the coercivity only with respect to the weaker $\mathbb{X}$-norm and the upper bound $C(\varphi)$ can depend on φ. We will also provide sufficient conditions for local q-superlinear and q-quadratic convergence. However, these conditions are only slightly more general than other conditions used in the literature. Finally we show that it is sufficient that j' is only semismooth rather than C^1, which is analog to the unconstrained semismooth Newton method.

4.7.1 Global convergence

We show global convergence in case the projected Newton's method is combined with Armijo backtracking along the Newton direction as described in Algorithm 4.1 for the choice $a_k = j''(\varphi_k)$.

For the global convergence of the projected Newton method we need the following assumption, which is slightly stronger that **(A7)**, since the property is demanded not only for $\varphi \in \Phi_{ad}$, but also in a neighborhood of Φ_{ad}.

(A7') There exists a $(\mathbb{X}\cap\mathbb{D})$-neighborhood U of Φ_{ad} such that for each $\varphi \in U$ and for each sequence $(\varphi_i)_i \subset \mathbb{X}\cap\mathbb{D}$ with $\varphi_i \to 0$ weakly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$ it holds $\langle j'(\varphi), \varphi_i\rangle \to 0$ as $i \to \infty$.

We start with an auxiliary lemma.

Lemma 4.36. *It holds that*

$$Y := \{l \in (\mathbb{X}\cap\mathbb{D})^* \mid \langle l, v_k\rangle \to 0 \text{ for any sequence } (v_k)_{k\in\mathbb{N}} \subset \mathbb{X}\cap\mathbb{D} \text{ with } v_k \to 0 \text{ weakly in } \mathbb{X} \text{ and weakly-* in } \mathbb{D}\}$$

endowed with the $(\mathbb{X}\cap\mathbb{D})^*$-norm is a Banach space.

Proof. It is clear that Y is a linear subspace of $(\mathbb{X} \cap \mathbb{D})^*$, since $0 \in Y$ and Y is closed under linear combinations. It remains to show that Y is a closed subspace of $(\mathbb{X} \cap \mathbb{D})^*$. Let $(l_i)_i \subset Y$ be a sequence with $l_i \to l$ in $(\mathbb{X} \cap \mathbb{D})^*$ for some $l \in (\mathbb{X} \cap \mathbb{D})^*$. We show $l \in Y$. Let $(v_k)_{k \in \mathbb{N}} \subset \mathbb{X} \cap \mathbb{D}$ with $v_k \to 0$ weakly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$. Then

$$\begin{aligned} |\langle l, v_k \rangle| &\leq |\langle l - l_i, v_k \rangle| + |\langle l_i, v_k \rangle| \leq \|l - l_i\|_{(\mathbb{X} \cap \mathbb{D})^*} \|v_k\|_{\mathbb{X} \cap \mathbb{D}} + |\langle l_i, v_k \rangle| \\ &\leq C \|l - l_i\|_{(\mathbb{X} \cap \mathbb{D})^*} + |\langle l_i, v_k \rangle| \end{aligned}$$

Let $\varepsilon > 0$. Choose i such that $\|l - l_i\|_{(\mathbb{X} \cap \mathbb{D})^*} < \varepsilon/(2C)$, then find $\overline{k}$, such that $|\langle l_i, v_k \rangle| < \varepsilon/2$ for all $k \geq \overline{k}$. Then we get $|\langle l, v_k \rangle| \leq \varepsilon$ for all $k \geq \overline{k}$. □

Lemma 4.37. *Let the assumptions **(A1)-(A7)** hold, together with **(A7')**. In addition let j be two times Fréchet differentiable in an open neighborhood of $\Phi_{ad} \subset \mathbb{X} \cap \mathbb{D}$.*
Then $j' : \mathbb{X} \cap \mathbb{D} \supset U \to Y$ is Fréchet differentiable in some open neighborhood U of $\Phi_{ad} \subset \mathbb{X} \cap \mathbb{D}$. In particular $j''(\varphi)[u, .] \in Y$ for any $\varphi \in U$ and $u \in \mathbb{X} \cap \mathbb{D}$.

Proof. Let U be an open neighborhood of Φ_{ad} in which **(A7')** holds and in which j is two times Fréchet differentiable. We have to show $j''(\varphi) \in \mathcal{L}(\mathbb{X} \cap \mathbb{D}, Y)$ for any $\varphi \in U$ and the estimate for the rest term in the Taylor expansion. For arbitrary $\varphi \in U$ and $u \in \mathbb{X} \cap \mathbb{D}$ it holds

$$\frac{j'(\varphi + tu) - j'(\varphi)}{t} \to j''(\varphi)[u, .] \text{ in } (\mathbb{X} \cap \mathbb{D})^* \text{ as } t \searrow 0.$$

For small t the difference quotient is an element of Y because of **(A7')**. Since Y is a closed subspace of $(\mathbb{X} \cap \mathbb{D})^*$, which was shown in Lemma 4.36, we get $j''(\varphi)[u, .] \in Y$. By assumption it holds that $j''(\varphi) \in \mathcal{L}(\mathbb{X} \cap \mathbb{D}, (\mathbb{X} \cap \mathbb{D})^*)$. Since Y is endowed with the same norm as $(\mathbb{X} \cap \mathbb{D})^*$, we also get $j''(\varphi) \in \mathcal{L}(\mathbb{X} \cap \mathbb{D}, Y)$. Now consider the rest term $r(h) = j'(\varphi + h) - j'(\varphi) - j''(\varphi)[h, .]$ for $\varphi \in U$ and $h \in \mathbb{X} \cap \mathbb{D}$. Then it holds

$$\frac{\|r(h)\|_Y}{\|h\|_{\mathbb{X} \cap \mathbb{D}}} = \frac{\|r(h)\|_{(\mathbb{X} \cap \mathbb{D})^*}}{\|h\|_{\mathbb{X} \cap \mathbb{D}}} \to 0 \text{ as } \|h\|_{\mathbb{X} \cap \mathbb{D}} \to 0,$$

since $j' : U \to (\mathbb{X} \cap \mathbb{D})^*$ is Fréchet differentiable by assumption. □

Theorem 4.38. *Let the assumptions **(A1)-(A7)** hold as well as **(A7')**. In addition let j be two times continuously Fréchet differentiable in an open neighborhood of $\Phi_{ad} \subset \mathbb{X} \cap \mathbb{D}$. Let there exist a positive constant m such that*

$$m \|u\|_{\mathbb{X}}^2 \leq j''(\varphi)[u, u] \tag{48}$$

for all $\varphi \in \Phi_{ad}$ and $u \in \mathbb{X} \cap \mathbb{D}$. Let also

$$\lambda_{min} \leq \lambda_k \leq \lambda_{max}$$

be fulfilled for all $k \in \mathbb{N}_0$.
*Then the standard assumptions **(A1)-(A13)** are fulfilled for $a_k := j''(\varphi_k)$. In particular the global convergence results of Theorem 4.14 can be applied.*

Proof. The inner product a_k is chosen point based, i.e. $a_k = a_{\varphi_k}$ with $a_\varphi = j''(\varphi)$. We show that a_φ fulfills the assumptions **(A8')**-**(A12')**. Then we apply Lemma 4.23.
(A8') and **(A9')** follow from (48). **(A10')** is fulfilled since $j''(\varphi) \in \mathcal{L}(\mathbb{X} \cap \mathbb{D}, (\mathbb{X} \cap \mathbb{D})^*)$ and **(A12')** because j'' is continuous. Finally, **(A11')** was shown in Lemma 4.37. □

A simple example which fulfills the weak assumptions in Theorem 4.38 can be found in [Trö09, Sec. 4.10.2]. Consider

$$\begin{aligned} \min j(\varphi) &= -\int_0^1 \cos(\varphi(x))\,\mathrm{d}x \\ -\frac{\pi}{2} + \varepsilon \le \varphi &\le \frac{\pi}{2} - \varepsilon \text{ a.e. in } [0,1] \end{aligned} \tag{49}$$

for some $\varepsilon > 0$. It is shown in [Trö09] that j is two times continuously differentiable in $L^\infty([0,1])$. Moreover, it holds

$$j''(\varphi)[u,u] = \int_0^1 \cos(\varphi)|u|^2\,\mathrm{d}x \ge c\|u\|_{L^2}^2.$$

Note that we choose the bound constraints on φ such that the above coercivity holds. Thus, for the choice $\mathbb{X} = L^2([0,1])$ and $\mathbb{D} = L^\infty([0,1])$ the assumptions of Theorem 4.38 are fulfilled, which can be shown easily. On the other hand it can be proved that j is *not* two times Fréchet differentiable in $L^2([0,1])$. Neither exists $c > 0$, such that $j''(\varphi)[u,u] \ge c\|u\|_{L^\infty}^2$. Thus the widely used assumption (47) is not fulfilled here, but only the weaker assumptions we used in Theorem 4.38. In fact, j is two times continuously differentiable in $L^p([0,1])$ if and only if $p > 2$, see [Trö09]. Thus the choice $\mathbb{D} = L^p([0,1])$ for $p > 2$ would also be possible.

A similar global convergence result is given in [Dun80, Thm. 4.1], where globalization is done using the Goldstein step length rule for α instead of Armijo backtracking, and λ_k is set to 1. Only a single Banach space is considered instead of the intersection $\mathbb{X} \cap \mathbb{D}$. The weaker condition $j''(\varphi)[u,u] \ge 0 \ \forall u, \varphi$ is assumed instead of (48). Therefore no well posedness of the projection type subproblem is shown in [Dun80], which we can prove using (48). Moreover, [Dun80] assumes $\|j''(\varphi)\| \le M$ for some M independent of φ and the uniform boundedness of the search directions v_k, which we both don't need as assumption. In fact we showed the uniform boundedness of v_k using (48). In Section 6.13.11 we give an example of a metric a_k, which fulfills $a_k(u,u) \ge 0 \ \forall u, k$, but which is not uniformly coercive with respect to the $\mathbb{X}$- norm. It turns out that the projection type subproblem doesn't have a solution for this example and thus the method is not well defined. Hence the condition $a_k(u,u) \ge 0$ assumed in [Dun80] for $j''(\varphi)$ is too weak for a practical algorithm.

4.7.2 Local convergence rates

To show local q-superlinear convergence of the projected Newton method we impose an additional continuity property on $j''(\varphi)$ at the minimizer $\overline{\varphi}$, see (51) below. The following result is a generalization of the convergence result of the projected Newton method in finite dimension [Ber99, Prop. 2.3.5.].

Theorem 4.39. *Let the assumptions **(A1)-(A7)** hold as well as **(A7')**. Let $\overline{\varphi} \in \Phi_{ad}$ be a local minimum of j in Φ_{ad}. For $\theta > 0$, let $B_\theta := \{\varphi \in \Phi_{ad} \mid \|\varphi - \overline{\varphi}\|_{\mathbb{X}} \le \theta\}$. In addition let j be two times Fréchet differentiable in an open neighborhood of $B_\theta \subset \mathbb{X} \cap \mathbb{D}$ for some $\theta > 0$. Let there exist a positive constant m such that*

$$m\|u\|_{\mathbb{X}}^2 \le j''(\varphi)[u,u] \tag{50}$$

for all $u \in \mathbb{X} \cap \mathbb{D}$ and all $\varphi \in B_\theta$ and let

$$j''(\varphi_k) \to j''(\overline{\varphi}) \text{ in } \mathcal{L}(\mathbb{X}, \mathbb{X}^*) \tag{51}$$

for any sequence $(\varphi_k)_k \subset B_\theta$ *with* $\varphi_k \to \overline{\varphi}$ *in* $\mathbb{X}$.
Then there exists some positive $\delta < \theta$ *such that for any initial guess* φ_0 *in* B_δ, *the sequence of iterates of the projected Newton method, i.e. the unglobalized VMPT method with* $a_k := j''(\varphi_k)$ *and* $\lambda_k := \alpha_k := 1$, *stays in* B_δ *and converges q-superlinearly to* $\overline{\varphi}$ *in* $\mathbb{X}$. *Moreover, if there exists some* $L > 0$ *such that the Lipschitz-type condition*

$$\|j''(\varphi) - j''(\overline{\varphi})\|_{\mathcal{L}(\mathbb{X},\mathbb{X}^*)} \le L\|\varphi - \overline{\varphi}\|_{\mathbb{X}} \tag{52}$$

holds for all $\varphi \in B_\theta$, *then the sequence converges q-quadratically in* $\mathbb{X}$.

Proof. The assumptions guarantee that the projection type subproblem is uniquely solvable, thus the method is well defined. Let $\varphi \in B_\theta$. We first show estimates for $y := \mathcal{P}_{j''(\varphi),1}(\varphi)$. By testing the corresponding variational inequality (22) by $\eta = \overline{\varphi}$ we get

$$j''(\varphi)[y - \varphi, \overline{\varphi} - y] + \langle j'(\varphi), \overline{\varphi} - y\rangle \ge 0.$$

We add $j''(\varphi)[y - \overline{\varphi}, y - \overline{\varphi}]$ to both sides, leading to

$$j''(\varphi)[\overline{\varphi} - \varphi, \overline{\varphi} - y] + \langle j'(\varphi), \overline{\varphi} - y\rangle \ge j''(\varphi)[y - \overline{\varphi}, y - \overline{\varphi}].$$

We apply the coercivity (50) and the fundamental theorem of calculus.

$$\begin{aligned}\|y - \overline{\varphi}\|_{\mathbb{X}}^2 &\le \frac{1}{m} j''(\varphi)[y - \overline{\varphi}, y - \overline{\varphi}] \le \frac{1}{m}\left(j''(\varphi)[\overline{\varphi} - \varphi, \overline{\varphi} - y] + \langle j'(\varphi), \overline{\varphi} - y\rangle\right) = \\ &= \frac{1}{m}\Big(j''(\varphi)[\overline{\varphi} - \varphi, \overline{\varphi} - y] + \int_0^1 j''(\overline{\varphi} + t(\varphi - \overline{\varphi}))[\overline{\varphi} - y, \varphi - \overline{\varphi}])\,\mathrm{d}t + \\ &\quad + \langle j'(\overline{\varphi}), \overline{\varphi} - y\rangle\Big).\end{aligned} \tag{53}$$

Note that the integral is finite for all φ near $\overline{\varphi}$ because of (51). Since $\overline{\varphi}$ is a local minimum we have the estimate $\langle j'(\overline{\varphi}), y - \overline{\varphi}\rangle \ge 0$, thus

$$\begin{aligned}\|y - \overline{\varphi}\|_{\mathbb{X}}^2 &\le \frac{1}{m}\int_0^1 (j''(\overline{\varphi} + t(\varphi - \overline{\varphi})) - j''(\varphi))[\overline{\varphi} - y, \varphi - \overline{\varphi}]\,\mathrm{d}t \\ &\le \frac{1}{m}\int_0^1 \|j''(\overline{\varphi} + t(\varphi - \overline{\varphi})) - j''(\varphi)\|_{\mathcal{L}(\mathbb{X},\mathbb{X}^*)}\,\mathrm{d}t \|\overline{\varphi} - y\|_{\mathbb{X}}\|\varphi - \overline{\varphi}\|_{\mathbb{X}}.\end{aligned}$$

Dividing by $\|y - \overline{\varphi}\|_{\mathbb{X}}$ yields

$$\|y - \overline{\varphi}\|_{\mathbb{X}} \le \frac{1}{m}\int_0^1 \|j''(\overline{\varphi} + t(\varphi - \overline{\varphi})) - j''(\varphi)\|_{\mathcal{L}(\mathbb{X},\mathbb{X}^*)}\,\mathrm{d}t \|\varphi - \overline{\varphi}\|_{\mathbb{X}}. \tag{54}$$

Given some $C_0 < 1$ we can choose $\delta < \theta$ so small that

$$\frac{1}{m}\int_0^1 \|j''(\overline{\varphi} + t(\eta - \overline{\varphi})) - j''(\eta)\|_{\mathcal{L}(\mathbb{X},\mathbb{X}^*)}\,\mathrm{d}t \le C_0$$

holds for all $\eta \in B_\delta$, which is due to the triangle inequality,

$$\begin{aligned}\|j''(\overline{\varphi} + t(\eta - \overline{\varphi})) - j''(\eta)\|_{\mathcal{L}(\mathbb{X},\mathbb{X}^*)} &\le \|j''(\overline{\varphi} + t(\eta - \overline{\varphi})) - j''(\overline{\varphi})\|_{\mathcal{L}(\mathbb{X},\mathbb{X}^*)} \\ &\quad + \|j''(\overline{\varphi}) - j''(\eta)\|_{\mathcal{L}(\mathbb{X},\mathbb{X}^*)},\end{aligned}$$

and the continuity (51). Note that $\overline{\varphi} + t(\eta - \overline{\varphi}) \in B_\delta$ for all $t \in [0,1]$ as soon as $\eta \in B_\delta$. For some initial guess $\varphi_0 \in B_\delta$ we get $\varphi_1 = \mathcal{P}_{j''(\varphi_0),1}(\varphi_0)$ by the definition of the method.

Equation (54) then yields $\|\varphi_1 - \overline{\varphi}\|_{\mathbb{X}} \leq C_0 \|\varphi_0 - \overline{\varphi}\|_{\mathbb{X}} < \delta$ and by induction

$$\|\varphi_k - \overline{\varphi}\|_{\mathbb{X}} \leq C_0^k \|\varphi_0 - \overline{\varphi}\|_{\mathbb{X}}.$$

Thus $\varphi_k \to \overline{\varphi}$ in $\mathbb{X}$. From (54) we also get

$$\|\varphi_{k+1} - \overline{\varphi}\|_{\mathbb{X}} \leq \frac{1}{m} \int_0^1 \|j''(\overline{\varphi} + t(\varphi_k - \overline{\varphi})) - j''(\varphi_k)\|_{\mathcal{L}(\mathbb{X},\mathbb{X}^*)} \,\mathrm{d}t \|\varphi_k - \overline{\varphi}\|_{\mathbb{X}}. \tag{55}$$

Due to (51), it holds $\|j''(\overline{\varphi} + t(\varphi_k - \overline{\varphi})) - j''(\varphi_k)\|_{\mathcal{L}(\mathbb{X},\mathbb{X}^*)} \to 0$ uniformly in t as $k \to \infty$ and thus the convergence rate is q-superlinear.
If in addition (52) holds, we get from (55)

$$\begin{aligned}
\|\varphi_{k+1} - \overline{\varphi}\|_{\mathbb{X}} &\leq \frac{1}{m} \int_0^1 \|j''(\overline{\varphi} + t(\varphi_k - \overline{\varphi})) - j''(\overline{\varphi})\|_{\mathcal{L}(\mathbb{X},\mathbb{X}^*)} + \|j''(\overline{\varphi}) - j''(\varphi_k)\|_{\mathcal{L}(\mathbb{X},\mathbb{X}^*)} \,\mathrm{d}t \\
&\quad \|\varphi_k - \overline{\varphi}\|_{\mathbb{X}} \\
&\leq \frac{1}{m} \int_0^1 Lt\|\varphi_k - \overline{\varphi}\|_{\mathbb{X}} + L\|\varphi_k - \overline{\varphi}\|_{\mathbb{X}} \,\mathrm{d}t \|\varphi_k - \overline{\varphi}\|_{\mathbb{X}} \\
&\leq \frac{3L}{2m} \|\varphi_k - \overline{\varphi}\|_{\mathbb{X}}^2.
\end{aligned} \tag{56}$$

Thus the rate of convergence is quadratic. □

Remark 4.40. For varying λ_k, the local q-superlinear convergence rate can also be maintained provided that $\lambda_k \to 1$. The integrand in (55) is in this case

$$\|\lambda_k j''(\overline{\varphi} + t(\varphi_k - \overline{\varphi})) - j''(\varphi_k)\|_{\mathcal{L}(\mathbb{X},\mathbb{X}^*)},$$

which also converges to zero.

Remark 4.41. A possible choice for the convergence radius δ is $\delta = \frac{2m}{3L}$ where L is the Lipschitz constant of j'' and m is the coercivity constant of j'', see (56). This constant can be improved to $\delta = \frac{2m}{L}$ if the Lipschitz continuity (52) is claimed in a neighborhood of $\overline{\varphi}$, since the triangle inequality is then not needed in the estimate (56).

Remark 4.42. We give some remarks about assumption (51). If we would have $j'(\varphi) \in \mathbb{X}^*$ for all φ near $\overline{\varphi}$ and assumption (51) would hold for all sequences $(\varphi_k)_{k\in\mathbb{N}}$ near $\overline{\varphi}$, it would follow that $j' : \mathbb{X} \to \mathbb{X}^*$ is Gâteaux differentiable near $\overline{\varphi}$ and Fréchet differentiable in $\overline{\varphi}$ (see e.g. [Wer07, Satz III.5.4(c)]). But in general we only have $j'(\varphi) \in (\mathbb{X} \cap \mathbb{D})^*$ and (51) only holds for sequences $(\varphi_k)_{k\in\mathbb{N}}$ in Φ_{ad}, thus assumption (51) is weaker than a differentiability condition. On the other hand, a necessary condition for the assumption (51) is that it holds $j'(\varphi) = j'(\overline{\varphi}) + r(\varphi)$ for some $r(\varphi) \in \mathbb{X}^*$ and all $\varphi \in B_\theta$ with θ sufficiently small. The reason is that from $j''(\varphi) \in \mathcal{L}(\mathbb{X}, \mathbb{X}^*)$ we get $j'(\varphi) - j'(\overline{\varphi}) \in \mathbb{X}^*$ (for an extension), since it holds

$$\begin{aligned}
\langle j'(\varphi) - j'(\overline{\varphi}), v\rangle &= \int_0^1 j''(\overline{\varphi} + t(\varphi - \overline{\varphi}))[v, \varphi - \overline{\varphi}] \,\mathrm{d}t \\
&\leq \int_0^1 \|j''(\overline{\varphi} + t(\varphi - \overline{\varphi}))\|_{\mathcal{L}(\mathbb{X},\mathbb{X}^*)} \,\mathrm{d}t \|v\|_{\mathbb{X}} \|\varphi - \overline{\varphi}\|_{\mathbb{X}}
\end{aligned}$$

for all $v \in \mathbb{X} \cap \mathbb{D}$. The integral is finite for all $\varphi \in B_\theta$ with θ sufficiently small due to (51). Thus, assumption (51) is only slightly weaker than Gâteaux differentiability near $\overline{\varphi}$ and Fréchet differentiability in $\overline{\varphi}$ with respect to the $\mathbb{X}$-norm.

Remark 4.43. We are able to show local convergence rates if $j''(\varphi)$ is uniformly $\mathbb{X}$-coercive

and bounded in $\mathbb{X}$ (Theorem 4.39). For the global convergence result (Theorem 4.38), a weaker assumption is made on $j''(\varphi)$, namely uniformly $\mathbb{X}$-coercivity and boundedness in $\mathbb{X} \cap \mathbb{D}$. However, under the latter weak assumptions no local convergence rates can be shown by this kind of proof, since instead of the equation (55) we would only get

$$\frac{\|\varphi_{k+1} - \overline{\varphi}\|_{\mathbb{X}}^2}{\|\varphi_{k+1} - \overline{\varphi}\|_{\mathbb{X}\cap\mathbb{D}}} \leq C \int_0^1 \|j''(\overline{\varphi} + t(\varphi_k - \overline{\varphi})) - j''(\varphi_k)\|_{\mathcal{L}(\mathbb{X}\cap\mathbb{D},(\mathbb{X}\cap\mathbb{D})^*)}\, \mathrm{d}t \|\varphi_k - \overline{\varphi}\|_{\mathbb{X}\cap\mathbb{D}},$$

which does not suffice to show q-superlinear convergence.

Remark 4.44. In the spirit of [Dun80] the coercivity assumption (50) can be weakened to the second order condition

$$\exists m > 0: \quad \frac{1}{2} j''(\overline{\varphi})[y - \overline{\varphi}, y - \overline{\varphi}] + \left\langle j'(\overline{\varphi}), y - \overline{\varphi} \right\rangle \geq \frac{m}{2} \|y - \overline{\varphi}\|_{\mathbb{X}}^2 \quad \forall y \in \Phi_{ad}, \tag{57}$$

where $\overline{\varphi} \in \Phi_{ad}$ is a minimizer of j in Φ_{ad}. In this case, one gets instead of (54) the estimate

$$\|y - \overline{\varphi}\|_{\mathbb{X}} \leq \frac{2\int_0^1 \|j''(\varphi + t(\overline{\varphi} - \varphi)) - j''(\varphi)\|_{\mathcal{L}(\mathbb{X},\mathbb{X}^*)}\, \mathrm{d}t}{m - \|j''(\varphi) - j''(\overline{\varphi})\|_{\mathcal{L}(\mathbb{X},\mathbb{X}^*)}} \|\varphi - \overline{\varphi}\|_{\mathbb{X}},$$

which can be proved by the techniques in [Dun80, Thm. 2.5]. Superlinear and quadratic convergence can then be proved as in Theorem 4.39. However, without the coercivity assumption (50) one cannot show the well posedness of the subproblem. Note that (57) is weaker than (50), since it holds $\langle j'(\overline{\varphi}), y - \overline{\varphi}\rangle \geq 0$ for all $y \in \Phi_{ad}$. Moreover, (57) is only assumed at $\overline{\varphi}$ and only in tangential directions $y - \overline{\varphi}$ and can also be fulfilled for nonconvex j. Also note that in [Dun80], the lower bound in (57) is assumed in the stronger norm $\|y - \overline{\varphi}\|_{\mathbb{X}\cap\mathbb{D}}^2$, thus our assumption (57) is weaker.
It is remarkable that [Dun80] does not need the convexity of Φ_{ad} in order to show local convergence. In this case the first order condition of the subproblem is not a variational inequality anymore and thus the Newton type method doesn't coincide with a Josephy-Newton method.

In the following we replace the Lipschitz condition (52) and the continuity assumption (51) by another condition which is probably more practical. However, we can then only show q-superlinear convergence instead of quadratic convergence. We assume that for all $M > 0$ there exists $L(M)$ such that the Lipschitz estimate

$$|(j''(\varphi + h) - j''(\varphi))[u_1, u_2]| \leq L(M)\|h\|_{\mathbb{D}}\|u_1\|_{\mathbb{X}}\|u_2\|_{\mathbb{X}}$$

holds for all φ, h, u_1, u_2 with $\|\varphi\|_{\mathbb{D}} \leq M$ and $\|h\|_{\mathbb{D}} \leq M$. In the case $\mathbb{X} = L^2$ and $\mathbb{D} = L^\infty$ this condition is fulfilled for the semilinear elliptic optimal control problem discussed in Section 4.12 (see [Trö09, Lem. 4.26]). Also for the example problem (49) the above estimate is fulfilled, since we have

$$\left|(j''(\varphi + h) - j''(\varphi))[u_1, u_2]\right| = \left|\int_0^1 (\cos(\varphi + h) - \cos(\varphi))u_1 u_2\right| \leq C\|h\|_{L^\infty}\|u_1\|_{L^2}\|u_2\|_{L^2}$$

due to the Lipschitz continuity of cos and the Hölder inequality. On the other hand the stronger continuity assumption (51) in Theorem 4.39 is not fulfilled.

Note that in the following corollary the balls are $(\mathbb{X} \cap \mathbb{D})$-balls instead of $\mathbb{X}$-balls.

Corollary 4.45. *Let the assumptions **(A1)-(A7)** hold as well as **(A7')**. Let $\overline{\varphi} \in \Phi_{ad}$ be*

a local minimum of j in Φ_{ad}. For $\theta > 0$, let $B_\theta := \{\varphi \in \Phi_{ad} \mid \|\varphi - \overline{\varphi}\|_{\mathbb{X}\cap\mathbb{D}} \le \theta\}$. In addition let j be two times Fréchet differentiable in an open neighborhood of $B_\theta \subset \mathbb{X} \cap \mathbb{D}$ for some $\theta > 0$. Let there exist a positive constant m such that

$$m\|u\|_{\mathbb{X}}^2 \le j''(\varphi)[u,u] \tag{58}$$

for all $u \in \mathbb{X} \cap \mathbb{D}$ and all $\varphi \in B_\theta$. For all $M > 0$ let there exist $L(M)$ such that

$$|(j''(\varphi + h) - j''(\varphi))[u_1, u_2]| \le L(M)\|h\|_{\mathbb{D}}\|u_1\|_{\mathbb{X}}\|u_2\|_{\mathbb{X}} \tag{59}$$

for all $u_1, u_2, h \in \mathbb{X} \cap \mathbb{D}$, $\varphi \in B_\theta$ with $\varphi + h \in B_\theta$, $\|\varphi\|_{\mathbb{D}} \le M$ and $\|h\|_{\mathbb{D}} \le M$. Let the iterates of the projected Newton method fulfill $\varphi_k \to \overline{\varphi}$ in $\mathbb{X} \cap \mathbb{D}$.
Then it holds $\varphi_k \to \overline{\varphi}$ q-superlinearly in $\mathbb{X}$.

Proof. Without loss of generality we assume that $\varphi_k \in B_\theta$ for all k. As in the proof of Theorem 4.39 we get that for any $\varphi \in B_\theta$ with $y := \mathcal{P}_{j''(\varphi),1}(\varphi)$ it holds

$$\|y - \overline{\varphi}\|_{\mathbb{X}}^2 \le \frac{1}{m}\int_0^1 (j''(\overline{\varphi} + t(\varphi - \overline{\varphi})) - j''(\varphi))[\overline{\varphi} - y, \varphi - \overline{\varphi}]\,\mathrm{d}t.$$

Applying (59) for $M = \|\overline{\varphi}\|_{\mathbb{D}} + \theta$ we get

$$\|\varphi_{k+1} - \overline{\varphi}\|_{\mathbb{X}} \le \frac{L(M)}{2m}\|\varphi_k - \overline{\varphi}\|_{\mathbb{D}}\|\varphi_k - \overline{\varphi}\|_{\mathbb{X}}.$$

Since $\|\varphi_k - \overline{\varphi}\|_{\mathbb{D}} \to 0$ we conclude q-superlinear convergence in $\mathbb{X}$. □

We turn to another corollary of Theorem 4.39. If $j''(\varphi)$ admits a stronger coercivity than in (50), convergence in $\mathbb{X} \cap \mathbb{D}$ can be shown. Since the stronger coercivity

$$\exists m > 0: \quad m\|u\|_{\mathbb{X}\cap\mathbb{D}}^2 \le j''(\varphi)[u,u] \quad \forall u \in \mathbb{X} \cap \mathbb{D}$$

together with $j''(\varphi)[u,u] \le C\|u\|_{\mathbb{X}\cap\mathbb{D}}^2$ results in $(\mathbb{X} \cap \mathbb{D}, j''(\varphi))$ being a Hilbert space, we restrict ourselves to the Hilbert space case.

Corollary 4.46. *Let $\mathbb{H}$ be a real Hilbert space. Let $\Phi_{ad} \subset \mathbb{H}$ be a convex, closed and nonempty subset. Let $\overline{\varphi} \in \Phi_{ad}$ be a local minimizer of j in Φ_{ad}. For $\theta > 0$, let $B_\theta := \{\varphi \in \Phi_{ad} \mid \|\varphi - \overline{\varphi}\|_{\mathbb{H}} \le \theta\}$. Let j be two times continuously differentiable near B_θ for some $\theta > 0$. Let there exist a positive constant m such that*

$$m\|u\|_{\mathbb{H}}^2 \le j''(\varphi)[u,u] \tag{60}$$

for all $u \in \mathbb{H}$ and all $\varphi \in B_\theta$.
Then there exists some positive $\delta < \theta$ such that for any initial guess φ_0 in B_δ, the sequence of iterates of the projected Newton method stays in B_δ and converges q-superlinearly to $\overline{\varphi}$ in $\mathbb{H}$. Moreover, if there exists some $L > 0$ such that the Lipschitz-type condition

$$\|j''(\varphi) - j''(\overline{\varphi})\|_{\mathcal{L}(\mathbb{H},(\mathbb{H})^*)} \le L\|\varphi - \overline{\varphi}\|_{\mathbb{H}}$$

holds for all $\varphi \in B_\theta$, then the sequence converges q-quadratically in $\mathbb{H}$.

Proof. We apply the proof of Theorem 4.39 for the choices $\mathbb{X} = \mathbb{D} = \mathbb{H}$. Note that the boundedness of Φ_{ad} **(A4)** is not needed for the well posedness of the subproblem in the Hilbert space setting, cf. Remark 4.21. Also the boundedness of j from below **(A6)** is not

needed here. Note that (51) follows from the continuous second order differentiability of j in $\mathbb{H}$. □

We note that the statements of Corollary 4.46 are well known, see e.g. [Dun80]. The superlinear convergence in Corollary 4.46 also follows from the more general result in [Don12, Thm. 3], where the condition (60) is relaxed to the strong metric subregularity of some linearized operator. It turns out that the assumptions of Corollary 4.46 also yield strong regularity of $\overline{\varphi}$ in the sense of Robinson, which we show in the following. Thus, superlinear convergence follows also from the results concerning the Josephy Newton method in Theorem 3.2.

Lemma 4.47. *Under the assumptions of Corollary 4.46, the local minimum $\overline{\varphi} \in \Phi_{ad}$ is a strongly regular solution (in the sense of Robinson) of the variational inequality*

$$\varphi \in \Phi_{ad}, \quad \left\langle j'(\varphi), \eta - \varphi \right\rangle \geq 0 \quad \forall \eta \in \Phi_{ad}.$$

Proof. We have to show that there exists some $\delta > 0$ such that the linearized variational inequality

$$\varphi \in \Phi_{ad}, \quad \left\langle j'(\overline{\varphi}) - p, \eta - \varphi \right\rangle + j''(\overline{\varphi})[\varphi - \overline{\varphi}, \eta - \varphi] \geq 0 \quad \forall \eta \in \Phi_{ad}.$$

has a locally unique solution $\varphi(p)$ for each $p \in \mathbb{H}^*$ with $\|p\|_{\mathbb{H}^*} \leq \delta$ and that the solution depends Lipschitz continuously on p. As in Theorem 4.6 it can be shown that this linearized VI even has a globally unique solution. To show Lipschitz continuity, let $p_1, p_2 \in \mathbb{H}^*$. We test the VI for p_1 by $\eta = \varphi(p_2)$ and vice versa and add up the VIs. We get

$$\langle p_1 - p_2, \varphi(p_1) - \varphi(p_2) \rangle \geq j''(\overline{\varphi})[\varphi(p_1) - \varphi(p_2), \varphi(p_1) - \varphi(p_2)].$$

Using (60) we end up with

$$m\|\varphi(p_1) - \varphi(p_2)\|_{\mathbb{H}}^2 \leq \|p_1 - p_2\|_{\mathbb{H}^*} \|\varphi(p_1) - \varphi(p_2)\|_{\mathbb{H}},$$

thus

$$\|\varphi(p_1) - \varphi(p_2)\|_{\mathbb{H}} \leq C\|p_1 - p_2\|_{\mathbb{H}^*}.$$

□

Note that assumption (60) is not necessary for the minimum $\overline{\varphi}$ to be strongly regular. Thus local superlinear convergence can be obtained also by weaker assumptions on j. In this case the Josephy-Newton method cannot be seen as a projection type method anymore, since $j''(\varphi)$ does not define an inner product in general.
Unlike Corollary 4.46, one cannot prove Theorem 4.39 and Corollary 4.45 by means of the Josephy-Newton method. The reason is that the differentiability of j' is given in $\mathbb{X} \cap \mathbb{D} \supset \Phi_{ad} \to (\mathbb{X} \cap \mathbb{D})^*$, but the Lipschitz continuity of the solution of the linearized variational inequality in Lemma 4.47 can only be shown in $\mathbb{X}^* \ni p \mapsto \varphi(p) \in \mathbb{X}$, which is due to (50). Also the strong metric subregularity condition in [Don12] is required with respect to a stronger norm than in the coercivity assumptions (50) and (58).

4.7.3 Semismooth projected Newton method

We showed that q-superlinear convergence can be achieved when using the inner product $a_k = j''(\varphi_k)$ in the VMPT method. In this case j has to be two times differentiable. We

now show that q-superlinear convergence can also be obtained if j' is only semismooth. For simplicity we assume in this section that $\mathbb{X} = \mathbb{D} = \mathbb{H}$ for some real Hilbert space $\mathbb{H}$. Consider the VMPT method using arbitrary inner products $(a_k)_k$ fulfilling the standard assumptions. As in the proof of Theorem 4.39 one can derive an estimate similar to (53)

$$\begin{aligned}\|\varphi_{k+1} - \overline{\varphi}\|_{\mathbb{H}}^2 &\le C(a_k(\varphi_k - \overline{\varphi}, \varphi_{k+1} - \overline{\varphi}) - \langle j'(\varphi_k), \varphi_{k+1} - \overline{\varphi}\rangle) + \langle j'(\overline{\varphi}), \varphi_{k+1} - \overline{\varphi}\rangle)\\ &\le C\|a_k(\varphi_k - \overline{\varphi}, .) - j'(\varphi_k) + j'(\overline{\varphi})\|_{\mathbb{H}^*}\|\varphi_{k+1} - \overline{\varphi}\|_{\mathbb{H}}.\end{aligned}$$

Thus

$$\|\varphi_{k+1} - \overline{\varphi}\|_{\mathbb{H}} \le C\|a_k(\varphi_k - \overline{\varphi}, .) - j'(\varphi_k) + j'(\overline{\varphi})\|_{\mathbb{H}^*} \tag{61}$$

and the convergence rate is q-superlinear if

$$\|a_k(\varphi_k - \overline{\varphi}, .) - j'(\varphi_k) + j'(\overline{\varphi})\|_{\mathbb{H}^*} = o(\|\varphi_k - \overline{\varphi}\|_{\mathbb{H}}) \quad \text{as } k \to \infty. \tag{62}$$

A possible choice for a_k is therefore a generalized differential of j' at φ_k in the sense of Definition 3.3.

Theorem 4.48. *Let $\mathbb{H}$ be a real Hilbert space. Let $\Phi_{ad} \subset \mathbb{H}$ be a convex, closed and nonempty subset and $j \in C^1(\mathbb{H})$. Let $\overline{\varphi} \in \Phi_{ad}$ be a local minimizer of j in Φ_{ad}. Let $\partial j' : \mathbb{H} \rightrightarrows \mathcal{L}(\mathbb{H}, \mathbb{H}^*)$ be a set-valued mapping and assume that j' is $\partial j'$-semismooth at $\overline{\varphi}$. Consider the point based choice of the inner product*

$$a_k := a_{\varphi_k} \text{ with } a_\varphi \in \partial j'(\varphi)$$

and assume that there exists $c > 0$ such that

$$c\|u\|_{\mathbb{H}}^2 \le a_\varphi(u, u) \quad \forall u \in \mathbb{H} \tag{63}$$

for all φ in a neighborhood of $\overline{\varphi}$.
Then the iterates of the unglobalized VMPT method (i.e. $\lambda_k = \alpha_k = 1$) converge q-superlinearly to $\overline{\varphi}$ provided that φ_0 is sufficiently close to $\overline{\varphi}$.

Proof. Under the assumptions the projection type subproblem admits a unique solution when starting near $\overline{\varphi}$. Thus the method is well defined. Assume without loss of generality $\varphi_k \neq \overline{\varphi}$ for all k. From the estimate (61) we get

$$\|\varphi_{k+1} - \overline{\varphi}\|_{\mathbb{H}} \le C\frac{\|a_{\varphi_k}(\varphi_k - \overline{\varphi}, .) - j'(\varphi_k) + j'(\overline{\varphi})\|_{\mathbb{H}^*}}{\|\varphi_k - \overline{\varphi}\|_{\mathbb{H}}}\|\varphi_k - \overline{\varphi}\|_{\mathbb{H}}$$

By definition of semismoothness, $\frac{\|a_{\varphi_k}(\varphi_k - \overline{\varphi}, .) - j'(\varphi_k) + j'(\overline{\varphi})\|_{\mathbb{H}^*}}{\|\varphi_k - \overline{\varphi}\|_{\mathbb{H}}}$ can be made arbitrarily small if $\|\varphi_k - \overline{\varphi}\|_{\mathbb{H}}$ is sufficiently small. Thus starting sufficiently close to $\overline{\varphi}$, we get as in the proof of Theorem that 4.39 $\varphi_k \to \overline{\varphi}$, and since $\frac{\|a_{\varphi_k}(\varphi_k - \overline{\varphi}, .) - j'(\varphi_k) + j'(\overline{\varphi})\|_{\mathbb{H}^*}}{\|\varphi_k - \overline{\varphi}\|_{\mathbb{H}}} \to 0$, the convergence rate is q-superlinear. □

In particular, if j is two times continuously differentiable at $\overline{\varphi}$, one can easily show that (62) holds if the inner products fulfill

$$\frac{\|a_{\varphi_k}(\varphi_k - \overline{\varphi}, .) - j''(\overline{\varphi})[\varphi_k - \overline{\varphi}, .]\|_{\mathbb{H}^*}}{\|\varphi_k - \overline{\varphi}\|_{\mathbb{H}}} \to 0, \tag{64}$$

i.e. a_{φ_k} converges to $j''(\overline{\varphi})$ for certain directions. Note that this condition is similar to the well known Dennis-Moré condition (see (65) below) for superlinear convergence of quasi-Newton methods [DM74]. Note that (64) is sufficient but not necessary for superlinear convergence. We refer to [GB84] and [Ber82], where superlinear convergence is shown for a similar method, where the scaling matrix corresponding to a_k equals the Hessian of j only in the linear subspace of active constraints. In the finite dimensional quasi-Newton method in [Rus84] the operators a_{φ_k} and $j''(\overline{\varphi})$ in the condition corresponding to (64) are replaced by a kind of projection of them onto the tangent space.
For a more general treatment of the Josephy-Newton method for semismooth generalized equations in finite dimension we refer to [IKS13]. Necessary and sufficient conditions of Dennis-Moré type for superlinear convergence in a Banach space setting are given in [Don12].

4.8 Quasi-Newton updates

In the previous section we showed that it is possible to take $a_k = j''(\varphi_k)$ as an inner product, which often leads to a good scaling of the problem. We gave sufficient conditions for superlinear convergence of the resulting method. However, in many cases it is not practical to take $a_k = j''(\varphi_k)$, since the solution of the subproblem may be too expensive. For instance in optimal control, the evaluation of the second order derivative of the reduced cost functional involves the solution of PDEs, which can be very expensive, see [Trö09]. In this case it can be better to take only an approximation of $j''(\varphi_k)$ instead of $j''(\varphi_k)$ itself. The resulting methods are then called quasi-Newton methods. Under certain conditions the superlinear rate of convergence can be maintained for quasi-Newton methods. However, we will only cover global convergence here. As already discussed, a sufficient condition on the inner products a_k for superlinear convergence is (62) or (64). Even if no superlinear convergence can be attained, the scaling of the problem by quasi-Newton methods often leads to an efficient method, cf. e.g. the results in Section 6.14.
In the following we concentrate on the BFGS update (see (69) below), which is the most commonly used quasi-Newton method.

Quasi-Newton methods are extensively covered in the literature and many results are available. For unconstrained optimization problems a necessary and sufficient condition for superlinear convergence is the Dennis-Moré condition

$$\frac{\|(B_k - \nabla^2 j(\overline{\varphi}))(\varphi_{k+1} - \varphi_k)\|}{\|\varphi_{k+1} - \varphi_k\|} \to 0, \tag{65}$$

where B_k is the operator approximating the Hessian $\nabla^2 j(\varphi_k)$. This is shown in [DM74] in finite dimension and can be found in [GS81] in Hilbert space. The Banach space case is covered in [Don12]. The finite dimensional BFGS method fulfills (65) if B_0 is sufficiently close to $\nabla^2 j(\overline{\varphi})$ and the initial guess φ_0 is sufficiently close to $\overline{\varphi}$, see [Kel99]. A global convergence result in Hilbert space is given in [GS81] under the assumption

$$\exists C, c > 0: \quad c\|u\|^2 \le j''(\varphi)[u,u] \le C\|u\|^2 \quad \forall \varphi, u. \tag{66}$$

In the special Banach space $L^\infty(0,T)$ convergence is shown in [MQ80] under the same assumption (66) using the L^∞-norm. Note that in finite dimension (66) is sufficient for the BFGS method (together with an efficient step length) to fulfill the Dennis-Moré condition (65), see [BN89]. A global convergence result for a modified BFGS iteration in

finite dimension without assuming the convexity of j can be found in [LF01]. However, for superlinear convergence they still assume (66) near the solution. A mesh independency result in Hilbert spaces can be found in [KS87].

For constrained optimization problems most of the literature deals with the finite dimensional case and little is known in infinite dimensions. For box constrained optimization problems the L-BFGS-B method [BLNZ95] is very popular. In this method the subproblem is only solved approximately, which can be done cheaply using the special structure of the admissible set by a combination of projected gradient method and an unconstrained method. Another quasi-Newton method for constrained optimization is a modification of the SQP method, where a quasi-Newton approximation of the Hessian of the Lagrangian is used [Han76, GPM76, Han77]. We mention this method here, since for linearly constrained problems the BFGS-SQP method coincides with the BFGS method discussed in this section. Superlinear convergence can be shown under standard second order conditions if the initial guess φ_0 and B_0 are close enough to $\overline{\varphi}$ and $\nabla^2 j(\overline{\varphi})$, respectively. For global convergence the condition

$$\exists C, c > 0 : \quad c\|u\|^2 \leq (B_k u, u) \leq C\|u\|^2 \quad \forall k, u. \tag{67}$$

is assumed amongst others. As already mentioned, local convergence theory using the Dennis-Moré condition in Banach space is covered in [Don12]. Of course there are also other methods of quasi-Newton type available which we don't mention here.

The goal of this section is to analyze the BFGS update for the variable metric a_k in our Banach space setting and give sufficient conditions for global convergence of the resulting VMPT method. There are also quasi-Newton updates other than BFGS available. However, for the VMPT method it is necessary that a_k is an inner product, thus the update should maintain positive definiteness. Other positive definite updates are e.g. the PSB or DFP update. We refer to [GS81] for an overview.

In a Hilbert space $\mathbb{H}$ the BFGS operator B_k is a linear operator in $\mathcal{L}(\mathbb{H}, \mathbb{H})$, see [GS81]. Since we cannot use the Riesz representative of $j'(\varphi)$ in the definition of the BFGS update in the Banach space setting, the operator B_k is here rather a bilinear form on $\mathbb{X} \cap \mathbb{D}$, or equivalently an operator in $\mathcal{L}(\mathbb{X} \cap \mathbb{D}, (\mathbb{X} \cap \mathbb{D})^*)$. We will use the latter space because of easier notation. Since the operator B_k is chosen to approximate $j''(\varphi_k)$, one claims the so called quasi-Newton or secant equation

$$B_{k+1}(\varphi_{k+1} - \varphi_k) = j'(\varphi_{k+1}) - j'(\varphi_k), \tag{68}$$

which is fulfilled by $j''(\xi)$ for some ξ between φ_k and φ_{k+1} due to the mean value theorem. The (scaled) BFGS update is defined recursively by

$$B_{k+1} = \rho_k \left(B_k - \frac{(B_k p_k) \otimes (B_k p_k)}{\langle B_k p_k, p_k \rangle} \right) + \frac{y_k \otimes y_k}{\langle y_k, p_k \rangle} \in \mathcal{L}(\mathbb{X} \cap \mathbb{D}, (\mathbb{X} \cap \mathbb{D})^*), \tag{69}$$

with an initialization B_0 and scaling parameters $\rho_k > 0$, $k \in \mathbb{N}_0$,

$$\begin{aligned} p_k &:= \varphi_{k+1} - \varphi_k \in \mathbb{X} \cap \mathbb{D} \text{ and} \\ y_k &:= j'(\varphi_{k+1}) - j'(\varphi_k) \in (\mathbb{X} \cap \mathbb{D})^*. \end{aligned}$$

By $y_k \otimes y_k$ we denote the bilinear form $(u, v) \mapsto \langle y_k, u \rangle \langle y_k, v \rangle$ and the same for $(B_k p_k) \otimes$

$(B_k p_k)$. Note that the operator B_k depends on all previous iterates $\varphi_0, \dots, \varphi_k$ and thus it is not point based. Obviously, the BFGS operator B_k fulfills the quasi-Newton equation (68) for all $k \geq 1$. As already mentioned it is required that B_k is positive definite for all $k \in \mathbb{N}_0$. A necessary condition therefor is the following standard assumption.

(A14) For the iterates $(\varphi_k)_k$ of the method it holds $\langle j'(\varphi_{k+1}) - j'(\varphi_k), \varphi_{k+1} - \varphi_k \rangle > 0$.

The necessity can be seen by (68) and

$$\left\langle j'(\varphi_{k+1}) - j'(\varphi_k), \varphi_{k+1} - \varphi_k \right\rangle = \langle B_{k+1}(\varphi_{k+1} - \varphi_k), \varphi_{k+1} - \varphi_k \rangle > 0$$

if $\varphi_{k+1} \neq \varphi_k$. A sufficient condition for **(A14)** is in turn

$$j''(\varphi)[\varphi_{k+1} - \varphi_k, \varphi_{k+1} - \varphi_k] > 0$$

for all φ on the line segment connecting φ_{k+1} and φ_k, since we have by the fundamental theorem of calculus that

$$\left\langle j'(\varphi_{k+1}) - j'(\varphi_k), \varphi_{k+1} - \varphi_k \right\rangle = \int_0^1 j''(\varphi_k + t(\varphi_{k+1} - \varphi_k))[\varphi_{k+1} - \varphi_k, \varphi_{k+1} - \varphi_k] \, \mathrm{d}t > 0. \tag{70}$$

We will assume **(A14)** throughout the analysis of the BFGS update. However, one cannot expect that **(A14)** holds in general if j is not convex. In this case it is common to skip the update, i.e. set $B_{k+1} = B_k$ (see [Che96]), or take some other positive definite operator for B_{k+1}, e.g. the identity in $\mathbb{R}^n$.

We consider now the VMPT method with the variable metric given by $a_k(u, v) = \langle B_k u, v \rangle$ for $u, v \in \mathbb{X} \cap \mathbb{D}$, $k \in \mathbb{N}_0$. Thus we get global convergence if the sequence of inner products $(a_k)_k$ fulfills the general assumptions **(A8)**-**(A12)**.

Lemma 4.49. *Let **(A1)**-**(A7)** and **(A14)** hold. Let B_0 be given, B_k, $k \in \mathbb{N}$ defined by the BFGS update* (69) *with $\rho_k > 0$ and $a_k(u, v) := \langle B_k u, v \rangle$ for $u, v \in \mathbb{X} \cap \mathbb{D}$, $k \in \mathbb{N}_0$. If a_0 fulfills **(A8)**,**(A10)**,**(A11)** then the whole sequence $(a_k)_k$ fulfills **(A8)**,**(A10)**,**(A11)**.*

Proof. The proof is by induction. a_0 fulfills **(A8)**,**(A10)**,**(A11)** by assumption. Now assume that a_{k-1} fulfills **(A8)**,**(A10)**,**(A11)**. We omit the index $(k-1)$ for ease of notation. By definition we have $a_k(u, v) = \rho\left(\langle Bu, v \rangle - \frac{\langle Bp, u \rangle \langle Bp, v \rangle}{\langle Bp, p \rangle}\right) + \frac{\langle y, u \rangle \langle y, v \rangle}{\langle y, p \rangle}$, thus a_k clearly is a symmetric bilinear form. The proof for positive definiteness is the same as in finite dimension: For $u \in \mathbb{X} \cap \mathbb{D}$, $u \neq 0$, we have

$$a_k(u, u) = \rho\left(\langle Bu, u \rangle - \frac{\langle Bp, u \rangle^2}{\langle Bp, p \rangle}\right) + \frac{\langle y, u \rangle^2}{\langle y, p \rangle}$$

Since, by induction hypothesis, B defines an inner product, it holds the Cauchy-Schwarz inequality $\langle Bp, u \rangle^2 \leq \langle Bp, p \rangle \langle Bu, u \rangle$ with equality if and only if p and u are linearly dependent. As a first case, assume that p and u are linearly independent. Then $\langle Bp, u \rangle^2 < \langle Bp, p \rangle \langle Bu, u \rangle$ and $\frac{\langle y, u \rangle^2}{\langle y, p \rangle} \geq 0$, thus $a_k(u, u) > 0$. Recall that $\langle y, p \rangle > 0$ holds by **(A14)**. If p and u are linearly dependent, then there exists some $\lambda \in \mathbb{R}$, $\lambda \neq 0$, with $u = \lambda p$. Thus,

$$a_k(u, u) = \frac{\langle y, u \rangle^2}{\langle y, p \rangle} = \lambda^2 \langle y, p \rangle > 0$$

and a_k fulfills **(A8)**. The boundedness **(A10)** follows from the estimate

$$a_k(u,v) \le \left(\rho\left(C_{k-1} + \frac{\|Bp\|^2_{(\mathbb{X}\cap\mathbb{D})^*}}{\langle Bp,p\rangle}\right) + \frac{\|y\|^2_{(\mathbb{X}\cap\mathbb{D})^*}}{\langle y,p\rangle}\right)\|u\|_{\mathbb{X}\cap\mathbb{D}}\|v\|_{\mathbb{X}\cap\mathbb{D}},$$

where $C_{k-1} > 0$ is an upper bound for the norm of B_{k-1}. To show **(A11)**, let $\varphi \in \Phi_{ad}$ and $(p_i)_i \subset \mathbb{X}\cap\mathbb{D}$ with $p_i \to 0$ weakly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$. Consider

$$a_k(\varphi,p_i) = \rho\left(\langle B\varphi,p_i\rangle - \frac{\langle Bp,\varphi\rangle\langle Bp,p_i\rangle}{\langle Bp,p\rangle}\right) + \frac{\langle y,\varphi\rangle\langle y,p_i\rangle}{\langle y,p\rangle}.$$

It holds $\langle B\varphi,p_i\rangle \to 0$ by induction hypothesis. Moreover $\langle Bp,p_i\rangle = \langle B\varphi_k,p_i\rangle - \langle B\varphi_{k-1},p_i\rangle \to 0$ as $i\to\infty$. By the assumption **(A7)**, it also holds that $\langle y,p_i\rangle = \langle j'(\varphi_k),p_i\rangle - \langle j'(\varphi_{k-1}),p_i\rangle \to 0$ as $i\to\infty$. Hence **(A11)** is shown. □

For global convergence it remains to show the uniform coercivity **(A9)** and **(A12)**. We note that in Hilbert space one can show coercivity of the operators B_k based on eigenvalue estimates, see [GS81]. However, the coercivity is in general not uniform in k. In the literature a condition like (67) is often assumed (rather than deduced) for quasi-Newton methods to show global convergence, see e.g. [GS81, Kel99, MQ80, HT77, Han77, Rus84, CPR14].
To circumvent this difficulty one can introduce a shift a_s fulfilling **(A8)**-**(A11)** and take

$$a_k(u,v) := \langle B_k u,v\rangle + r_k a_s(u,v) \tag{71}$$

for some $r_k > 0$. If the sequence $(r_k)_k$ is bounded away from zero then the uniform coercivity of a_k follows from the coercivity of a_s and the positivity of B_k. This idea can also be found in [Che96] and similarly in [LF01]. It is notable that superlinear convergence is still possible if $r_k \to 0$. In this case one can choose r_k to be bounded away from zero to get global convergence, and as soon as φ_k is near a local minimizer one can let $r_k \to 0$ to get fast convergence.

Finally, we give sufficient conditions for assumption **(A12)**. Therefor we need stronger assumptions similar to Theorem 4.39. In particular we assume the standard condition (66).

Lemma 4.50. *In addition to the assumptions of Lemma 4.49 let j be two times Fréchet differentiable in an open neighborhood of $\Phi_{ad} \subset \mathbb{X}\cap\mathbb{D}$. Let there exist positive constants m, M, such that*

$$m\|u\|^2_{\mathbb{X}} \le j''(\varphi)[u,u] \le M\|u\|^2_{\mathbb{X}} \tag{72}$$

holds for all $\varphi \in \Phi_{ad}$ and $u \in \mathbb{X}\cap\mathbb{D}$, and let there exist some $\overline{\rho}$, such that $0 < \rho_k \le \overline{\rho} < 1$ for all $k \in \mathbb{N}_0$. Moreover, let a_0 fulfill

$$a_0(u,u) \le C_0\|u\|^2_{\mathbb{X}} \quad \forall u \in \mathbb{X}\cap\mathbb{D}.$$

Then it holds

$$a_k(u,u) \le \max\left\{C_0, \frac{M^2}{m(1-\overline{\rho})}\right\}\|u\|^2_{\mathbb{X}} \quad \forall k \in \mathbb{N}_0,\ u \in \mathbb{X}\cap\mathbb{D}. \tag{73}$$

*and **(A12)** is fulfilled.*

Proof. We first show that $\frac{\langle y_k,u\rangle^2}{\langle y_k,p_k\rangle} \le \frac{M^2}{m}\|u\|_{\mathbb{X}}^2$ for all $u \in \mathbb{X}\cap\mathbb{D}$ and $k \in \mathbb{N}_0$. Note that (72) induces $j''(\varphi)[u,v] \le M\|u\|_{\mathbb{X}}\|v\|_{\mathbb{X}}$ for all $u,v \in \mathbb{X}\cap\mathbb{D}$, see Remark 4.3. By the fundamental theorem of calculus we have

$$\begin{aligned}\langle y_k,u\rangle = \left\langle j'(\varphi_{k+1}) - j'(\varphi_k), u\right\rangle = \int_0^1 j''(\varphi_k + t(\varphi_{k+1}-\varphi_k))[\varphi_{k+1}-\varphi_k,u]\,\mathrm{d}t \le \\ \le M\|\varphi_{k+1}-\varphi_k\|_{\mathbb{X}}\|u\|_{\mathbb{X}} = M\|p_k\|_{\mathbb{X}}\|u\|_{\mathbb{X}}\end{aligned}$$

and analogously

$$\langle y_k,p_k\rangle \ge m\|p_k\|_{\mathbb{X}}^2.$$

Now let $k \ge 1$. We get the estimate

$$a_k(u,u) = \rho_{k-1}\left(\underbrace{\langle B_{k-1}u,u\rangle}_{=a_{k-1}(u,u)} - \underbrace{\frac{\langle B_{k-1}p_{k-1},u\rangle^2}{\langle B_{k-1}p_{k-1},p_{k-1}\rangle}}_{\ge 0}\right) + \underbrace{\frac{\langle y_{k-1},u\rangle^2}{\langle y_{k-1},p_{k-1}\rangle}}_{\le \frac{M^2}{m}\|u\|_{\mathbb{X}}^2} \le \left(\overline{\rho}\|a_{k-1}\|_{\mathcal{L}(\mathbb{X},\mathbb{X}^*)} + \frac{M^2}{m}\right)\|u\|_{\mathbb{X}}^2$$

By induction we get

$$a_k(u,u) \le g^k(C_0)\|u\|_{\mathbb{X}}^2$$

with $g(x) := \overline{\rho}x + \frac{M^2}{m}$. Note that g is a contraction with fixed point $\frac{M^2}{m(1-\overline{\rho})}$. Thus, by the contraction mapping principle, the fixed point iteration $g^k(C_0)$ converges to the fixed point. Since we have $0 < g'(x) < 1$, the fixed point is strongly attractive and $g^k(C_0)$ converges monotonic (increasing if $C_0 < \frac{M^2}{m(1-\overline{\rho})}$ and decreasing if $C_0 > \frac{M^2}{m(1-\overline{\rho})}$), see e.g. [Bal99, Sec. 6.10]. Thus, $g^k(C_0) \le \max\left\{C_0, \frac{M^2}{m(1-\overline{\rho})}\right\}$, which yields (73).
As already discussed in Lemma 4.19, **(A12)** follows from (73) and **(A9)**. □

By combining Lemma 4.49, Lemma 4.50 and the shift (71) for $(r_k)_k$ uniformly bounded from above and away from zero, one can apply Theorem 4.14 and gets global convergence of the BFGS-VMPT method.
We note that global convergence of the unconstrained BFGS method in Hilbert space can be shown without the uniform coercivity assumption **(A9)**, using only (66), see [GS81]. This is due to the special structure of the BFGS method and the proof differs considerably from the global convergence proof for the VMPT method.

4.9 Discussion of the projection type subproblem

The VMPT method replaces the original optimization problem by a sequence of subproblems which should be easier to solve. At least the structure of the subproblem is favorable, since it is a strictly convex quadratic program. In general the same considerations about solvers for the subproblem and approximation errors apply to the projected gradient method as well as to the VMPT method. We discuss these in the following and finally explain a certain similarity of the subproblem to the generalized projection in Banach space.

As for the scaled projected gradient method it is important to have a good solver for the projection type subproblem. If the solution of the subproblem is as expensive as the solution of the overall optimization problem itself the application of the VMPT method

does not make sense.
For instance it is well known that L^2 projections on box constraints can be calculated pointwise, see [Trö09], which is very cheap. Similarly, the Euclidean projection on box constraints can be calculated coordinate-wise. Also projections onto special geometries like balls can be calculated easily, even on mixed p, q-balls, see [SvdBFM09] and the references therein. For $\Phi_{ad} = \{x \in \mathbb{R}^n \mid b^T x = r,\ l \le x \le u\}$ in finite dimension with $b, l, u \in \mathbb{R}^n$, $r \in \mathbb{R}$, the Euclidean projection is a continuous quadratic knapsack problem, for which efficient solvers like breakpoint searching, variable fixing, iterative projections, bracketing and bisection algorithms are available, see [Tav15] and references therein. If an additional sum constraint is present an alternating projection method for the computation of the Euclidean projection is given in [Tav15]. Note that the effort of these methods scale linearly in the number of variables.
Since the subproblem is equivalent to a linear coercive variational inequality, any numerical method for the solution thereof can be used, including multigrid methods [Kor94]. If no specialized solver for the projection type subproblem is available, then any black-box QP solver can be used. For instance we will use a primal dual active set method in Section 6.10. It is even possible to use solvers which converge only locally. In Corollary 4.33 we proved that $\mathcal{P}_{a,\lambda}(\varphi) \to \varphi$ in $\mathbb{X}$ as $\lambda \to 0$. This implies that φ can be used as an initial guess, which is arbitrarily near the solution $\mathcal{P}_{a,\lambda}(\varphi)$ if λ is chosen sufficiently small. Thus, by choosing λ small enough convergence can be obtained. Therefore it is also meaningful to use the VMPT method even if j itself is quadratic and convex. In this case the structure of the subproblem is the same as for the overall optimization problem, but the parameter λ can be controlled to guarantee convergence of local methods. In this sense the VMPT method can be used as a globalization of local methods. This idea is similar to proximal point methods [Roc76].

We note that a possible choice for the parameter λ_k can be obtained by the ideas of Barzilai and Borwein [BB88]. They choose the step length λ such that the quasi-Newton equation (68) is fulfilled in some direction. For the usual gradient method in finite dimension this amounts to $\lambda_{k+1} = (p_k, y_k)/\|y_k\|^2$ or $\lambda_{k+1} = \|p_k\|^2/(p_k, y_k)$ with $p_k := \varphi_{k+1} - \varphi_k$ and $y_k = \nabla j(\varphi_{k+1}) - \nabla j(\varphi_k)$. To derive a similar step length we first recall that the VMPT method using the metric a_k and scaling λ_k is equivalent to the VMPT method using the metric a_k/λ_k and scaling $\lambda_k = 1$. Thus we can use the ansatz $B_k = \frac{1}{\lambda_k} a_k$ for the quasi-Newton operator. The quasi-Newton equation (68) then reads

$$a_k(p_{k-1}, \eta) = \lambda_k \langle y_{k-1}, \eta \rangle \quad \forall \eta \in \mathbb{X} \cap \mathbb{D}.$$

The single unknown λ_k is determined by a concrete choice for η. If we take $\eta = p_{k-1}$ we get

$$\lambda_k = \frac{a_k(p_{k-1}, p_{k-1})}{\langle y_{k-1}, p_{k-1} \rangle},$$

which is analog to the second Barzilai-Borwein step length using a variable metric a_k. The first Barzilai-Borwein step length corresponds to the choice $\eta = y_{k-1}$. However, this is not possible here, since we have in our Banach space setting $y_k \in (\mathbb{X} \cap \mathbb{D})^*$ and $p_k \in \mathbb{X} \cap \mathbb{D}$, i.e. the spaces don't match. Note that one has in addition to ensure that the assumption **(A13)** holds. In general the optimal parameter λ_k is problem specific. In Section 6.12 we derive a scaling λ_k which performs well for the concretely given problem.

In practice the subproblem is often not solved exactly, but probably by a truncated iterative method. Thus approximation errors are involved. In this case the statements 1 and 2 of the global convergence Theorem 4.14 stay true if the search direction v_k is still a gradient related descent direction in the sense of Lemma 4.12 and if it holds $\varphi_k + v_k \in \Phi_{ad}$. The latter is a serious restriction if the solution of the subproblem is approximated from the exterior of Φ_{ad}, e.g. by using active-set or exterior penalty methods. Thus, interior-point methods can be advantageous when considering inexact methods. It is necessary that $\varphi_k + v_k \in \Phi_{ad}$ holds since otherwise φ_{k+1} may be unfeasible and thus v_{k+1} may not be a descent direction. Recall that we proved the descent property of v_{k+1} in Lemma 4.9 by testing the variational inequality (26) of the $(k+1)$th step by $\eta = \varphi_{k+1}$. If φ_{k+1} is not feasible this is not possible anymore. In the numerical experiments in the second part of the thesis the line search fails only if the iterates are close to the solution, see Section 6.11. Away from the solution the approximation errors play a minor role and the used active-set method works fine.
Error analysis for a general class of feasible descent methods in finite dimension can be found in [LT93], where errors e_k in the evaluation of the gradient $\nabla j(\varphi_k)$ are considered. However, the projection has to be evaluated exactly to obtain the feasibility $\varphi_k + v_k \in \Phi_{ad}$. Global convergence can be shown under the summability condition $\sum_{k=1}^{\infty} \|e_k\| < \infty$. Under stronger conditions on e_k a linear rate of convergence can be shown. Error analysis in the more general context of operator splitting methods is performed in [CPR14] and [CV14]. In the latter the feasibility condition $\varphi_k + v_k \in \Phi_{ad}$ is not needed. The analysis applies to convex functionals j in a Hilbert space where λ is chosen small enough and $\alpha_k = 1$ is used (i.e. no Armijo backtracking is considered). Convergence can be shown if the errors in the gradient $\nabla j(\varphi_k)$ and in the projection are absolutely summable.

The projection type subproblem is a generalization of the operation $\varphi \mapsto P_{\perp}(\varphi - \lambda \nabla_{\mathbb{H}} j(\varphi))$ in Hilbert spaces. As already mentioned, orthogonal projections don't exist in Banach spaces. However, if the Banach space X is uniformly convex and uniformly smooth a generalized projection $\pi_{\Phi_{ad}} : X^* \to X$ exists, see [Alb96]. The projection $y = \pi_{\Phi_{ad}}(z)$ for $z \in X^*$ is defined as solution of the minimization problem

$$\min_{y \in \Phi_{ad}} \|z\|_{X^*}^2 - 2\langle z, y \rangle + \|y\|_X^2.$$

This notion is extended to reflexive Banach spaces in [Li05], where the projection is in general not unique anymore. If we now choose the normed space $X = (\mathbb{X} \cap \mathbb{D}, \|.\|_a)$ and the functional $\langle z, \eta \rangle = a(\varphi, \eta) - \lambda \langle j'(\varphi), \eta \rangle$ for all $\eta \in \mathbb{X} \cap \mathbb{D}$, then we get the formal coincidence $\pi_{\Phi_{ad}}(z) = \mathcal{P}_{a,\lambda}(\varphi)$. Thus, the projection type subproblem can be seen as generalized projection with respect to the a-norm, if we identify $\varphi \in \mathbb{X} \cap \mathbb{D}$ with the functional $a(\varphi, .) \in (\mathbb{X} \cap \mathbb{D})^*$. However, this coincidence is only formally, since the space $X = (\mathbb{X} \cap \mathbb{D}, \|.\|_a)$ is in general not needed to be complete, reflexive, uniformly convex nor uniformly smooth. Moreover, we don't assume that $j'(\varphi) \in X^*$, i.e. that $j'(\varphi)$ is bounded in the a-norm, but only in the stronger $\mathbb{X} \cap \mathbb{D}$-norm.

4.10 Generalization to the minimization of the sum of a smooth and a nonsmooth convex functional

Until now we considered the convexly constrained optimization problem

$$\min_{\varphi \in \Phi_{ad}} j(\varphi).$$

By introducing the indicator function $\chi_{\Phi_{ad}} : \mathbb{X} \cap \mathbb{D} \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$, of the convex set Φ_{ad}, defined by

$$\chi_{\Phi_{ad}}(\varphi) = \begin{cases} 0 & \varphi \in \Phi_{ad} \\ \infty & \varphi \notin \Phi_{ad} \end{cases},$$

the optimization problem is equivalent to

$$\min_{\varphi \in \mathbb{X} \cap \mathbb{D}} j(\varphi) + \chi_{\Phi_{ad}}(\varphi).$$

Note that $\chi_{\Phi_{ad}}$ is convex and lower semi-continuous (l.s.c.) if $\Phi_{ad} \subset \mathbb{X} \cap \mathbb{D}$ is convex and closed. The idea is now to generalize the VMPT method by replacing $\chi_{\Phi_{ad}}$ by some arbitrary convex l.s.c. function, which is not necessarily the indicator function of some convex set. The generalized optimization problem we will consider in this section is

$$\min_{\varphi \in \mathbb{X} \cap \mathbb{D}} j(\varphi) + g(\varphi),$$

where $j : \mathbb{X} \cap \mathbb{D} \to \mathbb{R}$ is a smooth function and $g : \mathbb{X} \cap \mathbb{D} \to \overline{\mathbb{R}}$ is a convex nonsmooth function. Such problems arise e.g. in inverse problems, where g can be the nonsmooth L^1-norm or the total variation norm, see [Bre09]. In [Bre09] a numerical method is analyzed to solve such problems in Banach space for convex $(j + g)$. It turns out that there are certain similarities to the VMPT method, see Section 4.11.2. However, in [Bre09] no variable metric is allowed and j is assumed to be convex, which we do not assume here.

We use the following definitions and results, which can be found in [ET99]. A convex function $g : X \to \overline{\mathbb{R}}$ is called *proper*, if it nowhere takes the value $-\infty$ and is not identically equal to $+\infty$. The set $\operatorname{dom}(g) := \{\varphi \mid g(\varphi) < \infty\}$ is called the *effective domain* of g. For convex g also $\operatorname{dom}(g)$ is convex. We call $\varphi^* \in X^*$ *subgradient* of g at $\varphi \in X$ if $g(\varphi)$ is finite and

$$\langle \eta - \varphi, \varphi^* \rangle + g(\varphi) \leq g(\eta) \quad \forall \eta \in X.$$

The set of subgradients at φ is called *subdifferential* and denoted by $\partial g(\varphi) \subset X^*$. From the definition it directly follows that $\varphi \in X$ is a global minimizer of some convex proper functional g if and only if $0 \in \partial g(\varphi)$. If a convex function g is Gâteaux differentiable at $\varphi \in X$, we have $\partial g(\varphi) = \{g'(\varphi)\}$.

Analog to the smooth VMPT method we solve the following subproblem in each VMPT step.

$$\min_{y \in \mathbb{X} \cap \mathbb{D}} \quad \frac{1}{2}\|y - \varphi_k\|_{a_k}^2 + \lambda_k \left\langle j'(\varphi_k), y - \varphi_k \right\rangle + \lambda_k g(y) \tag{74}$$

If g is the indicator function of some convex set we recover the original subproblem (18). We denote the solution of (74) (leaving away the index k) by $\mathcal{P}_{a,\lambda}(\varphi)$ and let $\mathcal{P}_k := \mathcal{P}_{a_k,\lambda_k}$ as in the smooth case. The generalization of the VMPT method to the nonsmooth setting is given in Algorithm 4.4. The differences to the smooth algorithm is that Φ_{ad} is replaced by $\operatorname{dom}(g)$ and the Armijo condition is adapted appropriately. If g is the indicator function of some convex set and φ_k, $\varphi_k + v_k$ are feasible, then the original Armijo condition (20) is recovered.

Algorithm 4.4 VMPT method for nonsmooth functionals with line search

Choose $0 < \beta < 1$, $0 < \sigma < 1$ and $\varphi_0 \in \operatorname{dom}(g)$.
$k := 0$
while $k \leq k_{max}$ **do**
  Calculate the minimum $y_k = \mathcal{P}_k(\varphi_k)$ of the subproblem (74).
  Set the search direction $v_k := y_k - \varphi_k$
  if $\|v_k\|_{\mathbb{X}} \leq tol$ **then**
    return
  end if
  Determine the step length $\alpha_k := \beta^{m_k}$ with minimal $m_k \in \mathbb{N}_0$ such that the following Armijo type condition is fulfilled

$$(j+g)(\varphi_k + \alpha_k v_k) \leq (j+g)(\varphi_k) + \alpha_k \sigma(\langle j'(\varphi_k), v_k \rangle + g(\varphi_k + v_k) - g(\varphi_k)). \tag{75}$$

  Update $\varphi_{k+1} := \varphi_k + \alpha_k v_k$
  $k := k + 1$
end while

We will prove global convergence of the generalized method in the following. However, many steps in the proof are analog to the smooth case. Thus we will sketch the analog arguments of the proofs only briefly. We note that global convergence of a similar method (the variable metric forward-backward algorithm) is shown e.g. in [CPR14] under different assumptions. However, the analysis is carried out in finite dimension, whereas we allow Banach spaces here. We refer to Section 4.11.2 for an overview of existing results concerning forward-backward algorithms.
For the global convergence result we use the following assumptions.

(AG1) $\mathbb{X}$ is a real reflexive Banach space. $\mathbb{B}$ is a separable real Banach space and $\mathbb{D}$ is a real Banach space which is isometrically isomorphic to $\mathbb{B}^*$. Moreover, for any sequence $(\varphi_i)_i$ in $\mathbb{X} \cap \mathbb{D}$ with $\varphi_i \to \varphi$ weakly in $\mathbb{X}$ and $\varphi_i \to \overline{\varphi}$ weakly-* in $\mathbb{D}$ for some $\varphi \in \mathbb{X}$, $\overline{\varphi} \in \mathbb{D}$, it holds $\varphi = \overline{\varphi}$.

(AG2) $g : \mathbb{X} \cap \mathbb{D} \to \overline{\mathbb{R}}$ is convex, l.s.c., proper and the following weak lower semi-continuity holds: Let $(\varphi_i)_i \subset \operatorname{dom}(g)$ with $\varphi_i \to \varphi$ weakly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$ for some $\varphi \in \mathbb{X} \cap \mathbb{D}$ and let $g(\varphi_i)$ be uniformly bounded. Then $\liminf_i g(\varphi_i) \geq g(\varphi)$.

(AG3) g grows faster than $\|\varphi\|_{\mathbb{D}}$, i.e. $\frac{g(\varphi)}{\|\varphi\|_{\mathbb{D}}} \to \infty$ as $\|\varphi\|_{\mathbb{D}} \to \infty$.

(AG4) $j : \mathbb{X} \cap \mathbb{D} \to \mathbb{R}$ is continuously differentiable in a neighborhood of $\operatorname{dom}(g) \subset \mathbb{X} \cap \mathbb{D}$.

(AG5) $j(\varphi) \geq -C$ and $g(\varphi) \geq -C$ for some $C > 0$ and all $\varphi \in \operatorname{dom}(g)$.

(AG6) For each $\varphi \in \operatorname{dom}(g)$ and for each sequence $(\varphi_i)_i \subseteq \mathbb{X} \cap \mathbb{D}$ with $\varphi_i \to 0$ weakly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$ it holds $\langle j'(\varphi), \varphi_i \rangle \to 0$ as $i \to \infty$.

Moreover, we request for the parameters a_k and λ_k of the algorithm that:

(AG7) $(a_k)_k$ is a sequence of inner products on $\mathbb{X} \cap \mathbb{D}$.

(AG8) There exists $c_1 > 0$ such that $c_1 \|u\|_{\mathbb{X}}^2 \leq a_k(u,u)$ for all $u \in \mathbb{X} \cap \mathbb{D}$ and $k \in \mathbb{N}_0$.

(AG9) For all $k \in \mathbb{N}_0$ there exists $c_2(k)$ such that $a_k(p,v) \leq c_2(k) \|p\|_{\mathbb{X}\cap\mathbb{D}} \|v\|_{\mathbb{X}\cap\mathbb{D}}$ for all $p, v \in \mathbb{X} \cap \mathbb{D}$.

(AG10) For all $k \in \mathbb{N}_0$, $p \in \mathrm{dom}(g)$ and for each sequence $(y_i)_i \subseteq \mathrm{dom}(g)$ where there exists some $y \in \mathbb{X} \cap \mathbb{D}$ with $y_i \to y$ weakly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$ it holds $a_k(p, y_i) \to a_k(p, y)$ as $i \to \infty$.

(AG11) For each subsequence $(\varphi_{k_i})_i$ of the iterates given by Algorithm 4.4, which converges in $\mathbb{X} \cap \mathbb{D}$, the corresponding subsequence $(a_{k_i})_i$ has the property that $a_{k_i}(p_i, y_i) \to 0$ for any sequences $(p_i)_i, (y_i)_i \subseteq \mathbb{X} \cap \mathbb{D}$ with $p_i \to 0$ strongly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$ and $(y_i)_i$ converging in $\mathbb{X} \cap \mathbb{D}$.

(AG12) There exist $\lambda_{min}, \lambda_{max}$, s.t. $0 < \lambda_{min} \leq \lambda_k \leq \lambda_{max}$ for all $k \in \mathbb{N}_0$.

We note that $\mathbb{D}$ may also be a reflexive Banach space rather than a dual space. Then all results still hold in the sense of Theorem 4.18.
The assumptions above will replace the standard assumptions throughout Section 4.10.

If Φ_{ad} is non-empty, convex and closed in $\mathbb{X}$, then $g = \chi_{\Phi_{ad}}$ fulfills **(AG2)**. Thus, **(AG2)** is a relaxation of **(A2)**-**(A3)**. Moreover, if Φ_{ad} is bounded in $\mathbb{D}$, then $g = \chi_{\Phi_{ad}}$ fulfills **(AG3)**, which is therefore a relaxation of **(A4)**. Thus, assumptions **(A1)**-**(A13)** imply **(AG1)**-**(AG12)**, and therefore the nonsmooth VMPT method considered in this section is a proper generalization of its smooth counterpart.

Examples for g fulfilling **(AG2)** are $g(\varphi) = \|\varphi\|_{L^p(\Omega)}$ for $1 \leq p \leq \infty$, where $\Omega \subset \mathbb{R}^d$ is a bounded domain. Note that any norm is convex, continuous and thus l.s.c. and weakly-l.s.c. Moreover, norms in dual spaces are weakly-*-l.s.c. [Bre11, Prop. 3.13]. In the case $1 < p < \infty$ one can thus choose $\mathbb{D} = L^p(\Omega)$, which is reflexive. For $p = 1$ one can take $\mathbb{D} = L^{1+\varepsilon}(\Omega)$ for some $\varepsilon > 0$, which is again reflexive and for $p = \infty$ a possible choice is $\mathbb{D} = L^\infty(\Omega) \cong L^1(\Omega)^*$.

The growth condition **(AG3)** is surely not fulfilled for applications where $g(\varphi) = \|\varphi\|_{\mathbb{D}}$ is taken. However, the assumption is needed to show the existence of minimizers for the subproblem (74). Otherwise the subproblem may not be bounded from below. The reason is roughly speaking that the j' term in the subproblem decreases linearly as $\|y\|_{\mathbb{D}} \to \infty$ and thus the g term has to grow faster in order to get a lower bound. If j' is bounded in $\mathrm{dom}(g)$ then **(AG3)** can be weakened to

$$\exists M, \varepsilon > 0 \ \forall \varphi \in \mathbb{X} \cap \mathbb{D}, \ \|\varphi\|_{\mathbb{D}} \geq M : \quad \frac{g(\varphi)}{\|\varphi\|_{\mathbb{D}}} \geq \sup_{y \in \mathrm{dom}(g)} \|j'(y)\|_{(\mathbb{X}\cap\mathbb{D})^*} + \varepsilon,$$

thus in this case it is possible to take $g(\varphi) = C\|\varphi\|_{\mathbb{D}}$ for C large enough. In [Bre09] such an assumption is not needed since a stronger coercivity of the term corresponding to $\|y - \varphi_k\|_{a_k}^2$ in (74) is assumed, respectively a stronger differentiability condition for j.
In the case that j' is unbounded a possibility to define g fulfilling **(AG3)** is to take $g(\varphi) = \|\varphi\|_{\mathbb{D}}^{1+\varepsilon}$ for some $\varepsilon > 0$ instead of $g(\varphi) = \|\varphi\|_{\mathbb{D}}$. Another workaround is to take $g(\varphi) = \|\varphi\|_{\mathbb{D}}$ and introduce additional artificial constraints: If the lower bound C_0 on j is explicitly known, then a minimizer φ fulfills $(j + g)(\varphi) \leq (j + g)(\varphi_0)$ for any $\varphi_0 \in \mathrm{dom}(g)$, hence $\|\varphi\|_{\mathbb{D}} \leq (j + g)(\varphi_0) - C_0 := M$. One can thus introduce the artificial constraint $\|\varphi\|_{\mathbb{D}} \leq rM$ with $r \gg 1$ and append the corresponding indicator function to g, which then fulfills **(AG3)**.
Recall that in the smooth VMPT method the boundedness of Φ_{ad} is not needed in case of a Hilbert space setting. An analog observation holds also for the nonsmooth VMPT method: If $\mathbb{X} = \mathbb{D} = \mathbb{H}$ is a Hilbert space then the growth condition **(AG3)** can be dropped. An example therefor are sparse optimal control problems where j is differentiable in $\mathbb{H} = L^2(\Omega)$

and $g(\varphi) = c\|\varphi\|_{L^1}$ for some $c > 0$.

The first theorem proves that the operator $\mathcal{P}_{a,\lambda}$ is well defined.

Theorem 4.51. *Let a and λ fulfill the assumptions. Then the operator $\mathcal{P}_{a,\lambda} : \mathrm{dom}(g) \to \mathrm{dom}(g)$ is well defined, i.e. the corresponding subproblem (see (74)) with $\varphi \in \mathrm{dom}(g)$ is uniquely solvable. Moreover, $y = \mathcal{P}_{a,\lambda}(\varphi)$ is given as unique solution of the inclusion*

$$-a(y-\varphi,.) - \lambda j'(\varphi) \in \lambda \partial g(y) \quad \text{in } (\mathbb{X} \cap \mathbb{D})^*,$$

or equivalently of the inequality

$$-a(y-\varphi, \eta - y) - \lambda \left\langle j'(\varphi), \eta - y \right\rangle + \lambda g(y) \le \lambda g(\eta) \quad \forall \eta \in \mathbb{X} \cap \mathbb{D}. \tag{76}$$

Proof. Let

$$h(y) = \frac{1}{2}\|y-\varphi\|_a^2 + \lambda \left\langle j'(\varphi), y - \varphi \right\rangle + \lambda g(y)$$

be the functional of the corresponding subproblem. Then it holds

$$\begin{aligned} h(y) &\ge \frac{c_1}{2}\|y-\varphi\|_{\mathbb{X}}^2 - \lambda \|j'(\varphi)\|_{(\mathbb{X}\cap\mathbb{D})^*}(\|y-\varphi\|_{\mathbb{X}} + \|y-\varphi\|_{\mathbb{D}}) + \lambda g(y) \\ &\ge \lambda g(y) - C\|y-\varphi\|_{\mathbb{D}} - C \end{aligned} \tag{77}$$

using that $C\|y-\varphi\|_{\mathbb{X}}^2 - \tilde{C}\|y-\varphi\|_{\mathbb{X}}$ is bounded from below. Let $M > 0$ arbitrary. Then there exists $C_M > 0$ such that for all $y \in \mathbb{X} \cap \mathbb{D}$ with $\|y\|_{\mathbb{D}} \le M$ it holds $h(y) \ge -C_M$. If $\|y\|_{\mathbb{D}} > M$ we get for M large enough

$$\frac{h(y)}{\|y\|_{\mathbb{D}}} \ge \lambda \underbrace{\frac{g(y)}{\|y\|_{\mathbb{D}}}}_{\to\infty \text{ as } M\to\infty} - \underbrace{\left(C\frac{\|y\|_{\mathbb{D}} + \|\varphi\|_{\mathbb{D}}}{\|y\|_{\mathbb{D}}} + \frac{C}{\|y\|_{\mathbb{D}}} \right)}_{\le C \text{ for } M \ge M_0 > 0} \ge 0. \tag{78}$$

Hence h is bounded from below. We take a minimizing sequence $(y_i)_i \subseteq \mathrm{dom}(g)$ with $h(y_i) \to \inf_y h(y) > -\infty$. From estimate (78) we get that $\|y_i\|_{\mathbb{D}}$ is uniformly bounded. Thus (77) yields $h(y) \ge C\|y-\varphi\|_{\mathbb{X}}^2 - \tilde{C}\|y-\varphi\|_{\mathbb{X}} - \tilde{C}$ and therefore $\|y_i\|_{\mathbb{X}}$ is uniformly bounded. We extract a subsequence with $y_i \to y$ weakly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$ for some $y \in \mathbb{X} \cap \mathbb{D}$. Note that $g(y_i)$ is also uniformly bounded, since $h(y_i)$ is uniformly bounded and estimate (77) holds. We use the lower semi-continuity of g and a (see proof of Lemma 4.6) to obtain

$$\inf_y h(y) = \liminf_i h(y_i) \ge h(y) \ge \inf_y h(y).$$

Thus y is a global minimizer of h. From strict convexity of h we get that the minimizer is unique. Moreover, y can be characterized by $0 \in \partial h(y)$ and it holds $\partial h(y) = a(y-\varphi,.) + \lambda j'(\varphi) + \lambda \partial g(y)$, see [ET99], which yields (76). Note that the equivalence of the inclusion and the inequality is due to the definition of the subdifferential. □

Corollary 4.52. *Let $(\varphi_k)_k$ be the sequence generated by Algorithm 4.4. Then $y_k := \mathcal{P}_k(\varphi_k)$ is given as the unique solution of the inequality*

$$-a_k(y_k - \varphi_k, \eta - y_k) - \lambda_k \left\langle j'(\varphi_k), \eta - y_k \right\rangle + \lambda_k g(y_k) \le \lambda_k g(\eta) \quad \forall \eta \in \mathbb{X} \cap \mathbb{D}. \tag{79}$$

Definition 4.53. We call $\varphi \in \mathrm{dom}(g)$ a stationary point of $j+g$ if and only if

$$g(\eta) - g(\varphi) + \langle j'(\varphi), \eta - \varphi \rangle \geq 0 \quad \forall \eta \in \mathbb{X} \cap \mathbb{D}$$

or equivalently

$$-j'(\varphi) \in \partial g(\varphi).$$

It is easy to prove that this is a first order necessary condition for a local minimizer φ, see e.g [lR15]. If j is convex, this condition is also sufficient, since we have then $\partial(j+g)(\varphi) = \partial j(\varphi) + \partial g(\varphi) = j'(\varphi) + \partial g(\varphi)$ [ET99].

Lemma 4.54. *Let a and λ fulfill the assumptions. Then $\varphi \in \mathrm{dom}(g)$ is a stationary point of $j+g$ if and only if $P_{a,\lambda}(\varphi) = \varphi$.*

Proof. Per definition φ is stationary if and only if

$$g(\eta) - g(\varphi) + \langle j'(\varphi), \eta - \varphi \rangle \geq 0 \quad \forall \eta \in \mathbb{X} \cap \mathbb{D},$$

which is due to (76) equivalent to $P_{a,\lambda}(\varphi) = \varphi$. □

Lemma 4.55. *Let $(\varphi_k)_k$ be the sequence generated by Algorithm 4.4 and let $v_k := \mathcal{P}_k(\varphi_k) - \varphi_k$ as in the algorithm. Then it holds for all $k \in \mathbb{N}_0$*

$$g(\varphi_k + v_k) - g(\varphi_k) + \langle j'(\varphi_k), v_k \rangle \leq -\frac{1}{\lambda_k} \|v_k\|_{a_k}^2. \tag{80}$$

Proof. Test (79) by $\eta = \varphi_k$ and rearrange terms. □

Lemma 4.56. *Let φ_k and v_k as in Algorithm 4.4. If $v_k \neq 0$, then there exists $\bar{\alpha} > 0$ such that the Armijo condition is fulfilled for all $0 < \alpha \leq \bar{\alpha}$.*

Proof. It holds for all $0 < \alpha \leq 1$ using the convexity of g

$$\begin{aligned}\frac{1}{\alpha}\left((j+g)(\varphi_k + \alpha v_k) - (j+g)(\varphi_k) - \alpha\sigma\left(\langle j'(\varphi_k), v_k\rangle + g(\varphi_k + v_k) - g(\varphi_k)\right)\right)\\ \leq (1-\sigma)(g(\varphi_k + v_k) - g(\varphi_k)) + \frac{1}{\alpha}(j(\varphi_k + \alpha v_k) - j(\varphi_k)) - \sigma\langle j'(\varphi_k), v_k\rangle\end{aligned}$$

Letting $\alpha \to 0$, the right hand side converges to

$$(1-\sigma)\left(g(\varphi_k + v_k) - g(\varphi_k) + \langle j'(\varphi_k), v_k\rangle\right) < 0$$

by (80). Thus, there exists $\bar{\alpha} > 0$ such that

$$(j+g)(\varphi_k + \alpha v_k) - (j+g)(\varphi_k) - \alpha\sigma\left(\langle j'(\varphi_k), v_k\rangle + g(\varphi_k + v_k) - g(\varphi_k)\right) < 0$$

for all $0 < \alpha \leq \bar{\alpha}$. □

Lemma 4.57. *Let for a sequence $(\varphi_i)_i \subseteq \mathrm{dom}(g)$ hold $\varphi_i \to \varphi$ in $\mathbb{X} \cap \mathbb{D}$ for some $\varphi \in \mathbb{X} \cap \mathbb{D}$ and $g(\varphi_i)$ uniformly bounded. Then there exists $C > 0$ such that $\|\mathcal{P}_k(\varphi_i)\|_{\mathbb{X}\cap\mathbb{D}} \leq C$ and $g(\mathcal{P}_k(\varphi_i)) \leq C$ for all $i, k \in \mathbb{N}_0$.*

Proof. As above it can be shown that the inequality (80) holds also for φ_i and $v_i := \mathcal{P}_k(\varphi_i) - \varphi_i$. Using the coercivity of a_k and the boundedness of $(j'(\varphi_i))_i$ we obtain

$$\begin{aligned}\frac{c_1}{\lambda_{max}}\|v_i\|_{\mathbb{X}}^2 &\leq g(\varphi_i) - g(\varphi_i + v_i) - \langle j'(\varphi_i), v_i\rangle \\ &\leq g(\varphi_i) - g(\varphi_i + v_i) + C(\|v_i\|_{\mathbb{X}} + \|v_i\|_{\mathbb{D}}). \end{aligned} \tag{81}$$

Without loss of generality assume $\varphi_i + v_i = \mathcal{P}_k(\varphi_i) \neq 0$ for all k, i (for the other indices the statement is trivial). We divide the inequality by $\|\varphi_i + v_i\|_{\mathbb{D}}$ and use the uniform boundedness of $g(\varphi_i)$ to get

$$-\frac{C}{\|\varphi_i + v_i\|_{\mathbb{D}}} \leq \frac{\frac{c_1}{\lambda_{max}}\|v_i\|_{\mathbb{X}}^2 - C - C\|v_i\|_{\mathbb{X}}}{\|\varphi_i + v_i\|_{\mathbb{D}}} \leq -\frac{g(\varphi_i + v_i)}{\|\varphi_i + v_i\|_{\mathbb{D}}} + C\frac{\|v_i\|_{\mathbb{D}}}{\|\varphi_i + v_i\|_{\mathbb{D}}} \tag{82}$$

Assume $\|\mathcal{P}_k(\varphi_i)\|_{\mathbb{D}} = \|\varphi_i + v_i\|_{\mathbb{D}}$ is unbounded. Then we get $\|\varphi_i + v_i\|_{\mathbb{D}} \to \infty$ for a subsequence and, due to the uniform boundedness of $\|\varphi_i\|_{\mathbb{D}}$, also $\|v_i\|_{\mathbb{D}} \to \infty$. Thus,

$$\frac{\|v_i\|_{\mathbb{D}}}{\|\varphi_i + v_i\|_{\mathbb{D}}} \leq \frac{\|v_i\|_{\mathbb{D}}}{|\|\varphi_i\|_{\mathbb{D}} - \|v_i\|_{\mathbb{D}}|} \to 1$$

is uniformly bounded. Then the right hand side in (82) converges to $-\infty$ (for a subsequence) due to **(AG3)**, and the left hand side converges to 0, which is a contradiction. Thus, $\|\mathcal{P}_k(\varphi_i)\|_{\mathbb{D}}$ and $\|v_i\|_{\mathbb{D}}$ are uniformly bounded. From (81) we finally get, using that $g(\varphi_i)$ is bounded from above and g is bounded from below,

$$\frac{c_1}{\lambda_{max}}\|v_i\|_{\mathbb{X}}^2 \leq -\tilde{C} + C(\|v_i\|_{\mathbb{X}} + 1), \tag{83}$$

which gives the boundedness of $(\|v_i\|_{\mathbb{X}})_i$ and thus of $(\|\mathcal{P}_k(\varphi_i)\|_{\mathbb{X}})_{i,k}$. The uniform boundedness of $g(\mathcal{P}_k(\varphi_i)) = g(\varphi_i + v_i)$ then follows from (81). □

Lemma 4.58. *Let $(\varphi_k)_k$ be the sequence generated by Algorithm 4.4. If it holds for a subsequence that $\varphi_{k_i} \to \varphi$ strongly in $\mathbb{X} \cap \mathbb{D}$ for some $\varphi \in \mathbb{X} \cap \mathbb{D}$, $g(\varphi_{k_i})$ is uniformly bounded and $v_{k_i} \to 0$ strongly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$ then φ is a stationary point of $j + g$.*

Proof. Assume without loss of generality $\varphi_k \to \varphi$. We divide (79) by λ_k to obtain for arbitrary $\eta \in \mathbb{X} \cap \mathbb{D}$ and $y_k := \mathcal{P}_k(\varphi_k)$

$$\begin{aligned} g(\eta) &\geq -\frac{1}{\lambda_k}a_k(y_k - \varphi_k, \eta - y_k) - \langle j'(\varphi_k), \eta - y_k\rangle + g(y_k) \\ &= -\frac{1}{\lambda_k}a_k(v_k, \eta + v_k - y_k) + \frac{1}{\lambda_k}a_k(v_k, v_k) - \langle j'(\varphi_k), \eta - y_k\rangle + g(y_k) \\ &\geq -\frac{1}{\lambda_{min}}|a_k(v_k, \eta - \varphi_k)| - \langle j'(\varphi_k), \eta - \varphi_k - v_k\rangle + g(\varphi_k + v_k). \end{aligned}$$

Taking the $\liminf_k$ of the right hand side and using **(AG2)**, **(AG11)** and the uniform boundedness of $g(\varphi_k + v_k)$ due to Lemma 4.57 leads to

$$g(\eta) \geq -\langle j'(\varphi), \eta - \varphi\rangle + g(\varphi),$$

i.e. φ is stationary. □

Theorem 4.59. *Let $(\varphi_k)_k$ be the sequence generated by the nonsmooth VMPT method (Algorithm 4.4). Then:*

1. $\lim_{k\to\infty}(j+g)(\varphi_k)$ *exists.*
2. *Every accumulation point of* $(\varphi_k)_k$ *in* $\mathbb{X}\cap\mathbb{D}$ *is a stationary point of* $j+g$.
3. *For each subsequence with* $\varphi_{k_i}\to\varphi$ *in* $\mathbb{X}\cap\mathbb{D}$ *for some* $\varphi\in\mathbb{X}\cap\mathbb{D}$, *it holds* $v_{k_i}\to 0$ *in* $\mathbb{X}$.
4. *If additionally* $j\in C^{1,\gamma}(\mathrm{dom}(g))$ *for some* $0<\gamma\le 1$ *and* j' *is uniformly bounded on* $\mathrm{dom}(g)$, *then the whole sequence* $(v_k)_k$ *converges to zero in* $\mathbb{X}$.
5. *Let additionally* $j\in C^{1,\gamma}(\mathrm{dom}(g))$ *for some* $0<\gamma\le 1$ *and let* j *be convex. Moreover, let* a_k *be uniformly bounded in the sense that there exists* $C>0$ *such that*

$$|a_k(p,v)|\le C\|p\|_{\mathbb{X}}\|v\|_{\mathbb{X}}\quad\forall k\in\mathbb{N}_0,\ p,v\in\mathbb{X}\cap\mathbb{D}. \tag{84}$$

Then not only strong accumulation points but even weak accumulation points are stationary in the following sense: Let there exist $\varphi\in\mathbb{X}\cap\mathbb{D}$ *such that* $\varphi_{k_i}\to\varphi$ *weakly in* $\mathbb{X}$ *for a subsequence. Then* φ *is a global minimizer of* $j+g$.

Proof. Without loss of generality we assume $v_k\neq 0$ and thus $\alpha_k>0$.
1. Due to the Armijo condition and (80) we observe that $((j+g)(\varphi_k))_k$ is a decreasing sequence, which is bounded from below and thus converges. Moreover, we get that $g(\varphi_k)$ is uniformly bounded.
2. Let $\varphi_{k_i}\to\varphi$ in $\mathbb{X}\cap\mathbb{D}$ for a subsequence and some $\varphi\in\mathbb{X}\cap\mathbb{D}$. From the Armijo condition we get that

$$\alpha_k\sigma(\langle j'(\varphi_k),v_k\rangle+g(\varphi_k+v_k)-g(\varphi_k))\to 0.$$

As a first case, assume $\alpha_k\ge C$ for some $C>0$ and all k. Then $\langle j'(\varphi_k),v_k\rangle+g(\varphi_k+v_k)-g(\varphi_k)\to 0$ and due to (80) also $\|v_k\|_{\mathbb{X}}\to 0$. From Lemma 4.57 we get that $(v_k)_k$ is bounded in $\mathbb{D}$ and thus we get $v_k\to 0$ weakly-* in $\mathbb{D}$. Lemma 4.58 shows that φ is stationary. In the second case, we find a subsequence such that $\alpha_k\to 0$ and $\alpha_k/\beta\le 1$ for all k. Thus the Armijo condition for $\alpha=\alpha_k/\beta$ is not fulfilled, leading to

$$(j+g)(\varphi_k+\alpha_k/\beta v_k)-(j+g)(\varphi_k)>\alpha_k/\beta\sigma(\langle j'(\varphi_k),v_k\rangle+g(\varphi_k+v_k)-g(\varphi_k)) \tag{85}$$

We use the convexity of g on the left hand side and apply the mean value theorem for j and obtain for some $0\le\tilde\alpha_k\le\alpha_k/\beta$

$$\alpha_k/\beta(\langle j'(\varphi_k+\tilde\alpha_k v_k),v_k\rangle+g(\varphi_k+v_k)-g(\varphi_k))>\alpha_k/\beta\sigma(\langle j'(\varphi_k),v_k\rangle+g(\varphi_k+v_k)-g(\varphi_k)).$$

Rearranging terms gives

$$\langle j'(\varphi_k+\tilde\alpha_k v_k),v_k\rangle-\sigma\langle j'(\varphi_k),v_k\rangle+(1-\sigma)(g(\varphi_k+v_k)-g(\varphi_k))>0. \tag{86}$$

Due to Lemma 4.57 we see that v_k is uniformly bounded in $\mathbb{X}\cap\mathbb{D}$ and thus $\varphi_k+\tilde\alpha_k v_k\to\varphi$ strongly in $\mathbb{X}\cap\mathbb{D}$. Moreover, we can extract a subsequence such that $v_k\to\bar v$ weakly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$ for some $\bar v\in\mathbb{X}\cap\mathbb{D}$. Taking the $\liminf_k$ of the inequality (86) gives

$$\langle j'(\varphi),\bar v\rangle+\liminf_k(g(\varphi_k+v_k)-g(\varphi_k))\ge 0.$$

This $\liminf_k$ coincides with the $\liminf_k$ of the left hand side of (80), thus the $\liminf_k$ vanishes and we conclude from (80) that $v_k\to 0$ strongly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$. Finally,

Lemma 4.58 yields that φ is stationary.
3. Follows from the proof of 2. and the subsequence argument in Lemma 7.3.
4. We prove $\|v_k\|_{\mathbb{X}} \to 0$ by a subsequence argument. We take an arbitrary subsequence of $(v_k)_k$, which we denote the same. As in 2. the statement follows in the case $\alpha_k \geq C$. Otherwise, again as in 2., $\alpha_k \to 0$ for a subsequence such that α_k/β does not fulfill the Armijo condition. Instead of applying the mean value theorem on the left hand side of (85) we now use the Hölder estimate from Lemma 7.2 to obtain

$$\begin{aligned}\frac{L}{1+\gamma}(\alpha_k/\beta)^{1+\gamma}\|v_k\|_{\mathbb{X}\cap\mathbb{D}}^{1+\gamma} + \alpha_k/\beta\left(\left\langle j'(\varphi_k), v_k\right\rangle + g(\varphi_k+v_k) - g(\varphi_k)\right)\\ > \alpha_k/\beta\sigma\left(\left\langle j'(\varphi_k), v_k\right\rangle + g(\varphi_k+v_k) - g(\varphi_k)\right).\end{aligned}$$

We rearrange the inequality and get

$$(1-\sigma)\alpha_k/\beta\left(\left\langle j'(\varphi_k), v_k\right\rangle + g(\varphi_k+v_k) - g(\varphi_k)\right) > -\frac{L}{1+\gamma}(\alpha_k/\beta)^{1+\gamma}\|v_k\|_{\mathbb{X}\cap\mathbb{D}}^{1+\gamma}.$$

Applying (80) yields

$$\|v_k\|_{\mathbb{X}}^2 < C(\alpha_k)^{\gamma}\|v_k\|_{\mathbb{X}\cap\mathbb{D}}^{1+\gamma} \leq C(\alpha_k)^{\gamma}(\|v_k\|_{\mathbb{X}}^{1+\gamma} + \|v_k\|_{\mathbb{D}}^{1+\gamma}).$$

We divide by the left hand side to get (recall $v_k \neq 0$)

$$1 < C(\alpha_k)^{\gamma}\left(\|v_k\|_{\mathbb{X}}^{-1+\gamma} + \frac{\|v_k\|_{\mathbb{D}}^{1+\gamma}}{\|v_k\|_{\mathbb{X}}^2}\right).$$

Assume $\|v_k\|_{\mathbb{X}} \geq C$ for a subsequence. Then $\|v_k\|_{\mathbb{D}} \to \infty$ from above inequality and $\alpha_k \to 0$. We derive a contradiction. Since $g(\varphi_k)$ is uniformly bounded, we get that $\|\varphi_k\|_{\mathbb{D}}$ is uniformly bounded due to **(AG3)**. Thus $\|\varphi_k + v_k\|_{\mathbb{D}} \to \infty$. From (78) we deduce $h(\varphi_k + v_k) \to \infty$. Note that the functional h also depends on k and we used the uniform boundedness of $j'(\varphi_k)$ and **(AG12)** here. On the other hand, since φ_k+v_k is the minimizer of h, we get $h(\varphi_k + v_k) \leq h(\varphi_k) = \lambda_k g(\varphi_k) \leq C$, which is a contradiction. Thus, from any subsequence of $(v_k)_k$ we can extract another subsequence with $v_k \to 0$ in $\mathbb{X}$ and hence $v_k \to 0$ in $\mathbb{X}$ holds for the whole sequence, cf. Lemma 7.3.
5. Let $\varphi_{k_i} \to \varphi$ weakly in $\mathbb{X}$. From 1. we get the uniform boundedness of $g(\varphi_k)$ and thus of φ_k in $\mathbb{D}$ due to **(AG3)**. Hence we get $\varphi_{k_i} \to \varphi$ weakly-* in $\mathbb{D}$. The first step is to show

$$\left\langle j'(\varphi_{k_i}), v_{k_i}\right\rangle + g(\varphi_{k_i} + v_{k_i}) - g(\varphi_{k_i}) \to 0. \tag{87}$$

For ease of notation we replace the subsequence index k_i by the index k. We take an arbitrary subsequence and denote it again by index k. We note that $\|\varphi_k\|_{\mathbb{X}\cap\mathbb{D}}$ is uniformly bounded as well as

$$\begin{aligned}\|j'(\varphi_k)\|_{(\mathbb{X}\cap\mathbb{D})^*} &\leq \|j'(\varphi_k) - j'(\varphi_0)\|_{(\mathbb{X}\cap\mathbb{D})^*} + \|j'(\varphi_0)\|_{(\mathbb{X}\cap\mathbb{D})^*}\\ &\leq C\|\varphi_k - \varphi_0\|_{\mathbb{X}\cap\mathbb{D}}^{\gamma} + \|j'(\varphi_0)\|_{(\mathbb{X}\cap\mathbb{D})^*} \leq C.\end{aligned}$$

This implies the uniform boundedness of $\|\mathcal{P}_k(\varphi_k)\|_{\mathbb{X}\cap\mathbb{D}}$ and of $\|v_k\|_{\mathbb{X}\cap\mathbb{D}}$, which can be proved as in Lemma 4.57. The following arguments are as in the proof of 2.: If $\alpha_k \geq C$ for all k then (87) follows. Else $\alpha_k \to 0$ and $\alpha_k/\beta \leq 1$ holds for a subsequence, which does not fulfill the Armijo condition and we apply the mean value theorem to get (86) for some

$0 \le \tilde{\alpha}_k \le \alpha_k/\beta$. Thus,

$$\left\langle j'(\varphi_k + \tilde{\alpha}_k v_k), v_k \right\rangle - \left\langle j'(\varphi_k), v_k \right\rangle + (1-\sigma)(\left\langle j'(\varphi_k), v_k \right\rangle + g(\varphi_k + v_k) - g(\varphi_k)) > 0. \quad (88)$$

For the first two terms we get

$$\begin{aligned} |\left\langle j'(\varphi_k + \tilde{\alpha}_k v_k), v_k \right\rangle - \left\langle j'(\varphi_k), v_k \right\rangle| &\le \|j'(\varphi_k + \tilde{\alpha}_k v_k) - j'(\varphi_k)\|_{(\mathbb{X}\cap\mathbb{D})^*} \|v_k\|_{\mathbb{X}\cap\mathbb{D}} \\ &\le C(\tilde{\alpha}_k)^\gamma \|v_k\|^{1+\gamma}_{\mathbb{X}\cap\mathbb{D}} \to 0. \end{aligned}$$

Thus we can take the $\liminf_k$ in (88) to obtain

$$\liminf_k \left\langle j'(\varphi_k), v_k \right\rangle + g(\varphi_k + v_k) - g(\varphi_k) \ge 0.$$

On the other hand we get from (80) that $\limsup_k \langle j'(\varphi_k), v_k\rangle + g(\varphi_k + v_k) - g(\varphi_k) \le 0$ and thus (87) follows for the chosen subsequence. Since from any subsequence we can choose a subsequence converging to zero we get (87) from Lemma 7.3, and (80) yields $\|v_k\|_{\mathbb{X}} \to 0$. Since j is convex it holds $j(\eta) \ge j(\varphi_k) + \langle j'(\varphi_k), \eta - \varphi_k\rangle$ for all $\eta \in \mathbb{X} \cap \mathbb{D}$, thus

$$(j+g)(\eta) \ge (j+g)(\varphi_k) + \left\langle j'(\varphi_k), \eta - \varphi_k \right\rangle + g(\eta) - g(\varphi_k) \quad \forall \eta \in \mathbb{X} \cap \mathbb{D}. \quad (89)$$

We estimate $g(\eta)$ by the inequality (79), which yields

$$\left\langle j'(\varphi_k), \eta - \varphi_k \right\rangle + g(\eta) - g(\varphi_k) \ge -\frac{1}{\lambda_k} a_k(v_k, \eta - \mathcal{P}_k(\varphi_k)) + \left\langle j'(\varphi_k), v_k \right\rangle + g(\varphi_k + v_k) - g(\varphi_k)$$

Using (87), (84), **(AG12)**, $\|v_k\|_{\mathbb{X}} \to 0$ and the uniform boundedness of $\|\mathcal{P}_k(\varphi_k)\|_{\mathbb{X}}$ we obtain

$$\begin{aligned} \left\langle j'(\varphi_k), \eta - \varphi_k \right\rangle + g(\eta) - g(\varphi_k) \ge &-C\|v_k\|_{\mathbb{X}} \|\eta - \mathcal{P}_k(\varphi_k)\|_{\mathbb{X}} \\ &+ \left\langle j'(\varphi_k), v_k \right\rangle + g(\varphi_k + v_k) - g(\varphi_k) \to 0. \end{aligned}$$

We can take the $\liminf_k$ in (89) and get

$$(j+g)(\eta) \ge \lim_k (j+g)(\varphi_k) \quad \forall \eta \in \mathbb{X} \cap \mathbb{D},$$

thus $(\varphi_k)_k$ is a minimizing sequence for $j+g$. It remains to show $\lim_k (j+g)(\varphi_k) = (j+g)(\varphi)$. Analog to the proof of Theorem 4.14 this can be shown by the special lower semicontinuity of j, see Lemma 4.13, and of g, see **(AG2)**. □

4.11 Overlap with other numerical methods

In this section we point out overlaps with other numerical methods, i.e. we show that for certain optimization problems the VMPT method formally coincides with other numerical methods. This insight can be used to apply the global convergence theory of the VMPT method to other numerical methods. For instance in Section 6.8 we will use the global convergence theory of the VMPT method to show global convergence of certain pseudo time stepping methods available in the literature. Moreover, it will help us to develop an adaptive time stepping scheme and a sensible stopping criterion. Additionally, we thereby can explain the mesh dependent behavior of the projected L^2-gradient method in Section 6.13.11.

4.11.1 Pseudo time stepping

Pseudo time stepping is a variational approach for computing stationary states of an energy. The idea is to derive some gradient flow of the energy and compute the solution of the flow for large times. The time discrete gradient flow is then called pseudo time stepping.

We formally derive the pseudo time stepping method for the optimization problem

$$\min j(\varphi) \quad \text{s.t.} \quad \varphi \in \Phi_{ad} \tag{90}$$

where Φ_{ad} is a convex set and j is differentiable. To derive the gradient flow with respect to some given inner product $(.,.)_{\mathbb{H}}$ we first write the constrained smooth optimization problem as an unconstrained nonsmooth problem as in Section 4.10 by introducing the indicator function $\chi_{\Phi_{ad}}$ of Φ_{ad}. Thus we end up with the equivalent problem

$$\min j(\varphi) + \chi_{\Phi_{ad}}(\varphi).$$

The $\mathbb{H}$-gradient flow starting at $\varphi_0 \in \Phi_{ad}$ is then defined as the time-depending function $\varphi(t)$, which fulfills the evolution equation (or inclusion)

$$\begin{aligned} \partial_t \varphi &\in -(j'(\varphi) + \partial \chi_{\Phi_{ad}}(\varphi)) \quad \forall t \geq 0 \\ \varphi(0) &= \varphi_0, \end{aligned}$$

where $\partial \chi_{\Phi_{ad}} \subset \mathbb{H}^*$ denotes the subdifferential as defined in Section 4.10. Here, we identify $\partial_t \varphi \in \mathbb{H}$ with the functional $(\partial_t \varphi, .)_{\mathbb{H}} \in \mathbb{H}^*$ using the Riesz isomorphism. By definition of the subdifferential this can be equivalently written as the variational inequality

$$\begin{aligned} \varphi &\in \Phi_{ad} \quad \forall t \geq 0 \\ (\partial_t \varphi, \eta - \varphi)_{\mathbb{H}} + \left\langle j'(\varphi), \eta - \varphi \right\rangle &\geq 0 \quad \forall \eta \in \Phi_{ad},\ t \geq 0 \\ \varphi(0) &= \varphi_0, \end{aligned} \tag{91}$$

which can be seen as a weak formulation of the gradient flow. If there exists $t_0 > 0$ such that $\partial_t \varphi(t_0) = 0$, then we have by (91)

$$\left\langle j'(\varphi(t_0)), \eta - \varphi(t_0) \right\rangle \geq 0 \quad \forall \eta \in \Phi_{ad},$$

thus $\varphi(t_0)$ is a stationary point of j and vice versa. Similarly, if $\varphi(t) \to \varphi_0$ as $t \to \infty$ then, under certain assumptions, $\partial_t \varphi(t) \to 0$ and φ_0 is stationary. Moreover, one can derive an energy estimate by testing the variational inequality (91) by $\eta = \varphi(t-\varepsilon) \in \Phi_{ad}$ and dividing by $\varepsilon > 0$ to get

$$\left\langle j'(\varphi), \frac{\varphi(t) - \varphi(t-\varepsilon)}{\varepsilon} \right\rangle \leq -\left(\partial_t \varphi, \frac{\varphi(t) - \varphi(t-\varepsilon)}{\varepsilon} \right)_{\mathbb{H}} \quad \forall t \geq \varepsilon.$$

Taking the limit $\varepsilon \to 0$ we arrive at

$$\partial_t j(\varphi) = \left\langle j'(\varphi), \partial_t \varphi \right\rangle \leq -\|\partial_t \varphi\|_{\mathbb{H}}^2 \quad \forall t \geq 0.$$

Thus, the energy $j(\varphi)$ is decreasing in time. To compute a stationary point of j in Φ_{ad}, one can hence compute the gradient flow $\varphi(t)$ for large t, leading to decreasing energies, and stop if $\partial_t \varphi = 0$, which is then a stationary point.

We now discretize (91) in time by replacing $\partial_t \varphi$ by a difference quotient. To account for semi-implicit time discretization we split $j'(\varphi) = j_e'(\varphi) + j_i'(\varphi)$ in an explicit and an implicit term. Let $\tau_k > 0$ be the time step size in the kth time step and let $\varphi_0 \in \Phi_{ad}$ be given. Then the discretized gradient flow reads

$$\varphi_{k+1} \in \Phi_{ad}, \quad \frac{1}{\tau_k}(\varphi_{k+1} - \varphi_k, \eta - \varphi_{k+1})_{\mathbb{H}} + \left\langle j_e'(\varphi_k), \eta - \varphi_{k+1} \right\rangle + \left\langle j_i'(\varphi_{k+1}), \eta - \varphi_{k+1} \right\rangle \geq 0 \quad (92)$$

for all $\eta \in \Phi_{ad}$, $k \in \mathbb{N}_0$. Note that the inner product $(.,.)_{\mathbb{H}}$ of the pseudo time stepping cannot depend on time by definition of the gradient flow, since the time is purely artificial. However, a point based inner product in the sense of a Riemannian metric is possible.

Explicit time discretization The choice $j_i'(\varphi) = 0$ and $j_e'(\varphi) = j'(\varphi)$ leads to an explicit discretization in time. If we compare the variational inequality (92) of the pseudo time stepping with the variational inequality (22) of the subproblem of the VMPT method, we recognize that the solution of (92) fulfills $\varphi_{k+1} = \mathcal{P}_k(\varphi_k)$ with $a_k(x,y) = (x,y)_{\mathbb{H}}$ and $\lambda_k = \tau_k$. Thus, the explicitly discretized $\mathbb{H}$-gradient flow coincides with the projected $\mathbb{H}$-gradient method with step size $\lambda_k = \tau_k$ and without line search (i.e. $\alpha_k = 1$). In a Hilbert space setting one can thus show global convergence of this pseudo time stepping if the time step size τ_k is chosen appropriately, see Theorem 4.35. In the considered Banach space setting a backup line search has to be included to show global convergence, see Theorem 4.34.

Semi-implicit time discretization The variational inequality (22) in the VMPT method is always linear. Thus, the semi-implicitly discretized gradient flow (92) can only coincide with a VMPT method if j_i' is affine linear, say $j_i'(\varphi) = A\varphi + b$. In this case we can rewrite (92) to

$$\varphi_{k+1} \in \Phi_{ad}, \quad (\varphi_{k+1} - \varphi_k, \eta - \varphi_{k+1})_{\mathbb{H}} + \tau_k \left\langle A(\varphi_{k+1} - \varphi_k), \eta - \varphi_{k+1} \right\rangle + \tau_k \left\langle j'(\varphi_k), \eta - \varphi_{k+1} \right\rangle \geq 0$$

for all $\eta \in \Phi_{ad}$. Comparing this to (22), we get $\varphi_{k+1} = \mathcal{P}_k(\varphi_k)$ with $a_k(x,y) = (x,y)_{\mathbb{H}} + \tau_k \langle Ax, y \rangle$ and step length $\lambda_k = \tau_k$. Under certain assumptions on A and τ_k one can thus show global convergence.

Implicit time discretization The choice $j_i'(\varphi) = j'(\varphi)$ and $j_e'(\varphi) = 0$ in (92) leads to an implicit discretization in time. However, to get a linear variational inequality j' has to be affine linear, i.e. j itself has to be quadratic. In this case we solve with $j'(\varphi) = A\varphi + b$ the variational inequality

$$\varphi_{k+1} \in \Phi_{ad}, \quad (\varphi_{k+1} - \varphi_k, \eta - \varphi_{k+1})_{\mathbb{H}} + \tau_k \left\langle A(\varphi_{k+1} - \varphi_k), \eta - \varphi_{k+1} \right\rangle + \tau_k \left\langle j'(\varphi_k), \eta - \varphi_{k+1} \right\rangle \geq 0$$

for all $\eta \in \Phi_{ad}$, i.e. it again holds $\varphi_{k+1} = \mathcal{P}_k(\varphi_k)$ with $a_k(x,y) = (x,y)_{\mathbb{H}} + \tau_k \langle Ax, y \rangle$ and $\lambda_k = \tau_k$, which is the same as for the semi-implicit time discretization. Note that this variational inequality is a first order necessary condition for the optimization problem

$$\min_{\varphi \in \Phi_{ad}} j(\varphi) + \frac{1}{2\lambda_k} \|\varphi - \varphi_k\|_{\mathbb{H}}^2.$$

This procedure is also known as proximal point algorithm, for which convergence theory is available, see e.g. [Roc76].

4.11.2 Operator splitting methods

Operator splitting methods are a large class of numerical methods. In its general form an operator splitting method can be used to compute a solution of the inclusion

$$0 \in T(\varphi),$$

where T is some set-valued operator. Here we consider only the special case of variational inequalities. We further restrict ourselves to the special variational inequality

$$\varphi \in \Phi_{ad}, \quad \left\langle j'(\varphi), \eta - \varphi \right\rangle \geq 0 \quad \forall \eta \in \Phi_{ad},$$

which is a first order condition of the considered optimization problem (90). Following the outline in [CR97], an operator splitting method utilizes a decomposition $j'(\varphi) = j'_e(\varphi) + j'_i(\varphi)$ in two parts (splitting). The iterative procedure then solves in each step the subproblem

$$y \in \Phi_{ad}, \quad \left\langle j'_e(\varphi_k), \eta - y \right\rangle + \left\langle j'_i(y), \eta - y \right\rangle + \frac{1}{\lambda_k} \left\langle H_k(y - \varphi_k), \eta - y \right\rangle \geq 0 \quad \forall \eta \in \Phi_{ad}, \tag{93}$$

and performs the update $\varphi_{k+1} = y$. The parameter $\lambda_k > 0$ is a step size and the linear operator H_k is called implementation mapping, which does not necessarily define an inner product. We notice that for the special case of $\lambda_k = \tau_k$ and $\langle H_k x, y \rangle = (x, y)_{\mathbb{H}}$ the subproblem coincides with the variational inequality (92) of the pseudo time stepping method. Thus, the pseudo time stepping is a special case of an operator splitting method. All observations for the pseudo time stepping remain true for the iteration (93). In particular (93) can be seen as VMPT method if j'_i is affine linear and if H_k defines an inner product.

If H_k is invertible, the solution operator of (93) can be written as

$$y \in (I + \lambda_k H_k^{-1}(j'_i + \partial\chi_{\Phi_{ad}}))^{-1}(I - \lambda_k H_k^{-1} j'_e)(\varphi_k),$$

where the application of $(I - \lambda_k H_k^{-1} j'_e)$ is called the forward step and $(I + \lambda_k H_k^{-1}(j'_i + \partial\chi_{\Phi_{ad}}))^{-1}$ the backward step. In this sense the method is also called forward-backward splitting. Note that for the classical projected gradient method, i.e. $j'_e = j'$, $j'_i = 0$ and $H_k = I$, the forward step corresponds to $\varphi \mapsto \varphi - \lambda_k \nabla j(\varphi)$ whereas the backward step corresponds to the projection $\varphi \mapsto P_\perp(\varphi)$ on Φ_{ad}.

The following results are available in the literature, which we translate for the case $j'_e = j'$ and $j'_i = 0$, being of our interest. The results are often more general.
Global convergence with linear rate in finite dimension is shown in [CR97], amongst others assuming that j is convex (resp. j' monotone), that $0 < \lambda_{min} \leq \lambda_k \leq \lambda_{max}$ with λ_{max} small enough (no line search in α is necessary then) and assuming that H_k are symmetric positive definite matrices with $H_k \to H$ for some H. Global convergence of an inexact version of the method in finite dimension is studied in [CPR14]. A more general Hilbert space setting is considered in [CV14], where global convergence is shown assuming amongst others that j is convex (resp. j' cocoercive) and

$$\begin{aligned} 0 < \lambda_{min} \leq \lambda_k < \lambda_{max}, \\ 0 < c\|u\|_{\mathbb{H}}^2 \leq \|u\|_{H_k} \leq C\|u\|_{\mathbb{H}}^2, \end{aligned}$$

where λ_{min} and λ_{max} are not arbitrary, but depend on data like the Lipschitz constant

of j'. A generalization thereof to Banach spaces is given in [LMMWX12] for $H_k = I$ (no variable metric) and global convergence is proved. However, the generalized equation which they solve is $0 \in T(\varphi)$, where $T : X \rightrightarrows X$ maps X into X. On the other hand, for convexly constrained optimization problems one wants to solve $0 \in \partial(j + \chi_{\Phi_{ad}})$, where $\partial(j + \chi_{\Phi_{ad}}) : X \rightrightarrows X^*$ maps X into its dual X^*. Thus, the method presented in [LMMWX12] does not apply to our optimization problem. In fact the corollaries in [LMMWX12] regarding the projected gradient method are formulated in a Hilbert space setting. The same setting ($T : X \rightrightarrows X$) is treated in [Cho15]. The for us more interesting case $T : X \rightrightarrows X^*$ is studied in [OI07], where the transition from X to X^* is handled e.g. by the p-duality mapping $\partial\|.\|^p/p$. However, only the case $j_i' = j'$, $j_e' = 0$ is considered, i.e. the proximal point method, which is not a VMPT method in general.

A more interesting generalized operator splitting method in Banach space is given in [Bre09], which we discuss in detail. The same optimization problem is considered as in Section 4.10, i.e. the sum of a smooth and a convex nonsmooth functional,

$$\min_{\varphi \in X} j(\varphi) + g(\varphi),$$

posed in a reflexive Banach space X. In each step of the method the subproblem

$$\min_{y \in X} \quad \frac{1}{\gamma + 1} \|y - \varphi_k\|_X^{\gamma+1} + \lambda_k \left\langle j'(\varphi_k), y - \varphi_k \right\rangle + \lambda_k g(y) \tag{94}$$

is solved, where $\gamma \in (0, 1]$ has to be chosen such that it locally holds $j \in C^{1,\gamma}$. The next iterate is then $\varphi_{k+1} = y$, which corresponds to $\alpha_k = 1$ in the VMPT method. Note that the scaling parameters λ_k in [Bre09] are chosen such that the Armijo condition (75) is satisfied for $\alpha_k = 1$, which is shown in [Bre09, Prop. 7]. The needed bounds for λ_k depend on γ and the Hölder constant of j' (similar to [Gol64]), which are probably unknown a priori. We observe that the method in [Bre09] is very similar to the VMPT method given in Section 4.10. However, there are certain differences. First of all, Bredies needs a reflexive Banach space, which is not the case for the VMPT method ($\mathbb{D}$ can be e.g. L^∞). Bredies only considers convex functionals j, whereas j is not needed to be convex in the VMPT method. The regularizing norm in the subproblem (94) is a general norm, which is not assumed to stem from an inner product, whereas in the VMPT method the a_k-norm is always defined by an inner product. Therefore the VMPT subproblem is always quadratic (in the case $g = \chi_{\Phi_{ad}}$), which is not the case for the subproblem (94). A major difference is that the VMPT method can use a variable metric a_k, whereas in (94) the X-norm cannot depend on the iteration number k. Additionally, the solution of the VMPT subproblem is unique, which is not the case for (94). On the other hand, if g is a general convex function we need the growth condition **(AG3)** (due to the weak coercivity of a_k), which is not needed by Bredies.
We can conclude that the VMPT method and the method in [Bre09] have a common overlap (if X is a Hilbert space and $\gamma = 1$), but neither is a special case of the other.

To our knowledge there is no operator splitting method in Banach space available in the literature which can handle a variable metric.

4.11.3 Others

As already mentioned, the projected Newton's method described in Section 4.7, which is the VMPT method with metric $a_k = j''(\varphi_k)$, coincides with the Josephy-Newton method applied to the variational inequality $\langle j'(\varphi), \eta - \varphi \rangle \geq 0 \ \forall \eta \in \Phi_{ad}$. Of course this also holds for quasi-Newton variants thereof. If the constraints defining Φ_{ad} are linear, then the projected Newton's method coincides with the SQP method. This applies also to quasi-Newton methods like BFGS-SQP [Che96]. However, the SQP methods can also be used if Φ_{ad} is not convex. The non-smooth VMPT method described in Section 4.10 can be seen as generalization of the variable metric proximal gradient method (see e.g. [TDLC15]) to Banach spaces. In the special case that the cost functional is quadratic and convex, the proximal point method [Roc76] with variable metric H_k and step length c_k coincides with the VMPT method using the metric $a_k(u,v) = (u,v)_{H_k} + c_k j''(\varphi_k)[u,v]$ and $\lambda_k = c_k$.

It is interesting that also other specialized numerical methods can be seen as VMPT method. For instance in [BE91a] the optimization problem

$$\min_{\varphi \in \Phi_{ad}} j(\varphi) := b(\varphi,\varphi) - c(\varphi,\varphi) - 2l(\varphi)$$

is considered, where b and c are symmetric and bilinear, l is linear and Φ_{ad} is a convex closed non-empty subset of a Hilbert space $\mathbb{H}$. The numerical method they propose consists of the successive solution of the variational inequality

$$y \in \Phi_{ad}, \quad b(y, \eta - y) \geq c(\varphi_k, \eta - y) + l(\eta - \varphi_k) \quad \forall \eta \in \Phi_{ad}$$

and the update $\varphi_{k+1} = y$. Rearranging the variational inequality gives

$$y \in \Phi_{ad}, \quad 2b(y - \varphi_k, \eta - y) + \underbrace{2b(\varphi_k, \eta - y) - 2c(\varphi_k, \eta - y) - 2l(\eta - \varphi_k)}_{= \langle j'(\varphi_k), \eta - y \rangle} \geq 0 \quad \forall \eta \in \Phi_{ad},$$

which coincides with the variational inequality of the VMPT subproblem (26) with inner product $a_k(u,v) = 2b(u,v)$ and $\lambda_k = 1$. Since b is assumed to be coercive in [BE91a], the proposed method is a VMPT method without line search. In fact, since the metric is not variable, the method corresponds to the classical projected gradient method in the Hilbert space $(\mathbb{H}, 2b)$. As application a free boundary problem arising in the theory of liquid drops and plasma physics is considered in [BE91a] with the concrete choices of $\mathbb{H} = H_0^1(\Omega)$ and $b(u,v) = \gamma(\nabla u, \nabla v)_{L^2(\Omega)} + \kappa^2(1-\gamma)(u,v)_{L^2(\Omega)}$. The real parameters γ and κ are chosen such that b is H_0^1-coercive. Thus, the method corresponds to a (scaled) projected H_0^1-gradient method. For the case $\gamma = 0$, they choose $\mathbb{H} = L^2(\Omega)$ and $b(u,v) = \kappa^2(u,v)_{L^2(\Omega)}$, which corresponds to a (scaled) projected L^2-gradient method.

Another paper using a method similar to the VMPT method is [TP13]. For the considered topology optimization problem the proposed method coincides with the VMPT method for a special choice of the parameters using the data

$$a_k(f,g) = \frac{1}{\tau}(f,g)_{L^2(\Omega)} + \beta(\nabla f, \nabla g)_{L^2(\Omega)}, \quad \lambda_k = 1,$$

with $\beta, \tau > 0$. This will be discussed in more detail at the end of Section 6.13.6.

4.12 Application to a semilinear elliptic optimal control problem

In this section we discuss the application of the VMPT method with a_k being the L^2-inner product to a semilinear elliptic optimal control problem analyzed in [Trö09]. The resulting numerical method is well known, since it formally coincides with the projected L^2-gradient method. However, since the cost functional is only differentiable in L^∞ rather than L^2, the global convergence cannot be shown using projected gradient theory, but it follows from the analysis in this thesis.

We use the same notation as [Trö09], i.e. u denotes the control, y the state and p the adjoint state. We consider the optimal control problem

$$\begin{aligned} \min_{u,y} \int_\Omega \varphi(x, y(x))\,\mathrm{d}x + \int_\Omega \psi(x, u(x))\,\mathrm{d}x \\ -\Delta y + d(x, y) = u \quad \text{in } \Omega \\ \partial_\nu y = 0 \quad \text{on } \partial\Omega \\ u_a(x) \le u(x) \le u_b(x) \text{ a.e. in } \Omega, \end{aligned}$$

where $\Omega \subset \mathbb{R}^N$, $N \in \mathbb{N}$, is a bounded Lipschitz domain with outer normal ν, and φ, ψ and d are functions mapping $\Omega \times \mathbb{R}$ into $\mathbb{R}$. The bounds u_a and u_b are in $L^\infty(\Omega)$ and it holds $u_a \le u_b$ a.e. in Ω. Moreover, d is monotone in y and ψ is convex in u. For the precise assumptions on φ, ψ and d we refer to [Trö09]. In particular the assumptions imply that φ, ψ and d are as Nemytskii operators two times continuously differentiable in $L^\infty(\Omega)$. The following results can be found in [Trö09]. Under the assumptions the state equation is uniquely solvable, where the solution y is a continuous function. We can thus define the control-to-state operator $G : L^\infty(\Omega) \to H^1(\Omega) \cap C(\overline{\Omega})$, which maps u to the solution y of the state equation. Moreover, we define the reduced cost functional $j(u) := \int_\Omega \varphi(x, G(u)(x))\,\mathrm{d}x + \int_\Omega \psi(x, u(x))\,\mathrm{d}x$ for all $u \in \Phi_{ad} := \{u \in L^\infty(\Omega) \mid u_a(x) \le u(x) \le u_b(x) \text{ a.e. in } \Omega\}$. Under the assumptions $j : L^\infty(\Omega) \to \mathbb{R}$ is two times continuously differentiable. Using an adjoint approach for the first order derivative we get

$$\left\langle j'(u), h \right\rangle = \int_\Omega (p + \psi_u(x, u))h\,\mathrm{d}x \quad \forall u, h \in L^\infty(\Omega),$$

where $p \in H^1(\Omega)$ is the adjoint state, which is defined as solution of the elliptic PDE

$$\begin{aligned} -\Delta p + d_y(x, y)p = \varphi_y(x, y) \quad \text{in } \Omega \\ \partial_\nu p = 0 \quad \text{on } \partial\Omega. \end{aligned}$$

To solve the optimal control problem, we apply the VMPT method to the reduced problem using the spaces $\mathbb{X} = L^2(\Omega)$, $\mathbb{D} = L^\infty(\Omega)$, and the inner product $a_k(u, v) = (u, v)_{L^2(\Omega)}$. One easily shows that the assumptions **(A1)**-**(A12)** are fulfilled. In particular j is differentiable in $\mathbb{D}$ and thus also in $\mathbb{X} \cap \mathbb{D}$. Since it holds $\langle j'(u), h\rangle = \int_\Omega fh\,\mathrm{d}x$ for $f = p + \psi_u(x, u) \in L^1(\Omega)$ we get **(A7)**. The assumptions on the metric follow from Lemma 4.19. Hence we can apply Theorem 4.14 to obtain global convergence for appropriately chosen λ_k.

We briefly discuss which calculations have to be done for a single VMPT step. The subproblem in the kth VMPT step reads

$$\min_{u_a \le w \le u_b} \frac{1}{2}\|w - u_k\|_{L^2} + \lambda_k \int_\Omega (p_k + \psi_u(x, u_k))(w - u_k)\,\mathrm{d}x,$$

where p_k is the adjoint state corresponding to u_k. By the calculations in Section 4.1 we get that this is equivalent to the L^2-projection problem

$$\min_{u_a \le w \le u_b} \frac{1}{2} \|w - (u_k - \lambda_k(p_k + \psi_u(x, u_k)))\|_{L^2}^2.$$

It is well known that L^2-projections on box constraints can be calculated pointwise [Trö09]. Thus we obtain the closed-form expression

$$w(x) = P_{[u_a(x), u_b(x)]}(u_k(x) - \lambda_k(p_k(x) + \psi_u(x, u_k(x)))) \quad \text{a.e. in } \Omega. \tag{95}$$

In this application the cost for the solution of the subproblem can be neglected. However, the adjoint state p_k has to be calculated, which in turn depends on the state y_k. In each step of the Armijo backtracking in line 9 of Algorithm 4.1 the cost functional j has to be evaluated, which amounts to the solution of the state equation. The state variable from the final backtracking step can be recycled in the next VMPT step. Thus the following tasks have to be done by the VMPT method:

1. Solve the adjoint equation.
2. Calculate w_k by equation (95).
3. Check the stopping criterion $\|w_k - u_k\|_{L^2} \le tol$.
4. Armijo backtracking along the direction $w_k - u_k$, where in each step the state equation has to be solved.
5. Go to step 1.

In each VMPT iteration the adjoint equation has to be solved once and the state equation $K + 1$ times, where $K \in \mathbb{N}_0$ is the number of backtracking steps performed. Since the calculation of the projection is cheap in this case, also a curved search along the projection arc can be performed as in Algorithm 4.2 instead of a line search along $w_k - u_k$. The cost of the iteration is then the same.

Often the function ψ has the form $\psi(x, u) = \frac{\beta}{2} u^2$ for some $\beta > 0$. In this case one can choose $\lambda_k = \frac{1}{\beta}$, since (95) then simplifies to

$$w(x) = P_{[u_a(x), u_b(x)]}\left(-\frac{1}{\beta} p_k(x)\right) \quad \text{a.e. in } \Omega.$$

If $\alpha_k = 1$ is accepted by the backtracking, then the method can be seen as fixed point iteration on the well known optimality condition

$$\overline{u}(x) = P_{[u_a(x), u_b(x)]}\left(-\frac{1}{\beta} \overline{p}(x)\right) \quad \text{a.e. in } \Omega.$$

5 Introduction to topology optimization

The classical problem statement in structural topology optimization is the following: Let a design container Ω be given together with certain boundary conditions, e.g. that one side of the container is fixed to some wall. Moreover, let certain forces be given, which act in Ω and on $\partial\Omega$. The goal is to find an optimal material configuration within Ω such that a cost functional is minimized. This can in general form be written as

$$\begin{aligned} &\min f(D, \boldsymbol{u}(D)) \\ &D \in \mathcal{X}, \end{aligned} \tag{96}$$

where $\mathcal{X} \subset \mathcal{P}(\Omega)$ is a subset of the power set of Ω, which contains sufficiently regular subsets. The set $D \subset \Omega$ describes the portion of Ω which is filled with some elastic material. Taking the given boundary conditions into account, the area D containing material is deformed under the given forces, resulting in the displacement $\boldsymbol{u}(D)$, whereon the cost functional typically depends. In most of the literature $\boldsymbol{u}$ is given as solution of the equations of linearized elasticity

$$\begin{aligned} -\nabla \cdot (\boldsymbol{C}\mathcal{E}(\boldsymbol{u})) &= \boldsymbol{f} && \text{in } D, \\ (\boldsymbol{C}\mathcal{E}(\boldsymbol{u})) \cdot \boldsymbol{n} &= \boldsymbol{g} && \text{on } \Gamma_g, \\ \boldsymbol{u} &= 0 && \text{on } \Gamma_D, \\ (\boldsymbol{C}\mathcal{E}(\boldsymbol{u})) \cdot \boldsymbol{n} &= 0 && \text{on } \partial D \setminus (\Gamma_g \cup \Gamma_D), \end{aligned} \tag{97}$$

where the fourth order stiffness tensor $\boldsymbol{C}$ describes the material properties, $\mathcal{E}(\boldsymbol{u}) := \frac{1}{2}(D\boldsymbol{u} + D\boldsymbol{u}^T)$ is the linearized strain, $\boldsymbol{n}$ the outer normal on ∂D, $\boldsymbol{f}$ is the body force, $\boldsymbol{g}$ is the boundary traction acting on Γ_g and Γ_D is the portion of $\partial\Omega$ where homogeneous Dirichlet boundary conditions are imposed. Here, the boundaries Γ_g and Γ_D are part of ∂D and are not under optimization. The advantage of linearized elasticity over nonlinear elasticity is that the equations are much easier to solve numerically. Also the analysis simplifies, e.g. non-uniqueness of solutions does not appear. However, nonlinear effects such as buckling are not taken into account.

The space $\mathcal{X}$ of admissible shapes may contain certain constraints on the shape D itself, e.g. constraints on the volume of D or the perimeter of D, and also constraints on variables depending indirectly on D, e.g. constraints on the resulting displacement, on the strain or the stress. A typical objective functional, which is often considered in the literature is the mean compliance of the structure, which describes the stiffness of the structure under a given load.

In topology optimization the space of admissible shapes in usually very large, since one does not prescribe the position or number of holes in the structure, the number of components which compose the structure, the size or thickness of the structure or the position where it is assembled to the wall. Another discipline with an essentially smaller space of admissible shapes is the so called shape optimization. In shape optimization the topology of the shape is usually given a priori, e.g. it is prescribed how many holes there are in the structure. Therefore, only the shape of the holes and the boundary of the structure is optimized, whereas the fundamental appearance is unchanged. A possible method in shape optimization is the method of mappings or perturbation of identity [MS76], where D is parametrized over a reference domain D_{ref}, i.e. $D = (id + v)(D_{ref})$ for some vector field v and the optimization problem is formulated in terms of the unknown v. Another possibility is to write D as the area below the graph of some function, e.g. $D = \{(x, y) \mid 0 \le y \le g(x)\}$

and formulate the optimization problem in terms of the unknown function g [HM03]. There are also other disciplines for which the set of admissible shapes is even smaller. For instance in sizing optimization the topology, placement and orientation of the distinct parts of D are given, and only the thickness of the parts is optimized, see [BS03]. In this thesis we consider the most general discipline, namely topology optimization.
Often, the optimization problem is simplified by the so called **ersatz material** approach. The motivation is that a given approximation D of the optimal shape can have a complicated boundary and thus the state equation (97) corresponding to D may be difficult to solve, since D has to be discretized in some way. To overcome this problem, the void, i.e. $\Omega \setminus D$, is modelled as a very elastic or weak material. In this manner $\boldsymbol{C}$ is extended on the whole design domain Ω, e.g. by $\delta\boldsymbol{C}$ where $0 < \delta \ll 1$, and the state equation (97) is solved in Ω instead of D. The container Ω often is a simple shape, e.g. a square, and can be discretized more easily. Thereby, also the displacement $\boldsymbol{u}$ becomes well defined on whole Ω. The approach can be justified in the case of compliance minimization in the sense that the quasiconvexifications of the stress formulation of the topology optimization problem converge as the stiffness of the ersatz material tends to zero [All02, ch. 4.2.2]. Also a result in [BC03] is available showing that the ersatz material designs converge to the void design. However, as the stiffness of the ersatz material tends to zero, the condition number of the state equation explodes, which leads to high numerical errors in the state $\boldsymbol{u}$, which in turn can lead to numerical instabilities [DK10, Gou06]. The ersatz material approach is used in many methods, e.g. the homogenization method [All02], the ESO method [HX09], the level set method [AJT04] or the phase field approach [BFGS14]. Also the SIMP method is based on the ersatz material approach (see below). Throughout this thesis the void is modeled as an ersatz material. Thus, when we speak of 'void' we always refer to a very weak material.

The classical problem (96) can be extended to a **multi-material** problem, where one seeks to distribute N different homogeneous materials D_i with associated stiffness tensors $\boldsymbol{C}_i$, $i = 1, \dots N$, within Ω. In case that no void is present, or if it is approximated by an ersatz material, the state equation (97) can be solved on Ω by setting $\boldsymbol{C}(x) = \boldsymbol{C}_i$ for $x \in D_i$, $i = 1, \dots, N$. Since the distinct materials should define a partition of Ω, one usually imposes the constraints $D_i \cap D_j = \emptyset$, $i \neq j$, and $\bigcup_{i=1}^{N} D_i = \Omega$. This is of course more involved than the original problem and many methods have been develop to cope with multiple materials (see below). One of the first papers considering multiple materials is [Tho92]. In the present thesis also multiple materials are considered.

A major issue in the context of topology optimization is the matter of **ill-posedness**. If the set of admissible shapes or the cost functional is not appropriately chosen, it may happen that no minimizer exists, i.e. that no design D in $\mathcal{X}$ realizes the infimum of the problem (96). For instance it is well known that the problem of minimum compliance is not well posed without further restrictions on the admissible shapes or certain regularizations [SK86]. In a typical minimizing sequence $D_i \in \mathcal{X}$, $i \in \mathbb{N}$, with $f(D_i, \boldsymbol{u}(D_i)) \to \inf_{D \in \mathcal{X}} f(D, \boldsymbol{u}(D))$, microstructures develop, i.e. the structures within D_i get finer and finer. Due to the lack of compactness in an appropriate space, no existence of a minimizer can be shown. In the numerics, this ill posedness usually manifests in mesh dependent solutions [DS95, SP98]. To circumvent nonexistence of minimizers certain methods have been developed. For instance a restriction on the perimeter of the shapes prevents the formation of microstructures [Jog02, Pet99]. Other techniques include the penalization of the perimeter of D rather than a hard restriction [AB93], constraints

or penalization of other geometric quantities, such as the slope of the boundary of D [HM03], or the capacity of D [BZ95]. Moreover, Tikhonov regularization can be used to gain well posedness [TP13] as well as other methods that define a minimal length scale [Pou03, PS98]. In the numerics it is popular to apply a filtering to the densities [Bou01] or the sensitivities [SM12]. Filtering involves the convolution with a kernel, such that high frequencies are filtered out.

A difficulty in topology optimization is that in general the space of admissible shapes $\mathcal{X}$ is not a subset of a vector space. Thus, the classical optimization theory is not applicable. Other tools such as the shape calculus have been developed to be able to formulate optimality conditions [SZ92, DZ01]. The corresponding optimization methods often utilize shape gradients and topological derivatives [GGM00, SZ99].

Another approach to get an optimization problem posed in a vector space is the introduction of a fictitious material: The topology optimization problem (96) can also be written in terms of characteristic functions using the one-to-one correspondence between χ_D and $D = \{\chi_D = 1\}$, leading to

$$\begin{aligned} &\min f(\{\rho = 1\}, \boldsymbol{u}(\{\rho = 1\})) \\ &\rho : \Omega \to \{0, 1\}, \\ &\{\rho = 1\} \in \mathcal{X}. \end{aligned} \tag{98}$$

Since ρ can only attain the values 0 and 1, the topology optimization problem can be seen as discrete optimization problem. In the fictitious material approach the optimization problem is relaxed by allowing intermediate densities $\rho \in [0, 1]$. This leads to a continuous optimization problem for which derivatives (sensitivities) can be calculated and classical optimization methods, such as steepest descent, can be applied. A very popular fictitious material method is the so called **SIMP method** (single isotropic material with penalization) [Ben89, RZB92]. The stiffness tensor is interpolated as $\boldsymbol{C}(\rho) = \rho^p \boldsymbol{C}$, where $p \in \mathbb{N}$, $p \geq 2$, is a penalization parameter. Thus the stiffness corresponding to intermediate densities is very low if p is large. When minimizing the compliance of the structure together with a volume or mass constraint of the form $\int_\Omega \rho = m|\Omega|$, the interpolation acts as a kind of penalization of intermediate densities, since areas with $0 < \rho < 1$ don't increase the stiffness of the structure much, but consume much mass on the other hand. Often a continuation method is used, where the optimization is started with $p = 1$ and afterwards p is gradually increased [PS98]. However, the appearance of intermediate densities is still a problem in the SIMP method [SS01]. Only for special cases one can show that a 0-1 design is obtained for p large enough on the discrete level [Rie01]. Usually a positive lower bound for ρ is prescribed, i.e. $\rho \in [\overline{\rho}, 1]$ with $\overline{\rho} > 0$, in order to avoid degeneration of the stiffness tensor for $\rho = 0$. This corresponds to an ersatz material approach for the void. Note that the SIMP method does not address the ill-posedness of the optimization problem [BS03]. Thus it is known that without further actions the solutions of the SIMP method are mesh dependent, i.e. the minimal length scale of the obtained optimal shapes is determined by the mesh width. Frequently a filtering of the sensitivities is applied to get mesh independent solutions and to avoid checkerboard patterns. Recent studies show that this filtering is equivalent to considering an optimization problem in nonlocal elasticity and can in this sense be justified [SM12]. SIMP is also used for other cost functionals, e.g. for the compliant mechanism problem [Sig97]. It is an important method, which is implemented in many commercial software packages for topology optimization [Roz09]. An extensive discussion

of the SIMP approach can be found in the monograph [BS03]. Moreover, extensions of the SIMP method are available which can handle multiple materials. For instance in [BS99] the interpolation scheme $\boldsymbol{C}(\rho_1,\rho_2) = \rho_1^{p_1}(\rho_2^{p_2}\boldsymbol{C}_1 + (1-\rho_2^{p_2})\boldsymbol{C}_2)$ is used for two materials with stiffness tensors $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$, and void. Here, ρ_1 is the density of the non-void region (material 1 plus material 2) and ρ_2 is the density of material 1 within the non-void region and p_1, p_2 are the penalization parameters. For an extension of this idea to N materials, $N-1$ density functions are needed. Another material interpolation scheme where only a single design variable is needed for the description of N materials is given in [YA01]. They use Gaussian distributions to define the interpolation $\boldsymbol{C}(\rho) = \sum_{i=1}^{N-1} \exp(-\frac{(\rho-\mu_i)^2}{2\sigma_i^2})\boldsymbol{C}_i + \boldsymbol{C}_N$, where μ_i and σ_i are the mean and the standard deviation, respectively, for the ith material and $\boldsymbol{C}_N$ is the stiffness of the void. Thus, if σ_i is very small for $i = 1,\dots,N-1$, it holds approximately $\boldsymbol{C}(\rho) \approx \boldsymbol{C}_i$ for $\rho = \mu_i$, $i = 1,\dots,N-1$ and $\boldsymbol{C}(\rho) \approx \boldsymbol{C}_N$ else. This $\boldsymbol{C}$ is called a peak function. However, certain numerical difficulties are reported for this interpolation scheme and a careful choice for σ_i is necessary.

Another way to reformulate the original discrete topology optimization problem to a continuous one is by the use of the **homogenization method**. For instance in [BK88] it is assumed that in each point $x \in \Omega$ there is a unit cell defining a periodic microstructure consisting of a material with a square hole in it. The size of the hole and its angle are parametrized by functions $\alpha : \Omega \to \mathbb{R}$ and $\theta : \Omega \to \mathbb{R}$, respectively, which are used as continuous design variables. The respective elasticity tensor at $x \in \Omega$ is then computed by homogenization as a function of the design variables, $\boldsymbol{C}(x) = \boldsymbol{C}(\alpha(x),\theta(x))$. In [NFMK98, AKG94] this method is used to solve a compliant mechanism problem, where the width and height of the hole in the unit cell are treated as separate design variables, resulting in a total of three design variables. In [All02] another formulation is given, in which the unknowns are the stress, the homogenized stiffness tensor and the density of the material. It corresponds to the quasiconvexification of the original problem in stress formulation, which is a well posed problem. Contrary to the first mentioned homogenization method, the microstructure is not given, but also under optimization. The optimal microstructure is a rank-2 sequential laminate (in 2D) or a rank-3 sequential laminate (in 3D) with lamination directions given by the eigenvectors of the optimal stress. As in the SIMP method, checkerboard patterns also occur in the homogenization method. To overcome this problem one can again use filtering techniques for the density, which is done e.g. in [All02, NFMK98]. Another drawback is that the final design is not necessarily a 0-1 design, i.e. there are points $x \in \Omega$ in which there is a microstructure consisting not entirely of void or material. In [All02] a post-processing technique is applied, which penalizes intermediate densities and thus produces a 0- 1 design based on the unpenalized composite design. This technique is heuristic and mesh dependent. In [BK88] a lumping strategy based on a cut-off as in [CO82] is performed on the optimal composite to gain a 0-1 design which is near the optimal composite. The homogenization method is also capable of handling multiple materials. For instance in [Tho92] a rank-2 laminate is considered, consisting of two materials and void. The design variables are then the proportions of the distinct materials in the laminate (3 functions) and the angle of the lamination (1 function). The generation of composite material ('intermediate densities') in the final design can be avoided by choosing the lamination angle to *minimize* the stiffness.

Many numerical methods are used to solve the problems in the SIMP formulation and the homogenization formulation. A version of the **sequentially linear programming** (SLP) method is applied e.g. in [PS98, Sig97, NFMK98]. The linearized problem, consist-

ing of the linearized objective functional and the linearized constraints, is solved iteratively until convergence. Additionally a move limit is introduced to stabilize the method. A typical form for a move limit is $\Delta\rho_{min} \le \Delta\rho \le \Delta\rho_{max}$ (component wise), where $\Delta\rho \in \mathbb{R}^n$ is the change of the design variable ρ in the current iteration. The limits $\Delta\rho_{min}$ and $\Delta\rho_{max}$ can vary from iteration to iteration and they are chosen such that the next iterate fulfills the box constraints. Move limits are used very often in numerical methods for topology optimization problems and are somehow a substitute for the more established globalization methods such as line search or trust region methods. There is a clear similarity between move limits and weighted ℓ^∞ trust region constraints.

Another method used often is the **method of moving asymptotes** (MMA) [Sva87], which is formulated as a solver for general nonlinear mathematical programming problems in finite dimension including box constraints. Iteratively a convex subproblem is solved, where the cost functional as well as the constraints are replaced by 1st order approximations. Unlike other methods as SLP, projected gradient or conditional gradient methods, these 1st order approximations are not polynomials, but rather have $1/x$ singularities. Moreover, the approximation is separable in the design variables. Thus, if only box constraints are present, then the solution of the subproblem can be written down explicitly, which makes the MMA to a very cheap method. In case that other inequality constraints are present, the primal variables can be eliminated, such that only an optimization problem involving the dual variables has to be solved. In the case that there is only a volume constraint, this corresponds to an optimization problem in 1D. Also in the MMA method one uses move limits to stabilize convergence. Since the MMA method is a general optimization method, it can handle a variety of cost functionals and constraints. Unfortunately convergence of the method cannot be guaranteed and the method may cycle. However, in [Zil93] the MMA method is combined with an Armijo backtracking strategy involving a merit function, for which global convergence is shown. A special case of the MMA method, which is also often used is the CONLIN method [Fle89].

Also heuristic methods are used to solve topology optimization problems, such as genetic algorithms and modified controlled random search algorithms (for a description see [HM03]). Important and widely used heuristic methods are the **optimality criteria methods** (OC). Based on the discrete KKT system, update schemes for the design variable and the Lagrange multipliers are deduced. Usually these update schemes are of explicit form, which can be easily calculated. This is no general optimization method, since the update depends on the objective function and the constraints. To give an impression of the method, we describe the frequently used method in compliance minimization (see e.g. [BS03]). Let the discrete problem be given as

$$\min f(x), \quad h(x) \le 0, \quad x_{min} \le x \le x_{max},$$

where f is the compliance, x is the design variable and $h \le 0$ describes a mass inequality constraint. The gradient equations in the KKT system for the inactive components are reformulated to $D_i := -\frac{1}{\Lambda}\frac{\partial_i f}{\partial_i h} = 1$ for all i with $x_{min,i} < x_i < x_{max,i}$, where Λ is the Lagrange multiplier of the mass constraint. A typical update for x has the form of the fixed point iteration $x_i^+ = D_i^\eta x_i$ together with a componentwise projection on $[x_{min}, x_{max}]$ and a move limit. The parameter $\eta \in (0,1)$ is a tuning parameter. The Lagrange multiplier Λ is e.g. chosen such that mass equality is fulfilled, i.e. $h(x^+) = 0$. For the method it is assumed that $\Lambda > 0$, i.e. the mass inequality constraint is active and strict complementarity holds, and that $\frac{\partial_i f}{\partial_i h} < 0$. Both assumptions are usually fulfilled for compliance minimization.

However, for other applications such as eigenfrequency optimization these assumptions are not fulfilled and a shift for the Lagrange multiplier has to be introduced to ensure $D_i > 0$ [MKC95]. The OC method is e.g. used in [BK88] for the compliance minimization using a homogenization method and in [YA01, AKG94] for the design of compliant mechanisms. Moreover, most of the numerical results in the book [BS03] are obtained either by OC or the MMA method. An optimality criteria method for stress constraints and displacement constraints is given in [ZR92]. The drawbacks of OC methods are that it is often not straight forward to generalize the method to other problems, that the method is heuristic and thus convergence cannot be guaranteed, and that the final convergence is often slow. However, the updates are usually cheap, it is easy to implement and the results are mostly satisfactory [BS03].

The heuristic but widely used **ESO method** [XS97, XS93] directly solves the unrelaxed topology optimization problem (98). Therefor, ρ is discretized e.g. by assuming ρ to be piecewise constant on a quadrilateral triangulation of Ω. Beginning with $\rho \equiv 1$, i.e. material everywhere, an iteration removes elements from the material by a heuristic (not gradient based) criterion, e.g. the von Mises stress, until a stable state is reached. The drawbacks are that it is a heuristic method and thus no convergence proof of the method is available and it breaks down for certain examples [ZR01]. Improved methods based on ESO are also available, e.g. BESO (bidirectional ESO) [QSX98], where material can also be added in each iteration rather than only removed, a combination of ESO with material interpolation methods like SIMP and extensions for multiple materials [HX09]. Additional efforts are needed to overcome checkerboard patterns and mesh dependent solutions. A critical review of the ESO method and a comparison to SIMP can be found in [Roz09].

The preceding models and methods are based on the discrete optimization problem and no analysis in function space is performed. Nowadays the material interpolation and homogenization methods are referred to as classical. Modern methods for topology optimization are the level-set method and the phase field method.

In the **level-set method**, which is introduced in [OS88] for evolving hypersurfaces, the unknown shape is given as the level set of some function ψ, i.e. $D = \{x \in \Omega \mid \psi(x) \leq 0\}$. In particular, the interface $\Gamma := \partial D \cap \Omega$ is given as the zero set of ψ. Typically, ψ resembles the signed distance function of the free boundary Γ. As optimization method one usually considers a flow of Γ in normal direction, where the velocity field is given by the negative shape gradient or a descent direction in general. The flow equation for Γ translates into a Hamilton-Jacobi convection equation for the level set function ψ, which can be solved numerically after discretization in time and space. The flow is calculated for a small time and then the shape gradient is updated. The time has to be chosen so small that the objective function decreases, similar to conventional gradient methods. The procedure is iterated until convergence. Periodically, a reinitialization of the level-set function is necessary, which drives ψ towards the signed distance function of Γ by solving a first order PDE. The described method is applied in [AJT04]. Similar methods are used by the group of Allaire [ADDM14, AJ05, AJT04] and the group of Wang [LLC+08, WC09, WCWM05, WW04]. In [YINT10] another approach is used, being a mixture of phase field and level-set method together with a pseudo time stepping. It is not a trivial task to include constraints in this level-set framework. A possibility is to use a penalization instead of a hard constraint as it is done in [AJT04] for the mass constraint. Alternatively one can use an augmented Lagrangian type iteration, where the

constraints are appended to the cost functional using Lagrange multipliers. The multipliers are then updated together with the level-set function, such that the constraints are fulfilled in the limit. This is implemented e.g. in [ADDM14, LLC+08]. In [WW04] the shape gradient is projected on the tangent cone of the admissible set. However, the shape D is infeasible for positive time steps, but the constraints are 'nearly' fulfilled. As pointed out in [AJT04, WCWM05], a drawback of the level set method is that nucleation of new holes in the material is not possible due to a maximum principle for the Hamilton-Jacobi equation. This can be overcome by placing enough holes in the initial guess or by other techniques as described in [YINT10, DAK13, AJ08], including the usage of topological derivatives. Moreover, the necessary reinitialization of the level-set function is expensive and no convergence proof for the numerical method is available [WW04]. Advantages of the level-set method are that other topological changes like splitting, merging and cancellation of holes are possible. In addition, the interface Γ is explicitly given as zero of the level-set function as opposed to other methods as SIMP or the homogenization method. Also multiple materials can be handled by the level-set method by using e.g. the 'color' level sets introduced in [VC02] for image processing and applied in [WW04, ADDM14] for topology optimization. In this method a distinct material is given by a combination of signs of N level set functions. Thus, 2^N material phases can be described using N level-set functions. We note that level-set methods also have important applications beyond topology optimization, e.g. in the computation of geometric PDEs [DDE05] or in image processing [TYW01, VC02].

In this thesis we concentrate on the **phase-field approach** to structural topology optimization. Originally introduced in [CH58] for the modelling of phase transitions where ideas go back to van der Waals [vdW93], the phase-field concept has nowadays applications in many areas, e.g. grain growth [Lus99], image processing [LK11], geometric PDEs [DDE05] and many more. Bourdin and Chambolle [BC00, BC03] were the first who used a phase field formulation for a topology optimization problem. The idea of the phase field method is a relaxation of the original problem (98) where the jumps of the characteristic function are replaced by smooth transitions from one phase to the other. Typically the phase field φ attains an approximate value of -1 in one phase (e.g. material) and an approximate value of $+1$ in the other phase (e.g. void) with a smooth transition layer called diffuse interface between the phases, whose width is controlled by the phase field parameter ε. As ε approaches 0, also the width of the diffuse interface layer tends to 0 and the phase field variable φ converges to the characteristic function (up to a rescaling) of the original topology optimization problem. The characteristic form of the phase field φ is maintained by the Ginzburg-Landau energy

$$E(\varphi) = \int_\Omega \{\frac{\varepsilon}{2}|\nabla\varphi|^2 + \frac{1}{\varepsilon}\psi(\varphi)\},$$

where $\psi : \mathbb{R} \to \overline{\mathbb{R}}$ is the potential or homogeneous free energy density with local minima at ± 1. The first term ensures the smoothness of the phase transition and the second term forces the values of φ to ± 1. The ε-scaling of the terms in the Ginzburg-Landau energy establishes an interface thickness proportional to ε. It is a famous result that the Ginzburg-Landau energy converges to a multiple of the perimeter of the set $\{x \in \Omega \mid \varphi(x) = 1\}$ as $\varepsilon \to 0$ in the sense of Γ-convergence [MM77, Mod87]. Thus, the phase field model perfectly fits into perimeter penalized topology optimization. Including a perimeter penalization in

the original optimization problem (98) as suggested in [AB93] leads to the problem

$$\begin{aligned} &\min f(\{\varphi = 1\}, \boldsymbol{u}(\{\varphi = 1\})) + \tilde{\gamma} P(\{\varphi = 1\}) \\ &\varphi : \Omega \to \{\pm 1\}, \\ &\{\varphi = 1\} \in \mathcal{X}, \end{aligned} \tag{99}$$

where $\tilde{\gamma} > 0$ is the weight of the penalization and $P(A)$ denotes the perimeter of the set A, which is roughly speaking the length of the boundary ∂A within Ω (for the exact definition see Definition 6.21). A phase field relaxation of (99) could then read

$$\begin{aligned} &\min f(\varphi, \boldsymbol{u}(\varphi)) + \gamma E(\varphi) \\ &\varphi : \Omega \to \mathbb{R}, \\ &\varphi \in \tilde{\mathcal{X}}. \end{aligned} \tag{100}$$

Since the phase field can now attain values other than ± 1 the stiffness tensor $\boldsymbol{C}(\varphi)$ in the state equation (97) has to be defined appropriately for all values of φ. Also f, γ and $\tilde{\mathcal{X}}$ have to be chosen appropriately such that (100) is an approximation of problem (99). Because the phase field relaxation (100) gives rise to an optimal control problem posed in a vector space, any classical optimization method can be used to solve the problem. Since the pioneering work of Bourdin and Chambolle, many authors have considered phase field relaxations of topology optimization problems [BFGS14, BS06, DBH12, GP12, GH14, PRW12, TNK10, TM14, WIR15, WZ07]. As a numerical method for (100) many authors consider a pseudo time stepping method. The basis is a gradient flow of the cost functional, which is discretized in time. Taking the gradient flow with respect to the L^2 inner product leads to a modified Allen-Cahn equation [AC79]. An H^{-1} gradient flow results in a modified Cahn-Hilliard equation [CH58], where mass is automatically conserved. Methods which are motivated by a gradient flow are used e.g. in [BFGS14, BGS+12, BC03, DBH12, GP12, TNK10, Tav14, WIR15, WR12, WZ07]. Adaptivity for the time step sizes is only utilized in [BC03, DBH12], where in the former paper the time step size is gradually increased, but only if the objective function can be decreased. In the latter the adaptive time stepping scheme is based on that in [GCH09], developed for the Cahn-Hilliard equation, which stems from a comparison of two different time stepping schemes. In the remaining papers, only constant time step sizes are considered. Other authors discretize first and then apply a classical optimization method to the discretized problem [BS06, PRW12, TM14]. In [WZ04a] a half-quadratic regularization together with a two-step alternating algorithm is applied, where the inner problem is solved with the heuristic OC method (see above). In [WZ04b] the MMA method (see above) is used. To our knowledge there are no convergence results for any of the cited algorithms in function space.

The phase field method can easily be generalized to problems involving multiple materials. The different subsets $D_1, \dots, D_N$ of Ω describing the N materials are first of all replaced by N characteristic functions, which are then relaxed to allow for a smooth transition between 0 and 1 on a lengthscale proportional to ε. This leads to a vector valued phase field $\boldsymbol{\varphi} = (\varphi_i)_{i=1}^N$, where $\boldsymbol{\varphi}(x) \approx \boldsymbol{e}_i$ if $x \in D_i$ (away from the interface). If an obstacle potential is used, then equality holds $\boldsymbol{\varphi}(x) = \boldsymbol{e}_i$ in D_i. A typical constraint is then that $\boldsymbol{\varphi}$ defines a partition of unity, i.e. $\varphi_i \geq 0$ for all i and $\sum_i \varphi_i = 1$ [EL91], and the potential ψ attains the value ∞ if $\boldsymbol{\varphi}$ does not fulfill these constraints, respectively. The Ginzburg-Landau energy can be generalized to be an approximation of a weighted sum of the interface lengths, see [Bal90] or Section 6.4.

The drawback of the phase field model is that an additional regularization parameter ε is introduced, which has to be driven to zero to obtain a solution of the original perimeter penalized problem (99). Thus, possibly a sequence of optimization problems has to be solved, which can be a time consuming task. Another disadvantage is that the number of phase field variables grows linearly with the number of materials, leading to huge optimization problems. In contrast, the SIMP peak function method [YA01] only needs a single function for the description of N materials and in the 'color' level-set method [WW04] only $\log_2 N$ level-set functions are needed. However, there are other multiphase concepts which need less phase field variables, e.g. in [BGN08] 2 functions are needed for 3 phases and in [WZ04b, BC03] 1 function is needed for 3 phases. However, there is no obvious extension to N phases. On the other hand there are many advantages. If convergence of the relaxed problem (100) to the original problem (99) can be shown as e.g. in [BC03], the phase field method provides a rigorous tool for solving perimeter penalized topology optimization problems. We will demonstrate this in Section 6.4. As already mentioned, the relaxed problem (100) is a smooth optimization problem (in contrast to the sharp interface problem (99)), for which the VMPT method is a rigorous solver as we will see in Section 6.7. Also the generalization to N phases is straight forward compared to other methods. Topological changes are handled implicitly and need no special treatment. In contrast to the level-set method, also nucleation of new holes is possible, which of course depends on the optimization method used, and also no reinitialization of the phase field variable is needed. As opposed to the SIMP method a well posed problem is solved (see e.g. [BGHR15]), thus the solutions are not mesh dependent and no filtering of densities or sensitivities is needed. Also intermediate densities from which the SIMP and homogenization methods suffer, are not a problem for the phase field method, since a clear 0-1 design is obtained up to the small diffuse interface, if ε is small enough. Thus no post processing is needed. Checkerboard patters, which often occur in the SIMP method, are not observed using the phase field model. Moreover, the phase field model can handle a large variety of cost functionals. Finally, the class of admissible shapes in the approximated sharp interface problem (99) is very large (these are sets of finite perimeter, see Definition 6.21) and no smoothness assumption on the admissible designs is imposed. In fact, not even a Lipschitz boundary is needed.

6 Phase field approach to structural topology optimization

6.1 Problem formulation

We give a short introduction in linearized elasticity and describe the used vector valued phase field model. For more details on linear or nonlinear elasticity we refer to [EGK08, Cia93] and for details about vector valued phase field models we refer to [EL91, GNS99].
Let $\Omega \subset \mathbb{R}^d$ be a nonempty bounded domain with Lipschitz-boundary, and $\Gamma_D \subset \partial\Omega$ with positive $(d-1)$ dimensional measure. We assume $|\Omega| > 0$.
The goal is to find an optimal distribution of N elastic homogeneous materials (including void in the sense of ersatz materials) within Ω, each with a prescribed mass (resp. volume) $\mathfrak{m}_i > 0$, such that some functional F is minimized. For a given material distribution $\boldsymbol{C}(x)$ the displacement field $\boldsymbol{u} : \Omega \to \mathbb{R}^d$ under certain loads is modelled by the equations of linearized elasticity

$$\begin{aligned} -\nabla \cdot \boldsymbol{\sigma}(\boldsymbol{u}) &= \boldsymbol{f} && \text{in } \Omega \\ \boldsymbol{u} &= \mathbf{0} && \text{on } \boldsymbol{\Gamma_D} \\ \boldsymbol{\sigma}(\boldsymbol{u}) \cdot \boldsymbol{n} &= \boldsymbol{g} && \text{on } \Gamma_g \\ \boldsymbol{\sigma}(\boldsymbol{u}) \cdot \boldsymbol{n} &= \mathbf{0} && \text{on } \partial\boldsymbol{\Omega} \setminus \{\boldsymbol{\Gamma_g} \cup \boldsymbol{\Gamma_D}\}. \end{aligned}$$

The material is fixed at the Dirichlet boundary $\Gamma_D \subset \partial\Omega$. For simplicity we only consider homogeneous Dirichlet boundary conditions (otherwise the problem has to be translated appropriately, see Section 6.1.2). A force $\boldsymbol{f} : \Omega \to \mathbb{R}^d$ is acting in Ω, and a boundary traction $\boldsymbol{g} : \Gamma_g \to \mathbb{R}^d$ is acting on some part $\Gamma_g \subset \partial\Omega$ of the boundary. The stress tensor is defined as

$$\boldsymbol{\sigma}(\boldsymbol{u}) = \boldsymbol{C}\mathcal{E}(\boldsymbol{u})$$

where $\mathcal{E}(\boldsymbol{u})$ denotes the linearized strain tensor

$$\mathcal{E}(\boldsymbol{u}) = \frac{1}{2}(D\boldsymbol{u} + D\boldsymbol{u}^T).$$

The weak form of the equations of linearized elasticity can be written as

$$\int_\Omega \boldsymbol{C}\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g} \cdot \boldsymbol{\xi} \quad \forall \boldsymbol{\xi} \in H^1_D, \tag{101}$$

where we seek $\boldsymbol{u}$ in the space

$$H^1_D := \{\boldsymbol{u} \in H^1(\Omega)^d \mid \boldsymbol{u}|_{\Gamma_D} = \mathbf{0} \text{ a.e.}\}.$$

Under certain assumptions, the weak equation (101) has a unique solution. An important tool for proving this is Korn's inequality [CDN10, Gob62]:

Lemma 6.1. *There exists a constant $C > 0$, s.t.*

$$\|D\boldsymbol{u}\|^2_{L^2} \le C(\|\mathcal{E}(\boldsymbol{u})\|^2_{L^2} + \|\boldsymbol{u}\|^2_{L^2}) \quad \forall \boldsymbol{u} \in H^1(\Omega)^d.$$

Taking the boundary conditions into account and using the compact embedding $H^1 \hookrightarrow L^2$ gives (see e.g. [Zei88] for the case $d = 3$ or Theorem 7.1)

$$\|D\boldsymbol{u}\|_{L^2} \le C\|\mathcal{E}(\boldsymbol{u})\|_{L^2} \quad \forall \boldsymbol{u} \in H^1_D,$$

which is also referred to as Korn's inequality. Using the Poincaré inequality in H^1_D [Alt12], we conclude that $\|\mathcal{E}(\boldsymbol{u})\|_{L^2}$ defines a norm on H^1_D, which is equivalent to the H^1-norm. This property will be used very often in the following.

In the phase field ansatz, the material distribution in Ω is described by means of a vector valued phase field $\boldsymbol{\varphi} : \Omega \to \mathbb{R}^N$, which gives in each point $x \in \Omega$ the volume fraction of each material. The presence of the i'th material at $x \in \Omega$ is thus given by $\boldsymbol{\varphi}(x) = \boldsymbol{e}_i$, where $\boldsymbol{e}_i$ is the ith unit vector in $\mathbb{R}^N$. Between the materials there is a thin diffuse interface where $\boldsymbol{\varphi}$ changes its value rapidly but smoothly. The prescription of the masses of the materials is modelled as a constraint

$$\fint_\Omega \boldsymbol{\varphi} := \frac{1}{|\Omega|} \int_\Omega \boldsymbol{\varphi} = \mathfrak{m}$$

on the phase field. The sum of the volume fractions of each material in one point $x \in \Omega$ should be one, and each fraction should be nonnegative, so we impose

$$\sum_{i=1}^N \varphi_i \equiv 1, \quad \boldsymbol{\varphi} \geq 0 \quad \text{a.e. in } \Omega,$$

where the latter has to be understood component-wise. A compatibility requirement is then

$$\sum_{i=1}^N \mathfrak{m}_i = 1.$$

Because the values of the stiffness tensor $\boldsymbol{C}$ depend on the material, we describe it as a function of the phase field $\boldsymbol{C} : \mathbb{R}^N \to \mathbb{R}^{d\times d} \otimes (\mathbb{R}^{d\times d})^*$, $\boldsymbol{\varphi} \mapsto \boldsymbol{C}(\boldsymbol{\varphi})$. Let $\boldsymbol{C}_i$ be the homogeneous elasticity tensor of the ith material, then we assume $\boldsymbol{C}(\boldsymbol{e}_i) = \boldsymbol{C}_i$. In our setting we allow also the forces $\boldsymbol{f}$ and $\boldsymbol{g}$ to depend on the material. For instance the gravity force density depends on the mass density of the material. Thus we consider $\boldsymbol{f} : \Omega \times \mathbb{R}^N \to \mathbb{R}^d$ and $\boldsymbol{g} : \Gamma_g \times \mathbb{R}^N \to \mathbb{R}^d$. The weak form of the state equation is thus

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}(x))\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) = \int_\Omega \boldsymbol{f}(x, \boldsymbol{\varphi}(x)) \cdot \boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g}(x, \boldsymbol{\varphi}(x)) \cdot \boldsymbol{\xi} \quad \forall \boldsymbol{\xi} \in H^1_D,$$

The Ginzburg-Landau energy for vector valued phase fields is given by

$$E(\boldsymbol{\varphi}) := \int_\Omega \left\{ \frac{\varepsilon}{2} |\nabla \boldsymbol{\varphi}|^2 + \frac{1}{\varepsilon} \psi(\boldsymbol{\varphi}) \right\},$$

where the parameter $\varepsilon > 0$ controls the thickness of the interface. The potential $\psi : \mathbb{R}^N \to \overline{\mathbb{R}}$ is chosen such that it attains its minima at the standard basis vectors $\boldsymbol{e}_i$, $i = 1, \dots, N$. In this thesis we will consider obstacle potentials, which are of the form

$$\psi(\boldsymbol{\varphi}) := \begin{cases} \psi_0(\boldsymbol{\varphi}) & \boldsymbol{\varphi} \geq 0,\ \sum_{i=1}^N \varphi_i \equiv 1 \\ \infty & \text{else} \end{cases}$$

for some smooth real-valued function ψ_0. In Section 6.13.1 we will argue that an obstacle potential has certain advantages over a smooth potential from a numerical point of view. In the following we will impose $\boldsymbol{\varphi} \geq 0$ and $\sum_{i=1}^N \varphi_i \equiv 1$ as hard constraints and will extend ψ smoothly by ψ_0 to $\mathbb{R}^N$. The resulting admissible set Φ_{ad} is thus bounded in L^∞, which we will frequently use in the analysis.

The gradient term in the Ginzburg-Landau energy ensures a smooth transition from one phase to another, whereas the potential term promotes the values $\boldsymbol{e}_i$, $i, \dots, N$ in the pure phases aside from the interface. The energy favors phase fields which take the values $\boldsymbol{e}_i$, $i, \dots, N$ on large areas in Ω, which are separated by a thin interfacial transition layer whose width is proportional to ε. As the interfacial parameter ε approaches zero, the Ginzburg-Landau energy approaches a surface energy of the hypersurfaces separating the distinct materials within Ω [Bal90].

6.1.1 General objective for multiple phases

As discussed in the last section the final topology optimization problem reads

$$\min \gamma E(\boldsymbol{\varphi}) + F(\boldsymbol{\varphi}, \boldsymbol{u}) \tag{102}$$

$$\boldsymbol{\varphi} \in H^1(\Omega)^N, \quad \boldsymbol{u} \in H^1_D$$

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) = \int_\Omega \boldsymbol{f}(x, \boldsymbol{\varphi}) \cdot \boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g}(x, \boldsymbol{\varphi}) \cdot \boldsymbol{\xi} \quad \forall \boldsymbol{\xi} \in H^1_D \tag{103}$$

$$\boldsymbol{\varphi} \geq 0$$

$$\sum_{i=1}^{N} \varphi_i = 1$$

$$\fint \boldsymbol{\varphi} = \mathfrak{m},$$

which has the typical form of an optimal control problem with control constraints, where $\boldsymbol{\varphi}$ is the control and $\boldsymbol{u}$ the state variable (cf. [Trö09]). For the analysis of the optimization problem we make use of the following assumptions:

For the elasticity tensor $\boldsymbol{C} = (C_{ijkl})_{ijkl=1}^d$ we assume:

(AP1) $\boldsymbol{C} : \mathbb{R}^N \to \mathbb{R}^{d\times d} \otimes (\mathbb{R}^{d\times d})^*$.

(AP2) $\boldsymbol{C}_{ijkl} \in C^{2,1}(\mathbb{R}^N)$.

(AP3) $\boldsymbol{C}_{ijkl} = \boldsymbol{C}_{jikl} = \boldsymbol{C}_{klij}$.

(AP4) There exist $0 < a_0 < a_1$ s.t. $a_0|\boldsymbol{A}|^2 \leq \boldsymbol{C}(\boldsymbol{\varphi})\boldsymbol{A} : \boldsymbol{A} \leq a_1|\boldsymbol{A}|^2$ for all symmetric matrices $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ and for all $\boldsymbol{\varphi} \in \mathbb{R}^N$.

From these assumptions it follows that $\boldsymbol{C}(\boldsymbol{\varphi})$ defines an inner product on the space of symmetric matrices for all $\boldsymbol{\varphi} \in \mathbb{R}^N$. In particular the Cauchy-Schwartz inequality holds and from **(AP4)** we conclude that

$$\boldsymbol{C}(\boldsymbol{\varphi})\boldsymbol{A} : \boldsymbol{B} \leq a_1|\boldsymbol{A}||\boldsymbol{B}| \quad \forall \boldsymbol{A},\ \boldsymbol{B} \in \mathbb{R}^{d\times d} \text{ symmetric},\ \forall \boldsymbol{\varphi} \in \mathbb{R}^N. \tag{104}$$

From **(AP2)**, it follows that $\boldsymbol{C}$ and $\boldsymbol{C}'$ both are locally Lipschitz continuous on $\mathbb{R}^N$, i.e.

$$\forall M \exists C(M) : \quad |\boldsymbol{C}(\boldsymbol{\varphi}_1) - \boldsymbol{C}(\boldsymbol{\varphi}_2)| \leq C(M)|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2| \text{ for all } \boldsymbol{\varphi}_i \in \mathbb{R}^N \text{ with } |\boldsymbol{\varphi}_i| \leq M,\ i = 1, 2 \tag{105}$$

and the same for $\boldsymbol{C}'$, due to the local boundedness of $\boldsymbol{C}'$ and $\boldsymbol{C}''$ and the mean value theorem. We also get that $\boldsymbol{C}$, $\boldsymbol{C}'$ and $\boldsymbol{C}''$ define Nemytskii operators from L^∞ into L^∞ since from $|\boldsymbol{\varphi}(x)| \leq M$ almost everywhere in Ω we conclude that $|\boldsymbol{C}(\boldsymbol{\varphi}(x))| \leq C(M)$ almost

everywhere in Ω for some constant $C(M)$. The same holds for $\boldsymbol{C}'$ and $\boldsymbol{C}''$.

To simplify notation we write $\boldsymbol{f}(\boldsymbol{\varphi})$ and $\boldsymbol{g}(\boldsymbol{\varphi})$ for the function $x \mapsto \boldsymbol{f}(x, \boldsymbol{\varphi}(x))$ and $x \mapsto \boldsymbol{g}(x, \boldsymbol{\varphi}(x))$, respectively. For these we assume:

(AP5) $\boldsymbol{f} : \Omega \times \mathbb{R}^N \to \mathbb{R}^d$, $\boldsymbol{g} : \Gamma_g \times \mathbb{R}^N \to \mathbb{R}^d$ are Carathéodory functions.

(AP6) $\boldsymbol{f} \in C^2(L^\infty(\Omega)^N, L^2(\Omega)^d)$, $\boldsymbol{g} \in C^2(L^\infty(\Gamma_g)^N, L^2(\Gamma_g)^d)$.

(AP7) $\boldsymbol{f}$ and $\boldsymbol{g}$ and their derivatives are locally Lipschitz, i.e.
$\forall M \exists L(M) : \ \|D^l_{\boldsymbol{\varphi}} \boldsymbol{f}(\boldsymbol{\varphi}_1) - D^l_{\boldsymbol{\varphi}} \boldsymbol{f}(\boldsymbol{\varphi}_2)\|_{L^2(\Omega)} \le L(M) \|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|_{L^\infty(\Omega)}$ for all $\boldsymbol{\varphi}_i \in L^\infty(\Omega)^N$ with $\|\boldsymbol{\varphi}_i\|_{L^\infty} \le M$, $i = 1, 2$ and $l = 0, 1, 2$.
$\forall M \exists L(M) : \ \|D^l_{\boldsymbol{\varphi}} \boldsymbol{g}(\boldsymbol{\varphi}_1) - D^l_{\boldsymbol{\varphi}} \boldsymbol{g}(\boldsymbol{\varphi}_2)\|_{L^2(\Gamma_g)} \le L(M) \|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|_{L^\infty(\Gamma_g)}$ for all $\boldsymbol{\varphi}_i \in L^\infty(\Gamma_g)^N$ with $\|\boldsymbol{\varphi}_i\|_{L^\infty} \le M$, $i = 1, 2$ and $l = 0, 1, 2$

A Carathéodory function is by definition measurable in the first argument for any fixed second argument, and continuous in the second argument for almost every fixed first argument [Sho97]. A sufficient condition for **(AP6)** and **(AP7)** is the C^2 regularity of $\boldsymbol{f} : \Omega \times \mathbb{R}^N \to \mathbb{R}^d$ with respect to $\boldsymbol{\varphi}$, together with the local boundedness

$$\forall M > 0 \ \exists C(M), \ \forall \boldsymbol{\varphi} \in \mathbb{R}^N, \ |\boldsymbol{\varphi}| \le M : \ |\boldsymbol{f}_{\varphi,\varphi}(x, \boldsymbol{\varphi})| \le C(M) \text{ a.e. in } \Omega$$

and that $\boldsymbol{f}$ (and its derivatives $\boldsymbol{f}_\varphi$, $\boldsymbol{f}_{\varphi,\varphi}$) is locally Lipschitz in its second argument for almost every fixed first argument, and the Lipschitz constant does not depend on the first argument, see [GKT92]. In this case **(AP6)** and **(AP7)** even hold if L^2 is replaced by L^∞. Analogously for $\boldsymbol{g}$.

From **(AP7)** it follows the local boundedness

$$\|D^l_{\boldsymbol{\varphi}} \boldsymbol{f}(\boldsymbol{\varphi})\|_{L^2(\Omega)} \le C(M) \text{ for all } \boldsymbol{\varphi} \in L^\infty(\Omega)^N \text{ with } \|\boldsymbol{\varphi}\|_{L^\infty} \le M \text{ and } l = 0, 1, 2, \tag{106}$$

$$\|D^l_{\boldsymbol{\varphi}} \boldsymbol{g}(\boldsymbol{\varphi})\|_{L^2(\Gamma_g)} \le C(M) \text{ for all } \boldsymbol{\varphi} \in L^\infty(\Gamma_g)^N \text{ with } \|\boldsymbol{\varphi}\|_{L^\infty} \le M \text{ and } l = 0, 1, 2. \tag{107}$$

From the trace estimate in Lemma 7.4 we get that $\boldsymbol{g}$ is a well defined operator from $H^1(\Omega)^N \cap L^\infty(\Omega)^N$ into $L^2(\Gamma_g)^d$, which is two times continuously Fréchet differentiable, even if $H^1(\Omega)^N \cap L^\infty(\Omega)^N$ is equipped with the $L^\infty(\Omega)^N$ norm.

Taking the preceding assumptions into account, we prove in Theorem 6.6 that for each control $\boldsymbol{\varphi}$ there exists a unique state $\boldsymbol{u}$ solving the state equation (103). This gives rise to the *control-to-state operator*, which we denote in the following by S, i.e. $S(\boldsymbol{\varphi}) := \boldsymbol{u}$.

Denote the set of admissible controls by $\Phi_{ad} \subset H^1(\Omega)^N \cap L^\infty(\Omega)^N$.
The assumptions on the cost functional are:

(AP8) $F \in C^2((H^1(\Omega)^N \cap L^\infty(\Omega)^N) \times H^1_D)$.

(AP9) There exists $C > 0$ s.t. $F(\boldsymbol{\varphi}, \boldsymbol{u}) \ge -C$ for all $\boldsymbol{\varphi} \in \Phi_{ad}$, where $\boldsymbol{u} = S(\boldsymbol{\varphi})$ is the corresponding state.

(AP10) It holds the following lower semi-continuity: Let $\boldsymbol{\varphi}_n \in \Phi_{ad}$ be a sequence of controls and $\boldsymbol{u}_n \in H^1_D$ the corresponding sequence of states. Let $\boldsymbol{\varphi}_n \to \boldsymbol{\varphi}$ weakly in H^1 and $\boldsymbol{u}_n \to \boldsymbol{u}$ weakly in H^1 for some $\boldsymbol{\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$ and $\boldsymbol{u} \in H^1_D$, respectively. Then $\liminf_{n\to\infty} F(\boldsymbol{\varphi}_n, \boldsymbol{u}_n) \ge F(\boldsymbol{\varphi}, \boldsymbol{u})$.

(AP11) $\psi_0 \in C^{2,1}(\mathbb{R}^N)$.

The assumptions **(AP8)-(AP10)** are also true for the Ginzburg-Landau energy, which will be proved in Theorem 6.23 and Theorem 6.18.
Further compatibility assumptions are:

(AP12) $\mathfrak{m} \geq 0$.

(AP13) $\sum_{i=1}^{N} \mathfrak{m}_i = 1$.

We emphasize that we don't impose conditions on the space dimension d.
We introduce the $(N-1)$-dimensional standard simplex or Gibbs simplex

$$\Delta^{N-1} := \left\{ x \in \mathbb{R}^N \;\middle|\; \sum_{i=1}^{N} x_i = 1 \text{ and } x_i \geq 0 \text{ for all } i \right\}.$$

Note that $\boldsymbol{\varphi}(x) \in \Delta^{N-1}$ holds almost everywhere for all $\boldsymbol{\varphi} \in \Phi_{ad}$.

Analysis for special cases of the topology optimization problem (102) is performed in [BFGS14, BGHR15]. In [BFGS14] the cost functional F is the compliance or a tracking type functional. A special case of $\boldsymbol{f}$ is considered and $\boldsymbol{g}$ is independent of $\boldsymbol{\varphi}$. Well-posedness of the problem is shown as well as the Fréchet differentiability of the reduced cost functional. First order necessary conditions are deduced and sharp interface asymptotics are derived formally. In [BGHR15] the case $N = 2$ is considered, but the cost functional F can be arbitrary. Well-posedness is shown and first order optimality conditions are derived. The existence of a Lagrange multiplier for the mass inequality constraint is shown. Moreover, the sharp interface limit is established in the sense of Γ-convergence and convergence of the optimality condition is shown. Special cases of (102) are also considered in [WR12, BGS$^+$12, Sar10], but no analysis is performed.
In the following we generalize the results in [BFGS14] for arbitrary cost functional F and forces $\boldsymbol{f}(\boldsymbol{\varphi})$ and $\boldsymbol{g}(\boldsymbol{\varphi})$. We show not only Fréchet differentiability but C^2-regularity of the reduced cost functional by means of the implicit function theorem in contrast to [BFGS14]. We also generalize the results in [BGHR15] in the sense that we show also existence of Lagrange multipliers for the inequality and the sum constraints and not only for the mass constraint using a result of Zowe and Kurcyusz [ZK79]. Additionally we show uniqueness of Lagrange multipliers using ideas from [BGSS13a]. However, the pointwise formulation used in [BGSS13a] is not possible anymore in our abstract setting and thus we have to develop variational techniques to show uniqueness. We also transfer the arguments for Γ-convergence in [BGHR15] partially to the problem (102).

As pointed out in Section 5, most of the existing literature on phase field relaxations of topology optimization problems consider a pseudo time stepping scheme as numerical method. No convergence proof is yet available. In the following we will apply the VMPT method on the problem (102), for which global convergence in the continuous setting is given. We will check the abstract assumptions for global convergence in Section 6.7 and give many examples of possible inner products for the VMPT method. Moreover, it turns out that the pseudo time stepping approaches in [BGS$^+$12, BFGS14] are special instances of the VMPT method and thus we are able to show global convergence for these methods in Section 6.8, where we also propose an adaptivity strategy for the pseudo time step size based on the Armijo condition. In particular it is possible that the time step size tends to infinity without destroying global convergence.

In Section 6.10 we propose a primal dual active set method as solver for the projection type subproblem in the VMPT method. The convergence analysis is performed for the discretized problem only.
Many numerical results are presented in Section 6.14 and 6.13. Amongst others we will study the dependency of the VMPT method on various parameters, support the theoretical results, such as global convergence in the continuous setting, by numerical experiments, and compare our results to the literature. We refer to the introduction of the respective sections for an overview of the numerical results.

6.1.2 Examples

We give an example for a volume force $\boldsymbol{f}$. Let $\boldsymbol{f}_i \in L^2(\Omega)^d$ be a volume force which only acts on the ith material. Define

$$\boldsymbol{f}(x, \boldsymbol{\varphi}) := \sum_{i=1}^{N} \varphi_i \boldsymbol{f}_i(x) \tag{108}$$

for all $\boldsymbol{\varphi} \in \mathbb{R}^N$ and almost every $x \in \Omega$. Then $\boldsymbol{f}(x, \boldsymbol{\varphi}(x)) = \boldsymbol{f}_i(x)$ if x lies in the ith material and $\boldsymbol{f}$ interpolates the forces $\boldsymbol{f}_i$ linearly on the interface. This also includes the case when $\boldsymbol{f}$ is a force independent of $\boldsymbol{\varphi}$ and the case $\boldsymbol{f}(\boldsymbol{\varphi}) = (1 - \varphi_N)\tilde{\boldsymbol{f}} = \sum_{i=1}^{N-1} \varphi_i \tilde{\boldsymbol{f}}$ for $\boldsymbol{\varphi} \in \Delta^{N-1}$, where $\tilde{\boldsymbol{f}}$ is a force only acting on the material but not on the void phase defined by $\{\varphi_N = 1\}$ as discussed in [BFGS14]. When considering gravity, we have $\boldsymbol{f}_i = \rho_i \boldsymbol{g}$, where ρ_i is the mass density of the i-th material and $\boldsymbol{g}$ is the gravitational acceleration vector.
This $\boldsymbol{f}$ defined in (108) fulfills the assumptions. Since $\boldsymbol{f}$ is linear and continuous in $\boldsymbol{\varphi} \in L^\infty(\Omega)^N$, which follows from the estimate

$$\|\boldsymbol{f}(\boldsymbol{\varphi})\|_{L^2} \le \sum_{i=1}^{N} \|\varphi_i\|_{L^\infty} \|\boldsymbol{f}_i\|_{L^2} \le \sum_{i=1}^{N} \|\boldsymbol{f}_i\|_{L^2} \|\boldsymbol{\varphi}\|_{L^\infty},$$

we get $\boldsymbol{f} \in C^\infty(L^\infty(\Omega)^N, L^2(\Omega)^d)$ and in particular **(AP6)**. For linear functions, continuity and Lipschitz continuity is the same, thus also **(AP7)** is fulfilled for $l = 0$. The cases $l = 1$ and $l = 2$ are trivial.
The boundary traction $\boldsymbol{g}$ can also be a linear interpolation

$$\boldsymbol{g}(x, \boldsymbol{\varphi}) := \sum_{i=1}^{N} \varphi_i \boldsymbol{g}_i(x)$$

with $\boldsymbol{g}_i \in L^2(\Gamma_g)^d$. Thus we can handle body forces and boundary tractions which depend on the material phase. For a sharp interface formulation thereof in the context of level set functions we refer e.g. to [WW04]. Another application where the load vectors depend on the design variable is in the design of multiphysics actuators [Sig01]. Eigenstrain and design dependent loads are discussed in Remark 6.13.

An example for the stiffness tensor interpolation is as follows. Let $\boldsymbol{C}_i \in \mathbb{R}^{d\times d} \otimes (\mathbb{R}^{d\times d})^*$ be the stiffness tensor of the i-th material, fulfilling **(AP3)** and **(AP4)**. A possibility is to interpolate the tensors linearly, i.e. take

$$\boldsymbol{C}(\boldsymbol{\varphi}) = \sum_{i=1}^{N} \varphi_i \boldsymbol{C}_i \tag{109}$$

for $\boldsymbol{\varphi} \in \Delta^{N-1}$ with a suitable extension to $\mathbb{R}^N$. In [BFGS14] such an extension is given

such that $\boldsymbol{C} \in C^{1,1}(\mathbb{R}^N, \mathbb{R}^{d\times d} \otimes (\mathbb{R}^{d\times d})^*)$. This construction can easily be modified such that $\boldsymbol{C} \in C^{2,1}(\mathbb{R}^N, \mathbb{R}^{d\times d} \otimes (\mathbb{R}^{d\times d})^*)$ holds, i.e. we take $\boldsymbol{C}(\boldsymbol{\varphi}) = \sum_{i=1}^N w(P(\varphi)_i)\boldsymbol{C}_i$ where $w \in C^{2,1}(\mathbb{R})$ is monotone and it holds $w = id$ on $[0,1]$ and $-\delta < w < 1+\delta$ for some $\delta > 0$. Here, $P(\boldsymbol{\varphi}) = (\varphi_i - \frac{1}{N}((\sum_{j=1}^N \varphi_j) - 1))_i$ denotes the orthogonal projection onto the affine space $\{\boldsymbol{\varphi} \in \mathbb{R}^N \mid \sum_{i=1}^N \varphi_i = 1\}$. If δ is small enough, then $\boldsymbol{C}$ fulfills assumption **(AP4)** [BFGS14]. **(AP1)-(AP3)** are trivial.
Another possibility is to interpolate the stiffness tensors quadratically, e.g. take

$$\boldsymbol{C}(\boldsymbol{\varphi}) = \sum_{i,j=1}^N \varphi_i\varphi_j \boldsymbol{C}_{max\{i,j\}} \tag{110}$$

for $\boldsymbol{\varphi} \in \Delta^{N-1}$, where the stiffness tensors $\boldsymbol{C}_i$ are ordered from high stiffness to low stiffness. The interpolation is constructed such that the minimum of the interpolation along an edge of the Gibbs simplex is attained at the weaker phase, i.e. $\boldsymbol{C}'(\boldsymbol{e}_i)(\boldsymbol{e}_k - \boldsymbol{e}_i) = 0$ for $k \le i$. This interpolation will prove to be numerically advantageous compared to the linear interpolation for compliance minimization, see Section 6.13.2. Note that a similar quadratic interpolation in the special case of two phases is used in [WR12, PRW12, BC06]. Recall that also in the SIMP method higher order interpolations of the stiffness tensor is used to penalize intermediate densities.
A possible extension of (110) to $\mathbb{R}^N$ is $\boldsymbol{C}(\boldsymbol{\varphi}) = \sum_{i,j=1}^N w(P(\varphi)_i)w(P(\varphi)_j)\boldsymbol{C}_{max\{i,j\}}$, where w and P are as above.

Lemma 6.2. *Let w, P and $\boldsymbol{C}_i$, $i = 1,\dots,N$ as above. Then the interpolation*

$$\boldsymbol{C}(\boldsymbol{\varphi}) = \sum_{i,j=1}^N w(P(\varphi)_i)w(P(\varphi)_j)\boldsymbol{C}_{max\{i,j\}}, \quad \boldsymbol{\varphi} \in \mathbb{R}^N \tag{111}$$

*fulfills the assumption **(AP4)** if δ is small enough.*

Proof. Let $\theta > 0$ and $\Theta >$ denote the constants such that $\theta|B|^2 \le \boldsymbol{C}_iB : B \le \Theta|B|^2$ for all $i = 1,\dots,N$ and all symmetric matrices B. Let $A \in \mathbb{R}^{d\times d}$ be symmetric. We have

$$\boldsymbol{C}(\boldsymbol{\varphi})A : A = \sum_{i=1}^N w(P(\varphi)_i)^2\boldsymbol{C}_iA : A + 2\sum_{\substack{i=1\\ j<i}}^N w(P(\varphi)_i)w(P(\varphi)_j)\boldsymbol{C}_iA : A.$$

Since $\sum_{i=1}^N P(\varphi)_i = 1$ we get $P(\varphi)_j \ge \frac{1}{N}$ for some j. Since w is monotone (and positive on $[1/N, \infty)$) we get $w(P(\varphi)_j)^2 \ge w(\frac{1}{N})^2 = \left(\frac{1}{N}\right)^2$. We estimate $w(P(\varphi)_i)^2\boldsymbol{C}_iA : A \ge 0$ for all $i \ne j$ to get

$$\sum_{i=1}^N w(P(\varphi)_i)^2\boldsymbol{C}_iA : A \ge \theta\left(\frac{1}{N}\right)^2|A|^2.$$

In the second sum we can estimate $w(P(\varphi)_i)w(P(\varphi)_j)\boldsymbol{C}_iA : A \ge 0$ in the case that $w(P(\varphi)_i)w(P(\varphi)_j) \ge 0$. If $w(P(\varphi)_i)w(P(\varphi)_j) < 0$ we can assume that $-\delta \le w(P(\varphi)_i) < 0$ and $0 < w(P(\varphi)_j) \le 1+\delta$. We get

$$w(P(\varphi)_i)w(P(\varphi)_j)\boldsymbol{C}_iA : A \ge -\delta(1+\delta)\Theta|A|^2$$

and finally

$$\boldsymbol{C}(\boldsymbol{\varphi})A : A \geq \left(\theta\left(\frac{1}{N}\right)^2 - 2N^2\delta(1+\delta)\Theta\right)|A|^2.$$

We can choose $\delta > 0$ small enough, such that $\theta\left(\frac{1}{N}\right)^2 - 2N^2\delta(1+\delta)\Theta > 0$. The upper bound in **(AP4)** we get by

$$\boldsymbol{C}(\boldsymbol{\varphi})A : A \leq N^2(1+\delta)^2\Theta|A|^2.$$

□

We remark that the choice of the extension of $\boldsymbol{C}(\boldsymbol{\varphi})$ for $\varphi \notin \Delta^{N-1}$ does not influence the VMPT method. This is because all iterates are feasible, i.e. $\boldsymbol{\varphi}_k \in \Delta^{N-1}$ for all k. Also the local minima of the optimization problem don't depend on the extension of $\boldsymbol{C}(\boldsymbol{\varphi})$. However, we need the existence of an extension to show that the objective is differentiable in a neighborhood of Φ_{ad}.

A common choice for the potential is

$$\psi_0(\boldsymbol{\varphi}) = -\frac{1}{2}\boldsymbol{\varphi}^T A\boldsymbol{\varphi},$$

for some symmetric $A \in \mathbb{R}^{N\times N}$ having at least one positive eigenvalue, cf. the deep quench limit problem in [EL91]. Note that it is desired that for $\varepsilon \to 0$, the optimal control only has values in $\{\boldsymbol{e}_i \mid i = 1,\dots,N\}$, which means that no interface is present. Therefore one can choose $\psi_0 \geq 0$ in Δ^{N-1} with $\psi_0(\boldsymbol{\varphi}) = 0$ if and only if $\boldsymbol{\varphi} = \boldsymbol{e}_i$ for some $i \in \{1,\dots,N\}$, i.e. the global minima are at the corners of the Gibbs simplex. This is the case for $A = (a_{ij})_{ij}$ with $a_{ii} = 0$, $i = 1,\dots,N$ and $a_{ij} < 0$ for $i \neq j$. However, in the analysis for positive ε it is not important that ψ_0 has minima at $\boldsymbol{e}_i$ and therefore we do not list this as an assumption. Potentials of higher order can be found in [GNS99].

Finally we give examples for the cost functional F. First of all there is the widely used mean compliance functional

$$F(\boldsymbol{\varphi}, \boldsymbol{u}) = \int_\Omega \boldsymbol{f}(\boldsymbol{\varphi})\cdot\boldsymbol{u} + \int_{\Gamma_g} \boldsymbol{g}(\boldsymbol{\varphi})\cdot\boldsymbol{u}, \tag{112}$$

which is the work done by the applied forces $\boldsymbol{f}$ and $\boldsymbol{g}$. Minimizing the compliance is equivalent to maximizing the stiffness of the structure under the given loads.

Lemma 6.3. *The mean compliance functional* (112) *fulfills assumptions* ***(AP8)-(AP10)****.*

Proof. The differentiability **(AP8)** follows from the chain rule, noting that the mappings

$$L^2(\Omega)^d \times H^1_D \ni (\boldsymbol{f}, \boldsymbol{u}) \mapsto \int_\Omega \boldsymbol{f}\cdot\boldsymbol{u}$$
$$L^2(\Gamma_g)^d \times L^2(\Gamma_g)^d \ni (\boldsymbol{g}, \boldsymbol{u}) \mapsto \int_{\Gamma_g} \boldsymbol{g}\cdot\boldsymbol{u}$$

are bilinear and continuous and thus smooth, the trace theorem and Lemma 7.4.
To see that F is bounded from below for feasible pairs $(\boldsymbol{\varphi}, \boldsymbol{u})$, where $\boldsymbol{\varphi} \in \Phi_{ad}$ and $\boldsymbol{u} = S(\boldsymbol{\varphi})$,

we test the state equation by $\boldsymbol{\xi} = \boldsymbol{u} \in H^1_D$ and use **(AP4)** to get

$$F(\boldsymbol{\varphi}, \boldsymbol{u}) = \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{u}) \geq C\|\mathcal{E}(\boldsymbol{u})\|^2_{L^2} \geq 0. \tag{113}$$

It remains to show the lower semi-continuity **(AP10)**. Let $(\boldsymbol{\varphi}_n)_n \subset \Phi_{ad}$ with $\boldsymbol{\varphi}_n \to \boldsymbol{\varphi}$ weakly in H^1 for some $\boldsymbol{\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$ and let $\boldsymbol{u}_n := S(\boldsymbol{\varphi}_n) \to \boldsymbol{u}$ weakly in H^1 for some $\boldsymbol{u} \in H^1_D$. From compact embeddings we get $\boldsymbol{u}_n \to \boldsymbol{u}$ in $L^2(\Omega)$ and in $L^2(\partial\Omega)$ in the trace sense [Alt12, A6.13]. The same holds for the sequence $\boldsymbol{\varphi}_n$. After possibly choosing a subsequence we get $\boldsymbol{\varphi}_n \to \boldsymbol{\varphi}$ almost everywhere in Ω and almost everywhere in $\partial\Omega$ in the trace sense. Since $\boldsymbol{f}$ and $\boldsymbol{g}$ are Carathéodory, we get $\boldsymbol{f}(x, \boldsymbol{\varphi}_n(x)) \to \boldsymbol{f}(x, \boldsymbol{\varphi}(x))$ almost everywhere in Ω and $\boldsymbol{g}(x, \boldsymbol{\varphi}_n(x)) \to \boldsymbol{g}(x, \boldsymbol{\varphi}(x))$ almost everywhere in Γ_g. Since Φ_{ad} is bounded in L^∞ we get that $\boldsymbol{f}(\boldsymbol{\varphi}_n)$ as well as $\boldsymbol{g}(\boldsymbol{\varphi}_n)$ is uniformly bounded in L^2 (see (106)-(107) and Lemma 7.4). Therefore we get a weakly converging subsequence, and Egorov's theorem [Alt12, A1.18] yields that the weak limit coincides with the pointwise limit, i.e.

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{\varphi}_n) &\to \boldsymbol{f}(\boldsymbol{\varphi}) \quad \text{weakly in } L^2(\Omega), \\ \boldsymbol{g}(\boldsymbol{\varphi}_n) &\to \boldsymbol{g}(\boldsymbol{\varphi}) \quad \text{weakly in } L^2(\Gamma_g), \end{aligned}$$

thus

$$\begin{aligned} \int_\Omega \boldsymbol{f}(\boldsymbol{\varphi}_n) \cdot \boldsymbol{u}_n &\to \int_\Omega \boldsymbol{f}(\boldsymbol{\varphi}) \cdot \boldsymbol{u}, \\ \int_{\Gamma_g} \boldsymbol{g}(\boldsymbol{\varphi}_n) \cdot \boldsymbol{u}_n &\to \int_{\Gamma_g} \boldsymbol{g}(\boldsymbol{\varphi}) \cdot \boldsymbol{u}. \end{aligned}$$

The preceding holds for a subsequence. Since the limit is unique we get by Lemma 7.3 convergence of the whole sequence. We conclude

$$F(\boldsymbol{\varphi}_n, \boldsymbol{u}_n) \to F(\boldsymbol{\varphi}, \boldsymbol{u}).$$

We even get continuity and not only lower semi-continuity. □

Another choice for the cost functional is related to the so called compliant mechanism problem

$$F(\boldsymbol{\varphi}, \boldsymbol{u}) = \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{u}_\Omega\|^2_{L^2},$$

which is of tracking type. Here $\boldsymbol{u}_\Omega \in L^2(\Omega)^d$ is the desired displacement of the structure under the given loads. It is also possible to include a weighting factor $c \in L^\infty(\Omega)$ with $c \geq 0$ a.e.,

$$F(\boldsymbol{\varphi}, \boldsymbol{u}) = \frac{1}{2}\int_\Omega c|\boldsymbol{u} - \boldsymbol{u}_\Omega|^2.$$

This also includes the case where $\boldsymbol{u}$ is only tracked in some measurable subset $A \subset \Omega$ by choosing the characteristic function $c = \chi_A$. Moreover there are applications in which only the displacement of a certain material is to be tracked. In this case the weight c can depend on the values of the phase field $\boldsymbol{\varphi}$, which leads to

$$F(\boldsymbol{\varphi}, \boldsymbol{u}) = \frac{1}{2}\int_\Omega c(x, \boldsymbol{\varphi})|\boldsymbol{u} - \boldsymbol{u}_\Omega|^2. \tag{114}$$

Here we assume that $c : \Omega \times \mathbb{R}^N \to \mathbb{R}$ is a Carathéodory function which fulfills $c \in C^2(L^\infty(\Omega)^N, L^\infty(\Omega))$ in the sense of Nemytskii operators and $c(x, \boldsymbol{\varphi}) \geq 0$ for almost all $x \in \Omega$ and all $\boldsymbol{\varphi} \in \Delta^{N-1}$. Note that a necessary condition therefor is the local boundedness of c. In particular we have $|c(x, \boldsymbol{\varphi})| \leq C$ a.e. in Ω for all $\boldsymbol{\varphi} \in \Delta^{n-1}$ (see [GKT92, Thm. 3]). A special case of this function is discussed in [BFGS14], which is given by $c(x, \boldsymbol{\varphi}) = \overline{c}(x)(1 - \varphi_N)$ for some non-negative $\overline{c} \in L^\infty(\Omega)$. In [BFGS14] the set $\{\varphi_N = 1\}$ corresponds to the void phase, where the displacement is not tracked.

Lemma 6.4. *Let $\boldsymbol{u}_\Omega$ and c be as above. Then the compliant mechanism functional* (114) *fulfills the assumptions **(AP8)-(AP10)**.*

Proof. Differentiability of F follows from the chain rule, where we use that

$$L^\infty(\Omega) \times L^1(\Omega) \ni (c, u) \mapsto \int_\Omega cu \in \mathbb{R} \quad \text{and}$$
$$L^2(\Omega)^d \times L^2(\Omega)^d \ni (\boldsymbol{u}_1, \boldsymbol{u}_2) \mapsto \boldsymbol{u}_1 \cdot \boldsymbol{u}_2 \in L^1(\Omega)$$

are smooth functions.
By the assumption on c we have

$$F(\boldsymbol{\varphi}, \boldsymbol{u}) = \frac{1}{2} \int_\Omega c(\boldsymbol{\varphi}) |\boldsymbol{u} - \boldsymbol{u}_\Omega|^2 \geq 0$$

for all $\boldsymbol{\varphi} \in \Phi_{ad}$ and all $\boldsymbol{u} \in H^1_D$, thus **(AP9)** holds.
It remains to show the lower semi-continuity **(AP10)**. Let $(\boldsymbol{\varphi}_n)_n \subset \Phi_{ad}$ with $\boldsymbol{\varphi}_n \to \boldsymbol{\varphi}$ weakly in H^1 for some $\boldsymbol{\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$ and let $\boldsymbol{u}_n := S(\boldsymbol{\varphi}_n) \to \boldsymbol{u}$ weakly in H^1 for some $\boldsymbol{u} \in H^1_D$. After possibly choosing a subsequence we get $\boldsymbol{\varphi}_n \to \boldsymbol{\varphi}$ and $\boldsymbol{u}_n \to \boldsymbol{u}$ almost everywhere in Ω and thus

$$c(x, \boldsymbol{\varphi}_n(x)) |\boldsymbol{u}_n(x) - \boldsymbol{u}_\Omega(x)|^2 \to c(x, \boldsymbol{\varphi}(x)) |\boldsymbol{u}(x) - \boldsymbol{u}_\Omega(x)|^2 \quad \text{almost everywhere in } \Omega.$$

Since the integrands are non-negative we can apply Fatou's lemma to obtain

$$\liminf_{n \to \infty} \int_\Omega c(\boldsymbol{\varphi}_n) |\boldsymbol{u}_n - \boldsymbol{u}_\Omega|^2 \geq \int_\Omega c(\boldsymbol{\varphi}) |\boldsymbol{u} - \boldsymbol{u}_\Omega|^2. \tag{115}$$

It remains to show that (115) holds also for the whole sequence. We use arguments similar to Lemma 7.3. Assume $\liminf_{n\to\infty} \int_\Omega c(\boldsymbol{\varphi}_n) |\boldsymbol{u}_n - \boldsymbol{u}_\Omega|^2 < \int_\Omega c(\boldsymbol{\varphi}) |\boldsymbol{u} - \boldsymbol{u}_\Omega|^2$. We take a subsequence which converges to the $\liminf$. By the arguments above we get (115) for a subsequence of the subsequence, which is a contradiction.
We note that also continuity can be shown instead of lower semi-continuity, since $c(\boldsymbol{\varphi}_n) \to c(\boldsymbol{\varphi})$ weakly-* in L^∞, which follows from the uniform boundedness of $c(\boldsymbol{\varphi}_n)$ in L^∞, and $\boldsymbol{u}_n \to \boldsymbol{u}$ strongly in L^2. □

It is also possible to track the displacement only on the boundary in the sense of traces:

$$F(\boldsymbol{\varphi}, \boldsymbol{u}) = \frac{1}{2} \int_{\partial\Omega} c(\boldsymbol{\varphi}) |\boldsymbol{u} - \boldsymbol{u}_\Gamma|^2.$$

Another compliant mechanism functional given in [Sig97] and also used in [YINT10] is

$$F(\boldsymbol{\varphi}, \boldsymbol{u}) = - \int_{\Gamma_{out}} \boldsymbol{g}_{out} \cdot \boldsymbol{u}, \tag{116}$$

where $\Gamma_{out} \subset \partial\Omega$ is the output port and $\boldsymbol{g}_{out} \in L^2(\Gamma_{out})^d$ is a dummy traction on Γ_{out},

which describes the desired direction for $\boldsymbol{u}$ in Γ_{out}. Therefore, the displacement $\boldsymbol{u}$ is maximized at Γ_{out} in direction $\boldsymbol{g}_{out}$. It is straight forward to show **(AP8)**-**(AP10)**, where for **(AP9)** one needs $\|S(\boldsymbol{\varphi})\|_{H^1} \leq C$ for all $\boldsymbol{\varphi} \in \Phi_{ad}$, which follows from the a priori estimate (123) proved later. Of course, Γ_{out} can also be replaced by some measurable $\Omega_{out} \subset \Omega$ to be able to control u in the interior of Ω.

In addition to the above mentioned functionals one can add certain penalization terms in order to penalize unwanted solutions or structures. If the hard mass constraint is not present one can add the term

$$+ \beta \left| \int_\Omega \boldsymbol{\varphi} \, \mathrm{d}x - \mathfrak{m} \right|^2$$

to F, which penalizes the deviation from the desired mass $\mathfrak{m}$ with a penalization factor $\beta > 0$.
If one doesn't want that e.g. material 2 and material 3 meet at a common boundary one can add

$$+ \beta \int_\Omega \varphi_2 \varphi_3 \tag{117}$$

to the functional. One can think of material 3 being void and material 2 being prone to rust. Thus one wants that material 2 is put in the interior of the structure. To avoid that material 2 is put on the boundary $\partial\Omega$, one can add the term

$$+ \beta \int_{\partial\Omega} \varphi_2. \tag{118}$$

It can be easily shown that these three penalization terms fulfill the assumptions **(AP8)**-**(AP10)**. We refer to Section 6.13.6 for a numerical example using the previous two penalizations.
Although the VMPT method cannot handle a hard stress constraint $\sigma_{min} \leq \sigma \leq \sigma_{max}$ as e.g. in [BS06], since this results in a non-convex feasible set for $\boldsymbol{\varphi}$, we can nevertheless penalize large stresses $\sigma = \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u})$ in certain subsets of Ω by adding the term

$$+ \beta \int_\Omega c(\boldsymbol{\varphi}) |\boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u})|^2,$$

where c is as in the compliant mechanism functional (114). The minimal stress functional can also be used as stand alone functional rather than as penalization term as in [AJ08].

Lemma 6.5. *Let c be as above. Then the functional*

$$\int_\Omega c(\boldsymbol{\varphi}) |\boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u})|^2 \tag{119}$$

*fulfills the assumptions **(AP8)**-**(AP10)**.*

Proof. **(AP8)** follows from the chain rule, where the required regularity of $\boldsymbol{C}$ will be shown in Lemma 6.10. **(AP9)** is obvious. It remains **(AP10)**. Let $(\boldsymbol{\varphi}_n)_n \subset \Phi_{ad}$ with $\boldsymbol{\varphi}_n \to \boldsymbol{\varphi}$ weakly in H^1 for some $\boldsymbol{\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$ and let $\boldsymbol{u}_n := S(\boldsymbol{\varphi}_n) \to \boldsymbol{u}$ weakly in H^1 for some $\boldsymbol{u} \in H^1_D$. After possibly choosing a subsequence we get $\boldsymbol{\varphi}_n \to \boldsymbol{\varphi}$ almost everywhere in Ω and thus $c(\boldsymbol{\varphi}_n) \to c(\boldsymbol{\varphi})$ and $\boldsymbol{C}(\boldsymbol{\varphi}_n) \to \boldsymbol{C}(\boldsymbol{\varphi})$ a.e. in Ω. Note that we cannot apply Fatou's lemma, since $\mathcal{E}(\boldsymbol{u}_n)$ does not converge pointwise. However, we use

an estimate similar to (25):

$$\int_\Omega c(\varphi_n)|\boldsymbol{C}(\varphi_n)\mathcal{E}(\boldsymbol{u}_n)|^2 = \int_\Omega c(\varphi_n)|\boldsymbol{C}(\varphi_n)\mathcal{E}(\boldsymbol{u}_n) - \boldsymbol{C}(\varphi)\mathcal{E}(\boldsymbol{u})|^2 + 2\int_\Omega c(\varphi_n)(\boldsymbol{C}(\varphi_n)\mathcal{E}(\boldsymbol{u}_n)) : (\boldsymbol{C}(\varphi)\mathcal{E}(\boldsymbol{u})) - \int_\Omega c(\varphi_n)|\boldsymbol{C}(\varphi)\mathcal{E}(\boldsymbol{u})|^2$$

The first term is nonnegative. Under the assumptions on c it holds $0 \le c(\varphi_n) \le C$ a.e. in Ω uniformly in n (see [GKT92, Thm. 3]) and we can use dominated convergence to pass to the limit in the third term. For the middle term consider the integrand without $\mathcal{E}(\boldsymbol{u}_n)$, i.e. in coordinates $c(\varphi_n)\boldsymbol{C}_{ijkl}(\varphi_n)\boldsymbol{C}_{abkl}(\varphi)\mathcal{E}_{ab}(\boldsymbol{u})$. Since we have $|c(\varphi_n)\boldsymbol{C}_{ijkl}(\varphi_n)| \le C$ a.e. in Ω and uniformly in n, we can use dominated convergence to obtain

$$c(\varphi_n)\boldsymbol{C}_{ijkl}(\varphi_n)\boldsymbol{C}_{abkl}(\varphi)\mathcal{E}_{ab}(\boldsymbol{u}) \to c(\varphi)\boldsymbol{C}_{ijkl}(\varphi)\boldsymbol{C}_{abkl}(\varphi)\mathcal{E}_{ab}(\boldsymbol{u}) \quad \text{in } L^2(\Omega).$$

Due to $\mathcal{E}(\boldsymbol{u}_n) \to \mathcal{E}(\boldsymbol{u})$ weakly in $L^2(\Omega)^{d\times d}$ we get

$$\int_\Omega c(\varphi_n)(\boldsymbol{C}(\varphi_n)\mathcal{E}(\boldsymbol{u}_n)) : (\boldsymbol{C}(\varphi)\mathcal{E}(\boldsymbol{u})) \to \int_\Omega c(\varphi)|\boldsymbol{C}(\varphi)\mathcal{E}(\boldsymbol{u})|^2.$$

Thus we showed

$$\liminf_n \int_\Omega c(\varphi_n)|\boldsymbol{C}(\varphi_n)\mathcal{E}(\boldsymbol{u}_n)|^2 \ge \int_\Omega c(\varphi)|\boldsymbol{C}(\varphi)\mathcal{E}(\boldsymbol{u})|^2$$

for a subsequence. This holds also for the whole sequence due to the arguments used in Lemma 6.4. □

It is possible to consider additional control constraints, as long as Φ_{ad} remains convex and closed in H^1, which is needed by the VMPT method. It is often wanted to prescribe material in some regions or to forbid a material to be placed in some regions. This can be modelled by the constraints

$$\varphi_i = 0 \text{ a.e. in } S^i \quad i = 1, \dots, N,$$

where S^i is a measurable subset of Ω for all i. Prescribing the j'th material can be achieved by setting $\varphi_i = 0$ for all $i \neq j$ due to the sum constraint. One has to be careful to choose the sets S^i such that there is space for an interface with positive thickness between the pure phases. Otherwise Φ_{ad} is empty, since H^1-functions cannot have jumps across hypersurfaces.

Sometimes it is wanted to calculate an optimal design without mass constraint. We note that in this case the inner products of the VMPT method defined in Section 6.7 have to be slightly amended, since the Poincaré inequality for mass free functions cannot be applied. We refer to the discussion in Section 6.14. In this case the existence and uniqueness of Lagrange multipliers can still be shown, see Remark 6.39. Also the proof for global convergence of the VMPT method stays unchanged.

Instead of a mass equality constraint or a mass penalization term, one can also consider a mass inequality constraint as in [BGHR15]. However, especially for the mean compliance problem it can be expected that the mass constraint is active at the solution. In fact this is assumed by many optimality criteria methods (see Section 5), or by the derivation of the stress formulation in [All02]. For the existence of Lagrange multipliers for mass inequality constraints we refer to [BGHR15].

When considering inhomogeneous Dirichlet boundary conditions for $\boldsymbol{u}$, i.e. $\boldsymbol{u} = \boldsymbol{h}$ on Γ_D for some $\boldsymbol{h} \in H^{1/2}(\Gamma_D)^d$, one has to perform a translation of the state equation. Therefor, $\boldsymbol{h}$ is extended to $\tilde{\boldsymbol{h}} \in H^1(\Omega)^d$ and the state $\boldsymbol{u} - \tilde{\boldsymbol{h}}$ is considered instead of $\boldsymbol{u}$, since homogeneous Dirichlet boundary conditions can be imposed on $\boldsymbol{u} - \tilde{\boldsymbol{h}}$. The additional term $\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\tilde{\boldsymbol{h}}) : \mathcal{E}(\boldsymbol{\xi})$ appearing in the state equation can be put on the right hand side. However, in this case the right hand side is in $(H^1_D)^*$ rather than in $L^2(\Omega)^d$, which does not fit in our setting.

It is possible to have other boundary conditions on the displacement field $\boldsymbol{u}$, e.g. homogeneous Dirichlet boundary condition for the y component of $\boldsymbol{u}$ on some part of the boundary. These conditions can be put into the space H^1_D and the analysis does not change. The only needed assumption on H^1_D is that $\|\mathcal{E}(\boldsymbol{u})\|_{L^2}$ is defines a norm on H^1_D, which is equivalent to the H^1-norm. This can be the case even if $\Gamma_D = \varnothing$ and other appropriate boundary conditions are imposed. We refer to Theorem 7.1 for an example.

The presence of eigenstrain as in [BGHR15] can be handled if the tensors $\boldsymbol{C}_i$ and the eigenstrains are interpolated linearly, see Remark 6.13. At least the C^2-regularity of the control-to-state operator can be shown in this case. For the global convergence of the VMPT method one has in addition to ensure that **(A7)** is fulfilled.

We assume that the stiffness tensor of a single material $\boldsymbol{C}$ is homogeneous, i.e. it does not depend on $x \in \Omega$. However, it is also possible to consider inhomogeneous materials by using $\boldsymbol{C}(x, \varphi(x))$, see Remark 6.13.

It is straight forward to generalize the optimization problem for multiple load conditions as in [AJ05]. In this case m different state equations are present, and each corresponding control-to-state operator can be handled separately as in Section 6.2.

When considering the VMPT method it is not possible to drop the constraints $\boldsymbol{\varphi} \geq 0$, $\sum_i \varphi_i = 1$. For the VMPT method the set of admissible controls has to be bounded in $L^\infty(\Omega)^N$, which has to be ensured by appropriate constraints.

6.1.3 Problem reduction for two phases

In case of two phases (i.e. two materials or material and void), one usually eliminates one phase by means of the equation $\varphi_1 + \varphi_2 = 1$ and considers only the difference $\tilde{\varphi} := \varphi_1 - \varphi_2$. On the other hand, given a scalar valued phase field φ, one can recover the corresponding vector valued phase field by $\varphi_1 := \frac{1+\tilde{\varphi}}{2}$ and $\varphi_2 := \frac{1-\tilde{\varphi}}{2}$. Thus there is a one-to-one correspondence between a scalar valued phase field $\tilde{\varphi}$ and a vector valued phase field $\boldsymbol{\varphi}$ (this will be proved in Theorem 6.17). The constraints are transformed in the following way:

$$\left.\begin{aligned} \varphi_1 + \varphi_2 &= 1 \\ \varphi_1 &\geq 0 \\ \varphi_2 &\geq 0 \end{aligned}\right\} \quad \Longleftrightarrow \quad -1 \leq \tilde{\varphi} \leq 1,$$

$$\fint_\Omega \boldsymbol{\varphi} = \mathfrak{m} \quad \Longleftrightarrow \quad \fint_\Omega \tilde{\varphi} = \mathfrak{m}_1 - \mathfrak{m}_2 =: \tilde{\mathfrak{m}}.$$

We can write the optimization problem in terms of $\tilde{\varphi}$ by transforming the functions of $\boldsymbol{\varphi}$ appropriately, i.e. we set

$$\begin{aligned}\tilde{F}(\tilde{\varphi}, \boldsymbol{u}) &:= F(\boldsymbol{\varphi}, \boldsymbol{u})\\ \tilde{\boldsymbol{C}}(\tilde{\varphi}) &:= \boldsymbol{C}(\boldsymbol{\varphi})\\ \tilde{\boldsymbol{f}}(\tilde{\varphi}) &:= \boldsymbol{f}(\boldsymbol{\varphi})\\ \tilde{\boldsymbol{g}}(\tilde{\varphi}) &:= \boldsymbol{g}(\boldsymbol{\varphi})\end{aligned}$$

with $\boldsymbol{\varphi} := (\frac{1+\tilde{\varphi}}{2}, \frac{1-\tilde{\varphi}}{2})^T$. For the Ginzburg-Landau energy we use a special treatment, since it holds almost everywhere

$$\begin{aligned}|\nabla\boldsymbol{\varphi}|^2 &= |\partial_1\boldsymbol{\varphi}|^2 + |\partial_2\boldsymbol{\varphi}|^2 = |(\partial_1\frac{1+\tilde{\varphi}}{2}, \partial_1\frac{1-\tilde{\varphi}}{2})^T|^2 + |(\partial_2\frac{1+\tilde{\varphi}}{2}, \partial_2\frac{1-\tilde{\varphi}}{2})^T|^2\\ &= \frac{1}{4}(|(\partial_1\tilde{\varphi}, -\partial_1\tilde{\varphi})^T|^2 + |(\partial_2\tilde{\varphi}, -\partial_2\tilde{\varphi})^T|^2) = \frac{1}{4}(2|\partial_1\tilde{\varphi}|^2 + 2|\partial_2\tilde{\varphi}|^2) = \frac{1}{2}|\nabla\tilde{\varphi}|^2.\end{aligned}$$

Therefore, we set

$$\begin{aligned}\tilde{E}(\tilde{\varphi}) &:= 2E(\boldsymbol{\varphi}) = 2\int \frac{\varepsilon}{2}|\nabla\boldsymbol{\varphi}|^2 + \frac{1}{\varepsilon}\psi_0(\boldsymbol{\varphi}) = \int \frac{\varepsilon}{2}|\nabla\tilde{\varphi}|^2 + \frac{1}{\varepsilon}\tilde{\psi}_0(\tilde{\varphi}),\\ \tilde{\psi}_0(\tilde{\varphi}) &:= 2\psi_0(\boldsymbol{\varphi}).\end{aligned}$$

With these definitions we get with $\tilde{\gamma} := \frac{\gamma}{2}$ that

$$\tilde{\gamma}\tilde{E}(\tilde{\varphi}) + \tilde{F}(\tilde{\varphi}, \boldsymbol{u}) = \gamma E(\boldsymbol{\varphi}) + F(\boldsymbol{\varphi}, \boldsymbol{u}). \tag{120}$$

We denote the set of admissible scalar valued controls by

$$\widetilde{\Phi_{ad}} = \{\tilde{\varphi} \in H^1(\Omega) \mid -1 \le \tilde{\varphi} \le 1, \ \fint \tilde{\varphi} = \tilde{\mathfrak{m}}\}$$

In the following, we often omit the tilde when it is clear from the context that scalar valued phase fields are considered.
For two phases, the reduced problem thus reads

$$\min \gamma E(\varphi) + F(\varphi, \boldsymbol{u}) \tag{121}$$

$$\begin{aligned}\varphi \in H^1(\Omega), \quad \boldsymbol{u} \in H^1_D\\ \int_\Omega \boldsymbol{C}(\varphi)\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) = \int_\Omega \boldsymbol{f}(\varphi)\cdot\boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g}(\varphi)\cdot\boldsymbol{\xi} \quad \forall \boldsymbol{\xi} \in H^1_D\\ -1 \le \varphi \le 1\\ \fint \varphi = \mathfrak{m}.\end{aligned} \tag{122}$$

We will prove below that the reduced optimization problem (121) is equivalent to the vector valued phase field problem (102) for $N = 2$, see Theorem 6.17.

6.2 Analysis of the control-to-state operator

In the following we will show C^2-regularity of the control-to-state operator S, which maps a control $\boldsymbol{\varphi} \in H^1 \cap L^\infty$ to the corresponding state $\boldsymbol{u} \in H^1_D$, being the solution of the state equation.

In [BFGS14] a similar state equation is discussed. The difference is that the right hand side $\boldsymbol{f}$ and $\boldsymbol{g}$ in our model can depend on the phase field $\boldsymbol{\varphi}$ in a more general way. The boundary traction $\boldsymbol{g}$ in [BFGS14] does not depend on $\boldsymbol{\varphi}$ at all. Therefore they can show well posedness and Fréchet differentiability of the control-to-state operator for $\boldsymbol{\varphi}$ in $L^\infty(\Omega)^N$. This is not possible in our case since we need the trace of $\boldsymbol{\varphi}$, on which $\boldsymbol{g}$ depends and which does not exist in $L^\infty(\Omega)^N$. Hence we show well posedness and Fréchet differentiability for $\boldsymbol{\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$. This is no restriction since the Ginzburg-Landau energy in the cost functional is also defined on $H^1(\Omega)^N$. In fact we show all estimates for the control-to-state operator with respect to the $L^\infty(\Omega)^N$ norm rather than the $H^1(\Omega)^N \cap L^\infty(\Omega)^N$ norm, which is a stronger result, but which is not needed in the rest of the work. We remark that in the case that $\boldsymbol{g}$ does not depend on $\boldsymbol{\varphi}$, C^2 regularity of the control-to-state operator can be shown for $\boldsymbol{\varphi} \in L^\infty(\Omega)^N$ by the same proofs as shown in this work. We also drop the condition $|\boldsymbol{C}'(\boldsymbol{\varphi})| \leq C \ \forall \boldsymbol{\varphi} \in \mathbb{R}^N$, which is assumed in [BFGS14], since it is not needed in our proof.

6.2.1 Well posedness and local Lipschitz continuity

Many arguments used here are the same as in [BFGS14]. We therefore don't go much into detail in the proofs.

Theorem 6.6. *For each $\boldsymbol{\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$, the state equation given in its weak form by*

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) = \int_\Omega \boldsymbol{f}(\boldsymbol{\varphi}) \cdot \boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g}(\boldsymbol{\varphi}) \cdot \boldsymbol{\xi} \quad \forall \boldsymbol{\xi} \in H^1_D$$

has a unique weak solution $\boldsymbol{u} \in H^1_D$. It holds the a priori estimate

$$\|\boldsymbol{u}\|_{H^1_D} \leq c(\|\boldsymbol{f}(\boldsymbol{\varphi})\|_{L^2(\Omega)} + \|\boldsymbol{g}(\boldsymbol{\varphi})\|_{L^2(\Gamma_g)}). \tag{123}$$

for some $c > 0$ independent of $\boldsymbol{\varphi}$.

Proof. We show the statement by the Lax-Milgram theorem. Therefor, define the bilinear form $a : H^1_D \times H^1_D \to \mathbb{R}$ and the linear form $l \in (H^1_D)^*$ by

$$\begin{aligned} a(\boldsymbol{u}, \boldsymbol{\xi}) &:= \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) && \text{for all } \boldsymbol{u},\ \boldsymbol{\xi} \in H^1_D \text{ and} \\ \langle l, \boldsymbol{\xi} \rangle &:= \int_\Omega \boldsymbol{f}(\boldsymbol{\varphi}) \cdot \boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g}(\boldsymbol{\varphi}) \cdot \boldsymbol{\xi} && \text{for all } \boldsymbol{\xi} \in H^1_D. \end{aligned}$$

H^1_D-coercivity of a can be shown by Korn's inequality and **(AP4)**. Continuity of a and l can be shown using (104). See [BFGS14] for details. The statement then follows from the Lax-Milgram theorem. The constant c in the a priori estimate only depends on the coercivity constant of $\boldsymbol{C}$ and the continuity constants of the embedding $H^1(\Omega)^d \hookrightarrow L^2(\Omega)^d$ and the trace $H^1(\Omega)^d \hookrightarrow L^2(\partial\Omega)^d$, as well as the constant coming from Korn's inequality, which all are independent of $\boldsymbol{\varphi}$. □

Denote by $S : H^1(\Omega)^N \cap L^\infty(\Omega)^N \to H^1_D$, $\boldsymbol{\varphi} \mapsto \boldsymbol{u}$, the solution operator of the state equation.

Theorem 6.7. *Let $M > 0$. Then there exists a constant $C(M) > 0$, such that for all $\boldsymbol{\varphi}_i \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$ with $\|\boldsymbol{\varphi}_i\|_{L^\infty} \leq M$, $i = 1,2$ and $\boldsymbol{u}_i = S(\boldsymbol{\varphi}_i)$, $i = 1,2$ it holds*

$$\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_{H^1} \leq C(M)\|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|_{L^\infty}.$$

Proof. Let $\boldsymbol{\varphi}_i$, $\boldsymbol{u}_i$, $i = 1,2$ as in the statement. As in [BFGS14] we subtract the state equations for $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, test the equation by the difference $\boldsymbol{\xi} = \boldsymbol{u}_1 - \boldsymbol{u}_2 \in H^1_D$ and obtain

$$\begin{aligned}\int_\Omega (\boldsymbol{C}(\boldsymbol{\varphi}_1)\mathcal{E}(\boldsymbol{u}_1) - \boldsymbol{C}(\boldsymbol{\varphi}_2)\mathcal{E}(\boldsymbol{u}_2)) : \mathcal{E}(\boldsymbol{u}_1 - \boldsymbol{u}_2) = \int_\Omega (\boldsymbol{f}(\boldsymbol{\varphi}_1) - \boldsymbol{f}(\boldsymbol{\varphi}_2)) \cdot (\boldsymbol{u}_1 - \boldsymbol{u}_2) \\ + \int_{\Gamma_g} (\boldsymbol{g}(\boldsymbol{\varphi}_1) - \boldsymbol{g}(\boldsymbol{\varphi}_2)) \cdot (\boldsymbol{u}_1 - \boldsymbol{u}_2)\end{aligned}$$

Using **(AP4)** and Korn's inequality, we get (cf. [BFGS14])

$$\begin{aligned}\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|^2_{H^1_D} \leq C\Big(\Big|\int_\Omega (\boldsymbol{C}(\boldsymbol{\varphi}_1) - \boldsymbol{C}(\boldsymbol{\varphi}_2))\mathcal{E}(\boldsymbol{u}_2) : \mathcal{E}(\boldsymbol{u}_1 - \boldsymbol{u}_2)\Big| \\ + \Big|\int_\Omega (\boldsymbol{C}(\boldsymbol{\varphi}_1)\mathcal{E}(\boldsymbol{u}_1) - \boldsymbol{C}(\boldsymbol{\varphi}_2)\mathcal{E}(\boldsymbol{u}_2)) : \mathcal{E}(\boldsymbol{u}_1 - \boldsymbol{u}_2)\Big|\Big)\end{aligned}$$

The local boundedness of $\boldsymbol{f}$ and $\boldsymbol{g}$ (106)-(107) and the a priori estimate (123) yield $\|\boldsymbol{u}_2\|_{H^1} \leq C(M)$. Applying Hölder's inequality and the local Lipschitz continuity of $\boldsymbol{C}$ gives

$$\begin{aligned}\Big|\int_\Omega (\boldsymbol{C}(\boldsymbol{\varphi}_1) - \boldsymbol{C}(\boldsymbol{\varphi}_2))\mathcal{E}(\boldsymbol{u}_2) : \mathcal{E}(\boldsymbol{u}_1 - \boldsymbol{u}_2)\Big| &\leq C(M)\|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|_{L^\infty}\|\boldsymbol{u}_2\|_{H^1_D}\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_{H^1_D} \\ &\leq C(M)\|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|_{L^\infty}\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_{H^1_D}.\end{aligned}$$

The local Lipschitz continuity of $\boldsymbol{f}$ and $\boldsymbol{g}$ **(AP7)**, Hölder's inequality, the trace theorem and Lemma 7.4 yields

$$\begin{aligned}\Big|\int_\Omega (\boldsymbol{f}(\boldsymbol{\varphi}_1) - \boldsymbol{f}(\boldsymbol{\varphi}_2)) \cdot (\boldsymbol{u}_1 - \boldsymbol{u}_2) + \int_{\Gamma_g} (\boldsymbol{g}(\boldsymbol{\varphi}_1) - \boldsymbol{g}(\boldsymbol{\varphi}_2)) \cdot (\boldsymbol{u}_1 - \boldsymbol{u}_2)\Big| \\ \leq C \cdot L(M)\|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|_{L^\infty}\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_{H^1_D}\end{aligned}$$

Putting all together and dividing by $\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_{H^1_D}$ gives the statement. □

6.2.2 Fréchet differentiability of first order

Lemma 6.8. *It holds that $\boldsymbol{C} : L^\infty(\Omega)^N \to L^\infty(\Omega; \mathbb{R}^{d\times d} \otimes (\mathbb{R}^{d\times d})^*)$ is continuously Fréchet differentiable. The Fréchet derivative is given by*

$$(\boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{h})(x) = \boldsymbol{C}'(\boldsymbol{\varphi}(x))\boldsymbol{h}(x) \quad \text{for all } \boldsymbol{\varphi} \in L^\infty(\Omega)^N,\ \boldsymbol{h} \in L^\infty(\Omega)^N \text{ and a.e. in } \Omega.$$

Proof. For the proof concerning scalar valued functions we refer to [Trö09, Lemma 4.12 and Lemma 4.13]. In the case of $\boldsymbol{C} : L^\infty(\Omega)^N \to L^\infty(\Omega; \mathbb{R}^{d\times d} \otimes (\mathbb{R}^{d\times d})^*)$, the proof is similar: The mapping $\boldsymbol{h} \mapsto \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{h}$ is in $\mathcal{L}(L^\infty(\Omega)^N, L^\infty(\Omega; \mathbb{R}^{d\times d} \otimes (\mathbb{R}^{d\times d})^*)$ for each $\boldsymbol{\varphi} \in L^\infty(\Omega)^N$, which follows from the estimate $\|\boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{h}\|_{L^\infty} \leq \|\boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty}\|\boldsymbol{h}\|_{L^\infty}$. For the rest term estimate, let $\boldsymbol{\varphi}$, $\boldsymbol{h} \in L^\infty(\Omega)^N$ be arbitrary. We apply the fundamental theorem

of calculus (separately for almost every $x \in \Omega$) to get the following estimate.

$$\|\boldsymbol{C}(\boldsymbol{\varphi}+\boldsymbol{h}) - \boldsymbol{C}(\boldsymbol{\varphi}) - \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{h}\|_{L^\infty} = \left\| \int_0^1 (\boldsymbol{C}'(\boldsymbol{\varphi}+t\boldsymbol{h}) - \boldsymbol{C}'(\boldsymbol{\varphi}))\boldsymbol{h}\,\mathrm{d}t \right\|_{L^\infty}.$$

Using the local Lipschitz continuity of $\boldsymbol{C}'$ we get

$$\left\| \int_0^1 (\boldsymbol{C}'(\boldsymbol{\varphi}+t\boldsymbol{h}) - \boldsymbol{C}'(\boldsymbol{\varphi}))\boldsymbol{h}\,\mathrm{d}t \right\|_{L^\infty} \le \left\| \int_0^1 Lt|\boldsymbol{h}|^2\,\mathrm{d}t \right\|_{L^\infty} = \frac{1}{2}L\,\|\boldsymbol{h}\|_{L^\infty}^2.$$

Thus, $\|\boldsymbol{C}(\boldsymbol{\varphi}+\boldsymbol{h}) - \boldsymbol{C}(\boldsymbol{\varphi}) - \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{h}\|_{L^\infty} = o(\|\boldsymbol{h}\|_{L^\infty})$ as $\|\boldsymbol{h}\|_{L^\infty} \to 0$ and Fréchet differentiability follows. To prove continuity of $\boldsymbol{C}'$, let $\boldsymbol{\varphi}_i \to \boldsymbol{\varphi}$ in $L^\infty(\Omega)^N$ and let $\boldsymbol{h} \in L^\infty(\Omega)^N$. We have due to local Lipschitz continuity of $\boldsymbol{C}'$

$$\|\boldsymbol{C}'(\boldsymbol{\varphi}_i)h - \boldsymbol{C}'(\boldsymbol{\varphi})h\|_{L^\infty} \le \big\|L|\boldsymbol{\varphi}_i - \boldsymbol{\varphi}|\cdot|\boldsymbol{h}|\big\|_{L^\infty} \le L\|\boldsymbol{\varphi}_i - \boldsymbol{\varphi}\|_{L^\infty}\|\boldsymbol{h}\|_{L^\infty},$$

thus $\boldsymbol{C}'(\boldsymbol{\varphi}_i) \to \boldsymbol{C}'(\boldsymbol{\varphi})$ in $\mathcal{L}(L^\infty(\Omega)^N, L^\infty(\Omega;\mathbb{R}^{d\times d}\otimes(\mathbb{R}^{d\times d})^*)$ and we even get Lipschitz continuity of $\boldsymbol{C}'$ in L^∞. □

We show the continuous Fréchet differentiability of S by the implicit function theorem [Zei85, Theorem 4.B].

Theorem 6.9. *The control-to-state operator $S : H^1(\Omega)^N \cap L^\infty(\Omega)^N \to H^1_D$ is continuously Fréchet differentiable. The Fréchet derivative at $\boldsymbol{\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$ in direction $\boldsymbol{\delta\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$ is given by $S'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi} = \boldsymbol{\delta u}$, where $\boldsymbol{\delta u} \in H^1_D$ is the unique solution of the linearized state equation*

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{\delta u}) : \mathcal{E}(\boldsymbol{\xi}) = -\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) + \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\cdot\boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\cdot\boldsymbol{\xi}$$
$$\forall \boldsymbol{\xi} \in H^1_D. \quad (124)$$

Moreover, it holds the a priori estimate

$$\|\boldsymbol{\delta u}\|_{H^1_D} \le C\left(\|\boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty}\|\boldsymbol{u}\|_{H^1_D} + \|\boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\|_{L^2} + \|\boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\|_{L^2}\right)\|\boldsymbol{\delta\varphi}\|_{L^\infty}, \quad (125)$$

where $C > 0$ is independent of $\boldsymbol{\varphi}$, $\boldsymbol{u}$ and $\boldsymbol{\delta\varphi}$.

Proof. We write the weak formulation of the state equation as

$$G(\boldsymbol{\varphi}, \boldsymbol{u}) = 0$$

where $G : (H^1(\Omega)^N \cap L^\infty(\Omega)^N) \times H^1_D \to (H^1_D)^*$ is defined by

$$\langle G(\boldsymbol{\varphi}, \boldsymbol{u}), \boldsymbol{\xi}\rangle_{(H^1_D)^*, H^1_D} := \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) - \int_\Omega \boldsymbol{f}(\boldsymbol{\varphi})\cdot\boldsymbol{\xi} - \int_{\Gamma_g} \boldsymbol{g}(\boldsymbol{\varphi})\cdot\boldsymbol{\xi}.$$

It has already been proved that $G(\boldsymbol{\varphi}, \boldsymbol{u}) = 0$ if and only if $\boldsymbol{u} = S(\boldsymbol{\varphi})$.
We have to show that G is C^1 and that $G_{\boldsymbol{u}}(\boldsymbol{\varphi}, \boldsymbol{u}) \in \mathcal{L}(H^1_D, (H^1_D)^*)$ is bijective for all $(\boldsymbol{\varphi}, \boldsymbol{u}) \in (H^1(\Omega)^N \cap L^\infty(\Omega)^N) \times H^1_D$. Then the statement of the theorem follows from the implicit function theorem. We prove $G \in C^1$ by showing that G is partially Fréchet differentiable with respect to $\boldsymbol{\varphi}$ and $\boldsymbol{u}$ and that the partial derivatives are continuous [Zei85, Proposition 4.14].
Note that in the following all estimates have to be uniformly in $\boldsymbol{\xi} \in H^1_D$.

i) G is partially Fréchet differentiable with respect to $\boldsymbol{u}$ for all $(\boldsymbol{\varphi}, \boldsymbol{u}) \in (H^1(\Omega)^N \cap L^\infty(\Omega)^N) \times H^1_D$:
G is affine linear in $\boldsymbol{u}$ and continuous in $\boldsymbol{u}$, which follows from the estimate

$$\left| \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}) \mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) \right| \leq C \|\boldsymbol{u}\|_{H^1_D} \|\boldsymbol{\xi}\|_{H^1_D},$$

where we used (104) and Hölder's inequality. Thus G is smooth with respect to $\boldsymbol{u}$ and the Fréchet derivative is given by

$$\langle G_{\boldsymbol{u}}(\boldsymbol{\varphi}, \boldsymbol{u}) \boldsymbol{h}, \boldsymbol{\xi} \rangle_{(H^1_D)^*, H^1_D} = \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}) \mathcal{E}(\boldsymbol{h}) : \mathcal{E}(\boldsymbol{\xi}).$$

ii) $G_{\boldsymbol{u}}(\boldsymbol{\varphi}, \boldsymbol{u}) \in \mathcal{L}(H^1_D, (H^1_D)^*)$ is bijective for all $(\boldsymbol{\varphi}, \boldsymbol{u}) \in (H^1(\Omega)^N \cap L^\infty(\Omega)^N) \times H^1_D$:
This can be proved by the Lax-Milgram theorem. The proof is almost identical to the proof of Theorem 6.6, except that the right hand side is an arbitrary functional in $(H^1_D)^*$.

iii) G is partially Fréchet differentiable with respect to $\boldsymbol{\varphi}$ for all $(\boldsymbol{\varphi}, \boldsymbol{u}) \in (H^1(\Omega)^N \cap L^\infty(\Omega)^N) \times H^1_D$:
The candidate for the Fréchet derivative is

$$\langle G_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}, \boldsymbol{u}) \boldsymbol{h}, \boldsymbol{\xi} \rangle_{(H^1_D)^*, H^1_D} = \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}) \boldsymbol{h} \mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) - \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}) \boldsymbol{h} \cdot \boldsymbol{\xi} - \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}) \boldsymbol{h} \cdot \boldsymbol{\xi}.$$

We show $G_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}, \boldsymbol{u}) \in \mathcal{L}(H^1(\Omega)^N \cap L^\infty(\Omega)^N, (H^1_D)^*)$. Linearity in $\boldsymbol{h}$ is obvious. For the boundedness consider the estimate

$$\begin{aligned} &\left| \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}) \boldsymbol{h} \mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) - \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}) \boldsymbol{h} \cdot \boldsymbol{\xi} - \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}) \boldsymbol{h} \cdot \boldsymbol{\xi} \right| \\ &\qquad \leq C \left(\|\boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty} \|\boldsymbol{u}\|_{H^1_D} + \|\boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\|_{L^2} + \|\boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\|_{L^2} \right) \|\boldsymbol{h}\|_{L^\infty} \|\boldsymbol{\xi}\|_{H^1_D}, \end{aligned} \tag{126}$$

where we used Lemma 6.8, **(AP6)**, Hölder's inequality and the trace theorem. We finally show the rest term estimate.

$$\begin{aligned} &\langle G(\boldsymbol{\varphi} + \boldsymbol{h}, \boldsymbol{u}) - G(\boldsymbol{\varphi}, \boldsymbol{u}) - G_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}, \boldsymbol{u}) \boldsymbol{h}, \boldsymbol{\xi} \rangle_{(H^1_D)^*, H^1_D} \\ &= \int_\Omega (\boldsymbol{C}(\boldsymbol{\varphi} + \boldsymbol{h}) - \boldsymbol{C}(\boldsymbol{\varphi}) - \boldsymbol{C}'(\boldsymbol{\varphi}) \boldsymbol{h}) \mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) - \int_\Omega (\boldsymbol{f}(\boldsymbol{\varphi} + \boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{\varphi}) - \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}) \boldsymbol{h}) \cdot \boldsymbol{\xi} \\ &\quad - \int_{\Gamma_g} (\boldsymbol{g}(\boldsymbol{\varphi} + \boldsymbol{h}) - \boldsymbol{g}(\boldsymbol{\varphi}) - \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}) \boldsymbol{h}) \cdot \boldsymbol{\xi} \\ &\leq C (\|\boldsymbol{C}(\boldsymbol{\varphi} + \boldsymbol{h}) - \boldsymbol{C}(\boldsymbol{\varphi}) - \boldsymbol{C}'(\boldsymbol{\varphi}) \boldsymbol{h}\|_{L^\infty} \|\boldsymbol{u}\|_{H^1_D} + \|\boldsymbol{f}(\boldsymbol{\varphi} + \boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{\varphi}) - \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}) \boldsymbol{h}\|_{L^2} \\ &\quad + \|\boldsymbol{g}(\boldsymbol{\varphi} + \boldsymbol{h}) - \boldsymbol{g}(\boldsymbol{\varphi}) - \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}) \boldsymbol{h}\|_{L^2}) \|\boldsymbol{\xi}\|_{H^1_D} \\ &= o(\|\boldsymbol{h}\|_{L^\infty}) \|\boldsymbol{\xi}\|_{H^1_D} \quad \text{as } \|\boldsymbol{h}\|_{L^\infty} \to 0, \end{aligned}$$

since $\boldsymbol{C}$, $\boldsymbol{f}$ and $\boldsymbol{g}$ are Fréchet differentiable in the right spaces (**(AP6)** and Lemma 6.8).

iv) $G_{\boldsymbol{u}}$ is continuous:
Let $\boldsymbol{\varphi}_n \to \boldsymbol{\varphi}$ in $H^1(\Omega)^N \cap L^\infty(\Omega)^N$, $\boldsymbol{u}_n \to \boldsymbol{u}$ in H^1_D and $\boldsymbol{h}, \boldsymbol{\xi} \in H^1_D$ be arbitrary. We

estimate using the continuity of $\boldsymbol{C}$, see Lemma 6.8,

$$\begin{aligned}\left|\langle (G_{\boldsymbol{u}}(\boldsymbol{\varphi}_n,\boldsymbol{u}_n)-G_{\boldsymbol{u}}(\boldsymbol{\varphi},\boldsymbol{u}))\boldsymbol{h},\boldsymbol{\xi}\rangle_{(H^1_D)^*,H^1_D}\right| &= \left|\int_\Omega (\boldsymbol{C}(\boldsymbol{\varphi}_n)-\boldsymbol{C}(\boldsymbol{\varphi}))\mathcal{E}(\boldsymbol{h}):\mathcal{E}(\boldsymbol{\xi})\right| \\ &\le C\underbrace{\|\boldsymbol{C}(\boldsymbol{\varphi}_n)-\boldsymbol{C}(\boldsymbol{\varphi})\|_{L^\infty}}_{\to 0}\|\boldsymbol{h}\|_{H^1_D}\|\boldsymbol{\xi}\|_{H^1_D}.\end{aligned}$$

v) $G_{\boldsymbol{\varphi}}$ is continuous:
Let $\boldsymbol{\varphi}_n \to \boldsymbol{\varphi}$ in $H^1(\Omega)^N \cap L^\infty(\Omega)^N$, $\boldsymbol{u}_n \to \boldsymbol{u}$ in H^1_D and $\boldsymbol{h} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$, $\boldsymbol{\xi} \in H^1_D$ be arbitrary.

$$\begin{aligned}\langle (G_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_n,\boldsymbol{u}_n)-G_{\boldsymbol{\varphi}}(\boldsymbol{\varphi},\boldsymbol{u}))\boldsymbol{h},\boldsymbol{\xi}\rangle_{(H^1_D)^*,H^1_D} &= \int_\Omega (\boldsymbol{C}'(\boldsymbol{\varphi}_n)\boldsymbol{h}\mathcal{E}(\boldsymbol{u}_n)-\boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{h}\mathcal{E}(\boldsymbol{u})):\mathcal{E}(\boldsymbol{\xi}) \\ &\quad -\int_\Omega (\boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_n)-\boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}))\boldsymbol{h}\cdot\boldsymbol{\xi} \\ &\quad -\int_{\Gamma_g} (\boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_n)-\boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}))\boldsymbol{h}\cdot\boldsymbol{\xi}.\end{aligned}$$

We show that the three terms converge to 0 (uniformly in $\boldsymbol{h}$ and $\boldsymbol{\xi}$).

$$\begin{aligned}&\left|\int_\Omega (\boldsymbol{C}'(\boldsymbol{\varphi}_n)\boldsymbol{h}\mathcal{E}(\boldsymbol{u}_n)-\boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{h}\mathcal{E}(\boldsymbol{u})):\mathcal{E}(\boldsymbol{\xi})\right| \\ &\qquad \le \left|\int_\Omega (\boldsymbol{C}'(\boldsymbol{\varphi}_n)-\boldsymbol{C}'(\boldsymbol{\varphi}))\boldsymbol{h}\mathcal{E}(\boldsymbol{u}_n):\mathcal{E}(\boldsymbol{\xi})\right| + \left|\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{h}\mathcal{E}(\boldsymbol{u}_n-\boldsymbol{u}):\mathcal{E}(\boldsymbol{\xi})\right| \\ &\qquad \le (\underbrace{\|\boldsymbol{C}'(\boldsymbol{\varphi}_n)-\boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty}}_{\to 0}\underbrace{\|\boldsymbol{u}_n\|_{H^1_D}}_{\le C} + \|\boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty}\underbrace{\|\boldsymbol{u}_n-\boldsymbol{u}\|_{H^1_D}}_{\to 0})\|\boldsymbol{h}\|_{L^\infty}\|\boldsymbol{\xi}\|_{H^1_D}\end{aligned}$$

For the second term we get

$$\int_\Omega (\boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_n)-\boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}))\boldsymbol{h}\cdot\boldsymbol{\xi} \le C\underbrace{\|\boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_n)-\boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\|_{L^2}}_{\to 0}\|\boldsymbol{h}\|_{L^\infty}\|\boldsymbol{\xi}\|_{H^1_D},$$

where we used the continuity of $\boldsymbol{f}_{\boldsymbol{\varphi}}$, **(AP6)**. The estimate for the third term is analogous.

Thus, all assumptions of the implicit function theorem are fulfilled and we get $S \in C^1(H^1(\Omega)^N \cap L^\infty(\Omega)^N, H^1_D)$. Moreover we get the formula for the derivative $G_{\boldsymbol{u}}(\boldsymbol{\varphi},S(\boldsymbol{\varphi}))S'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi} = -G_{\boldsymbol{\varphi}}(\boldsymbol{\varphi},S(\boldsymbol{\varphi}))\boldsymbol{\delta\varphi}$ in $(H^1_D)^*$. If we test the equation by $\boldsymbol{\xi}$ we end up with the linearized state equation (124). The a priori estimate follows from the Lax-Milgram theorem together with the estimate (126). □

Note that we used only the regularity $\boldsymbol{C} \in C^{1,1}$ and the C^1-regularity of $\boldsymbol{f}$ and $\boldsymbol{g}$ in the previous proof.

6.2.3 Fréchet differentiability of second order

Lemma 6.10. *It holds that $\boldsymbol{C}: L^\infty(\Omega)^N \to L^\infty(\Omega;\mathbb{R}^{d\times d}\otimes(\mathbb{R}^{d\times d})^*)$ is two times continuously Fréchet differentiable. The second order derivative is given by*

$$\boldsymbol{C}''(\boldsymbol{\varphi})[\boldsymbol{h},\boldsymbol{v}](x) = \boldsymbol{C}''(\boldsymbol{\varphi}(x))[\boldsymbol{h}(x),\boldsymbol{v}(x)] \quad \text{for all } \boldsymbol{\varphi},\ \boldsymbol{h},\ \boldsymbol{v} \in L^\infty(\Omega)^N \text{ and a.e. in } \Omega.$$

Proof. Similar to the differentiability of first order (Lemma 6.8) the real valued case follows from [Trö09, Theorem 4.22]. The proof for the case that $\boldsymbol{\varphi}$ is a vector is similar: We have to prove that $\boldsymbol{C}' : L^\infty(\Omega)^N \to \mathcal{L}(L^\infty(\Omega)^N, L^\infty(\Omega; \mathbb{R}^{d\times d} \otimes (\mathbb{R}^{d\times d})^*))$ is Fréchet differentiable. Thus it has to hold $\boldsymbol{C}''(\boldsymbol{\varphi}) \in \mathcal{L}(L^\infty(\Omega)^N, \mathcal{L}(L^\infty(\Omega)^N, L^\infty(\Omega; \mathbb{R}^{d\times d} \otimes (\mathbb{R}^{d\times d})^*)))$ for all $\boldsymbol{\varphi} \in L^\infty(\Omega)^N$. This follows from the estimate

$$\|\boldsymbol{C}''(\boldsymbol{\varphi})[\boldsymbol{h}, \boldsymbol{v}]\|_{L^\infty} \le \|\boldsymbol{C}''(\boldsymbol{\varphi})\|_{L^\infty} \|\boldsymbol{h}\|_{L^\infty} \|\boldsymbol{v}\|_{L^\infty}$$

which holds for all $\boldsymbol{\varphi}$, $\boldsymbol{h}$ and $\boldsymbol{v} \in L^\infty(\Omega)^N$. For the rest term estimate we first do the calculation pointwise. Let $\boldsymbol{\varphi}$, $\boldsymbol{h}$ and $\boldsymbol{v} \in \mathbb{R}^N$. Then it holds using the fundamental theorem of calculus and the Lipschitz continuity of $\boldsymbol{C}''$ **(AP2)** that

$$\begin{aligned} |(\boldsymbol{C}'(\boldsymbol{\varphi}+\boldsymbol{h}) - \boldsymbol{C}'(\boldsymbol{\varphi}) - \boldsymbol{C}''(\boldsymbol{\varphi})\boldsymbol{h})\boldsymbol{v}| &= \left| \int_0^1 (\boldsymbol{C}''(\boldsymbol{\varphi}+t\boldsymbol{h})\boldsymbol{h} - \boldsymbol{C}''(\boldsymbol{\varphi})\boldsymbol{h})\boldsymbol{v} \,\mathrm{d}t \right| \\ &\le \int_0^1 Lt|\boldsymbol{h}|^2|\boldsymbol{v}| \,\mathrm{d}t = \frac{L}{2}|\boldsymbol{h}|^2|\boldsymbol{v}|. \end{aligned}$$

Thus it holds for $\boldsymbol{\varphi}$, $\boldsymbol{h}$ and $\boldsymbol{v} \in L^\infty(\Omega)^N$

$$\|(\boldsymbol{C}'(\boldsymbol{\varphi}+\boldsymbol{h}) - \boldsymbol{C}'(\boldsymbol{\varphi}) - \boldsymbol{C}''(\boldsymbol{\varphi})\boldsymbol{h})\boldsymbol{v}\|_{L^\infty} \le \frac{L}{2}\|\boldsymbol{h}\|_{L^\infty}^2 \|\boldsymbol{v}\|_{L^\infty} = \mathrm{o}\left(\|\boldsymbol{h}\|_{L^\infty}\right) \|\boldsymbol{v}\|_{L^\infty} \text{ as } \|\boldsymbol{h}\|_{L^\infty} \to 0$$

and C' is Fréchet differentiable. Note that the estimate has to be uniform in $\boldsymbol{v}$. It remains to show that $C'' : L^\infty(\Omega)^N \to \mathcal{L}(L^\infty(\Omega)^N, \mathcal{L}(L^\infty(\Omega)^N, L^\infty(\Omega; \mathbb{R}^{d\times d} \otimes (\mathbb{R}^{d\times d})^*)))$ is continuous. Let $\boldsymbol{\varphi}_i \to \boldsymbol{\varphi}$ in $L^\infty(\Omega)^N$ and let $\boldsymbol{h}$ and $\boldsymbol{v} \in L^\infty(\Omega)^N$ be arbitrary. We estimate using the Lipschitz continuity of $\boldsymbol{C}''$ **(AP2)**

$$\begin{aligned} \|(\boldsymbol{C}''(\boldsymbol{\varphi}_i) - \boldsymbol{C}''(\boldsymbol{\varphi}))[\boldsymbol{h}, \boldsymbol{v}]\|_{L^\infty} &\le \|\boldsymbol{C}''(\boldsymbol{\varphi}_i) - \boldsymbol{C}''(\boldsymbol{\varphi})\|_{L^\infty} \|\boldsymbol{h}\|_{L^\infty} \|\boldsymbol{v}\|_{L^\infty} \\ &\le L \underbrace{\|\boldsymbol{\varphi}_i - \boldsymbol{\varphi}\|_{L^\infty}}_{\to 0} \|\boldsymbol{h}\|_{L^\infty} \|\boldsymbol{v}\|_{L^\infty} \end{aligned}$$

uniformly in $\boldsymbol{h}$ and $\boldsymbol{v}$. Thus $\boldsymbol{C}''(\boldsymbol{\varphi}_i) \to \boldsymbol{C}''(\boldsymbol{\varphi})$ in $\mathcal{L}(L^\infty(\Omega)^N, \mathcal{L}(L^\infty(\Omega)^N, L^\infty(\Omega; \mathbb{R}^{d\times d} \otimes (\mathbb{R}^{d\times d})^*)))$. □

Theorem 6.11. *The control-to-state operator $S : H^1(\Omega)^N \cap L^\infty(\Omega)^N \to H^1_D$ is two times continuously Fréchet differentiable. The second order Fréchet derivative at $\boldsymbol{\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$ in directions $\boldsymbol{\delta\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$ and $\boldsymbol{\tau\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$ is given by $S''(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi}, \boldsymbol{\tau\varphi}] = \boldsymbol{z}$, where $\boldsymbol{z} \in H^1_D$ is the unique solution of the equation*

$$\begin{aligned} \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{z}) : \mathcal{E}(\boldsymbol{\xi}) = &- \int_\Omega \boldsymbol{C}''(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi}, \boldsymbol{\tau\varphi}]\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) + \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi},\boldsymbol{\varphi}}(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi}, \boldsymbol{\tau\varphi}] \cdot \boldsymbol{\xi} \\ &+ \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi},\boldsymbol{\varphi}}(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi}, \boldsymbol{\tau\varphi}] \cdot \boldsymbol{\xi} - \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\tau\varphi}\mathcal{E}(\boldsymbol{\delta u}) : \mathcal{E}(\boldsymbol{\xi}) \\ &- \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\mathcal{E}(\boldsymbol{\tau u}) : \mathcal{E}(\boldsymbol{\xi}) \quad \forall \boldsymbol{\xi} \in H^1_D, \end{aligned} \tag{127}$$

where $\boldsymbol{u} = S(\boldsymbol{\varphi})$ and $\boldsymbol{\delta u} = S'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}$ and $\boldsymbol{\tau u} = S'(\boldsymbol{\varphi})\boldsymbol{\tau\varphi}$ are given as in (124).

Proof. As in the proof of Theorem 6.9, we use the implicit function theorem. Recall the function $G : (H^1(\Omega)^N \cap L^\infty(\Omega)^N) \times H^1_D \to (H^1_D)^*$ defined in the proof of Theorem 6.9

$$\langle G(\boldsymbol{\varphi}, \boldsymbol{u}), \boldsymbol{\xi} \rangle_{(H^1_D)^*, H^1_D} := \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) - \int_\Omega \boldsymbol{f}(\boldsymbol{\varphi}) \cdot \boldsymbol{\xi} - \int_{\Gamma_g} \boldsymbol{g}(\boldsymbol{\varphi}) \cdot \boldsymbol{\xi}$$

and its Fréchet derivative

$$\begin{aligned}\left\langle G'(\boldsymbol{\varphi},\boldsymbol{u})(\boldsymbol{\delta\varphi},\boldsymbol{\delta u}),\boldsymbol{\xi}\right\rangle_{(H^1_D)^*,H^1_D} &= \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\mathcal{E}(\boldsymbol{u}):\mathcal{E}(\boldsymbol{\xi}) - \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\cdot\boldsymbol{\xi}\\ &\quad - \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\cdot\boldsymbol{\xi} + \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{\delta u}):\mathcal{E}(\boldsymbol{\xi})\end{aligned}$$

for all $(\boldsymbol{\varphi},\boldsymbol{u}) \in (H^1(\Omega)^N \cap L^\infty(\Omega)^N)\times H^1_D$, $(\boldsymbol{\delta\varphi},\boldsymbol{\delta u}) \in (H^1(\Omega)^N \cap L^\infty(\Omega)^N)\times H^1_D$ and $\boldsymbol{\xi} \in H^1_D$. All we have to show is $G' \in C^1((H^1(\Omega)^N \cap L^\infty(\Omega)^N)\times H^1_D, \mathcal{L}((H^1(\Omega)^N \cap L^\infty(\Omega)^N)\times H^1_D, (H^1_D)^*))$. We again show that G' is partially Fréchet differentiable with respect to $\boldsymbol{\varphi}$ and $\boldsymbol{u}$ and that the partial Fréchet derivatives are continuous. Note that all following estimates have to be uniform in $(\boldsymbol{\delta u},\boldsymbol{\delta\varphi})$ and in $\boldsymbol{\xi}$.

i) G' is partially Fréchet differentiable with respect to $\boldsymbol{u}$ for all $(\boldsymbol{\varphi},\boldsymbol{u}) \in (H^1(\Omega)^N \cap L^\infty(\Omega)^N)\times H^1_D$: G' is affine linear in $\boldsymbol{u}$ and continuous, since

$$\begin{aligned}\left|\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\mathcal{E}(\boldsymbol{u}):\mathcal{E}(\boldsymbol{\xi})\right| &\le \|\boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty}\|\boldsymbol{\delta\varphi}\|_{L^\infty}\|\boldsymbol{u}\|_{H^1_D}\|\boldsymbol{\xi}\|_{H^1_D}\\ &\le C\|(\boldsymbol{\delta\varphi},\boldsymbol{\delta u})\|_{L^\infty\times H^1_D}\|\boldsymbol{u}\|_{H^1_D}\|\boldsymbol{\xi}\|_{H^1_D}.\end{aligned}$$

Thus, G' is partially Fréchet differentiable with respect to $\boldsymbol{u}$ with derivative

$$\left\langle (D_{\boldsymbol{u}}G'(\boldsymbol{\varphi},\boldsymbol{u})\boldsymbol{h})(\boldsymbol{\delta\varphi},\boldsymbol{\delta u}),\boldsymbol{\xi}\right\rangle = \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\mathcal{E}(\boldsymbol{h}):\mathcal{E}(\boldsymbol{\xi})$$

for all $(\boldsymbol{\varphi},\boldsymbol{u}) \in (H^1(\Omega)^N \cap L^\infty(\Omega)^N)\times H^1_D$, $\boldsymbol{h} \in H^1_D$, $(\boldsymbol{\delta\varphi},\boldsymbol{\delta u}) \in (H^1(\Omega)^N \cap L^\infty(\Omega)^N)\times H^1_D$ and $\boldsymbol{\xi} \in H^1_D$.

ii) $D_{\boldsymbol{u}}G'$ is continuous in all $(\boldsymbol{\varphi},\boldsymbol{u}) \in (H^1(\Omega)^N \cap L^\infty(\Omega)^N)\times H^1_D$:
Let $\boldsymbol{\varphi}_n \to \boldsymbol{\varphi}$ in $H^1(\Omega)^N \cap L^\infty(\Omega)^N$ and $\boldsymbol{u}_n \to \boldsymbol{u}$ in H^1_D. Let $\boldsymbol{h} \in H^1_D$, $(\boldsymbol{\delta\varphi},\boldsymbol{\delta u}) \in (H^1(\Omega)^N \cap L^\infty(\Omega)^N)\times H^1_D$ and $\boldsymbol{\xi} \in H^1_D$. We estimate

$$\begin{aligned}&\left|\left\langle((D_{\boldsymbol{u}}G'(\boldsymbol{\varphi}_n,\boldsymbol{u}_n) - D_{\boldsymbol{u}}G'(\boldsymbol{\varphi},\boldsymbol{u}))\boldsymbol{h})(\boldsymbol{\delta\varphi},\boldsymbol{\delta u}),\boldsymbol{\xi}\right\rangle\right|\\ &= \left|\int_\Omega (\boldsymbol{C}'(\boldsymbol{\varphi}_n) - \boldsymbol{C}'(\boldsymbol{\varphi}))\boldsymbol{\delta\varphi}\mathcal{E}(\boldsymbol{h}):\mathcal{E}(\boldsymbol{\xi})\right| \le \|\boldsymbol{C}'(\boldsymbol{\varphi}_n) - \boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty}\|\boldsymbol{\delta\varphi}\|_{L^\infty}\|\boldsymbol{h}\|_{H^1_D}\|\boldsymbol{\xi}\|_{H^1_D}\\ &\le \underbrace{\|\boldsymbol{C}'(\boldsymbol{\varphi}_n) - \boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty}}_{\to 0}\|\boldsymbol{h}\|_{H^1_D}\|(\boldsymbol{\delta\varphi},\boldsymbol{\delta u})\|_{L^\infty\times H^1_D}\|\boldsymbol{\xi}\|_{H^1_D},\end{aligned}$$

using that $\boldsymbol{C}'$ is continuous (Lemma 6.8). Thus, $D_{\boldsymbol{u}}G'(\boldsymbol{\varphi}_n,\boldsymbol{u}_n) \to D_{\boldsymbol{u}}G'(\boldsymbol{\varphi},\boldsymbol{u})$ in $\mathcal{L}(H^1_D, \mathcal{L}((H^1(\Omega)^N \cap L^\infty(\Omega)^N)\times H^1_D, (H^1_D)^*))$.

iii) G' is partially Fréchet differentiable with respect to $\boldsymbol{\varphi}$ for all $(\boldsymbol{\varphi},\boldsymbol{u}) \in (H^1(\Omega)^N \cap L^\infty(\Omega)^N)\times H^1_D$: By formally differentiating G' with respect to $\boldsymbol{\varphi}$ we get the candidate for the Fréchet derivative:

$$\begin{aligned}&\langle(D_{\boldsymbol{\varphi}}G'(\boldsymbol{\varphi},\boldsymbol{u})\boldsymbol{h})(\boldsymbol{\delta\varphi},\boldsymbol{\delta u}),\boldsymbol{\xi}\rangle_{(H^1_D)^*,H^1_D}\\ &\qquad = \int_\Omega \boldsymbol{C}''(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi},\boldsymbol{h}]\mathcal{E}(\boldsymbol{u}):\mathcal{E}(\boldsymbol{\xi}) - \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}\boldsymbol{\varphi}}(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi},\boldsymbol{h}]\cdot\boldsymbol{\xi}\\ &\qquad\quad - \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi}\boldsymbol{\varphi}}(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi},\boldsymbol{h}]\cdot\boldsymbol{\xi} + \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{h}\mathcal{E}(\boldsymbol{\delta u}):\mathcal{E}(\boldsymbol{\xi})\end{aligned}$$

The first thing to prove is
$D_{\boldsymbol{\varphi}}G'(\boldsymbol{\varphi},\boldsymbol{u}) \in \mathcal{L}(H^1(\Omega)^N \cap L^\infty(\Omega)^N, \mathcal{L}((H^1(\Omega)^N \cap L^\infty(\Omega)^N)\times H^1_D, (H^1_D)^*))$. Lin-

earity is obvious. For the boundedness consider the estimate

$$\begin{aligned}
&\left|\left\langle (D_\varphi G'(\varphi,u)h)(\delta\varphi,\delta u),\xi\right\rangle_{(H^1_D)^*,H^1_D}\right| \\
&\le \|C''(\varphi)\|_{L^\infty}\|\delta\varphi\|_{L^\infty}\|h\|_{L^\infty}\|u\|_{H^1_D}\|\xi\|_{H^1_D} + \|f_{\varphi\varphi}(\varphi)\|_{L^2}\|\delta\varphi\|_{L^\infty}\|h\|_{L^\infty}\|\xi\|_{H^1_D} \\
&\quad + C\|g_{\varphi\varphi}(\varphi)\|_{L^2}\|\delta\varphi\|_{L^\infty}\|h\|_{L^\infty}\|\xi\|_{H^1_D} + \|C'(\varphi)\|_{L^\infty}\|h\|_{L^\infty}\|\delta u\|_{H^1_D}\|\xi\|_{H^1_D} \\
&\le C\|h\|_{L^\infty}\|(\delta\varphi,\delta u)\|_{L^\infty\times H^1_D}\|\xi\|_{H^1_D},
\end{aligned}$$

where we used Lemma 6.10 and **(AP6)**. Finally we have to estimate the rest term.

$$\begin{aligned}
&\left|\left\langle (G'(\varphi+h,u)-G'(\varphi,u)-D_\varphi G'(\varphi,u)h)(\delta\varphi,\delta u),\xi\right\rangle_{(H^1_D)^*,H^1_D}\right| \\
&\qquad\le \left|\int_\Omega (C'(\varphi+h)-C'(\varphi)-C''(\varphi)h)\delta\varphi\mathcal{E}(u):\mathcal{E}(\xi)\right| \\
&\qquad\quad + \left|\int_\Omega (f_\varphi(\varphi+h)-f_\varphi(\varphi)-f_{\varphi\varphi}(\varphi)h)\delta\varphi\cdot\xi\right| \\
&\qquad\quad + \left|\int_\Omega (g_\varphi(\varphi+h)-g_\varphi(\varphi)-g_{\varphi\varphi}(\varphi)h)\delta\varphi\cdot\xi\right| \\
&\qquad\quad + \left|\int_\Omega (C(\varphi+h)-C(\varphi)-C'(\varphi)h)\mathcal{E}(\delta u):\mathcal{E}(\xi)\right|
\end{aligned}$$

We treat the four terms separately. For the first term we get

$$\begin{aligned}
&\left|\int_\Omega (C'(\varphi+h)-C'(\varphi)-C''(\varphi)h)\delta\varphi\mathcal{E}(u):\mathcal{E}(\xi)\right| \\
&\qquad\le \underbrace{\|C'(\varphi+h)-C'(\varphi)-C''(\varphi)h\|_{L^\infty}}_{=o(\|h\|_{L^\infty})}\|\delta\varphi\|_{L^\infty}\|u\|_{H^1_D}\|\xi\|_{H^1_D},
\end{aligned}$$

due to Lemma 6.10. Now consider the second term.

$$\begin{aligned}
&\left|\int_\Omega (f_\varphi(\varphi+h)-f_\varphi(\varphi)-f_{\varphi\varphi}(\varphi)h)\delta\varphi\cdot\xi\right| \\
&\qquad\le \underbrace{\|f_\varphi(\varphi+h)-f_\varphi(\varphi)-f_{\varphi\varphi}(\varphi)h\|_{L^2}}_{=o(\|h\|_{L^\infty})}\|\delta\varphi\|_{L^\infty}\|\xi\|_{H^1_D},
\end{aligned}$$

since f is two times Fréchet differentiable **(AP6)**. An analogous estimate we get for the third term. For the last term we get

$$\begin{aligned}
&\left|\int_\Omega (C(\varphi+h)-C(\varphi)-C'(\varphi)h)\mathcal{E}(\delta u):\mathcal{E}(\xi)\right| \\
&\qquad\le \underbrace{\|C(\varphi+h)-C(\varphi)-C'(\varphi)h\|_{L^\infty}}_{=o(\|h\|_{L^\infty})}\|\delta u\|_{H^1_D}\|\xi\|_{H^1_D},
\end{aligned}$$

since C is Fréchet differentiable (Lemma 6.8). Summarizing, we proved

$$\begin{aligned}
&\left|\left\langle (G'(\varphi+h,u)-G'(\varphi,u)-D_\varphi G'(\varphi,u)h)(\delta\varphi,\delta u),\xi\right\rangle_{(H^1_D)^*,H^1_D}\right| \\
&\qquad = o(\|h\|_{L^\infty})\|(\delta\varphi,\delta u)\|_{L^\infty\times H^1_D}\|\xi\|_{H^1_D}
\end{aligned}$$

iv) $D_\varphi G'$ is continuous in all $(\varphi, u) \in (H^1(\Omega)^N \cap L^\infty(\Omega)^N) \times H^1_D$:
Let $\varphi_n \to \varphi$ in $H^1(\Omega)^N \cap L^\infty(\Omega)^N$ and $u_n \to u$ in H^1_D. Let $h \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$, $(\delta\varphi, \delta u) \in (H^1(\Omega)^N \cap L^\infty(\Omega)^N) \times H^1_D$ and $\xi \in H^1_D$. We estimate

$$\begin{aligned}
&\left|\left\langle((D_\varphi G'(\varphi_n, u_n) - D_\varphi G'(\varphi, u))h)(\delta\varphi, \delta u), \xi\right\rangle\right| \\
&\qquad \le \left|\int_\Omega C''(\varphi_n)[\delta\varphi, h]\mathcal{E}(u_n) : \mathcal{E}(\xi) - \int_\Omega C''(\varphi)[\delta\varphi, h]\mathcal{E}(u) : \mathcal{E}(\xi)\right| \\
&\qquad\quad + \left|\int_\Omega (f_{\varphi\varphi}(\varphi_n) - f_{\varphi\varphi}(\varphi))[\delta\varphi, h] \cdot \xi\right| \\
&\qquad\quad + \left|\int_{\Gamma_g} (g_{\varphi\varphi}(\varphi_n) - g_{\varphi\varphi}(\varphi))[\delta\varphi, h] \cdot \xi\right| \\
&\qquad\quad + \left|\int_\Omega (C'(\varphi_n) - C'(\varphi))h\mathcal{E}(\delta u) : \mathcal{E}(\xi)\right|
\end{aligned}$$

We again treat the four terms separately. The first term gives

$$\begin{aligned}
&\left|\int_\Omega C''(\varphi_n)[\delta\varphi, h]\mathcal{E}(u_n) : \mathcal{E}(\xi) - \int_\Omega C''(\varphi)[\delta\varphi, h]\mathcal{E}(u) : \mathcal{E}(\xi)\right| \\
&\le \left|\int_\Omega (C''(\varphi_n) - C''(\varphi))[\delta\varphi, h]\mathcal{E}(u_n) : \mathcal{E}(\xi)\right| + \left|\int_\Omega C''(\varphi)[\delta\varphi, h]\mathcal{E}(u_n - u) : \mathcal{E}(\xi)\right| \\
&\le (\underbrace{\|C''(\varphi_n) - C''(\varphi)\|_{L^\infty}}_{\to 0} \underbrace{\|u_n\|_{H^1_D}}_{\le C} + \|C''(\varphi)\|_{L^\infty} \underbrace{\|u_n - u\|_{H^1_D}}_{\to 0})\|\delta\varphi\|_{L^\infty}\|h\|_{L^\infty}\|\xi\|_{H^1_D},
\end{aligned}$$

since C'' is continuous (Lemma 6.10). For the second term we get

$$\left|\int_\Omega (f_{\varphi\varphi}(\varphi_n) - f_{\varphi\varphi}(\varphi))[\delta\varphi, h] \cdot \xi\right| \le \underbrace{\|f_{\varphi\varphi}(\varphi_n) - f_{\varphi\varphi}(\varphi)\|_{L^2}}_{\to 0}\|\delta\varphi\|_{L^\infty}\|h\|_{L^\infty}\|\xi\|_{H^1_D}$$

due to **(AP6)** and analogously for the third term. Consider the last term:

$$\left|\int_\Omega (C'(\varphi_n) - C'(\varphi))h\mathcal{E}(\delta u) : \mathcal{E}(\xi)\right| \le \underbrace{\|C'(\varphi_n) - C'(\varphi)\|_{L^\infty}}_{\to 0}\|h\|_{L^\infty}\|\delta u\|_{H^1_D}\|\xi\|_{H^1_D}.$$

Thus it holds $D_\varphi G'(\varphi_n, u_n) \to D_\varphi G'(\varphi, u)$ in $\mathcal{L}(H^1(\Omega)^N \cap L^\infty(\Omega)^N, \mathcal{L}((H^1(\Omega)^N \cap L^\infty(\Omega)^N) \times H^1_D, (H^1_D)^*))$.

The implicit function theorem also gives a formula for the second order derivative, which is [Zei85]

$$\begin{aligned}
&G_{\varphi,\varphi}(\varphi, S(\varphi))[\delta\varphi, \tau\varphi] + G_{\varphi,u}(\varphi, S(\varphi))[S'(\varphi)\delta\varphi, \tau\varphi] \\
&+G_{u,\varphi}(\varphi, S(\varphi))[\delta\varphi, S'(\varphi)\tau\varphi] + G_{u,u}(\varphi, S(\varphi))[S'(\varphi)\delta\varphi, S'(\varphi)\tau\varphi] \\
&\qquad +G_u(\varphi, S(\varphi))S''(\varphi)[\delta\varphi, \tau\varphi] = 0 \quad \text{in } (H^1_D)^*
\end{aligned}$$

for any $\varphi, \delta\varphi, \tau\varphi \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$. Testing the equation by $\xi \in H^1_D$ gives (127). □

In the case that only two phases are present, i.e. $N = 2$, the conclusions of this section also hold for the scalar valued phase field $\varphi = \varphi_1 - \varphi_2$.

Theorem 6.12. *Consider the state equation* (122) *defined for the scalar valued phase field* $\varphi \in H^1(\Omega) \cap L^\infty(\Omega)$. *Then it holds:*

The state equation (122) *has for every* $\varphi \in H^1(\Omega) \cap L^\infty(\Omega)$ *a unique weak solution* $u \in H^1_D$. *It holds the a priori estimate*

$$\|\boldsymbol{u}\|_{H^1_D} \leq c(\|\boldsymbol{f}(\varphi)\|_{L^2(\Omega)} + \|\boldsymbol{g}(\varphi)\|_{L^2(\Gamma_g)}) \tag{128}$$

for some $c > 0$ *independent of* φ. *Thus the control-to-state operator* $S : H^1(\Omega) \cap L^\infty(\Omega) \to H^1_D$ *is well defined.*
Moreover, for any $M > 0$ *there exists a constant* $C(M) > 0$, *such that for all* $\varphi_i \in H^1(\Omega) \cap L^\infty(\Omega)$ *with* $\|\varphi_i\|_{L^\infty} \leq M$, $i = 1, 2$, *and* $\boldsymbol{u}_i = S(\varphi_i)$, $i = 1, 2$, *it holds*

$$\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_{H^1} \leq C(M)\|\varphi_1 - \varphi_2\|_{L^\infty}. \tag{129}$$

The control-to-state operator is two times continuously Fréchet differentiable. The first order Fréchet derivative is given as $S'(\varphi)\delta\varphi = \boldsymbol{\delta u}$, *where* $\boldsymbol{\delta u}$ *is the solution of*

$$\int_\Omega \boldsymbol{C}(\varphi)\mathcal{E}(\boldsymbol{\delta u}) : \mathcal{E}(\boldsymbol{\xi}) = -\int_\Omega \boldsymbol{C}'(\varphi)\delta\varphi\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) + \int_\Omega \boldsymbol{f}_\varphi(\varphi)\delta\varphi \cdot \boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g}_\varphi(\varphi)\delta\varphi \cdot \boldsymbol{\xi} \quad \forall \boldsymbol{\xi} \in H^1_D. \tag{130}$$

The second order Fréchet derivative is given as $S''(\varphi)[\delta\varphi, \tau\varphi] = \boldsymbol{z}$, *where* $\boldsymbol{z}$ *is the solution of*

$$\begin{aligned}\int_\Omega \boldsymbol{C}(\varphi)\mathcal{E}(\boldsymbol{z}) : \mathcal{E}(\boldsymbol{\xi}) = &-\int_\Omega \boldsymbol{C}''(\varphi)[\delta\varphi, \tau\varphi]\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) + \int_\Omega \boldsymbol{f}_{\varphi,\varphi}(\varphi)[\delta\varphi, \tau\varphi] \cdot \boldsymbol{\xi} \\ &+ \int_{\Gamma_g} \boldsymbol{g}_{\varphi,\varphi}(\varphi)[\delta\varphi, \tau\varphi] \cdot \boldsymbol{\xi} - \int_\Omega \boldsymbol{C}'(\varphi)\tau\varphi\mathcal{E}(\boldsymbol{\delta u}) : \mathcal{E}(\boldsymbol{\xi}) \\ &- \int_\Omega \boldsymbol{C}'(\varphi)\delta\varphi\mathcal{E}(\boldsymbol{\tau u}) : \mathcal{E}(\boldsymbol{\xi}) \quad \forall \boldsymbol{\xi} \in H^1_D,\end{aligned} \tag{131}$$

where $\boldsymbol{\delta u} = S'(\varphi)\delta\varphi$ *and* $\boldsymbol{\tau u} = S'(\varphi)\tau\varphi$.

Proof. As in Section 6.1.3 we add a tilde to all functions involving the scalar valued phase field. The phase fields variables themselves can be distinguished by the typeface (normal for scalar and bold for vector).
From the definitions of $\tilde{\boldsymbol{C}}(\varphi)$, $\tilde{\boldsymbol{f}}(\varphi)$ and $\tilde{\boldsymbol{g}}(\varphi)$, we observe that the state equation (122) is equivalent to the state equation (103) for $\boldsymbol{\varphi} = (\frac{1+\varphi}{2}, \frac{1-\varphi}{2})^T$. Thus existence and uniqueness, as well as the a priori estimate (128) follows and it holds $\tilde{S}(\varphi) = S(\boldsymbol{\varphi})$.
Denote by T the function transforming a scalar valued phase field into a vector valued phase field, i.e. $T : H^1(\Omega) \cap L^\infty(\Omega) \to H^1(\Omega)^2 \cap L^\infty(\Omega)^2$, $T(\varphi) = (\frac{1+\varphi}{2}, \frac{1-\varphi}{2})^T$. Since T is affine linear and continuous we can conclude that T is a smooth transformation. Moreover it holds

$$\|T(\varphi_1) - T(\varphi_2)\|_{L^\infty} = \frac{1}{2}\|\varphi_1 - \varphi_2\|_{L^\infty}$$

for all $\varphi_i \in H^1(\Omega) \cap L^\infty(\Omega)$, $i = 1, 2$. Hence, from the local Lipschitz continuity of S we conclude (129).
Since T is smooth we get from the differentiability of $S, \boldsymbol{C}, \boldsymbol{f}$ and $\boldsymbol{g}$ also the differentiability of $\tilde{S} = S \circ T$, $\tilde{\boldsymbol{C}}$, $\tilde{\boldsymbol{f}}$ and $\tilde{\boldsymbol{g}}$ by the chain rule, where we use that $H^1(\Omega)^2 \cap L^\infty(\Omega)^2 \hookrightarrow L^\infty(\Omega)^2$, as well as Lemma 7.4.
To get the equations for the first and second order derivatives, we observe that tangential

directions transform as $\boldsymbol{\delta\varphi} := T'(\varphi)\delta\varphi = (\frac{1}{2}\delta\varphi, -\frac{1}{2}\delta\varphi)^T$. Thus by chain rule we get

$$\tilde{S}'(\varphi)\delta\varphi = S'(T(\varphi))T'(\varphi)\delta\varphi = S'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}.$$

The derivatives of $\boldsymbol{C}$, $\boldsymbol{f}$ and $\boldsymbol{g}$ transform in the same way. Thus the linearized equation for the scalar valued phase field (130) is a consequence of the linearized equation for the vector valued phase field (124).
The second order derivatives transform like $\tilde{S}''(\varphi)[\delta\varphi, \tau\varphi] = S''(T(\varphi))[T'(\varphi)\delta\varphi, T'(\varphi)\tau\varphi] + S'(T(\varphi))T''(\varphi)[\delta\varphi, \tau\varphi]$, where the second term vanishes since T is affine linear. Thus we have

$$\tilde{S}''(\varphi)[\delta\varphi, \tau\varphi] = S''(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi}, \boldsymbol{\tau\varphi}].$$

This again holds also for $\boldsymbol{C}$, $\boldsymbol{f}$ and $\boldsymbol{g}$ and hence (131) follows from (127). □

Remark 6.13. The proof of well-posedness and C^2-regularity of $S : H^1(\Omega)^N \cap L^\infty(\Omega)^N \to H^1_D$ also works if $\boldsymbol{C}$, $\boldsymbol{f}$ and $\boldsymbol{g}$ are only abstract operators rather than Nemytskii operators. The only property needed is that the operators

$$\begin{aligned}
\boldsymbol{C} &: H^1(\Omega)^N \cap L^\infty(\Omega)^N \to L^\infty(\Omega; \mathbb{R}^{d\times d} \otimes (\mathbb{R}^{d\times d})^*),\\
\boldsymbol{f} &: H^1(\Omega)^N \cap L^\infty(\Omega)^N \to L^2(\Omega)^d \text{ and}\\
\boldsymbol{g} &: H^1(\Omega)^N \cap L^\infty(\Omega)^N \to L^2(\Gamma_g)^d
\end{aligned}$$

are two times continuously Fréchet differentiable and that $\boldsymbol{C}$ fulfills the structural assumptions **(AP3)** and **(AP4)** for all $\boldsymbol{\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$. In this case some norms in the proof have to be changed, e.g.

$$\|\boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty}$$

becomes

$$\|\boldsymbol{C}'(\boldsymbol{\varphi})\|_{\mathcal{L}(H^1\cap L^\infty, L^\infty)}$$

and so on. By the techniques in this section it can also be proved that S is C^k if the operators $\boldsymbol{C}$, $\boldsymbol{f}$ and $\boldsymbol{g}$ are C^k for any $k > 2$.
A possible application for abstract operators would be the presence of eigenstrain $\overline{\mathcal{E}}_i$ of the i-th material, cf. [BGHR15, Hec14], i.e. we consider the state equation

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})(\mathcal{E}(\boldsymbol{u}) - \overline{\mathcal{E}}(\boldsymbol{\varphi})) : \mathcal{E}(\boldsymbol{\xi}) = \int_\Omega \boldsymbol{f}(\boldsymbol{\varphi}) \cdot \boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g}(\boldsymbol{\varphi}) \cdot \boldsymbol{\xi} \quad \forall \boldsymbol{\xi} \in H^1_D,$$

where $\overline{\mathcal{E}}(\boldsymbol{\varphi})$ is a suitable interpolation of the values $\overline{\mathcal{E}}_i$, assuming $\overline{\mathcal{E}}_i = (\overline{\mathcal{E}}_i)^T$. Let the stiffness tensors and the eigenstrains be interpolated linearly on the interface, i.e.. let it hold $\boldsymbol{C}(\boldsymbol{\varphi}) = \sum_{i=1}^N \varphi_i \boldsymbol{C}_i$ and $\overline{\mathcal{E}}(\boldsymbol{\varphi}) = \sum_{i=1}^N \varphi_i \overline{\mathcal{E}}_i$ for $\boldsymbol{\varphi} \in \Delta^{n-1}$. Then we can put the eigenstrain term on the right hand side into the operators $\boldsymbol{f}$ and $\boldsymbol{g}$ as can be seen as

follows. Integration by parts and using the symmetry of $\boldsymbol{C}_m\overline{\mathcal{E}}_n$ leads to

$$\begin{aligned}\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\overline{\mathcal{E}}(\boldsymbol{\varphi}) : \mathcal{E}(\boldsymbol{\xi}) &= \int_\Omega \sum_{m,n=1}^{N} (\varphi_m\varphi_n \boldsymbol{C}_m\overline{\mathcal{E}}_n) : \mathcal{E}(\boldsymbol{\xi}) \\ &= -\int_\Omega \sum_{m,n=1}^{N} ((\boldsymbol{C}_m\overline{\mathcal{E}}_n)\nabla(\varphi_m\varphi_n))\cdot\boldsymbol{\xi} + \int_{\partial\Omega\setminus\Gamma_D} \sum_{m,n=1}^{N} ((\varphi_m\varphi_n\boldsymbol{C}_m\overline{\mathcal{E}}_n)\nu)\cdot\boldsymbol{\xi}\end{aligned}$$

for all $\boldsymbol{\xi} \in H^1_D$, where ν is the outer normal of Ω. The term $-\sum_{m,n=1}^{N}((\boldsymbol{C}_m\overline{\mathcal{E}}_n)\nabla(\varphi_m\varphi_n))$ can be put into $\boldsymbol{f}$ and the term $\sum_{m,n=1}^{N}((\varphi_m\varphi_n\boldsymbol{C}_m\overline{\mathcal{E}}_n)\nu)$ into $\boldsymbol{g}$ by defining $\Gamma_g := \partial\Omega\setminus\Gamma_D$. In this case $\boldsymbol{f}$ is no Nemytskii operator, since it does not depend pointwise on $\boldsymbol{\varphi}$ due to the gradient operator. However, the required smoothness of the terms can be seen easily by noting that the terms are quadratic in $\boldsymbol{\varphi}$ and that it holds $\|\nabla(\varphi_m\varphi_n)\|_{L^2} \le 2\|\varphi_m\|_{H^1\cap L^\infty}\|\varphi_n\|_{H^1\cap L^\infty}$ and $\|\varphi_m\varphi_n\|_{L^2(\partial\Omega)} \le \|\varphi_m\|_{H^1\cap L^\infty}\|\varphi_n\|_{H^1\cap L^\infty}$, cf. Lemma 7.4 and Theorem 7.5.
Another application would be the presence of design dependent loads considered in [BC03]. There, a part of $\boldsymbol{f}$ is given as $p\nabla(L(\varphi))$, where $p : \overline{\Omega} \to \mathbb{R}$ is a smooth pressure and $L : \mathbb{R} \to [0,1]$ is a smooth function, which is bounded with bounded derivatives (describing the liquid phase). The required smoothness of $\boldsymbol{f}$ can again be easily shown.

Remark 6.14. We cannot prove that S is differentiable in $\boldsymbol{\varphi} \in H^1(\Omega)^N$ by standard methods. The reason is that $\boldsymbol{\varphi}$ enters the coefficients in the second order term $\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi})$. Since $\mathcal{E}(\boldsymbol{u})$ and $\mathcal{E}(\boldsymbol{\xi})$ are in general not better than L^2, $\boldsymbol{C}(\boldsymbol{\varphi})$ has to be in L^∞. To show the differentiability of $\boldsymbol{C} : H^1 \to L^\infty$, one would use the Sobolev embedding $H^1 \hookrightarrow L^6$ for $d \le 3$ and prove that $\boldsymbol{C} : L^6 \to L^\infty$ is differentiable. But this can only hold if the Nemytskii operator $\boldsymbol{C}$ does not depend on $\boldsymbol{\varphi}$ [KZPP76, Thm 20.1], which is not the case. An alternative would be to introduce a smoothing operator G, as it is done in [CK15] for a similar problem, and consider the operator $\boldsymbol{C}(G(\boldsymbol{\varphi}))$ (and similarly $\boldsymbol{f}(G(\boldsymbol{\varphi}))$ and $\boldsymbol{g}(G(\boldsymbol{\varphi}))$). This corresponds to the new smoothed control-to-state operator $S \circ G$. If G is differentiable in $H^1(\Omega)^N \to H^1(\Omega)^N \cap L^\infty(\Omega)^N$, then $S\circ G$ is differentiable in $\boldsymbol{\varphi} \in H^1(\Omega)^N$. However, such a smoothening is not necessary since the VMPT method can handle the space $L^\infty(\Omega)^N$.
We also refer to [CJ12], where a similar state equation is considered, namely a second order elliptic equation (no system) with control in the second order coefficient on a domain $\Omega \subset \mathbb{R}^2$. They can show differentiability in $\varphi \in H^1(\Omega)$, since the state u has the regularity $u \in W^{1,p}(\Omega)$ for some $p > 2$. Therefor a smooth enough boundary of Ω is assumed. Moreover the Sobolev embedding $H^1(\Omega) \hookrightarrow L^q(\Omega)$ for any $q < \infty$ can be used due to the two dimensional domain. Since differentiability in L^∞ is sufficient for the VMPT method, we don't need to assume a smooth boundary and we can allow any space dimension.
Similarly, in [IK96] the embedding $H^2(\Omega) \hookrightarrow L^\infty(\Omega)$ in 2D is used to show differentiability of the control-to-state operator in $\varphi \in H^2(\Omega)$. Again, we don't need a Hilbert space and thus can take L^∞ without restrictions on the space dimension.

Remark 6.15. The regularity $\boldsymbol{\varphi} \in H^1(\Omega)^N$ is only needed since the boundary traction $\boldsymbol{g}$ depends on $\boldsymbol{\varphi}$ and thus the trace of $\boldsymbol{\varphi}$ has to be well defined. In the case that $\boldsymbol{g}$ is independent of $\boldsymbol{\varphi}$ the control-to-state operator S is differentiable from $L^\infty(\Omega)^N$ into H^1_D, which is a stronger property, see [BFGS14].

6.3 Existence of minimizers

We showed that the state equation has a unique solution $\boldsymbol{u}$ for every $\boldsymbol{\varphi}$. Thus $\boldsymbol{u}$ can be eliminated in the optimization problem and we can define the reduced cost functional.

Definition 6.16. The reduced cost functional is defined by

$$j(\boldsymbol{\varphi}) := \gamma E(\boldsymbol{\varphi}) + F(\boldsymbol{\varphi}, S(\boldsymbol{\varphi})).$$

For scalar valued phase fields we define analogously

$$\tilde{j}(\varphi) := \tilde{\gamma}\tilde{E}(\varphi) + \tilde{F}(\varphi, \tilde{S}(\varphi)).$$

Since the transformation between scalar valued and vector valued phase fields is a homeomorphism, which we show in the following, we get that the optimization problem concerning the scalar valued phase field is equivalent to the optimization problem concerning the vector valued phase field.

Theorem 6.17. *Let $N = 2$. Let φ be a local minimizer of $\tilde{j}$ in $\widetilde{\Phi_{ad}}$. Then $(\frac{1+\varphi}{2}, \frac{1-\varphi}{2})^T$ is a local minimizer of j in Φ_{ad}. Conversely, let $\boldsymbol{\varphi}$ be a local minimizer of j in Φ_{ad}. Then $\varphi_1 - \varphi_2$ is a local minimizer of $\tilde{j}$ in $\widetilde{\Phi_{ad}}$.*

Proof. Let $T : H^1(\Omega) \cap L^\infty(\Omega) \to H^1(\Omega)^2 \cap L^\infty(\Omega)^2$, $T(\varphi) = (\frac{1+\varphi}{2}, \frac{1-\varphi}{2})^T$. We first show that T is a homeomorphism between $\widetilde{\Phi_{ad}}$ and Φ_{ad}, i.e. T is bijective, continuous and its inverse is continuous.

- Let $\varphi \in \widetilde{\Phi_{ad}}$. We show $T(\varphi) \in \Phi_{ad}$. From $-1 \le \varphi \le 1$ we get $\frac{1\pm\varphi}{2} \ge 0$ and it also holds $\frac{1+\varphi}{2} + \frac{1-\varphi}{2} = 1$. Moreover, $\mathcal{f}\,\frac{1+\varphi}{2} = \frac{1}{2}(1 + \mathcal{f}\,\varphi) = \frac{1}{2}(1 + \tilde{\mathfrak{m}}) = \frac{1}{2}(1 + \mathfrak{m}_1 - \mathfrak{m}_2) = \frac{1}{2}(\mathfrak{m}_1 + \mathfrak{m}_2 + \mathfrak{m}_1 - \mathfrak{m}_2) = \mathfrak{m}_1$. By the same way we get $\mathcal{f}\,\frac{1-\varphi}{2} = \mathfrak{m}_2$.
- T is injective: Let $\varphi, \overline{\varphi} \in \widetilde{\Phi_{ad}}$ with $T(\varphi) = T(\overline{\varphi})$. Then it holds $\varphi = T_1(\varphi) - T_2(\varphi) = T_1(\overline{\varphi}) - T_2(\overline{\varphi}) = \overline{\varphi}$.
- T is surjective: Let $\boldsymbol{\varphi} \in \Phi_{ad}$. Then it holds for $\varphi := \varphi_1 - \varphi_2$ that $T(\varphi) = (\frac{1+\varphi_1-\varphi_2}{2}, \frac{1-(\varphi_1-\varphi_2)}{2})^T = (\frac{\varphi_1+\varphi_2+\varphi_1-\varphi_2}{2}, \frac{\varphi_1+\varphi_2-(\varphi_1-\varphi_2)}{2})^T = (\varphi_1, \varphi_2)^T = \boldsymbol{\varphi}$. It remains to show $\varphi \in \widetilde{\Phi_{ad}}$. From $\boldsymbol{\varphi} \in \Phi_{ad}$ we get $0 \le \varphi_1 \le 1$ and $0 \le \varphi_2 \le 1$. Thus $-1 \le \varphi \le 1$. Moreover it holds $\mathcal{f}\,\varphi = \mathcal{f}(\varphi_1 - \varphi_2) = \mathfrak{m}_1 - \mathfrak{m}_2 = \tilde{\mathfrak{m}}$.
- T is continuous: Let $\varphi_i \to \varphi$ in $H^1(\Omega) \cap L^\infty(\Omega)$. Then also $\frac{1\pm\varphi_i}{2} \to \frac{1\pm\varphi}{2}$ in $H^1(\Omega) \cap L^\infty(\Omega)$ and thus $T(\varphi_i) \to T(\varphi)$ in $H^1(\Omega)^2 \cap L^\infty(\Omega)^2$.
- By the same argument one shows that $T^{-1}(\boldsymbol{\varphi}) = \varphi_1 - \varphi_2$ is continuous.

We are now able to show the statement of the theorem. From equation (120) and from $\tilde{S}(\varphi) = S(T(\varphi))$ (cf. Theorem 6.12), we get $\tilde{j}(\varphi) = j(T(\varphi))$. Let now φ be a local minimizer of $\tilde{j}$ in $\widetilde{\Phi_{ad}}$. Then there exists a neighborhood $\tilde{U}$ of φ in $\widetilde{\Phi_{ad}}$, such that

$$\begin{aligned} \tilde{j}(\varphi) \le \tilde{j}(\eta) \quad \forall \eta \in \tilde{U} \text{ and} \\ j(T(\varphi)) \le j(T(\eta)) \quad \forall \eta \in \tilde{U}, \end{aligned}$$

respectively. Let $U := T(\tilde{U})$. Since T is a homeomorphism, we get that U is a neighborhood of $T(\varphi)$ in Φ_{ad}. Thus it holds

$$j(T(\varphi)) \le j(\boldsymbol{\eta})) \quad \forall \boldsymbol{\eta} \in U,$$

which shows that $T(\varphi)$ is a local minimizer of j in Φ_{ad}. The other direction can be proved in the same way, noting that $j(\boldsymbol{\varphi}) = \tilde{j}(T^{-1}(\boldsymbol{\varphi}))$ and T^{-1} is again a homeomorphism. □

Theorem 6.18. *There exists a global minimizer of problem* (102).

Proof. The proof for the special case of the mean compliance and compliant mechanism problem can be found in [BFGS14]. Our problem is different since we have a general objective and the forces $\boldsymbol{f}$ and $\boldsymbol{g}$ may depend on the phase field $\boldsymbol{\varphi}$.
The proof is by the direct method in the calculus of variations. The gradient term in the Ginzburg-Landau energy $\frac{\varepsilon}{2}\int_\Omega |\nabla\boldsymbol{\varphi}|^2$ is nonnegative. Since ψ_0 is continuous, it is bounded from below on the compact set Δ^{N-1}, thus the potential term $\frac{1}{\varepsilon}\int_\Omega \psi_0(\boldsymbol{\varphi})$ is bounded from below. By assumption, also $F(\boldsymbol{\varphi}, S(\boldsymbol{\varphi}))$ is bounded from below on Φ_{ad}. Hence, the reduced cost functional is bounded from below on Φ_{ad} and we can choose a minimizing sequence $(\boldsymbol{\varphi}_n)_n \subset \Phi_{ad}$ with $j(\boldsymbol{\varphi}_n) \to \inf_{\boldsymbol{\varphi}\in\Phi_{ad}} j(\boldsymbol{\varphi})$. Since all terms in the reduced cost functional are bounded from below there exists some $C > 0$ such that

$$j(\boldsymbol{\varphi}_n) \geq \gamma\frac{\varepsilon}{2}\int_\Omega |\nabla\boldsymbol{\varphi}_n|^2 - C.$$

From the convergence of $j(\boldsymbol{\varphi}_n)$ we get that $(\nabla\boldsymbol{\varphi}_n)_n$ is bounded in $L^2(\Omega)^{N\times d}$. Since Φ_{ad} is bounded in L^∞, we also get that $(\boldsymbol{\varphi}_n)_n$ is bounded in L^∞ and thus in L^2. This gives the boundedness of $(\boldsymbol{\varphi}_n)_n$ in $H^1(\Omega)^N$. We can extract a subsequence, again denoted by $\boldsymbol{\varphi}_n$, such that $\boldsymbol{\varphi}_n \to \boldsymbol{\varphi}$ weakly in H^1 for some $\boldsymbol{\varphi} \in H^1$. The admissible set Φ_{ad} is closed in H^1 and convex, thus $\boldsymbol{\varphi} \in \Phi_{ad}$ [Trö09]. From the compact embedding we get $\boldsymbol{\varphi}_n \to \boldsymbol{\varphi}$ strongly in $L^2(\Omega)$ and strongly in $L^2(\partial\Omega)$ in the trace sense [Alt12] and by possibly choosing another subsequence $\boldsymbol{\varphi}_n \to \boldsymbol{\varphi}$ almost everywhere in Ω and almost everywhere in $\partial\Omega$ in the trace sense. Since the H^1-half-norm is weakly lower semi continuous (see also the estimate (25)), we get

$$\liminf_{n\to\infty} \gamma\frac{\varepsilon}{2}\int_\Omega |\nabla\boldsymbol{\varphi}_n|^2 \geq \gamma\frac{\varepsilon}{2}\int_\Omega |\nabla\boldsymbol{\varphi}|^2.$$

Since $\psi_0(\boldsymbol{\varphi}_n)$ is uniformly bounded in L^∞ we get from the dominated convergence theorem

$$\lim_{n\to\infty} \frac{\gamma}{\varepsilon}\int_\Omega \psi_0(\boldsymbol{\varphi}_n) = \frac{\gamma}{\varepsilon}\int_\Omega \psi_0(\boldsymbol{\varphi}).$$

From the a priori estimate (123),

$$\|\boldsymbol{u}_n\|_{H^1_D} \leq c(\|\boldsymbol{f}(\boldsymbol{\varphi}_n)\|_{L^2(\Omega)} + \|\boldsymbol{g}(\boldsymbol{\varphi}_n)\|_{L^2(\Gamma_g)}),$$

we get the boundedness of $\boldsymbol{u}_n := S(\boldsymbol{\varphi}_n)$ in H^1 since by (106)-(107), the right hand side is bounded. Thus $\boldsymbol{u}_n \to \boldsymbol{u}$ weakly in H^1 and strongly in L^2 for a subsequence and for some $\boldsymbol{u} \in H^1_D$. We show $\boldsymbol{u} = S(\boldsymbol{\varphi})$. It holds

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_n)\mathcal{E}(\boldsymbol{u}_n) : \mathcal{E}(\boldsymbol{\xi}) = \int_\Omega \boldsymbol{f}(\boldsymbol{\varphi}_n)\cdot\boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g}(\boldsymbol{\varphi}_n)\cdot\boldsymbol{\xi} \quad \forall \boldsymbol{\xi} \in H^1_D$$

for all n. Let $\boldsymbol{\xi} \in C^\infty(\overline{\Omega})^d$. Since $\boldsymbol{C}$ is continuous on the compact set Δ^{N-1}, we have that $\boldsymbol{C}(\boldsymbol{\varphi}_n)\mathcal{E}(\boldsymbol{\xi})$ is bounded in $L^\infty(\Omega)$ uniformly in n. Again by the dominated convergence theorem we get $\boldsymbol{C}(\boldsymbol{\varphi}_n)\mathcal{E}(\boldsymbol{\xi}) \to \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{\xi})$ in $L^2(\Omega)^{d\times d}$ and thus

$$\begin{aligned}\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_n)\mathcal{E}(\boldsymbol{u}_n) : \mathcal{E}(\boldsymbol{\xi}) &= \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_n)\mathcal{E}(\boldsymbol{\xi}) : \mathcal{E}(\boldsymbol{u}_n) \\ &\to \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{\xi}) : \mathcal{E}(\boldsymbol{u}) = \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}).\end{aligned}$$

By (106)-(107) we get that $\boldsymbol{f}(\boldsymbol{\varphi}_n)$ and $\boldsymbol{g}(\boldsymbol{\varphi}_n)$ are bounded in L^2. Moreover, due to **(AP5)**, we have that $\boldsymbol{f}(x, \boldsymbol{\varphi}_n(x)) \to \boldsymbol{f}(x, \boldsymbol{\varphi}(x))$ almost everywhere in Ω and $\boldsymbol{g}(x, \boldsymbol{\varphi}_n(x)) \to$

$\boldsymbol{g}(x, \boldsymbol{\varphi}(x))$ almost everywhere in Γ_g. As a consequence of Vitali's convergence theorem [Alt12, 1.23], where the required equiintegrability follows from the uniform boundedness in L^2, we get that $\boldsymbol{f}(\boldsymbol{\varphi}_n) \to \boldsymbol{f}(\boldsymbol{\varphi})$ and $\boldsymbol{g}(\boldsymbol{\varphi}_n) \to \boldsymbol{g}(\boldsymbol{\varphi})$ in L^1, thus

$$\int_\Omega \boldsymbol{f}(\boldsymbol{\varphi}_n) \cdot \boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g}(\boldsymbol{\varphi}_n) \cdot \boldsymbol{\xi} \to \int_\Omega \boldsymbol{f}(\boldsymbol{\varphi}) \cdot \boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g}(\boldsymbol{\varphi}) \cdot \boldsymbol{\xi} \quad \forall \boldsymbol{\xi} \in C^\infty(\overline{\Omega})^d.$$

Hence we have

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}) \mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) = \int_\Omega \boldsymbol{f}(\boldsymbol{\varphi}) \cdot \boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g}(\boldsymbol{\varphi}) \cdot \boldsymbol{\xi} \quad \forall \boldsymbol{\xi} \in C^\infty(\overline{\Omega})^d \cap H^1_D$$

and using that $C^\infty(\overline{\Omega})^d \cap H^1_D$ is dense in H^1_D [EG91], we get that $\boldsymbol{u} = S(\boldsymbol{\varphi})$.
By assumption **(AP10)** we now get

$$\liminf_{n\to\infty} F(\boldsymbol{\varphi}_n, S(\boldsymbol{\varphi}_n)) \geq F(\boldsymbol{\varphi}, S(\boldsymbol{\varphi}))$$

and thus

$$\inf_{\boldsymbol{\eta} \in \Phi_{ad}} j(\boldsymbol{\eta}) = \liminf_{n\to\infty} j(\boldsymbol{\varphi}_n) \geq j(\boldsymbol{\varphi}) \geq \inf_{\boldsymbol{\eta} \in \Phi_{ad}} j(\boldsymbol{\eta}).$$

We conclude that equality holds and that $\boldsymbol{\varphi}$ is a global minimizer of j in Φ_{ad}. □

Due to Theorem 6.17 the same result holds for the scalar valued problem (121).

6.4 Γ-convergence result

In this section we show that for certain cost functionals the optimization problem involving two phases approximates a sharp interface problem as $\varepsilon \to 0$ in the sense of Γ-convergence. This has been shown in [BGHR15] under marginally different assumptions.

The following definition of Γ-convergence can be found in [DM93, Prop. 8.1].

Definition 6.19. Let X be a first countable topological space (e.g. a metric space) and let $f_i : X \to \overline{\mathbb{R}}$, $i \in \mathbb{N}$ be a sequence of functions on X, where $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$. We say f_i Γ-converges to $f : X \to \overline{\mathbb{R}}$ as $i \to \infty$ if

1. For each $x \in X$ and each sequence $(x_i)_i \subset X$ with $x_i \to x$ in X it holds

$$\liminf_i f_i(x_i) \geq f(x).$$

2. For each $x \in X$ there exists a sequence $(x_i)_i \subset X$ with $x_i \to x$ in X and

$$\lim_i f_i(x_i) = f(x).$$

The benefit of Γ-convergence in optimization is that minimizers converge to minimizers as is stated by the next Theorem [DM93, Cor. 7.20].

Theorem 6.20. *Let f_i Γ-converge to f and let x_i be a (global) minimizer of f_i for all $i \in \mathbb{N}$. Then it holds*

1. Each accumulation point x of $(x_i)_i$ is a (global) minimizer of f and

$$\limsup_i f_i(x_i) = f(x).$$

2. If $x_i \to x$ in X then

$$\lim_i f_i(x_i) = f(x).$$

It turns out that the right space for the Γ-limit problem is the space of functions with bounded variations. The following definition of the space $BV(\Omega)$ can be found in [EG91] and [BE91b].

Definition 6.21. Let $\Omega \subset \mathbb{R}^d$ be an open set. For $f \in L^1(\Omega)$ we define

$$\|Df\| := \sup\{ \int_\Omega f \nabla \cdot \boldsymbol{\varphi} \mid \boldsymbol{\varphi} \in C_0^1(\Omega)^d,\ \|\boldsymbol{\varphi}\|_\infty \le 1\}.$$

The space of functions with bounded variation in Ω is then defined by

$$BV(\Omega) := \{f \in L^1(\Omega) \mid \|Df\| < \infty\} \quad \text{and}$$
$$BV(\Omega, \{\pm 1\}) := \{f \in BV(\Omega) \mid f \in \{\pm 1\} \text{ a.e. in } \Omega\}.$$

For a bounded measurable subset $E \subset \Omega$ we define the perimeter of E in Ω

$$P_\Omega(E) := \|D\chi_E\|,$$

where χ_E is the characteristic function of E.

The following Γ-convergence result is based on the techniques used in [BC03, BGHR15, Hec14]. The difference to [BGHR15] is that we consider an arbitrary cost functional F here, which is not necessarily in integral form. However, the same proof can be applied. We add the index ε to the Ginzburg Landau energy E_ε to indicate the ε-dependency. Note that the functional F may not depend on ε.

Theorem 6.22. *Consider the scalar valued case for two phases and $d \in \{2,3\}$. Let $\boldsymbol{f}$ and $\boldsymbol{g}$ be independent of φ and let $\psi_0(\varphi) = \frac{1}{2}(1-\varphi^2)$. Let F and S be independent of ε and let $F(\varphi, \boldsymbol{u})$ be well defined for all $\varphi \in L^1(\Omega)$ with $|\varphi| \le 1$. Moreover, let for any sequence $(\varphi_i)_i \subset L^1(\Omega)$ with $|\varphi| \le 1$ and $\varphi_i \to \varphi$ in L^1 for some $\varphi \in L^1(\Omega)$ and for any sequence $(\boldsymbol{u}_i)_i \subset H^1_D$ with $\boldsymbol{u}_i \to \boldsymbol{u}$ weakly in H^1 for some $\boldsymbol{u} \in H^1_D$ hold $F(\varphi_i, \boldsymbol{u}_i) \to F(\varphi, \boldsymbol{u})$. Then the functionals*

$$\begin{cases} \gamma E_\varepsilon(\varphi) + F(\varphi, S(\varphi)) & \varphi \in H^1(\Omega),\ \fint_\Omega \varphi = \mathfrak{m},\ |\varphi| \le 1 \text{ a.e. in } \Omega \\ \infty & \text{else} \end{cases}$$

Γ-converge as $\varepsilon \to 0$ in $L^1(\Omega)$ to the functional

$$\begin{cases} \gamma c_0 P_\Omega(\{\varphi = 1\}) + F(\varphi, S(\varphi)) & \varphi \in BV(\Omega, \{\pm 1\}),\ \fint_\Omega \varphi = \mathfrak{m} \\ \infty & \text{else} \end{cases}$$

where $c_0 = \int_{-1}^1 \sqrt{2\psi_0} = \frac{\pi}{2}$. Note that S is well defined for $\varphi \in L^\infty(\Omega)$ since $\boldsymbol{g}$ is independent of φ, see also [BFGS14, BGHR15].

Proof. In [BGHR15] it is proved that $S(\varphi_i) \to S(\varphi)$ weakly in H^1 for any sequence $(\varphi_i)_i \subset L^1(\Omega)$ with $|\varphi_i| \le 1$ and $\varphi_i \to \varphi$ in L^1 for some $\varphi \in L^1(\Omega)$. Thus we get that $\varphi \mapsto F(\varphi, S(\varphi))$ is continuous in $\{\varphi \in L^1(\Omega) \mid |\varphi| \le 1\}$. We can extend this map continuously on whole $L^1(\Omega)$ without changing the functionals above. By a result of [Mod87, BE91b] we get

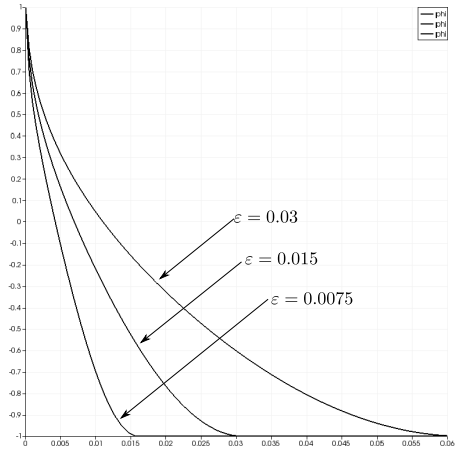


Figure 2: Counterexample for Γ-conv.

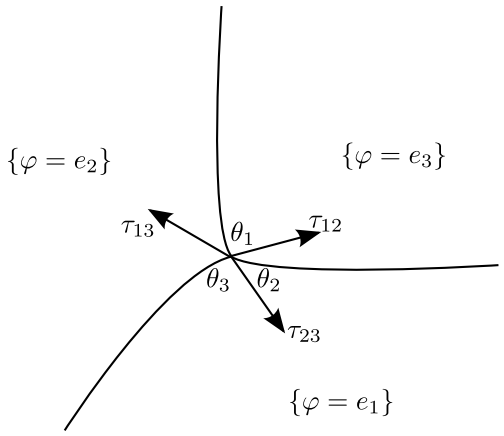


Figure 3: Angles at a triple junction.

the Γ-convergence of the Ginzburg-Landau energy including the mass constraint and the inequality constraints. Since Γ-convergence is invariant under continuous perturbations (see e.g. [DM93, Prop. 6.21]), the assumption follows.

□

If $\boldsymbol{f}$ depends on φ then the above Γ-convergence can also be shown if one assumes e.g. that $|\boldsymbol{f}(x,\varphi)| \leq C$ for all $\varphi \in [-1,1]$ and almost every $x \in \Omega$. However, $\boldsymbol{g}$ cannot depend on φ pointwise, since traces are not continuous in L^1 . This can be also seen in the numerics: If in the cantilever beam experiment (see Example 6.83) the boundary traction $\boldsymbol{g}$ is chosen smaller in the weak phase than in the strong phase then the minimizers fulfill $\varphi_\varepsilon|_{\Gamma_g} = 1$ (weak phase) for all $\varepsilon > 0$. However, for the L^1-limit φ_0 it holds $\varphi_0|_{\Gamma_g} = -1$ (strong phase), resulting in an upward jump of the compliance in the limit, which is a contradiction to the lim inf-criterion in the definition of Γ-convergence. The profiles of φ_ε are shown in Figure 2, where Γ_g corresponds to the point on the left hand side. The respective profile of φ_0 is constantly -1.
An idea to circumvent this possibly is to define $\boldsymbol{g}(x)$ depending on a local average of $\boldsymbol{\varphi}$ around $x \in \Gamma_g$, instead of $\boldsymbol{\varphi}(x)$ pointwise. However, further research is necessary in this direction.
Note that the Γ-convergence result also holds true if the mass constraint is dropped [MM77].

By the same arguments and under appropriate assumptions, Γ-convergence could also be shown in the multiphase case. The corresponding Γ-convergence of a vector valued Ginzburg-Landau energy with a potential defined on $\{\boldsymbol{\varphi} \geq 0\}$ can be found in [Bal90]. However, this would be out of the scope of this work. We note that the Γ-limit of the multiphase Ginzburg-Landau energy is [Bal90, Gar00]

$$\begin{cases} \sum_{\substack{k,l=1 \\ k<l}}^{N} \sigma_{kl} \mathcal{H}^{d-1}(\partial^*\{\boldsymbol{\varphi} = \boldsymbol{e}_k\} \cap \partial^*\{\boldsymbol{\varphi} = \boldsymbol{e}_l\}) & \boldsymbol{\varphi} \in BV(\Omega)^N,\ \fint_\Omega \boldsymbol{\varphi} = \mathfrak{m},\ \boldsymbol{\varphi} \in \{\boldsymbol{e}_i\}_{i=1}^N \text{ a.e.} \\ \infty & \text{else} \end{cases}$$

with the surface tensions

$$\sigma_{kl} := d(\boldsymbol{e}_k, \boldsymbol{e}_l) := \inf\Big\{\sqrt{2}\int_{-1}^{1}\sqrt{\psi_0(\boldsymbol{\gamma}(t))}|\boldsymbol{\gamma}'(t)|\,\mathrm{d}t \mid \boldsymbol{\gamma}\in C^{0,1}([-1,1])^N,\ \boldsymbol{\gamma}(-1)=\boldsymbol{e}_k,\ \boldsymbol{\gamma}(1)=\boldsymbol{e}_l,\ \sum_{i=1}^{N}\gamma_i=1,\ \boldsymbol{\gamma}\geq 0\Big\}. \tag{132}$$

Here, $\mathcal{H}^{d-1}$ denotes the $(d-1)$ dimensional Hausdorff measure and $\partial^*\{\boldsymbol{\varphi}=\boldsymbol{e}_k\}$ denotes the reduced boundary [EG91] of the set $\{x\in\Omega \mid \boldsymbol{\varphi}(x)=\boldsymbol{e}_k\}$, which is defined in a measure theoretic way. Note that there is a factor of $\sqrt{2}$ appearing in the definition of the surface tension, since we have a factor of $\frac{1}{2}$ in front of the gradient term in the Ginzburg-Landau energy as opposed to [Gar00]. Minimizers of the above Γ-limit have the property that the interfaces intersect with $\partial\Omega$ orthogonally and that a certain angle condition holds in the triple junctions [Gar00]. A triple junction is a point in Ω where three interfaces meet. This angle condition is given by Young's law

$$\tau_{12}\sigma_{12}+\tau_{13}\sigma_{13}+\tau_{23}\sigma_{23}=0,$$

where τ_{ij} denotes the outer normal of $\partial^*\{\boldsymbol{\varphi}=\boldsymbol{e}_i\}\cap\partial^*\{\boldsymbol{\varphi}=\boldsymbol{e}_j\}$, see Figure 3 for the 2D case. The respective angles θ_i then fulfill

$$\frac{\sigma_{12}}{\sin(\theta_3)}=\frac{\sigma_{13}}{\sin(\theta_2)}=\frac{\sigma_{23}}{\sin(\theta_1)},$$

where θ_1 is the angle between τ_{12} and τ_{13}, θ_2 is the angle between τ_{12} and τ_{23} and θ_3 is the angle between τ_{13} and τ_{23}.

A sharp interface analysis in the sense of formal asymptotics is performed in [BFGS14] for the mean compliance functional and the tracking type functional in the case that $\boldsymbol{f}$ and $\boldsymbol{g}$ are independent of $\boldsymbol{\varphi}$, whereas $\boldsymbol{C}(\boldsymbol{\varphi})$ depends on ε. The angle condition above is deduced. We note that the angle condition doesn't necessarily prescribe the optimal shape on a macroscopic scale, since the condition holds only in the triple point and the angles can be different away from the triple point. In the numerical experiments we observed that it is possible that the angle condition is satisfied in a small neighborhood of the triple junction, whereas in a larger neighborhood other angles can arise. For instance see Figure 34, where the interface does not seem to intersect $\partial\Omega$ perpendicularly. However, when zooming into the intersection, a 90° angle can be observed. Also in Figure 36a the two triple junctions within the beam seem to have different angles.

Examples for functionals F, which fulfill the assumptions of Theorem 6.22 and thus Γ-convergence is obtained, are the mean compliance functional (112)

$$F(\varphi,\boldsymbol{u})=\int_\Omega \boldsymbol{f}\cdot\boldsymbol{u}+\int_{\Gamma_g}\boldsymbol{g}\cdot\boldsymbol{u},$$

with $\boldsymbol{f}$ and $\boldsymbol{g}$ independent of φ or the compliant mechanism functional (114)

$$F(\varphi,\boldsymbol{u})=\frac{1}{2}\int_\Omega c(x,\varphi)|\boldsymbol{u}-\boldsymbol{u}_\Omega|^2.$$

For the latter one exploits that $c(x,\varphi_i(x))\to c(x,\varphi(x))$ weakly-* in $L^\infty(\Omega)$ (see the arguments in Lemma 6.3) and that $\boldsymbol{u}_i\to\boldsymbol{u}$ strongly in $L^2(\Omega)^d$. Another example is the

linear compliant mechanism functional (116)

$$F(\varphi, \boldsymbol{u}) = -\int_{\Gamma_{out}} \boldsymbol{g}_{out} \cdot \boldsymbol{u}.$$

An example for a functional, where the above result cannot be applied is the stress minimization with

$$F(\varphi, \boldsymbol{u}) = \int_\Omega c(\varphi)|\boldsymbol{C}(\varphi)\mathcal{E}(\boldsymbol{u})|^2,$$

since the assumed continuity is not given. Only lower semi-continuity can be shown for strong convergence of φ in L^1 and weak convergence of $\boldsymbol{u}$ in H^1_D as in Lemma 6.5.

Γ-convergence ensures that the solutions of the phase field relaxation approximate a solution of the sharp interface problem. However, this holds only for global minimizers. In Section 6.14 we give a numerical example where this approximation property is not given for local minimizers (see Figure 82).

6.5 First order optimality conditions

In this section we calculate the derivative j' by the usual adjoint approach and show the existence and uniqueness of Lagrange multipliers in dual spaces.

Theorem 6.23. *The Ginzburg-Landau energy $E(\boldsymbol{\varphi})$ is two times continuously Fréchet differentiable on $H^1(\Omega)^N \cap L^\infty(\Omega)^N$.*

Proof. Since $(\boldsymbol{v}, \boldsymbol{w}) \to \int_\Omega \nabla \boldsymbol{v} \cdot \nabla \boldsymbol{w}$ defines a bounded bilinear form on $H^1(\Omega)^N \times H^1(\Omega)^N$, we conclude that the first term

$$\frac{\varepsilon}{2}\int_\Omega |\nabla \boldsymbol{\varphi}|^2$$

in the Ginzburg-Landau energy is smooth on $H^1(\Omega)^N$. With the same arguments as in the proof of Lemma 6.10 we get from **(AP11)** that $\psi_0 \in C^2(L^\infty(\Omega)^N; L^\infty(\Omega))$, thus the second term

$$\frac{1}{\varepsilon}\int_\Omega \psi_0(\boldsymbol{\varphi})$$

is two times continuously Fréchet differentiable on $L^\infty(\Omega)^N$, where we use the embedding $L^\infty(\Omega) \to L^1(\Omega)$ and use that the integral is linear and continuous on $L^1(\Omega)$. □

Remark 6.24. It is possible to show $E \in C^2(H^1(\Omega)^N)$ by imposing stricter assumptions on ψ_0 and by assuming $d \le 3$ to get Sobolev embeddings, but we don't need this here.

Theorem 6.25. *The reduced cost functional*

$$j(\boldsymbol{\varphi}) := \gamma E(\boldsymbol{\varphi}) + F(\boldsymbol{\varphi}, S(\boldsymbol{\varphi}))$$

is two times continuously Fréchet differentiable on $H^1(\Omega)^N \cap L^\infty(\Omega)^N$. The first order

Fréchet derivative is given by

$$\begin{aligned}\left\langle j'(\boldsymbol{\varphi}), \boldsymbol{\delta\varphi}\right\rangle = \gamma\varepsilon \int_\Omega \nabla\boldsymbol{\varphi}\cdot\nabla\boldsymbol{\delta\varphi} + \frac{\gamma}{\varepsilon}\int_\Omega \psi_0'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi} \\ + \langle F_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}, S(\boldsymbol{\varphi})), \boldsymbol{\delta\varphi}\rangle + \left\langle F_{\boldsymbol{u}}(\boldsymbol{\varphi}, S(\boldsymbol{\varphi})), S'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\right\rangle_{(H^1_D)^*, H^1_D}\end{aligned}$$

for all $\boldsymbol{\varphi},\ \boldsymbol{\delta\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$.

Proof. Follows from Theorem 6.23, Theorem 6.11, **(AP8)** and the chain rule. □

We reformulate the last term in the derivative by

$$\left\langle F_{\boldsymbol{u}}(\boldsymbol{\varphi}, S(\boldsymbol{\varphi})), S'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\right\rangle_{(H^1_D)^*, H^1_D} = \left\langle S'(\boldsymbol{\varphi})^* F_{\boldsymbol{u}}(\boldsymbol{\varphi}, S(\boldsymbol{\varphi})), \boldsymbol{\delta\varphi}\right\rangle$$

and calculate $S'(\boldsymbol{\varphi})^* F_{\boldsymbol{u}}(\boldsymbol{\varphi}, S(\boldsymbol{\varphi}))$ by introducing an adjoint state $\boldsymbol{p}$.

Definition 6.26. For given $\boldsymbol{\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$ and $\boldsymbol{u} \in H^1_D$ we define the adjoint state $\boldsymbol{p} \in H^1_D$ as the solution of the equation

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{p}) : \mathcal{E}(\boldsymbol{\xi}) = \langle F_{\boldsymbol{u}}(\boldsymbol{\varphi}, \boldsymbol{u}), \boldsymbol{\xi}\rangle_{(H^1_D)^*, H^1_D} \quad \text{for all } \boldsymbol{\xi} \in H^1_D. \tag{133}$$

Remark 6.27. In case that it holds

$$F(\boldsymbol{\varphi}, \boldsymbol{u}) = \int_\Omega \alpha(x, \boldsymbol{\varphi}(x), \boldsymbol{u}(x))\,\mathrm{d}x + \int_{\partial\Omega} \beta(x, \boldsymbol{\varphi}(x), \boldsymbol{u}(x))\,\mathrm{d}x$$

for some functions α and β, then equation (133) is the weak formulation of the following adjoint PDE:

$$\begin{aligned} -\nabla\cdot(\boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{p})) &= \nabla_u \alpha(\boldsymbol{\varphi}, \boldsymbol{u}) \quad \text{in } \Omega \\ \boldsymbol{p} &= \boldsymbol{0} \quad \text{on } \Gamma_D \\ \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{p})\cdot\boldsymbol{n} &= \nabla_u \beta(\boldsymbol{\varphi}, \boldsymbol{u}) \quad \text{on } \partial\Omega \setminus \Gamma_D. \end{aligned}$$

For the mean compliance problem (112) it holds $\boldsymbol{p} = S(\boldsymbol{\varphi})$, since the corresponding adjoint equation is independent of $\boldsymbol{u}$ and coincides with the state equation.

Theorem 6.28. *For any* $\boldsymbol{\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$ *and* $\boldsymbol{u} \in H^1_D$*, the adjoint equation* (133) *has a unique solution* $\boldsymbol{p} \in H^1_D$*. It holds the a priori estimate*

$$\|\boldsymbol{p}\|_{H^1} \le c\|F_{\boldsymbol{u}}(\boldsymbol{\varphi}, \boldsymbol{u})\|_{(H^1_D)^*}. \tag{134}$$

Proof. This can be proved by the Lax-Milgram theorem as in Theorem 6.6. The right hand side $F_{\boldsymbol{u}}(\boldsymbol{\varphi}, \boldsymbol{u})$ is in $(H^1_D)^*$ by definition. □

Lemma 6.29. *For any* $\boldsymbol{\varphi}, \boldsymbol{\delta\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$ *it holds*

$$\begin{aligned}\left\langle F_{\boldsymbol{u}}(\boldsymbol{\varphi}, \boldsymbol{u}), S'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\right\rangle_{(H^1_D)^*, H^1_D} = -\int_\Omega (\nabla\boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{p}))\cdot\boldsymbol{\delta\varphi} + \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})^T\boldsymbol{p}\cdot\boldsymbol{\delta\varphi} \\ + \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})^T\boldsymbol{p}\cdot\boldsymbol{\delta\varphi},\end{aligned}$$

where $\boldsymbol{u} = S(\boldsymbol{\varphi})$ *and* $\boldsymbol{p}$ *is the adjoint state corresponding to* $\boldsymbol{\varphi}$ *and* $\boldsymbol{u}$*. The application of* $\nabla\boldsymbol{C}(\boldsymbol{\varphi})$ *has to be understood componentwise, i.e.*

$$(\nabla\boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{p}))_i = \partial_i\boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{p}) \quad i = 1, \dots, N.$$

Proof. We test the adjoint equation by $\boldsymbol{\xi} = S'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi} \in H^1_D$ and the linearized state equation (124) by $\boldsymbol{\xi} = \boldsymbol{p} \in H^1_D$ and get

$$\begin{aligned}\left\langle F_{\boldsymbol{u}}(\boldsymbol{\varphi},\boldsymbol{u}), S'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\right\rangle_{(H^1_D)^*,H^1_D} &= \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{p}) : \mathcal{E}(S'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}) \\ &= -\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{p}) + \int_\Omega \boldsymbol{f}_\varphi(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\cdot\boldsymbol{p} \\ &\quad + \int_{\Gamma_g} \boldsymbol{g}_\varphi(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\cdot\boldsymbol{p} \\ &= -\int_\Omega (\nabla\boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{p}))\cdot\boldsymbol{\delta\varphi} + \int_\Omega \boldsymbol{f}_\varphi(\boldsymbol{\varphi})^T\boldsymbol{p}\cdot\boldsymbol{\delta\varphi} \\ &\quad + \int_{\Gamma_g} \boldsymbol{g}_\varphi(\boldsymbol{\varphi})^T\boldsymbol{p}\cdot\boldsymbol{\delta\varphi},\end{aligned}$$

noting that $(\boldsymbol{u},\boldsymbol{v}) \mapsto \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{v})$ is symmetric. □

By means of Lemma 6.29, we can write down a nice expression for the derivative of the reduced cost functional.

Proposition 6.30. *It holds*

$$\begin{aligned}\left\langle j'(\boldsymbol{\varphi}), \boldsymbol{\delta\varphi}\right\rangle &= \gamma\varepsilon\int_\Omega \nabla\boldsymbol{\varphi} : \nabla\boldsymbol{\delta\varphi} + \frac{\gamma}{\varepsilon}\int_\Omega \psi_0'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi} \\ &\quad + \langle F_{\boldsymbol{\varphi}}(\boldsymbol{\varphi},\boldsymbol{u}), \boldsymbol{\delta\varphi}\rangle - \int_\Omega (\nabla\boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{p}))\cdot\boldsymbol{\delta\varphi} + \int_\Omega \boldsymbol{f}_\varphi(\boldsymbol{\varphi})^T\boldsymbol{p}\cdot\boldsymbol{\delta\varphi} \\ &\quad + \int_{\Gamma_g} \boldsymbol{g}_\varphi(\boldsymbol{\varphi})^T\boldsymbol{p}\cdot\boldsymbol{\delta\varphi}\end{aligned}$$

for all $\boldsymbol{\varphi},\ \boldsymbol{\delta\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$*, where* $\boldsymbol{u}$ *is state corresponding to the control* $\boldsymbol{\varphi}$ *and* $\boldsymbol{p}$ *is the adjoint state corresponding to* $\boldsymbol{\varphi}$ *and* $\boldsymbol{u}$*.* □

From a computational point of view, the adjoint representation of $j'(\boldsymbol{\varphi})$ in Proposition 6.30 is advantageous compared to the representation in Theorem 6.25 using sensitivities, since for the latter, the linearized state equation has to be solved for each $\boldsymbol{\delta\varphi}$ to compute $S'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}$. For the adjoint approach, the adjoint state equation has to be solved once to compute $\boldsymbol{p}$ and therewith $\langle j'(\boldsymbol{\varphi}), \boldsymbol{\delta\varphi}\rangle$ can be computed for different $\boldsymbol{\delta\varphi}$ by just computing integrals (assuming that $\langle F_{\boldsymbol{\varphi}}(\boldsymbol{\varphi},\boldsymbol{u}), \boldsymbol{\delta\varphi}\rangle$ can be computed by evaluating an integral) and no additional PDE has to be solved.

Remark 6.31. If $\boldsymbol{\varphi} \in H^2$, $\partial_\nu\boldsymbol{\varphi} = 0$, $F_{\boldsymbol{\varphi}}(\boldsymbol{\varphi},\boldsymbol{u}) \in L^1(\Omega)^N \subset (H^1(\Omega)^N \cap L^\infty(\Omega)^N)^*$ and $\boldsymbol{g}_{\boldsymbol{\varphi}} = 0$, then we have the additional regularity $j'(\boldsymbol{\varphi}) \in L^1(\Omega)^N$.

Since the given constraints are convex and j is Fréchet differentiable in $H^1(\Omega)^N \cap L^\infty(\Omega)^N$ we can formulate a first order optimality criterion.

Lemma 6.32. *Let* $\overline{\boldsymbol{\varphi}} \in \Phi_{ad}$ *be a local minimizer of* j *in* Φ_{ad}*. Then it holds*

$$\left\langle j'(\overline{\boldsymbol{\varphi}}), \boldsymbol{\eta} - \overline{\boldsymbol{\varphi}}\right\rangle \geq 0 \quad \forall \boldsymbol{\eta} \in \Phi_{ad}.$$

Proof. See [Trö09, Lemma 2.21]. □

Next we want to prove the existence of Lagrange multipliers for the pointwise bound and sum constraints as well as for the nonlocal integral constraint. This is only possible under the assumption $\mathfrak{m} > 0$, i.e. $\mathfrak{m}_i > 0$ for all i. In the case that $\mathfrak{m}_i = 0$ for some i, all feasible controls satisfy $\varphi_i = 0$. Thus the problem can be reduced to another problem with

less phases until it holds $\mathfrak{m} > 0$ and we can assume $\mathfrak{m} > 0$ without loss of generality.
Another difficulty is that the constraints are not independent. Let $\boldsymbol{\varphi}$ fulfill $\sum_{i=1}^{N} \varphi_i = 1$ and $\int_\Omega \varphi_i = \mathfrak{m}_i$ for $i = 1, \dots, N-1$. Then it automatically holds $\int_\Omega \varphi_N = \int_\Omega (1 - \sum_{i=1}^{N-1} \varphi_i) = 1 - \sum_{i=1}^{N-1} \mathfrak{m}_i = \mathfrak{m}_N$. Thus the constraint $\int_\Omega \varphi_N = \mathfrak{m}_N$ is redundant and therefore we drop it during the investigation of Lagrange multipliers. Another possibility is to keep the redundant constraint and introduce another constraint on the Lagrange multipliers, namely $\sum_{i=1}^{N} \lambda_i = 0$ where λ_i is the Lagrange multiplier for the mass constraint $\int_\Omega \varphi_i = \mathfrak{m}_i$. This approach is used e.g. in [BGSS13a]. We choose the former approach, i.e. we drop the redundant constraint, since this gives a slightly sparser linear system when we solve the projection type subproblem by a PDAS method, see Section 6.10.4.

We note that the following results about existence and uniqueness of Lagrange multipliers are valid for any cost functional j. We do not use here that j is the reduced cost functional in the topology optimization problem (102). The only assumption that is needed is that $j : H^1(\Omega)^N \cap L^\infty(\Omega)^N \to \mathbb{R}$ is continuously Fréchet differentiable. A consequence is that we also get existence and uniqueness of Lagrange multipliers for the corresponding projection type subproblem (18).

Theorem 6.33. *Let $j : H^1(\Omega)^N \cap L^\infty(\Omega)^N \to \mathbb{R}$ be an arbitrary cost functional which is continuously Fréchet differentiable and assume $\mathfrak{m} > 0$. Let $\overline{\boldsymbol{\varphi}} \in \Phi_{ad}$ be a local minimizer of j in Φ_{ad}. Then there exist Lagrange multipliers $\boldsymbol{\lambda} \in \mathbb{R}^{N-1}$, $\Lambda \in (H^1(\Omega) \cap L^\infty(\Omega))^*$ and $\boldsymbol{\mu} \in (L^\infty(\Omega)^N)^*$ such that the following KKT system is fulfilled.*

$$\begin{aligned}
\int_\Omega \overline{\varphi}_i &= \mathfrak{m}_i |\Omega| \quad i = 1, \dots, N-1, \\
\sum_{i=1}^{N} \overline{\varphi}_i &= 1, \\
\overline{\boldsymbol{\varphi}} &\geq 0,
\end{aligned}$$

$$\begin{aligned}
\langle j'(\overline{\boldsymbol{\varphi}}), \boldsymbol{\eta} \rangle - \sum_{i=1}^{N-1} \int_\Omega \eta_i \lambda_i \\
- \left\langle \Lambda, \sum_{i=1}^{N} \eta_i \right\rangle - \langle \boldsymbol{\mu}, \boldsymbol{\eta} \rangle_{(L^\infty)^*, L^\infty} = 0 \quad \forall \boldsymbol{\eta} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N, \qquad (135)
\end{aligned}$$

$$\langle \boldsymbol{\mu}, \boldsymbol{\eta} \rangle_{(L^\infty)^*, L^\infty} \geq 0 \quad \forall \boldsymbol{\eta} \in L^\infty(\Omega)^N, \boldsymbol{\eta} \geq 0, \qquad (136)$$

$$\langle \boldsymbol{\mu}, \overline{\boldsymbol{\varphi}} \rangle_{(L^\infty)^*, L^\infty} = 0. \qquad (137)$$

Proof. We show that the regularity condition of Robinson is satisfied at the solution $\overline{\boldsymbol{\varphi}}$, which is equivalent to the regularity condition of Zowe and Kurcyusz and is in finite dimension equivalent to the MFCQ constraint qualification [ZK79]. In the notation of [ZK79], the data of our problem is

$$\begin{aligned}
C &= X = H^1(\Omega)^N \cap L^\infty(\Omega)^N \\
Y &= \mathbb{R}^{N-1} \times (H^1(\Omega) \cap L^\infty(\Omega)) \times L^\infty(\Omega)^N \\
K &= \{0\} \times \{0\} \times \{\boldsymbol{\varphi} \in L^\infty(\Omega)^N \mid \boldsymbol{\varphi} \geq 0 \text{ a.e. in } \Omega\} \subset Y \\
g(\boldsymbol{\varphi}) &= \begin{pmatrix} (\int_\Omega \varphi_i - \mathfrak{m}_i |\Omega|)_{i=1}^{N-1} \\ \sum_{i=1}^{N} \varphi_i - 1 \\ \boldsymbol{\varphi} \end{pmatrix} \in Y.
\end{aligned}$$

Then our problem can be written as

$$\min j(\boldsymbol{\varphi}),\ \boldsymbol{\varphi} \in C,\ g(\boldsymbol{\varphi}) \in K.$$

The choice of Y is crucial for the regularity of the Lagrange multipliers and will be discussed in the following remark.
For the existence result in [ZK79] we need that $j : X \to \mathbb{R}$ is Fréchet differentiable, which holds by assumption, and that $g : X \to Y$ is continuously Fréchet differentiable, which can be seen easily, since g is affine linear and continuous. Obviously, $K \subset Y$ is a closed convex cone with vertex at the origin and $C \subset X$ is a non-empty closed convex subset. Y becomes a Banach space with the norm $\|(f, F, \boldsymbol{F})^T\|_Y := \max\{\|f\|_{\ell^\infty}, \|F\|_{H^1 \cap L^\infty}, \|\boldsymbol{F}\|_{L^\infty}\}$ for $(f, F, \boldsymbol{F})^T \in Y$. For the regularity condition of Robinson we have to show that

$$0 \in \operatorname{int}\{A\}, \quad A := g(\overline{\boldsymbol{\varphi}}) + g'(\overline{\boldsymbol{\varphi}})(C - \overline{\boldsymbol{\varphi}}) - K \subset Y$$

where $\operatorname{int}\{A\}$ denotes the interior of A in Y. Therefore we show that there exist a $\delta > 0$, such that for all $(f, F, \boldsymbol{F})^T \in Y$ with $\|(f, F, \boldsymbol{F})^T\|_Y \leq \delta$ it holds $(f, F, \boldsymbol{F})^T \in A$. Let $\mathfrak{m}_{min} := \min_i \mathfrak{m}_i > 0$, $\delta := \min\{\frac{\mathfrak{m}_{min}}{4(N-1)}|\Omega|, \frac{\mathfrak{m}_{min}}{4}\}$ and $(f, F, \boldsymbol{F})^T \in Y$ with $\|(f, F, \boldsymbol{F})^T\|_Y \leq \delta$. We notice that the equality

$$A = \left\{ \begin{pmatrix} (\int_\Omega \varphi_i - \mathfrak{m}_i|\Omega|)_{i=1}^{N-1} \\ \sum_{i=1}^N \varphi_i - 1 \\ \boldsymbol{\varphi} - \boldsymbol{\eta} \end{pmatrix} \middle| \ \boldsymbol{\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N,\ \boldsymbol{\eta} \in L^\infty(\Omega)^N,\ \boldsymbol{\eta} \geq 0 \right\}$$

holds. With the choices

$$\begin{aligned} \varphi_i &:= \mathfrak{m}_i + \frac{f_i}{|\Omega|} \text{ for } i = 1, \dots, N-1, \\ \varphi_N &:= F + 1 - \sum_{i=1}^{N-1} \varphi_i \text{ and} \\ \boldsymbol{\eta} &:= \boldsymbol{\varphi} - \boldsymbol{F}. \end{aligned}$$

it holds

$$\begin{pmatrix} (\int_\Omega \varphi_i - \mathfrak{m}_i|\Omega|)_{i=1}^{N-1} \\ \sum_{i=1}^N \varphi_i - 1 \\ \boldsymbol{\varphi} - \boldsymbol{\eta} \end{pmatrix} = \begin{pmatrix} f \\ F \\ \boldsymbol{F} \end{pmatrix}.$$

Obviously we have $\boldsymbol{\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N,\ \boldsymbol{\eta} \in L^\infty(\Omega)^N$ and it remains to show $\boldsymbol{\eta} \geq 0$. For $i = 1, \dots, N-1$ we get $\varphi_i \geq \mathfrak{m}_{min} - \frac{\delta}{|\Omega|} \geq \mathfrak{m}_{min} - \frac{\mathfrak{m}_{min}}{2} = \frac{\mathfrak{m}_{min}}{2}$. For $i = N$ we estimate for almost every $x \in \Omega$

$$\begin{aligned} \varphi_N(x) &= F(x) + 1 - \sum_{i=1}^{N-1} \varphi_i(x) \geq -\delta + 1 - \left(\sum_{i=1}^{N-1} \mathfrak{m}_i + \frac{f_i}{|\Omega|} \right) = -\delta + 1 - \left(1 - \mathfrak{m}_N + \sum_{i=1}^{N-1} \frac{f_i}{|\Omega|} \right) \\ &\geq -\delta + \mathfrak{m}_N - \frac{(N-1)\delta}{|\Omega|} \geq -\frac{\mathfrak{m}_{min}}{4} + \mathfrak{m}_{min} - \frac{\mathfrak{m}_{min}}{4} = \frac{\mathfrak{m}_{min}}{2}, \end{aligned}$$

where we used that $\sum_{i=1}^N \mathfrak{m}_i = 1$. Thus we get for $i = 1, \dots, N$ and almost every $x \in \Omega$

$$\boldsymbol{\eta}_i(x) = \boldsymbol{\varphi}_i(x) - \boldsymbol{F}_i(x) \geq \frac{\mathfrak{m}_{min}}{2} - \delta \geq \frac{\mathfrak{m}_{min}}{2} - \frac{\mathfrak{m}_{min}}{4} = \frac{\mathfrak{m}_{min}}{4} \geq 0.$$

We can apply [ZK79, Theorem 4.1] and get the existence of Lagrange multipliers $(\boldsymbol{\lambda}, \Lambda, \boldsymbol{\mu})^T$ fulfilling

$$(\boldsymbol{\lambda}, \Lambda, \boldsymbol{\mu})^T \in K^+ \tag{138}$$

$$\left\langle (\boldsymbol{\lambda}, \Lambda, \boldsymbol{\mu})^T, g(\overline{\boldsymbol{\varphi}}) \right\rangle_{Y^*,Y} = 0 \tag{139}$$

$$j'(\overline{\boldsymbol{\varphi}}) - (\boldsymbol{\lambda}, \Lambda, \boldsymbol{\mu})^T \circ g'(\overline{\boldsymbol{\varphi}}) \in C(\overline{\boldsymbol{\varphi}})^+ \tag{140}$$

where $K^+ = \{l \in Y^* \mid \langle l, x\rangle_{Y^*,Y} \geq 0 \ \forall x \in K\}$ is the polar cone of K and $C(\overline{\boldsymbol{\varphi}})^+ = \{0\} \subset (H^1(\Omega)^N \cap L^\infty(\Omega)^N)^*$ in our case. Equation (138) is equivalent to (136), equation (139) corresponds to (137) and testing (140) by $\boldsymbol{\eta} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$ gives (135). □

Lemma 6.34. *Let $(\overline{\boldsymbol{\varphi}}, \boldsymbol{\lambda}, \Lambda, \boldsymbol{\mu}) \in \Phi_{ad} \times \mathbb{R}^{N-1} \times (H^1(\Omega) \times L^\infty(\Omega))^* \cap (L^\infty(\Omega)^N)^*$ be a solution of the KKT system* (135)-(137)*. Then $\overline{\boldsymbol{\varphi}}$ is a stationary point of j in the sense of the variational inequality* (27)*.*

Proof. Let $\boldsymbol{\eta} \in \Phi_{ad}$ be arbitrary. From (135) we get

$$\left\langle j'(\overline{\boldsymbol{\varphi}}), \boldsymbol{\eta} - \overline{\boldsymbol{\varphi}} \right\rangle - \sum_{i=1}^{N-1} \underbrace{\int_\Omega (\eta_i - \overline{\varphi}_i)}_{=|\Omega|(\mathfrak{m}_i - \mathfrak{m}_i)=0} \lambda_i - \left\langle \Lambda, \underbrace{\sum_{i=1}^{N} \eta_i - \overline{\varphi}_i}_{=1-1=0} \right\rangle - \langle \boldsymbol{\mu}, \boldsymbol{\eta} - \overline{\boldsymbol{\varphi}} \rangle_{(L^\infty)^*,L^\infty} = 0$$

From (136) and (137) we conclude $\langle \boldsymbol{\mu}, \boldsymbol{\eta} - \overline{\boldsymbol{\varphi}} \rangle_{(L^\infty)^*,L^\infty} \geq 0$ and thus

$$\left\langle j'(\overline{\boldsymbol{\varphi}}), \boldsymbol{\eta} - \overline{\boldsymbol{\varphi}} \right\rangle \geq 0.$$

□

We want to discuss the regularity of the Lagrange multipliers. Let Y be the space in the proof of Theorem 6.33. It is desirable to choose Y as large as possible. Since the Lagrange multipliers are in Y^*, more regularity is obtained if Y is chosen larger. The natural choice would be $Y = \mathbb{R}^{N-1} \times (H^1(\Omega) \cap L^\infty(\Omega)) \times (H^1(\Omega)^N \cap L^\infty(\Omega)^N)$. In this case one would obtain only $\boldsymbol{\mu} \in (H^1(\Omega)^N \cap L^\infty(\Omega)^N)^*$. But since we proved regularity in the larger space $Y = \mathbb{R}^{N-1} \times (H^1(\Omega) \cap L^\infty(\Omega)) \times L^\infty(\Omega)^N$, we obtain better regularity for $\boldsymbol{\mu}$, namely $\boldsymbol{\mu} \in (L^\infty(\Omega)^N)^*$. On the other hand, nonnegative functionals in $(H^1(\Omega)^N \cap L^\infty(\Omega)^N)^*$ are always continuous in L^∞, cf. [EG91], which follows from the estimate

$$0 \leq \langle \mu, \|\eta\|_{L^\infty} - \eta \rangle = \|\eta\|_{L^\infty} \langle \mu, 1 \rangle - \langle \mu, \eta \rangle ,$$

thus $\langle \mu, \eta \rangle \leq \|\eta\|_{L^\infty} \langle \mu, 1 \rangle$ for all $\eta \in H^1(\Omega) \cap L^\infty(\Omega)$, where 1 is the constant function.
We cannot prove higher regularity by the proof above, e.g. $\boldsymbol{\mu} \in (L^p(\Omega)^N)^*$, since the set $\{\boldsymbol{\varphi} \in L^p(\Omega)^N \mid \boldsymbol{\varphi} \geq 0 \text{ a.e. in } \Omega\}$ doesn't have an interior for $p < \infty$. This is a general difficulty for box constrained optimal control problems, see e.g. [Trö09].
Unfortunately the space for the sum constraint cannot be chosen larger than $H^1(\Omega) \cap L^\infty(\Omega)$, since for equality constraints it is necessary to have that $g_2'(\overline{\boldsymbol{\varphi}})$ is surjective in order to show regularity of the solution [ZK79], where we define

$$\begin{aligned} g_2 : H^1(\Omega)^N \cap L^\infty(\Omega)^N &\to H^1(\Omega) \cap L^\infty(\Omega) \\ \boldsymbol{\varphi} &\mapsto \sum_{i=1}^{N} \varphi_i - 1. \end{aligned}$$

It holds that $g_2'(\overline{\boldsymbol{\varphi}})\boldsymbol{\eta} = \sum_{i=1}^N \eta_i$ is not surjective into a larger space than $H^1(\Omega) \cap L^\infty(\Omega)$. Thus, one only obtains $\Lambda \in (H^1(\Omega) \cap L^\infty(\Omega))^*$. Of course there are other ways to show regularity of Lagrange multipliers, e.g. by regularizing the control constraints, see [BGSS13a]. However, one cannot expect that the Lagrange multipliers are functions in general, which can be seen by the following argumentation. Consider the mean compliance problem (112) with $\boldsymbol{f} = 0$ a.e. and $\boldsymbol{g}$ being linear in $\boldsymbol{\varphi}$ as in (108). Moreover, assume that $\overline{\boldsymbol{\varphi}} \in H^2(\Omega)^N$, $\boldsymbol{\mu} \in L^1(\Omega)^N$ and $\Lambda \in L^1(\Omega)$. The gradient equation in the KKT system then reads

$$\begin{aligned}\gamma\varepsilon \int_\Omega \nabla\overline{\boldsymbol{\varphi}} : \nabla\boldsymbol{\eta} + \frac{\gamma}{\varepsilon}\int_\Omega \psi_0'(\overline{\boldsymbol{\varphi}})\boldsymbol{\eta} - \int_\Omega (\nabla \boldsymbol{C}(\overline{\boldsymbol{\varphi}})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{u}))\cdot\boldsymbol{\eta} + 2\int_{\Gamma_g} \boldsymbol{g}(\boldsymbol{\eta})\cdot\boldsymbol{u} \\ - \int_\Omega \boldsymbol{\lambda}\cdot\boldsymbol{\eta} - \int_\Omega \Lambda\left(\sum_{i=1}^N \eta_i\right) - \int_\Omega \boldsymbol{\mu}\cdot\boldsymbol{\eta} = 0 \quad \forall \boldsymbol{\eta} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N,\end{aligned}$$

where we used $\boldsymbol{g}_{\boldsymbol{\varphi}}(\overline{\boldsymbol{\varphi}})\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\eta})$ because of linearity and we set $\lambda_N := 0$. We use integration by parts to obtain

$$\gamma\varepsilon \int_\Omega \nabla\overline{\boldsymbol{\varphi}} : \nabla\boldsymbol{\eta} = -\gamma\varepsilon \int_\Omega \Delta\overline{\boldsymbol{\varphi}}\cdot\boldsymbol{\eta} + \gamma\varepsilon \int_{\partial\Omega} \partial_\nu\overline{\boldsymbol{\varphi}}\cdot\boldsymbol{\eta},$$

where $\boldsymbol{\nu}$ denotes the outer normal of Ω. We test with functions $\boldsymbol{\eta} \in C_0^\infty(\Omega)$ to obtain that

$$-\gamma\varepsilon\Delta\overline{\boldsymbol{\varphi}} + \frac{\gamma}{\varepsilon}\nabla\psi_0(\overline{\boldsymbol{\varphi}}) - \nabla\boldsymbol{C}(\overline{\boldsymbol{\varphi}})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{u}) - \boldsymbol{\lambda} - \Lambda\boldsymbol{e} - \boldsymbol{\mu} = 0 \quad \text{a.e. in } \Omega$$

where $\boldsymbol{e} := (1,\dots,1)^T$, and thus

$$\gamma\varepsilon \int_{\partial\Omega} \partial_\nu\overline{\boldsymbol{\varphi}}\cdot\boldsymbol{\eta} + 2\int_{\Gamma_g} \boldsymbol{g}(\boldsymbol{\eta})\cdot\boldsymbol{u} = 0 \quad \forall \boldsymbol{\eta} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N.$$

Using test functions with $\boldsymbol{\eta}|_{\Gamma_g} = 0$, we conclude $\partial_\nu\overline{\boldsymbol{\varphi}}|_{\partial\Omega\setminus\Gamma_g} = 0$. In the binary case (material and void), numerical experiments for the mean compliance problem indicate that the optimal $\overline{\boldsymbol{\varphi}}$ is constant around Γ_g, since material is put in this region. Thus we can assume that $\partial_\nu\overline{\boldsymbol{\varphi}}|_{\Gamma_g} = 0$, and hence $\int_{\Gamma_g} \boldsymbol{g}(\boldsymbol{\eta})\cdot\boldsymbol{u} = 0$ for all $\boldsymbol{\eta}$. Taking $\boldsymbol{\eta} = \overline{\boldsymbol{\varphi}}$ we obtain that the compliance $\int_{\Gamma_g} \boldsymbol{g}(\overline{\boldsymbol{\varphi}})\cdot\boldsymbol{u}$ vanishes in the minimum, which is very improbable. In Section 6.13.10 we present a numerical experiment, where the Lagrange multiplier $\boldsymbol{\mu}$ also includes boundary measures and thus is not a function.

Because of the previous consideration one cannot assume that the Lagrange multipliers are functions in general. To be able to write down a strong formulation of the KKT system we therefore assume that the Lagrange multipliers also include measures, which are concentrated on the boundary of Ω, i.e. we assume that there exist $\Lambda^d \in L^1(\Omega)$, $\Lambda^b \in L^1(\partial\Omega)$, $\boldsymbol{\mu}^d \in L^1(\Omega)^N$ and $\boldsymbol{\mu}^b \in L^1(\partial\Omega)^N$, such that

$$\begin{aligned}\langle \Lambda, \eta\rangle &= \int_\Omega \Lambda^d\eta + \int_{\partial\Omega} \Lambda^b\eta, \\ \langle \boldsymbol{\mu}, \boldsymbol{\eta}\rangle_{(L^\infty)^*, L^\infty} &= \int_\Omega \boldsymbol{\mu}^d\cdot\boldsymbol{\eta} + \int_{\partial\Omega} \boldsymbol{\mu}^b\cdot\boldsymbol{\eta}\end{aligned}$$

for all $\boldsymbol{\eta}$. If we assume that F is given by

$$F(\boldsymbol{\varphi}, \boldsymbol{u}) = \int_\Omega \alpha(x, \boldsymbol{\varphi}(x), \boldsymbol{u}(x))\,\mathrm{d}x + \int_{\partial\Omega} \beta(x, \boldsymbol{\varphi}(x), \boldsymbol{u}(x))\,\mathrm{d}x,$$

for some functions α and β, we can write down the KKT system in a strong formulation:

$$\begin{aligned}
-\gamma\varepsilon\Delta\overline{\boldsymbol{\varphi}} + \frac{\gamma}{\varepsilon}\nabla\psi_0(\overline{\boldsymbol{\varphi}}) + \nabla_\varphi\alpha(\overline{\boldsymbol{\varphi}},\boldsymbol{u}) - \nabla\boldsymbol{C}(\overline{\boldsymbol{\varphi}})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{p}) & \\
+\nabla\boldsymbol{f}(\overline{\boldsymbol{\varphi}})\boldsymbol{p} - \boldsymbol{\lambda} - \Lambda^d\boldsymbol{e} - \boldsymbol{\mu}^d = 0 \quad & \text{in } \Omega \\
\gamma\varepsilon\partial_\nu\overline{\boldsymbol{\varphi}} + \nabla_\varphi\beta(\boldsymbol{\varphi},\boldsymbol{u}) + \chi_{\Gamma_g}\nabla_\varphi\boldsymbol{g}(\overline{\boldsymbol{\varphi}})\boldsymbol{p} - \Lambda^b\boldsymbol{e} - \boldsymbol{\mu}^b = 0 \quad & \text{on } \partial\Omega \\
\boldsymbol{\mu}^d \geq 0 \quad & \text{in } \Omega \\
\boldsymbol{\mu}^b \geq 0 \quad & \text{on } \partial\Omega \\
\boldsymbol{\mu}^d \cdot \overline{\boldsymbol{\varphi}} = 0 \quad & \text{in } \Omega \\
\boldsymbol{\mu}^b \cdot \overline{\boldsymbol{\varphi}} = 0 \quad & \text{on } \partial\Omega \\
\int_\Omega \overline{\varphi}_i = \mathfrak{m}_i|\Omega| \quad & i = 1,\dots,N-1, \\
\sum_{i=1}^N \overline{\varphi}_i = 1 \quad & \text{in } \Omega, \\
\overline{\boldsymbol{\varphi}} \geq 0 \quad & \text{in } \Omega,
\end{aligned} \tag{141}$$

together with the state equation

$$\begin{aligned}
-\nabla\cdot\boldsymbol{C}(\overline{\boldsymbol{\varphi}})\mathcal{E}(\boldsymbol{u}) = \boldsymbol{f}(\overline{\boldsymbol{\varphi}}) \quad & \text{in } \Omega \\
\boldsymbol{u} = 0 \quad & \text{on } \Gamma_D \\
\boldsymbol{C}(\overline{\boldsymbol{\varphi}})\mathcal{E}(\boldsymbol{u})\cdot\boldsymbol{n} = \chi_{\Gamma_g}\boldsymbol{g}(\overline{\boldsymbol{\varphi}}) \quad & \text{on } \partial\Omega\setminus\Gamma_D
\end{aligned}$$

and the adjoint equation

$$\begin{aligned}
-\nabla\cdot\boldsymbol{C}(\overline{\boldsymbol{\varphi}})\mathcal{E}(\boldsymbol{p}) = \nabla_u\alpha(\overline{\boldsymbol{\varphi}},\boldsymbol{u}) \quad & \text{in } \Omega \\
\boldsymbol{p} = \boldsymbol{0} \quad & \text{on } \Gamma_D \\
\boldsymbol{C}(\overline{\boldsymbol{\varphi}})\mathcal{E}(\boldsymbol{p})\cdot\boldsymbol{n} = \nabla_u\beta(\overline{\boldsymbol{\varphi}},\boldsymbol{u}) \quad & \text{on } \partial\Omega\setminus\Gamma_D.
\end{aligned}$$

Note that this system is derived only formally. However, it should give a better understanding of the multipliers.
We note that it is reasonable to assume that the Lagrange multipliers contain measures concentrated on $\partial\Omega$, if the corresponding cost functional F contains integrals over $\partial\Omega$. In the case that F contains integrals over a lower dimensional object D which is not part of $\partial\Omega$, one thus should assume that the Lagrange multipliers contain measures concentrated on D. An example is the linear compliant mechanism functional (116) used by Sigmund, in case that the output port Γ_{out} is a lower dimensional manifold in the interior of Ω.

For the vector valued Allen-Cahn variational inequality with mass constraints one gets a system similar to (135)-(137). However, in this case $H^2(\Omega)$-regularity of the solution is given together with $\partial_\nu\overline{\boldsymbol{\varphi}} = 0$, and the term corresponding to $j'(\overline{\boldsymbol{\varphi}})$ is an $L^2(\Omega)^N$-function. Thus one is able to prove that the Lagrange multipliers are in L^2, see [BGSS13a].

It is challenging to show uniqueness of Lagrange multipliers if they are no functions but only functionals in some dual space. We will show uniqueness in Theorem 6.37. To point out the difficulties that arise in case of low regularity of Lagrange multipliers, we give the proof first for the case that the Lagrange multipliers are in $L^1(\Omega)$, which is much simpler, since pointwise arguments can be used. The arguments in the following theorem are the

same as for the vector valued Allen-Cahn variational inequality with mass constraint, see [BGSS13a, Theorem 2.4], where all functions are even in $L^2(\Omega)$.
In the following we identify $L^1(\Omega)^N$ functions by functionals in $(H^1(\Omega)^N \cap L^\infty(\Omega)^N)^*$ in the standard way, i.e. $\langle \boldsymbol{f}, \boldsymbol{\eta} \rangle := \int_\Omega \boldsymbol{f} \cdot \boldsymbol{\eta}$. Note that this embedding $L^1(\Omega)^N \hookrightarrow (H^1(\Omega)^N \cap L^\infty(\Omega)^N)^*$ is injective, since we have $C_0^\infty(\Omega) \subset H^1(\Omega) \cap L^\infty(\Omega)$.

Theorem 6.35. *Assume $\mathfrak{m} > 0$. Let $\overline{\boldsymbol{\varphi}} \in \Phi_{ad}$ be a local minimizer of j in Φ_{ad}. Assume for the Lagrange multipliers the regularity $\boldsymbol{\lambda} \in \mathbb{R}^{N-1}$, $\Lambda \in L^1(\Omega)$ and $\boldsymbol{\mu} \in L^1(\Omega)^N$, as well as $F_\varphi(\overline{\boldsymbol{\varphi}}, S(\overline{\boldsymbol{\varphi}})) \in L^1(\Omega)$. Then the multipliers are unique.*

Proof. We only give a brief sketch of the proof. For details we refer to [BGSS13a]. From (135) we get that $j'(\overline{\boldsymbol{\varphi}})$ is in $L^1(\Omega)^N$, i.e. we find $\boldsymbol{F} \in L^1(\Omega)^N$ such that

$$\left\langle j'(\overline{\boldsymbol{\varphi}}), \boldsymbol{\eta} \right\rangle = \int_\Omega \boldsymbol{F} \cdot \boldsymbol{\eta} \quad \forall \boldsymbol{\eta} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N.$$

Thus (135) holds pointwise almost everywhere, i.e.

$$F_i - \lambda_i - \Lambda - \mu_i = 0 \quad \forall i = 1, \dots, N, \tag{142}$$

where we set $\lambda_N := 0$. From (136) we get $\boldsymbol{\mu} \geq 0$ a.e. in Ω [EG91]. Now we pick an arbitrary representative of the equivalence class of $\overline{\boldsymbol{\varphi}}$ and define the inactive sets $I_i := \{x \in \Omega \mid \overline{\boldsymbol{\varphi}}_i(x) > 0\}$ as well as $I_{ij} := I_i \cap I_j$. From (137) we get that $\mu_i|_{I_i} = 0$ a.e. for all i. Subtracting the equations (142) for i and j, and integrating over I_{ij} we get

$$\lambda_i - \lambda_j = \frac{\int_{I_{ij}} F_i - F_j \, \mathrm{d}x}{|I_{ij}|} \quad \forall i, j = 1, \dots, N \text{ with } |I_{ij}| > 0. \tag{143}$$

To show the uniqueness of $\boldsymbol{\lambda}$, we define the graph $\mathcal{G}$ consisting of the nodes $\{1, \dots, N\}$ with an edge between i and j if and only if $|I_{ij}| > 0$. We prove that $\mathcal{G}$ is connected, which together with $\lambda_N = 0$ and (143) shows the uniqueness of $\boldsymbol{\lambda}$. Assume that $\mathcal{G}$ is not connected. Then there exists a partition $\mathcal{M} \sqcup \mathcal{L} = \{1, \dots, N\}$ with $\mathcal{M} \neq \varnothing$ and $\mathcal{L} \neq \varnothing$, where $\mathcal{M}$ and $\mathcal{L}$ are not connected. We define the functions $v = \sum_{i \in \mathcal{M}} \overline{\varphi}_i$ and $w = \sum_{i \in \mathcal{L}} \overline{\varphi}_i$. From the constraints on $\overline{\boldsymbol{\varphi}}$ we get $v, w \in H^1(\Omega)$, $v \geq 0$, $w \geq 0$ and $v + w = 1$ a.e. in Ω. Since $\mathcal{M}$ and $\mathcal{L}$ are not connected we get that the set $\{x \mid v(x) > 0,\ w(x) > 0\}$ has measure zero. Thus it holds $v \in \{0, 1\}$ a.e. in Ω. On the other hand v cannot be identical 1 or identical 0, since we have $\fint_\Omega v > 0$ and $\fint_\Omega w > 0$ due to the assumption $\mathfrak{m} > 0$. This is a contradiction since H^1 functions with finitely many values have to be constant. Thus, $\mathcal{G}$ is connected and $\boldsymbol{\lambda}$ is unique.
Restricting (142) to I_i we get

$$\Lambda|_{I_i} = F_i|_{I_i} - \lambda_i \text{ a.e.}, \tag{144}$$

and since $\Omega = \bigcup_{i=1}^N I_i$ up to a set with measure zero (which follows from the sum constraint), we conclude uniqueness of Λ. From (142) we finally see that $\boldsymbol{\mu}$ is unique. □

The reason why we need the regularity of the Lagrange multipliers to show uniqueness is that we want to test (135) by $\boldsymbol{\eta} = \chi_{I_{ij}}(\boldsymbol{e}_i - \boldsymbol{e}_j)$ to obtain (143). But this is in general not possible since $\boldsymbol{\eta}$ has to be in $H^1(\Omega)^N \cap L^\infty(\Omega)^N$, which is not satisfied by a characteristic function. However, it would be possible to use more regular test functions with support in I_{ij}, but it is not obvious that such functions exist. Since I_{ij} could have empty interior, one cannot use smooth cutoff functions with support in I_{ij}. Although I_i is not an open set,

because $\overline{\boldsymbol{\varphi}}$ is not continuous, it is as a level set of an $H^1(\Omega)$-function *quasi-open*, which means that it is open up to an open set of arbitrarily small capacity, see [KM92].
Another difficulty is that we have to test (135) by $\boldsymbol{\eta} = \chi_{I_i}\boldsymbol{e}_i$ in order to get the equation (144) for Λ. This is again not possible due to the lack of regularity of characteristic functions. However, we are able to solve all these problems, which can be seen in the proof of Theorem 6.37. First we need some lemmas.

We can prove a result, which is similar to the fact that $\mu_i|_{I_i} = 0$ a.e. in the case that $\boldsymbol{\mu}$ is a function. Roughly speaking the following lemma states that $\langle \mu_i, \eta\rangle = 0$ for each η with support in I_i, which vanishes at ∂I_i.

Lemma 6.36. *Let $\overline{\boldsymbol{\varphi}} \in \Phi_{ad}$ be a local minimizer of j in Φ_{ad} and $\boldsymbol{\mu} \in (L^\infty(\Omega)^N)^*$ an associated Lagrange multiplier. Then it holds for any $\eta \in L^\infty(\Omega)$ that*

$$\langle \boldsymbol{\mu}, \eta\overline{\varphi}_i\boldsymbol{e}_i\rangle = 0 \quad \forall i = 1, \dots, N.$$

Proof. We first show $\langle \boldsymbol{\mu}, \overline{\varphi}_i\boldsymbol{e}_i\rangle = 0$ for all i. From (137) we have

$$0 = \langle \boldsymbol{\mu}, \overline{\boldsymbol{\varphi}}\rangle = \sum_{i=1}^{N} \langle \boldsymbol{\mu}, \overline{\varphi}_i\boldsymbol{e}_i\rangle$$

and (136) gives $\langle \boldsymbol{\mu}, \overline{\varphi}_i\boldsymbol{e}_i\rangle \geq 0$, since $\overline{\varphi}_i\boldsymbol{e}_i \geq 0$. This shows $\langle \boldsymbol{\mu}, \overline{\varphi}_i\boldsymbol{e}_i\rangle = 0$ for all i.
Without loss of generality we assume $\|\eta\|_{L^\infty} > 0$. Let

$$f := \overline{\varphi}_i\left(1 - \frac{\eta}{\|\eta\|_{L^\infty}}\right).$$

From $\overline{\varphi}_i \geq 0$ we also get $f \geq 0$ a.e. in Ω. Now consider

$$\left\langle \boldsymbol{\mu}, \frac{\eta}{\|\eta\|_{L^\infty}}\overline{\varphi}_i\boldsymbol{e}_i\right\rangle = \langle \boldsymbol{\mu}, (\overline{\varphi}_i - f)\boldsymbol{e}_i\rangle = -\langle \boldsymbol{\mu}, f\boldsymbol{e}_i\rangle \leq 0,$$

where we used $\langle \boldsymbol{\mu}, \overline{\varphi}_i\boldsymbol{e}_i\rangle = 0$ and (136). The same estimate holds if η is replaced by $-\eta$, thus

$$\left\langle \boldsymbol{\mu}, \frac{\eta}{\|\eta\|_{L^\infty}}\overline{\varphi}_i\boldsymbol{e}_i\right\rangle \geq 0,$$

which shows the statement. □

Note that in general it does not hold that μ_i vanishes on the inactive set in the sense that $\langle \boldsymbol{\mu}, \chi_{I_i}\boldsymbol{e}_i\rangle = 0$, which can only be shown if $\boldsymbol{\mu}$ is a function.

The key lemma which we will use is that $H^1(\Omega) \cap L^\infty(\Omega)$ is an algebra, which means that products of valid test functions for (135) are again valid test functions, see Theorem 7.5.

Now we are able to prove uniqueness in the general case that the Lagrange multipliers are no functions.

Theorem 6.37. *Assume $\mathfrak{m} > 0$. Let $\overline{\boldsymbol{\varphi}} \in \Phi_{ad}$ be a local minimizer of j in Φ_{ad}. Then the Lagrange multipliers $\boldsymbol{\lambda}$ and Λ are unique and $\boldsymbol{\mu}|_{H^1\cap L^\infty}$ is unique.*

Proof. Let $i, j \in \{1, \dots, N\}$. We test equation (135) by $\boldsymbol{\eta} = \overline{\varphi}_i \overline{\varphi}_j (\boldsymbol{e}_i - \boldsymbol{e}_j)$, which is in $H^1(\Omega)^N \cap L^\infty(\Omega)^N$ by Theorem 7.5. This leads to

$$\left\langle j'(\overline{\boldsymbol{\varphi}}), \overline{\varphi}_i \overline{\varphi}_j (\boldsymbol{e}_i - \boldsymbol{e}_j) \right\rangle - \int_\Omega \overline{\varphi}_i \overline{\varphi}_j (\lambda_i - \lambda_j) - \left\langle \boldsymbol{\mu}, \overline{\varphi}_i \overline{\varphi}_j (\boldsymbol{e}_i - \boldsymbol{e}_j) \right\rangle_{(L^\infty)^*, L^\infty} = 0$$

By Lemma 6.36 we get that $\left\langle \boldsymbol{\mu}, \overline{\varphi}_i \overline{\varphi}_j (\boldsymbol{e}_i - \boldsymbol{e}_j) \right\rangle_{(L^\infty)^*, L^\infty} = 0$. As in the proof of Theorem 6.35, we introduce the inactive sets $I_i := \{x \in \Omega \mid \overline{\varphi}_i(x) > 0\}$ (for some representative $\overline{\varphi}_i$ of the equivalence class) as well as $I_{ij} := I_i \cap I_j$. One easily checks that $\int_\Omega \overline{\varphi}_i \overline{\varphi}_j > 0$ if and only if $|I_{ij}| > 0$. Thus we get

$$\lambda_i - \lambda_j = \frac{\left\langle j'(\overline{\boldsymbol{\varphi}}), \overline{\varphi}_i \overline{\varphi}_j (\boldsymbol{e}_i - \boldsymbol{e}_j) \right\rangle}{\int_\Omega \overline{\varphi}_i \overline{\varphi}_j} \quad \forall i, j = 1, \dots, N \text{ with } |I_{ij}| > 0, \tag{145}$$

which is analog to (143). As in the proof of Theorem 6.35 we conclude that λ_i, $i = 1, \dots, N-1$ is unique. To prove the uniqueness of Λ, we exploit that $\overline{\boldsymbol{\varphi}}$ defines a partition of unity. Let $\xi \in H^1(\Omega) \cap L^\infty(\Omega)$ be arbitrary. We observe that

$$\langle \Lambda, \xi \rangle = \left\langle \Lambda, \xi \sum_{i=1}^N \overline{\varphi}_i \right\rangle = \left\langle \Lambda, \sum_{i=1}^N \xi \overline{\varphi}_i \right\rangle.$$

Thus we have to test equation (135) by $\boldsymbol{\eta} = \xi \overline{\boldsymbol{\varphi}} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$, where we again use Theorem 7.5. We get

$$\langle \Lambda, \xi \rangle = \left\langle j'(\overline{\boldsymbol{\varphi}}), \xi \overline{\boldsymbol{\varphi}} \right\rangle - \sum_{i=1}^{N-1} \int_\Omega \xi \overline{\varphi}_i \lambda_i - \langle \boldsymbol{\mu}, \xi \overline{\boldsymbol{\varphi}} \rangle_{(L^\infty)^*, L^\infty}$$

Using Lemma 6.36 we see that

$$\langle \boldsymbol{\mu}, \xi \overline{\boldsymbol{\varphi}} \rangle_{(L^\infty)^*, L^\infty} = \left\langle \boldsymbol{\mu}, \sum_{i=1}^N \xi \overline{\varphi}_i \boldsymbol{e}_i \right\rangle_{(L^\infty)^*, L^\infty} = 0.$$

Thus Λ is uniquely defined by the equation

$$\langle \Lambda, \xi \rangle = \left\langle j'(\overline{\boldsymbol{\varphi}}), \xi \overline{\boldsymbol{\varphi}} \right\rangle - \sum_{i=1}^{N-1} \int_\Omega \xi \overline{\varphi}_i \lambda_i, \tag{146}$$

which is analog to (144) by formally choosing ξ as a Dirac measure and noting that formally it holds $\Lambda = \sum_{i=1}^N \overline{\varphi}_i \Lambda|_{I_i}$. The remaining Lagrange multiplier $\boldsymbol{\mu}|_{H^1 \cap L^\infty}$ is then uniquely defined by equation (135). □

For certain optimal control problems the regularity of the Lagrange multipliers can be deduced from the regularity of $j'(\overline{\boldsymbol{\varphi}})$, see [Trö09]. This can also be done in this case. Assume $j'(\overline{\boldsymbol{\varphi}}) \in L^1(\Omega)^N$, see Remark 6.31 for sufficient conditions therefor. Then, by (146), we get $\Lambda \in L^1(\Omega)$ and by (135) also $\boldsymbol{\mu}|_{H^1 \cap L^\infty} \in L^1(\Omega)^N$.

It is no restriction that $\boldsymbol{\mu}$ is only uniquely defined on the subspace $H^1(\Omega)^N \cap L^\infty(\Omega)^N$, since only those values of $\boldsymbol{\mu}$ are relevant for the KKT system. Moreover, if $\boldsymbol{\mu}$ is a function then it is unique since $C_0^\infty(\Omega) \subset H^1(\Omega) \cap L^\infty(\Omega)$.

In case that the Lagrange multipliers are functions one can integrate the equation (142)

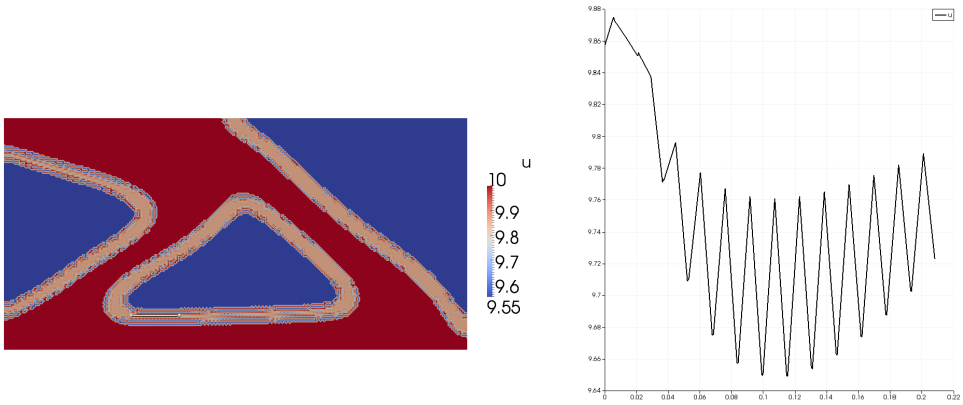


Figure 4: Discrete L^2-gradient $F_1 - F_2$ at the solution $\overline{\boldsymbol{\varphi}}$. The picture on the right shows the plot along the line on the interface shown in white on the left.

over a ball $B_\delta(x) \subset I_{ij}$ and subtract the resulting equations for i and j to obtain

$$\lambda_i - \lambda_j = \frac{\int_{B_\delta(x)} F_i - F_j \,\mathrm{d}x}{|B_\delta(x)|}.$$

By taking the limit $\delta \to 0$, one gets by the Lebesgue-Besicovitch theorem that

$$F_i - F_j \equiv \lambda_i - \lambda_j \text{ a.e. in the interior of } I_{ij}. \tag{147}$$

It turns out that equation (145), i.e.

$$\lambda_i - \lambda_j = \frac{\left\langle j'(\overline{\boldsymbol{\varphi}}), \overline{\varphi}_i \overline{\varphi}_j (\boldsymbol{e}_i - \boldsymbol{e}_j) \right\rangle}{\int_\Omega \overline{\varphi}_i \overline{\varphi}_j} \quad \forall i, j = 1, \dots, N \text{ with } |I_{ij}| > 0,$$

for computing λ is numerically more robust than equation (147). For example for a cantilever beam setting with $N = 2$ (see Example 6.83), the Lagrange multiplier for the discrete KKT system is $\lambda = \lambda_1 - \lambda_2 \approx 9.8428$. Figure 4 shows the discrete L^2-gradient $F = F_1 - F_2$, which we computed by the $L^2(\Omega)$ projection of $j'(\overline{\varphi})$ on the finite element space, i.e. $F = \sum_i \overline{g}_i \chi_i$ with $\overline{g} = M^{-1}(\langle j'(\overline{\varphi}), \chi_i \rangle)_i$, where $M = (\int_\Omega \chi_i \chi_j)_{ij}$ is the mass matrix and χ_i are the piecewise linear finite element basis functions, see Section 6.11 (we consider scalar valued phase fields here). Note that the color range in Figure 4 is rescaled to better show the values on the interface, i.e. the values in the red region are ≥ 10 and the values in the blue region are ≤ 9.55. It can be observed in the figure that F is not constant on the interface, as demanded by (147). There even are oscillations on a length scale of the mesh parameter h as seen on the right hand side of Figure 4. Thus (147) gives different values for λ depending on the point on the interface where the gradient is evaluated. We suspect that the Laplacian $\Delta\overline{\varphi}$ within $j'(\overline{\varphi})$ is responsible for these oscillations, since $\Delta\overline{\varphi}$ is not well defined for a piecewise linear finite element function $\overline{\varphi}$. Perhaps higher order finite elements would resolve this issue.
On the other hand, equation (145) gives the right value for λ up to an error of 10^{-10}. This is especially important for numerical methods like the semismooth Newton method (see Section 6.10), where a good initial guess for the Lagrange multipliers is needed.

Also equation (143) for computing $\lambda_i - \lambda_j$ by

$$\lambda_i - \lambda_j = \frac{\int_{I_{ij}} F_i - F_j \,\mathrm{d}x}{|I_{ij}|} \quad \forall i,j = 1,\dots,N \text{ with } |I_{ij}| > 0.$$

is disadvantageous from a computational point of view, since the sets I_{ij} are not given explicitly and integrals over these sets are hard to compute. For example in the discretized problem the set I_{ij} is not necessarily the union of triangles. On the other hand, in the formulation (145) only integrals over the whole domain Ω and $\partial\Omega$ appear (when ignoring F_φ), which can be computed by standard finite element methods. See also the discussion after Proposition 6.30.

The results of this section also apply to the scalar valued case. In the KKT system the Lagrange multiplier Λ drops out since the corresponding equality constraint $\varphi_1 + \varphi_2 = 1$ is eliminated.

Theorem 6.38. *Consider the problem for the scalar valued phase field* (121)*. It holds: The reduced cost functional $j : H^1(\Omega) \cap L^\infty(\Omega) \to \mathbb{R}$, $j(\varphi) = \gamma E(\varphi) + F(\varphi, S(\varphi))$ is two times continuously Fréchet differentiable. Its first order derivative is given by*

$$\begin{aligned}\big\langle j'(\varphi), \delta\varphi\big\rangle = \gamma\varepsilon \int_\Omega \nabla\varphi\cdot\nabla\delta\varphi + \frac{\gamma}{\varepsilon}\int_\Omega \psi_0'(\varphi)\delta\varphi \\ + \langle F_\varphi(\varphi,\boldsymbol{u}),\delta\varphi\rangle - \int_\Omega \boldsymbol{C}'(\varphi)\mathcal{E}(\boldsymbol{u}):\mathcal{E}(\boldsymbol{p})\delta\varphi + \int_\Omega \boldsymbol{f}_\varphi(\varphi)\cdot\boldsymbol{p}\,\delta\varphi \\ + \int_{\Gamma_g} \boldsymbol{g}_\varphi(\varphi)\cdot\boldsymbol{p}\,\delta\varphi,\end{aligned} \tag{148}$$

where $\boldsymbol{u} = S(\varphi)$ and the adjoint state $\boldsymbol{p}$ is the solution of

$$\int_\Omega \boldsymbol{C}(\varphi)\mathcal{E}(\boldsymbol{p}):\mathcal{E}(\boldsymbol{\xi}) = \langle F_{\boldsymbol{u}}(\varphi,\boldsymbol{u}),\boldsymbol{\xi}\rangle_{(H^1_D)^*,H^1_D} \qquad \text{for all } \boldsymbol{\xi}\in H^1_D. \tag{149}$$

Let now $j : H^1(\Omega)^N \cap L^\infty(\Omega)^N$ be an arbitrary continuously Fréchet differentiable cost functional. Let $-1 < \mathfrak{m} < 1$ and let $\overline{\varphi}$ be a local minimizer of j in Φ_{ad}. Then there exist Lagrange multipliers $\lambda\in\mathbb{R}$, $\mu_1 \in (L^\infty(\Omega))^*$ and $\mu_2 \in (L^\infty(\Omega))^*$, such that the KKT system

$$\begin{aligned} \fint_\Omega \overline{\varphi} &= \mathfrak{m}, \\ -1 \le \overline{\varphi} &\le 1, \\ \big\langle j'(\overline{\varphi}),\eta\big\rangle - \lambda\int_\Omega \eta - \langle \mu_1-\mu_2,\eta\rangle_{(L^\infty)^*,L^\infty} &= 0 \quad \forall \eta\in H^1(\Omega)\cap L^\infty(\Omega), \\ \langle \mu_1,\eta\rangle_{(L^\infty)^*,L^\infty} &\ge 0 \quad \forall \eta\in L^\infty(\Omega),\eta\ge 0, \\ \langle \mu_2,\eta\rangle_{(L^\infty)^*,L^\infty} &\ge 0 \quad \forall \eta\in L^\infty(\Omega),\eta\ge 0, \\ \langle \mu_1, 1+\overline{\varphi}\rangle_{(L^\infty)^*,L^\infty} &= 0, \\ \langle \mu_2, 1-\overline{\varphi}\rangle_{(L^\infty)^*,L^\infty} &= 0 \end{aligned} \tag{150, 151}$$

holds. Moreover, λ, $\mu_1|_{H^1\cap L^\infty}$ and $\mu_2|_{H^1\cap L^\infty}$ are unique.

Proof. As in Section 6.1.3 we add a tilde to all functions involving the scalar valued phase field.
Denote by T the function transforming a scalar valued phase field into a vector valued phase field, i.e. $T : H^1(\Omega)\cap L^\infty(\Omega) \to H^1(\Omega)^2 \cap L^\infty(\Omega)^2$, $T(\varphi) = (\frac{1+\varphi}{2}, \frac{1-\varphi}{2})^T$. The

differentiability of $\tilde{j} = j \circ T$ follows from the differentiability of j and the chain rule. By choosing $\boldsymbol{\varphi} = T(\varphi)$ and $\boldsymbol{\delta\varphi} = T'(\varphi)\delta\varphi$ in (6.30) and applying the chain rule as in Theorem 6.12, one ends up with (148) with $\boldsymbol{u} = S(T(\varphi)) = \tilde{S}(\varphi)$. The adjoint equation (133) for $\boldsymbol{\varphi} = T(\varphi)$ is equivalent to the adjoint equation (149). In particular the existence and uniqueness of the adjoint state $\boldsymbol{p}$ follows.
Now consider the KKT system. For the volume constraint, the condition $-1 < \tilde{\mathfrak{m}} < 1$ is equivalent to the condition $\mathfrak{m} > 0$ needed for the existence of Lagrange multipliers in the vector valued case. Recall that $\tilde{\mathfrak{m}} = \mathfrak{m}_1 - \mathfrak{m}_2$ and that it holds $\mathfrak{m}_1 + \mathfrak{m}_2 = 1$ and $\mathfrak{m} \geq 0$. Let $\overline{\varphi}$ be a local minimizer of $\tilde{j}$ in $\widetilde{\Phi_{ad}}$. Then $T(\overline{\varphi})$ is also a local minimizer of j in Φ_{ad}, see Theorem 6.17. Thus by Theorem 6.33 there exist Lagrange multipliers $\lambda_1 \in \mathbb{R}$, $\Lambda \in (H^1(\Omega) \cap L^\infty(\Omega))^*$ and $\boldsymbol{\mu} \in (L^\infty(\Omega)^2)^*$ for the vector valued optimization problem such that it holds

$$\begin{aligned}\left\langle j'(T(\overline{\varphi})), \boldsymbol{\eta}\right\rangle - \int_\Omega \eta_1\lambda_1 - \langle \Lambda, \eta_1 + \eta_2\rangle - \langle \boldsymbol{\mu}, \boldsymbol{\eta}\rangle_{(L^\infty)^*,L^\infty} = 0 \quad & \forall \boldsymbol{\eta} \in H^1(\Omega)^2 \cap L^\infty(\Omega)^2 \\ \langle \boldsymbol{\mu}, \boldsymbol{\eta}\rangle_{(L^\infty)^*,L^\infty} \geq 0 \quad & \forall \boldsymbol{\eta} \in L^\infty(\Omega)^2, \boldsymbol{\eta} \geq 0 \\ \langle \boldsymbol{\mu}, T(\overline{\varphi})\rangle_{(L^\infty)^*,L^\infty} = 0.\end{aligned}$$

Let now $\eta \in H^1(\Omega) \cap L^\infty(\Omega)$ be arbitrary. By testing the first equation in the KKT system by $\boldsymbol{\eta} = T'(\overline{\varphi})\eta = (\frac{1}{2}\eta, -\frac{1}{2}\eta)^T$ and applying the chain rule we get

$$\left\langle \tilde{j}'(\overline{\varphi}), \eta\right\rangle - \int_\Omega \frac{1}{2}\eta\lambda_1 - \left\langle \boldsymbol{\mu}, \left(\frac{1}{2}\eta, -\frac{1}{2}\eta\right)^T\right\rangle_{(L^\infty)^*,L^\infty} = 0 \quad \forall \eta \in H^1(\Omega) \cap L^\infty(\Omega).$$

With the definitions $\lambda := \frac{1}{2}\lambda_1$ and $\langle \mu_i, \eta\rangle := \frac{1}{2}\langle \boldsymbol{\mu}, \eta\boldsymbol{e}_i\rangle$ for all $\eta \in L^\infty(\Omega)$, $i = 1, 2$ we get

$$\left\langle \tilde{j}'(\overline{\varphi}), \eta\right\rangle - \lambda \int_\Omega \eta - \langle \mu_1 - \mu_2, \eta\rangle_{(L^\infty)^*,L^\infty} = 0 \quad \forall \eta \in H^1(\Omega) \cap L^\infty(\Omega).$$

Now test the second equation in the KKT system by $\boldsymbol{\eta} = \frac{1}{2}\eta\boldsymbol{e}_i$ to obtain

$$\langle \mu_i, \eta\rangle_{(L^\infty)^*,L^\infty} \geq 0 \quad \forall \eta \in L^\infty(\Omega), \eta \geq 0,\ i = 1, 2.$$

From the third equation in the KKT system we finally get

$$0 = \langle \boldsymbol{\mu}, T(\overline{\varphi})\rangle_{(L^\infty)^*,L^\infty} = \langle \mu_1, 1 + \overline{\varphi}\rangle_{(L^\infty)^*,L^\infty} + \langle \mu_2, 1 - \overline{\varphi}\rangle_{(L^\infty)^*,L^\infty}.$$

Since both summands are non negative, they vanish both. Thus the existence of Lagrange multipliers follows.
To show uniqueness of the Lagrange multipliers, we prove that the KKT system for the vector valued problem is indeed equivalent to the KKT system for the scalar valued problem. So let $\overline{\varphi}$ be a local minimizer of $\tilde{j}$ in $\widetilde{\Phi_{ad}}$ and let $\lambda \in \mathbb{R}$, $\mu_1 \in (L^\infty(\Omega))^*$ and $\mu_2 \in (L^\infty(\Omega))^*$ be Lagrange multipliers fulfilling the KKT system (150)-(151). We show that λ_1, Λ and $\boldsymbol{\mu}$, defined as

$$\begin{aligned}\lambda_1 &:= 2\lambda \\ \langle \boldsymbol{\mu}, \boldsymbol{\eta}\rangle_{(L^\infty)^*,L^\infty} &:= 2\langle \mu_1, \eta_1\rangle_{(L^\infty)^*,L^\infty} + 2\langle \mu_2, \eta_2\rangle_{(L^\infty)^*,L^\infty} \quad \forall \boldsymbol{\eta} \in L^\infty(\Omega)^2 \\ \langle \Lambda, \eta\rangle &:= \left\langle j'(T(\overline{\varphi})), (\eta, 0)^T\right\rangle - \int_\Omega \eta\lambda_1 - \left\langle \boldsymbol{\mu}, (\eta, 0)^T\right\rangle_{(L^\infty)^*,L^\infty} \quad \forall \eta \in H^1(\Omega) \cap L^\infty(\Omega)\end{aligned} \tag{152}$$

are a solution of the vector valued KKT system. From the uniqueness of λ_1, Λ and $\boldsymbol{\mu}$ then also uniqueness of λ, μ_1 and μ_2 follows, since the relationship is one-to-one.
From (150), the chain rule and the definitions of λ_1 and $\boldsymbol{\mu}$ we get

$$\left\langle j'(T(\overline{\varphi})), \left(\frac{1}{2}\eta, -\frac{1}{2}\eta\right)^T \right\rangle - \int_\Omega \frac{1}{2}\eta\lambda_1 - \left\langle \boldsymbol{\mu}, \left(\frac{1}{2}\eta, -\frac{1}{2}\eta\right)^T \right\rangle_{(L^\infty)^*, L^\infty} = 0 \quad \forall \eta \in H^1(\Omega) \cap L^\infty(\Omega).$$

Multiplying this equation by 2 and subtracting it from (152) leads to

$$\langle \Lambda, \eta \rangle = \left\langle j'(T(\overline{\varphi})), (0,\eta)^T \right\rangle - \left\langle \boldsymbol{\mu}, (0,\eta)^T \right\rangle_{(L^\infty)^*, L^\infty} \quad \forall \eta \in H^1(\Omega) \cap L^\infty(\Omega). \tag{153}$$

Let $\boldsymbol{\eta} \in H^1(\Omega)^2 \cap L^\infty(\Omega)^2$ be arbitrary. Test equation (152) by η_1, equation (153) by η_2 and add the equations up to get

$$\left\langle j'(T(\overline{\varphi})), \boldsymbol{\eta} \right\rangle - \int_\Omega \eta_1\lambda_1 - \langle \Lambda, \eta_1 + \eta_2 \rangle - \langle \boldsymbol{\mu}, \boldsymbol{\eta} \rangle_{(L^\infty)^*, L^\infty} = 0 \quad \forall \boldsymbol{\eta} \in H^1(\Omega)^2 \cap L^\infty(\Omega)^2,$$

which is the first equation in the vector valued KKT system. Moreover, we get

$$\langle \boldsymbol{\mu}, \boldsymbol{\eta} \rangle_{(L^\infty)^*, L^\infty} = 2\langle \mu_1, \eta_1 \rangle_{(L^\infty)^*, L^\infty} + 2\langle \mu_2, \eta_2 \rangle_{(L^\infty)^*, L^\infty} \geq 0 \quad \forall \boldsymbol{\eta} \in L^\infty(\Omega)^2,\ \boldsymbol{\eta} \geq 0,$$

as well as

$$\langle \boldsymbol{\mu}, T(\overline{\varphi}) \rangle_{(L^\infty)^*, L^\infty} = \langle \mu_1, 1 + \overline{\varphi} \rangle_{(L^\infty)^*, L^\infty} + \langle \mu_2, 1 - \overline{\varphi} \rangle_{(L^\infty)^*, L^\infty} = 0.$$

□

Remark 6.39. In case that no mass constraint is present one can show the existence and uniqueness of Lagrange multipliers by the same techniques. In this case all statements remain valid when setting $\lambda_i := 0$, $i = 1, \dots, N-1$. The assumption $\int_\Omega \overline{\boldsymbol{\varphi}} > 0$ is not needed, thus it is also possible that one phase vanishes in the minimum. In the proof of Theorem 6.33 one can choose $\delta = \min\{\frac{1}{2(N-1)}, \frac{1}{4}\}$, $\varphi_i \equiv \frac{1}{2(N-1)}$ for $i = 1, \dots, N-1$, $\varphi_N = F + 1 - \sum_{i=1}^{N-1} \varphi_i$ and $\boldsymbol{\eta} = \boldsymbol{\varphi} - \boldsymbol{F}$. The proof for uniqueness is the same as for Theorem 6.37.

6.6 Second order derivatives

In this section we compute the second order derivatives of j using an adjoint approach. This will be used later for the SQP method (Section 6.9). In Section 6.7 we also propose a second order metric for the VMPT method, which depends on the linearized adjoint state defined here.

The second order derivative $j''(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi}, \boldsymbol{\tau\varphi}]$ can be computed by differentiating $\langle j'(\boldsymbol{\varphi}), \boldsymbol{\delta\varphi} \rangle$ (see Theorem 6.25) in direction $\boldsymbol{\tau\varphi}$.

Theorem 6.40. *The reduced cost functional j is two times continuously Fréchet differentiable on $H^1(\Omega)^N \cap L^\infty(\Omega)^N$. The second order derivative is given by*

$$\begin{aligned} j''(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi}, \boldsymbol{\tau\varphi}] = \gamma\varepsilon \int_\Omega \nabla\boldsymbol{\tau\varphi} : \nabla\boldsymbol{\delta\varphi} + \frac{\gamma}{\varepsilon} \int_\Omega \psi_0''(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi}, \boldsymbol{\tau\varphi}] \\ + F_{\boldsymbol{\varphi},\boldsymbol{\varphi}}(\boldsymbol{\varphi}, S(\boldsymbol{\varphi}))[\boldsymbol{\tau\varphi}, \boldsymbol{\delta\varphi}] + F_{\boldsymbol{\varphi},\boldsymbol{u}}(\boldsymbol{\varphi}, S(\boldsymbol{\varphi}))[S'(\boldsymbol{\varphi})\boldsymbol{\tau\varphi}, \boldsymbol{\delta\varphi}] \\ + F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}, S(\boldsymbol{\varphi}))[\boldsymbol{\tau\varphi}, S'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}] + F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi}, S(\boldsymbol{\varphi}))[S'(\boldsymbol{\varphi})\boldsymbol{\tau\varphi}, S'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}] \\ + \left\langle F_{\boldsymbol{u}}(\boldsymbol{\varphi}, S(\boldsymbol{\varphi})), S''(\boldsymbol{\varphi})[\boldsymbol{\tau\varphi}, \boldsymbol{\delta\varphi}] \right\rangle_{(H_D^1)^*, H_D^1} \end{aligned} \tag{154}$$

for all $\boldsymbol{\varphi},\ \boldsymbol{\delta\varphi},\ \boldsymbol{\tau\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$.

Proof. It is already proved that E and S are two times continuously Fréchet differentiable, see Theorem 6.23 and Theorem 6.11, respectively. F has this property by assumption. Thus, j is C^2 and the formula for j'' can be calculated using the chain rule. □

Similar to the first order derivative we want to introduce an adjoint approach to render $j''(\boldsymbol{\varphi})[.,\boldsymbol{\tau\varphi}]$ computable. Therefore we have to differentiate the adjoint state equation to get the linearized adjoint state equation.

Definition 6.41. For given $\boldsymbol{\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$, $\boldsymbol{u} \in H^1_D$, $\boldsymbol{p} \in H^1_D$, $\boldsymbol{\delta u} \in H^1_D$ and $\boldsymbol{\delta\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$ we define the linearized adjoint state $\boldsymbol{\delta p} \in H^1_D$ as the solution of the equation

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{\delta p}) : \mathcal{E}(\boldsymbol{\xi}) = F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi},\boldsymbol{u})[\boldsymbol{\delta\varphi},\boldsymbol{\xi}] + F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi},\boldsymbol{u})[\boldsymbol{\delta u},\boldsymbol{\xi}] - \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\mathcal{E}(\boldsymbol{p}) : \mathcal{E}(\boldsymbol{\xi}) \tag{155}$$

for all $\boldsymbol{\xi} \in H^1_D$.

Remark 6.42. If F is given by

$$F(\boldsymbol{\varphi},\boldsymbol{u}) = \int_\Omega \alpha(x,\boldsymbol{\varphi}(x),\boldsymbol{u}(x))\,\mathrm{d}x + \int_{\partial\Omega} \beta(x,\boldsymbol{\varphi}(x),\boldsymbol{u}(x))\,\mathrm{d}x,$$

for some functions α and β, then (155) is the weak formulation of the following PDE:

$$\begin{aligned} -\nabla\cdot(\boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{\delta p})) &= \alpha_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi},\boldsymbol{u})\boldsymbol{\delta\varphi} + \alpha_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi},\boldsymbol{u})\boldsymbol{\delta u} + \nabla\cdot(\boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\mathcal{E}(\boldsymbol{p})) \quad \text{in } \Omega \\ \boldsymbol{\delta p} &= \boldsymbol{0} \quad \text{on } \Gamma_D \\ \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{\delta p})\cdot\boldsymbol{n} &= \beta_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi},\boldsymbol{u})\boldsymbol{\delta\varphi} + \beta_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi},\boldsymbol{u})\boldsymbol{\delta u} - \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\mathcal{E}(\boldsymbol{p})\cdot\boldsymbol{n} \quad \text{on } \partial\Omega\setminus\Gamma_D. \end{aligned}$$

Here, $\alpha_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi},\boldsymbol{u})$ has to be interpreted as matrix in $\mathbb{R}^{d\times N}$ (rather than a bilinear form) and similarly for the other expressions.

We have seen that for the mean compliance problem (112) the solution of the adjoint equation fulfills $\boldsymbol{p} = S(\boldsymbol{\varphi})$. Moreover, we get for the mean compliance problem $F_{\boldsymbol{u},\boldsymbol{u}} = 0$ and $F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi},\boldsymbol{u})[\boldsymbol{\delta\varphi},\boldsymbol{\xi}] = \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\cdot\boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\cdot\boldsymbol{\xi}$, thus the corresponding linearized adjoint equation is independent of $\boldsymbol{u}$ and $\boldsymbol{\delta u}$, and it coincides with the linearized state equation (124) with $\boldsymbol{u}$ replaced by $\boldsymbol{p}$. Thus, if one chooses $\boldsymbol{p}$ as the solution of the adjoint equation it holds $\boldsymbol{\delta p} = S'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}$.

Theorem 6.43. *For any* $\boldsymbol{\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$, $\boldsymbol{u} \in H^1_D$, $\boldsymbol{p} \in H^1_D$, $\boldsymbol{\delta u} \in H^1_D$ *and* $\boldsymbol{\delta\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$, *the linearized adjoint equation* (155) *has a unique solution* $\boldsymbol{\delta p} \in H^1_D$. *It holds the a priori estimate*

$$\begin{aligned} \|\boldsymbol{\delta p}\|_{H^1} \le C(&\|F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi},\boldsymbol{u})\|_{\mathcal{L}(H^1\cap L^\infty,(H^1)^*)}\|\boldsymbol{\delta\varphi}\|_{H^1\cap L^\infty} \\ &+ \|F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi},\boldsymbol{u})\|_{\mathcal{L}(H^1,(H^1)^*)}\|\boldsymbol{\delta u}\|_{H^1} + \|\boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty}\|\boldsymbol{\delta\varphi}\|_{L^\infty}\|\boldsymbol{p}\|_{H^1}) \end{aligned} \tag{156}$$

where $C > 0$ *is independent of* $\boldsymbol{\varphi}$, $\boldsymbol{u}$, $\boldsymbol{p}$, $\boldsymbol{\delta u}$ *and* $\boldsymbol{\delta\varphi}$.

Proof. This can be proved by the Lax-Milgram theorem as in Theorem 6.6. The right hand side as a function of $\boldsymbol{\xi}$ is in $(H^1_D)^*$, since $F \in C^2((H^1(\Omega)^N \cap L^\infty(\Omega)^N) \times H^1_D)$ by assumption **(AP8)**. □

The adjoint representation of $j''(\varphi)$ can be obtained by formally differentiating the adjoint representation of $j'(\varphi)$ (see Theorem 6.25). Similar to the first order derivative we are able to justify this approach rigorously.

Theorem 6.44. *It holds*

$$\begin{aligned} j''(\varphi)[\tau\varphi,\delta\varphi] = {} & \gamma\varepsilon\int_\Omega \nabla\tau\varphi:\nabla\delta\varphi + \frac{\gamma}{\varepsilon}\int_\Omega \psi_0''(\varphi)\delta\varphi\tau\varphi \\ & + F_{\varphi,\varphi}(\varphi,u)[\tau\varphi,\delta\varphi] + F_{\varphi,u}(\varphi,u)[\tau u,\delta\varphi] \\ & - \int_\Omega (C''(\varphi)\tau\varphi\mathcal{E}(u):\mathcal{E}(p))\cdot\delta\varphi - \int_\Omega (\nabla C(\varphi)\mathcal{E}(\tau u):\mathcal{E}(p))\cdot\delta\varphi \\ & - \int_\Omega (\nabla C(\varphi)\mathcal{E}(u):\mathcal{E}(\tau p))\cdot\delta\varphi + \int_\Omega f_{\varphi,\varphi}(\varphi)[\tau\varphi,\delta\varphi]\cdot p \\ & + \int_\Omega f_\varphi(\varphi)^T\tau p\cdot\delta\varphi + \int_{\Gamma_g} g_{\varphi,\varphi}(\varphi)[\tau\varphi,\delta\varphi]\cdot p \\ & + \int_{\Gamma_g} g_\varphi(\varphi)^T\tau p\cdot\delta\varphi \end{aligned}$$

for all $\varphi,\ \tau\varphi,\ \delta\varphi \in H^1(\Omega)^N\cap L^\infty(\Omega)^N$, *where* $u = S(\varphi)$, p *is the adjoint state corresponding to the data* (φ,u), $\tau u = S'(\varphi)\tau\varphi$ *and* τp *is the linearized adjoint state corresponding to the data* $(\varphi,u,p,\tau u,\tau\varphi)$.

Proof. We test the linearized adjoint equation (155) by $\xi = \delta u := S'(\varphi)\delta\varphi \in H^1_D$ to obtain

$$\begin{aligned} & \int_\Omega C(\varphi)\mathcal{E}(\tau p):\mathcal{E}(\delta u) \\ & \qquad = F_{u,\varphi}(\varphi,u)[\tau\varphi,\delta u] + F_{u,u}(\varphi,u)[\tau u,\delta u] - \int_\Omega C'(\varphi)\tau\varphi\mathcal{E}(p):\mathcal{E}(\delta u). \end{aligned} \tag{157}$$

Next we test the equation (127) for $S''(\varphi)[\delta\varphi,\tau\varphi]$ by $\xi = p \in H^1_D$ to get

$$\begin{aligned} & \int_\Omega C(\varphi)\mathcal{E}(S''(\varphi)[\tau\varphi,\delta\varphi]):\mathcal{E}(p) \\ & \qquad = -\int_\Omega C''(\varphi)[\delta\varphi,\tau\varphi]\mathcal{E}(u):\mathcal{E}(p) + \int_\Omega f_{\varphi,\varphi}(\varphi)[\delta\varphi,\tau\varphi]\cdot p \\ & \qquad\quad + \int_{\Gamma_g} g_{\varphi,\varphi}(\varphi)[\delta\varphi,\tau\varphi]\cdot p - \int_\Omega C'(\varphi)\tau\varphi\mathcal{E}(\delta u):\mathcal{E}(p) \\ & \qquad\quad - \int_\Omega C'(\varphi)\delta\varphi\mathcal{E}(\tau u):\mathcal{E}(p). \end{aligned} \tag{158}$$

Finally, we have to test the equation (124) for $S'(\varphi)\delta\varphi$ by $\xi = \tau p \in H^1_D$. We gain

$$\begin{aligned} & \int_\Omega C(\varphi)\mathcal{E}(\delta u):\mathcal{E}(\tau p) \\ & \qquad = -\int_\Omega C'(\varphi)\delta\varphi\mathcal{E}(u):\mathcal{E}(\tau p) + \int_\Omega f_\varphi(\varphi)\delta\varphi\cdot\tau p + \int_{\Gamma_g} g_\varphi(\varphi)\delta\varphi\cdot\tau p. \end{aligned} \tag{159}$$

Now solve equation (158) for $\int_\Omega C'(\varphi)\tau\varphi\mathcal{E}(\delta u):\mathcal{E}(p)$ and insert the result into (157)

and also insert (159) into (157). This yields

$$\begin{aligned}
&-\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\mathcal{E}(\boldsymbol{u}):\mathcal{E}(\boldsymbol{\tau p})+\int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\cdot\boldsymbol{\tau p}+\int_{\Gamma_g}\boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\cdot\boldsymbol{\tau p}\\
&=F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi},\boldsymbol{u})[\boldsymbol{\tau\varphi},\boldsymbol{\delta u}]+F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi},\boldsymbol{u})[\boldsymbol{\tau u},\boldsymbol{\delta u}]+\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(S''(\boldsymbol{\varphi})[\boldsymbol{\tau\varphi},\boldsymbol{\delta\varphi}]):\mathcal{E}(\boldsymbol{p})\\
&\quad+\int_\Omega \boldsymbol{C}''(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi},\boldsymbol{\tau\varphi}]\mathcal{E}(\boldsymbol{u}):\mathcal{E}(\boldsymbol{p})-\int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi},\boldsymbol{\varphi}}(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi},\boldsymbol{\tau\varphi}]\cdot\boldsymbol{p}\\
&\quad-\int_{\Gamma_g}\boldsymbol{g}_{\boldsymbol{\varphi},\boldsymbol{\varphi}}(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi},\boldsymbol{\tau\varphi}]\cdot\boldsymbol{p}+\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\mathcal{E}(\boldsymbol{\tau u}):\mathcal{E}(\boldsymbol{p}).
\end{aligned}$$

Solve this equation for the $F_{\boldsymbol{u},\boldsymbol{\varphi}}+F_{\boldsymbol{u},\boldsymbol{u}}$ term and insert the result into equation (154).

$$\begin{aligned}
j''(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi},\boldsymbol{\tau\varphi}]=\;&\gamma\varepsilon\int_\Omega \nabla\boldsymbol{\tau\varphi}:\nabla\boldsymbol{\delta\varphi}+\frac{\gamma}{\varepsilon}\int_\Omega \psi_0''(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi},\boldsymbol{\tau\varphi}]\\
&+F_{\boldsymbol{\varphi},\boldsymbol{\varphi}}(\boldsymbol{\varphi},\boldsymbol{u})[\boldsymbol{\tau\varphi},\boldsymbol{\delta\varphi}]+F_{\boldsymbol{\varphi},\boldsymbol{u}}(\boldsymbol{\varphi},\boldsymbol{u})[\boldsymbol{\tau u},\boldsymbol{\delta\varphi}]\\
&-\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\mathcal{E}(\boldsymbol{u}):\mathcal{E}(\boldsymbol{\tau p})+\int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\cdot\boldsymbol{\tau p}+\int_{\Gamma_g}\boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\cdot\boldsymbol{\tau p}\\
&-\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(S''(\boldsymbol{\varphi})[\boldsymbol{\tau\varphi},\boldsymbol{\delta\varphi}]):\mathcal{E}(\boldsymbol{p})\\
&-\int_\Omega \boldsymbol{C}''(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi},\boldsymbol{\tau\varphi}]\mathcal{E}(\boldsymbol{u}):\mathcal{E}(\boldsymbol{p})+\int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi},\boldsymbol{\varphi}}(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi},\boldsymbol{\tau\varphi}]\cdot\boldsymbol{p}\\
&+\int_{\Gamma_g}\boldsymbol{g}_{\boldsymbol{\varphi},\boldsymbol{\varphi}}(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi},\boldsymbol{\tau\varphi}]\cdot\boldsymbol{p}-\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\mathcal{E}(\boldsymbol{\tau u}):\mathcal{E}(\boldsymbol{p})\\
&+\left\langle F_{\boldsymbol{u}}(\boldsymbol{\varphi},\boldsymbol{u}),S''(\boldsymbol{\varphi})[\boldsymbol{\tau\varphi},\boldsymbol{\delta\varphi}]\right\rangle_{(H_D^1)^*,H_D^1}.
\end{aligned}$$

Finally test the adjoint equation (133) by $\boldsymbol{\xi}=S''(\boldsymbol{\varphi})[\boldsymbol{\tau\varphi},\boldsymbol{\delta\varphi}]\in H_D^1$ to see that the term

$$-\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{p}):\mathcal{E}(S''(\boldsymbol{\varphi})[\boldsymbol{\tau\varphi},\boldsymbol{\delta\varphi}])+\left\langle F_{\boldsymbol{u}}(\boldsymbol{\varphi},\boldsymbol{u}),S''(\boldsymbol{\varphi})[\boldsymbol{\tau\varphi},\boldsymbol{\delta\varphi}]\right\rangle_{(H_D^1)^*,H_D^1}$$

vanishes. □

With this adjoint representation, it can be very cheap to evaluate $j''(\boldsymbol{\varphi})[\boldsymbol{\tau\varphi},\boldsymbol{\delta\varphi}]$ for different $\boldsymbol{\delta\varphi}$. For given $\boldsymbol{\varphi}$ and $\boldsymbol{\tau\varphi}$ one first has to compute $\boldsymbol{u}$, $\boldsymbol{p}$, $\boldsymbol{\tau u}$ and $\boldsymbol{\tau p}$ by solving the state equation, adjoint equation, linearized state equation and linearized adjoint equation. Assume now that $F_{\boldsymbol{\varphi},\boldsymbol{\varphi}}(\boldsymbol{\varphi},S(\boldsymbol{\varphi}))[\boldsymbol{\tau\varphi},\boldsymbol{\delta\varphi}]+F_{\boldsymbol{\varphi},\boldsymbol{u}}(\boldsymbol{\varphi},S(\boldsymbol{\varphi}))[S'(\boldsymbol{\varphi})\boldsymbol{\tau\varphi},\boldsymbol{\delta\varphi}]$ can be cheaply evaluated in $\boldsymbol{\delta\varphi}$ (e.g. by computing an integral), then the whole second order derivative $j''(\boldsymbol{\varphi})[\boldsymbol{\tau\varphi},\boldsymbol{\delta\varphi}]$ can be computed for different $\boldsymbol{\delta\varphi}$ by computing only integrals. The variables $\boldsymbol{u}$, $\boldsymbol{p}$, $\boldsymbol{\tau u}$ and $\boldsymbol{\tau p}$ do not depend on $\boldsymbol{\delta\varphi}$ and thus don't have to be recomputed. As an example, consider the boundary term

$$\int_{\Gamma_g}\boldsymbol{g}_{\boldsymbol{\varphi},\boldsymbol{\varphi}}(\boldsymbol{\varphi})[\boldsymbol{\tau\varphi},\boldsymbol{\delta\varphi}]\cdot\boldsymbol{p}=\int_{\Gamma_g}\sum_{i,j=1}^{N}\sum_{k=1}^{d}\boldsymbol{g}^k_{\boldsymbol{\varphi}_i,\boldsymbol{\varphi}_j}(x,\boldsymbol{\varphi}(x))\boldsymbol{\tau\varphi}^i(x)\boldsymbol{\delta\varphi}^j(x)\boldsymbol{p}^k(x)\,\mathrm{d}x.$$

Here one can see that for varying $\boldsymbol{\delta\varphi}$ only a single integral has to be computed, which can be done by standard finite element techniques.

The next Theorem summarizes the results concerning second order derivatives for the scalar valued problem.

Theorem 6.45. *Consider the problem for the scalar valued phase field* (121)*. It holds*

$$\begin{aligned} j''(\varphi)[\tau\varphi,\delta\varphi] = \gamma\varepsilon \int_\Omega \nabla\tau\varphi\cdot\nabla\delta\varphi + \frac{\gamma}{\varepsilon}\int_\Omega \psi_0''(\varphi)\delta\varphi\tau\varphi \\ + F_{\varphi,\varphi}(\varphi,\boldsymbol{u})[\tau\varphi,\delta\varphi] + F_{\varphi,\boldsymbol{u}}(\varphi,\boldsymbol{u})[\boldsymbol{\tau u},\delta\varphi] \\ - \int_\Omega \boldsymbol{C}''(\varphi)\mathcal{E}(\boldsymbol{u}):\mathcal{E}(\boldsymbol{p})\tau\varphi\delta\varphi - \int_\Omega \boldsymbol{C}'(\varphi)\mathcal{E}(\boldsymbol{\tau u}):\mathcal{E}(\boldsymbol{p})\delta\varphi \\ - \int_\Omega \boldsymbol{C}'(\varphi)\mathcal{E}(\boldsymbol{u}):\mathcal{E}(\boldsymbol{\tau p})\delta\varphi + \int_\Omega \boldsymbol{f}_{\varphi,\varphi}(\varphi)\cdot\boldsymbol{p}\,\tau\varphi\,\delta\varphi \\ + \int_\Omega \boldsymbol{f}_\varphi(\varphi)\cdot\boldsymbol{\tau p}\,\delta\varphi + \int_{\Gamma_g} \boldsymbol{g}_{\varphi,\varphi}(\varphi)\cdot\boldsymbol{p}\,\tau\varphi\,\delta\varphi \\ + \int_{\Gamma_g} \boldsymbol{g}_\varphi(\varphi)\cdot\boldsymbol{\tau p}\,\delta\varphi \end{aligned}$$

for all $\varphi,\ \tau\varphi,\ \delta\varphi \in H^1(\Omega)\cap L^\infty(\Omega)$*, where* $\boldsymbol{u} = S(\varphi)$*,* $\boldsymbol{p}$ *is the adjoint state corresponding to the data* $(\varphi,\boldsymbol{u})$*, see* (149)*,* $\boldsymbol{\tau u} = S'(\varphi)\tau\varphi$ *and* $\boldsymbol{\tau p}$ *is the linearized adjoint state, which is the solution of the equation*

$$\int_\Omega \boldsymbol{C}(\varphi)\mathcal{E}(\boldsymbol{\tau p}):\mathcal{E}(\boldsymbol{\xi}) = F_{\boldsymbol{u},\varphi}(\varphi,\boldsymbol{u})[\tau\varphi,\boldsymbol{\xi}] + F_{\boldsymbol{u},\boldsymbol{u}}(\varphi,\boldsymbol{u})[\boldsymbol{\tau u},\boldsymbol{\xi}] - \int_\Omega \boldsymbol{C}'(\varphi)\tau\varphi\mathcal{E}(\boldsymbol{p}):\mathcal{E}(\boldsymbol{\xi})$$

for all $\boldsymbol{\xi}\in H^1_D$*.*

Proof. As before, we add a tilde to functions concerning the scalar valued problem and let $T(\varphi) = (\frac{1+\varphi}{2}, \frac{1-\varphi}{2})^T$. From $\tilde{j}(\varphi) = j(T(\varphi))$, we get by chain rule

$$\tilde{j}''(\varphi)[\tau\varphi,\delta\varphi] = j''(T(\varphi))[T'(\varphi)\tau\varphi, T'(\varphi)\delta\varphi],$$

which we computed in Theorem 6.44. The linearized adjoint equation for the scalar valued problem can be obtained by replacing $\boldsymbol{\tau\varphi}$ by $T'(\varphi)\tau\varphi$ in the linearized adjoint equation for the vector valued problem and applying the chain rule. □

6.7 Global convergence of variable metric projection type methods

In this section we show that the general structural topology optimization problem fulfills the assumptions that guarantee global convergence of the VMPT method. We also give many examples for inner products which can be used for the VMPT method. Therefor we use slightly stronger assumptions:

(PGC1) For each $\boldsymbol{\varphi}\in\Phi_{ad}$, $\boldsymbol{u}\in H^1_D$ and for each sequence $(\boldsymbol{\varphi}_i)_i \subset H^1(\Omega)^N \cap L^\infty(\Omega)^N$ with $\boldsymbol{\varphi}_i \to \boldsymbol{0}$ weakly in $H^1(\Omega)^N$ and weakly-* in $L^\infty(\Omega)^N$ it holds

$$\langle F_{\boldsymbol{\varphi}}(\boldsymbol{\varphi},\boldsymbol{u}),\boldsymbol{\varphi}_i\rangle \to 0.$$

(PGC2) It exists $C>0$ such that $|\boldsymbol{g}_{\boldsymbol{\varphi}}(x,y)|\le C$ for almost every $x\in\Gamma_g$ and every $y\in\Delta^{N-1}$.

Assumption **(PGC2)** is in fact not necessary and all results also hold without assuming **(PGC2)**. This can be shown by the arguments used in the proof of Lemma 6.3. However, to simplify the proofs we will assume **(PGC2)** in the following.
Assumption **(PGC1)** is not really a restriction because of the following lemma.

Lemma 6.46. *Let F be given as*

$$F(\boldsymbol{\varphi}, \boldsymbol{u}) = \int_\Omega \alpha(x, \boldsymbol{\varphi}(x), \boldsymbol{u}(x))\,\mathrm{d}x + \int_{\partial\Omega} \beta(x, \boldsymbol{\varphi}(x), \boldsymbol{u}(x))\,\mathrm{d}x \tag{160}$$

*with $\alpha : \Omega \times \mathbb{R}^N \times \mathbb{R}^d \to \mathbb{R}$, $\beta : \partial\Omega \times \mathbb{R}^N \times \mathbb{R}^d \to \mathbb{R}$ such that $\nabla_\varphi \alpha(x, \boldsymbol{\varphi}(x), \boldsymbol{u}(x)) \in L^1(\Omega)^N$ and $\nabla_\varphi \beta(x, \boldsymbol{\varphi}(x), \boldsymbol{u}(x)) \in L^1(\partial\Omega)^N$ for all $\boldsymbol{\varphi} \in \Phi_{ad}$ and $\boldsymbol{u} \in H^1_D$. Then **(PGC1)** is fulfilled.*

Proof. We have

$$\langle F_\varphi(\boldsymbol{\varphi}, \boldsymbol{u}), \boldsymbol{\varphi}_i \rangle = \int_\Omega \nabla_\varphi \alpha(x, \boldsymbol{\varphi}(x), \boldsymbol{u}(x)) \cdot \boldsymbol{\varphi}_i(x)\,\mathrm{d}x + \int_{\partial\Omega} \nabla_\varphi \beta(x, \boldsymbol{\varphi}(x), \boldsymbol{u}(x)) \cdot \boldsymbol{\varphi}_i(x)\,\mathrm{d}x.$$

Due to $\nabla_\varphi \alpha(x, \boldsymbol{\varphi}(x), \boldsymbol{u}(x)) \in L^1(\Omega)^N$ we can pass to the limit in the first term since $\boldsymbol{\varphi}_i \to \mathbf{0}$ weakly-* in L^∞. For the second term we observe that $\boldsymbol{\varphi}_i \to \mathbf{0}$ strongly in $L^2(\partial\Omega)^N$ and thus pointwise (for some subsequence). Due to $\boldsymbol{\varphi}_i \to \mathbf{0}$ weakly-* in L^∞ and Lemma 7.4 we get $\|\boldsymbol{\varphi}_i\|_{L^\infty(\partial\Omega)^N} \le C$. We pass to the limit in the second term using dominated convergence. □

Note that it is not necessary that α and β in the lemma above depend on $\boldsymbol{u}$ pointwise. For instance the same proof applies to the stress functional (119). The remaining cost functionals discussed in Section 6.1.2 are of the form (160) except the expression $\left|\int_\Omega \boldsymbol{\varphi}\,\mathrm{d}x - \mathfrak{m}\right|^2$, for which **(PGC1)** is trivially fulfilled.

We apply the VMPT method on the reduced problem, i.e. we eliminate $\boldsymbol{u}$ from the optimization problem, since the method requires convex constraints. Since the control-to-state operator S is nonlinear, we don't have convexity of the admissible set for the unreduced problem.

Our goal is to apply the VMPT method in the space $H^1(\Omega)^N \cap L^\infty(\Omega)^N$, using amongst others the inner product $a_k(\boldsymbol{p}, \boldsymbol{v}) = \int_\Omega \nabla \boldsymbol{p} : \nabla \boldsymbol{v}$. Since a_k does not define a positive bilinear form on the whole space $H^1(\Omega)^N \cap L^\infty(\Omega)^N$, we first have to choose an appropriate subspace. Therefor we translate the problem by the vector $\mathfrak{m} \in \Phi_{ad}$ and put the resulting constraint $⨍ \boldsymbol{\varphi} = \mathbf{0}$ into the space

$$H^1_{(0)}(\Omega)^N := \left\{ \boldsymbol{\varphi} \in H^1(\Omega)^N \;\middle|\; ⨍_\Omega \boldsymbol{\varphi} = \mathbf{0} \right\}.$$

The Poincaré inequality for functions with vanishing mean value then guarantees that a_k is positive. Moreover, we define the H^{-1} inner product to be the standard inner product in $(H^1_{(0)}(\Omega)^N)^*$, i.e. for $\boldsymbol{u}, \boldsymbol{v} \in (H^1_{(0)}(\Omega)^N)^*$ we have

$$(\boldsymbol{u}, \boldsymbol{v})_{H^{-1}} := \int_\Omega (\nabla(-\Delta_N)^{-1}\boldsymbol{u}) : (\nabla(-\Delta_N)^{-1}\boldsymbol{v}), \tag{161}$$

where $\boldsymbol{w} = (-\Delta_N)^{-1}\boldsymbol{u} \in H^1_{(0)}(\Omega)^N$ is the unique solution of $\int_\Omega \nabla \boldsymbol{w} : \nabla \boldsymbol{\eta} = \langle \boldsymbol{u}, \boldsymbol{\eta} \rangle \;\forall \eta \in H^1_{(0)}(\Omega)^N$. If $\boldsymbol{u}$ is a function, then $\boldsymbol{w}$ is the unique weak solution of the Neumann problem

$$-\Delta \boldsymbol{w} = \boldsymbol{u} - ⨍_\Omega \boldsymbol{u} \;\text{ in } \Omega, \quad ⨍_\Omega \boldsymbol{w} = \mathbf{0}, \quad \partial_\nu \boldsymbol{w} = \mathbf{0} \;\text{ on } \partial\Omega.$$

We note that the translation of the problem by the vector $\mathfrak{m}$ does not change the VMPT method since the method is translation invariant, cf. Theorem 4.27. The translated

reduced problem reads

$$\begin{aligned} \min\ & j(\boldsymbol{\varphi}+\mathfrak{m}) \\ & \boldsymbol{\varphi} \in H^1_{(0)}(\Omega)^N \\ & \boldsymbol{\varphi} \geq -\mathfrak{m} \\ & \sum_{i=1}^N \varphi_i = 0. \end{aligned} \tag{162}$$

Note that the constraint $\fint \boldsymbol{\varphi} = \mathbf{0}$ is now handled implicitly by the condition $\boldsymbol{\varphi} \in H^1_{(0)}(\Omega)^N$. The admissible set for the translated problem is tangential to the original admissible set Φ_{ad} (concerning the mass and the sum constraint), thus we define

$$\begin{aligned} \Phi_{ad}^{tan} &:= \left\{ \boldsymbol{\varphi} \in H^1_{(0)}(\Omega)^N \;\middle|\; \boldsymbol{\varphi} \geq -\mathfrak{m},\ \sum_{i=1}^N \varphi_i = 0 \right\} \\ &= \left\{ \boldsymbol{\varphi} \in H^1(\Omega)^N \;\middle|\; \boldsymbol{\varphi} \geq -\mathfrak{m},\ \sum_{i=1}^N \varphi_i = 0 \ \fint \boldsymbol{\varphi} = \mathbf{0} \right\}. \end{aligned}$$

Thus we apply the VMPT method in the spaces

$$\begin{aligned} &\mathbb{X} = H^1_{(0)}(\Omega)^N \text{ and} \\ &\mathbb{D} = L^\infty(\Omega)^N. \end{aligned} \tag{163}$$

In the following we equip $\mathbb{X}$ with the full H^1 norm $\|\boldsymbol{f}\|_{\mathbb{X}} = \|\boldsymbol{f}\|_{H^1} = \|\boldsymbol{f}\|_{L^2} + \|\nabla \boldsymbol{f}\|_{L^2}$, which is equivalent to the H^1 seminorm on $\mathbb{X}$. The seminorm will be denoted by $\|\boldsymbol{f}\|_{H^1_0} = \|\nabla \boldsymbol{f}\|_{L^2}$. We first have to check the assumptions on the spaces $\mathbb{X}$ and $\mathbb{D}$.

Lemma 6.47. *$\mathbb{X}$ and $\mathbb{D}$ fulfill the assumptions on the spaces **(A1)**.*

Proof. First of all, $H^1_{(0)}(\Omega)^N$ is a real Hilbert space and thus a reflexive Banach space. Moreover, it holds $L^\infty(\Omega)^N \cong \mathbb{B}^*$, where $\mathbb{B} = L^1(\Omega)^N$ is a separable Banach space. Let $(\boldsymbol{\varphi}_i)_i \subset H^1_{(0)}(\Omega)^N \cap L^\infty(\Omega)^N$ be a sequence, which converges to some $\boldsymbol{\varphi} \in H^1_{(0)}(\Omega)^N$ weakly in $H^1_{(0)}(\Omega)^N$ and to some $\tilde{\boldsymbol{\varphi}} \in L^\infty(\Omega)^N$ weakly-* in $L^\infty(\Omega)^N$. We have to show $\boldsymbol{\varphi} = \tilde{\boldsymbol{\varphi}}$. Since we have the embedding $H^1_{(0)}(\Omega)^N \hookrightarrow L^2(\Omega)^N$ we conclude that $\boldsymbol{\varphi}_i \to \boldsymbol{\varphi}$ and $\boldsymbol{\varphi}_i \to \tilde{\boldsymbol{\varphi}}$, both in the sense of distributions. Since this limit is unique, we get $\boldsymbol{\varphi} = \tilde{\boldsymbol{\varphi}}$. □

Lemma 6.48. *In addition to the standard assumptions in Section 6.1.1 let the assumptions **(PGC1)** and **(PGC2)** hold. Then the translated j together with Φ_{ad}^{tan} fulfills the assumptions on the problem **(A2)**-**(A7)**.*

Proof. By definition, it holds $\Phi_{ad}^{tan} \subset \mathbb{X}$. To see that it holds $\Phi_{ad}^{tan} \subset \mathbb{D}$, we use that

$$\mathbf{0} \leq \boldsymbol{\varphi} \leq \mathbf{1} \quad \forall \boldsymbol{\varphi} \in \Phi_{ad},$$

and thus

$$-\mathbf{1} \leq -\mathfrak{m} \leq \boldsymbol{\varphi} \leq \mathbf{1} - \mathfrak{m} \leq \mathbf{1} \quad \forall \boldsymbol{\varphi} \in \Phi_{ad}^{tan} \tag{164}$$

almost everywhere in Ω. Thus it holds $\Phi_{ad}^{tan} \subset \mathbb{X} \cap \mathbb{D}$. Moreover one easily checks that Φ_{ad}^{tan} is convex and closed in $L^2(\Omega)^N$ and thus also closed in $\mathbb{X} \hookrightarrow L^2(\Omega)^N$. Since $\mathbf{0} \in \Phi_{ad}^{tan}$, we get that Φ_{ad}^{tan} is non-empty. Thus **(A2)** and **(A3)** hold. From the estimate (164) we

also see that **(A4)** holds.
For **(A6)** we have to show that the translated j is bounded from below on Φ_{ad}^{tan}, i.e. that the original j is bounded from below on Φ_{ad}. First consider the Ginzburg-Landau energy

$$E(\boldsymbol{\varphi}) = \int_\Omega \frac{\varepsilon}{2}|\nabla\boldsymbol{\varphi}|^2 + \frac{1}{\varepsilon}\psi_0(\boldsymbol{\varphi})\,\mathrm{d}x.$$

The first term obviously is non-negative. Since ψ_0 is continuous, it is bounded from below on the compact set $\Delta^{N-1} \subset \mathbb{R}^N$. Thus $\psi_0(\boldsymbol{\varphi})$ is bounded from below for all $\boldsymbol{\varphi} \in \Phi_{ad}$. The remaining term $F(\boldsymbol{\varphi}, S(\boldsymbol{\varphi}))$ in the reduced cost functional is bounded from below by assumption **(AP9)**.
In Theorem 6.25 we proved that j is two times Fréchet differentiable on the whole space $H^1(\Omega)^N \cap L^\infty(\Omega)^N$, hence in particular the translated j fulfills **(A5)**.
To see that the last assumption **(A7)** is fulfilled, let $\boldsymbol{\varphi} \in \Phi_{ad}^{tan}$ be arbitrary and let $(\boldsymbol{\varphi}_i)_i \subset \mathbb{X} \cap \mathbb{D}$ be a sequence. Let $\boldsymbol{\varphi}_i \to 0$ weakly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$. We have to show

$$\left\langle j'(\boldsymbol{\varphi} + \mathfrak{m}), \boldsymbol{\varphi}_i \right\rangle \to 0.$$

From Proposition 6.30 we know that

$$\begin{aligned}\left\langle j'(\boldsymbol{\varphi} + \mathfrak{m}), \boldsymbol{\varphi}_i \right\rangle = \gamma\varepsilon \int_\Omega \nabla(\boldsymbol{\varphi} + \mathfrak{m}) : \nabla\boldsymbol{\varphi}_i + \frac{\gamma}{\varepsilon}\int_\Omega \psi_0'(\boldsymbol{\varphi} + \mathfrak{m})\boldsymbol{\varphi}_i \\ + \langle F_{\boldsymbol{\varphi}}(\boldsymbol{\varphi} + \mathfrak{m}, \boldsymbol{u}), \boldsymbol{\varphi}_i\rangle - \int_\Omega (\nabla \boldsymbol{C}(\boldsymbol{\varphi} + \mathfrak{m})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{p})) \cdot \boldsymbol{\varphi}_i \\ + \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi} + \mathfrak{m})^T \boldsymbol{p} \cdot \boldsymbol{\varphi}_i + \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi} + \mathfrak{m})^T \boldsymbol{p} \cdot \boldsymbol{\varphi}_i,\end{aligned}$$

where $\boldsymbol{u} = S(\boldsymbol{\varphi}+\mathfrak{m})$ and $\boldsymbol{p}$ is the adjoint state corresponding to $\boldsymbol{\varphi}+\mathfrak{m}$ and $\boldsymbol{u}$. We can pass to the limit in the first two terms since $\boldsymbol{\varphi}_i \to \mathbf{0}$ weakly in $H^1(\Omega)^N$ and weakly in $L^2(\Omega)^N$. The third term converges due to assumption **(PGC1)**. Since $\nabla \boldsymbol{C}(\boldsymbol{\varphi}+\mathfrak{m})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{p}) \in L^1(\Omega)^N$, $\boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi} + \mathfrak{m})^T\boldsymbol{p} \in L^1(\Omega)^N$ and $\boldsymbol{\varphi}_i \to \mathbf{0}$ weakly-* in $(L^1(\Omega)^N)^*$, we can pass to the limit in the fourth and fifth term. For the last term we observe that $\boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi} + \mathfrak{m}) \in L^\infty(\Gamma_g)^{d\times N}$ due to assumption **(PGC2)**, thus $\boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi} + \mathfrak{m})^T\boldsymbol{p} \in L^2(\Gamma_g)^N$. From $\boldsymbol{\varphi}_i \to \mathbf{0}$ weakly in $H^1(\Omega)^N$, we get for the traces that $\boldsymbol{\varphi}_i \to \mathbf{0}$ weakly in $L^2(\partial\Omega)^N$ and hence we can pass to the limit in the last term. □

The same holds for the topology optimization problem using scalar valued phase fields. Again one has to translate the admissible set to

$$\widetilde{\Phi_{ad}^{tan}} := \left\{\varphi \in H^1(\Omega) \;\middle|\; -1 - \mathfrak{m} \le \varphi \le 1 - \mathfrak{m} \text{ a.e. in } \Omega,\ \int_\Omega \varphi = 0\right\}$$

in order to use the spaces

$$\begin{aligned}\tilde{\mathbb{X}} &:= \left\{\varphi \in H^1(\Omega) \;\middle|\; \int_\Omega \varphi = 0\right\} \\ \tilde{\mathbb{D}} &:= L^\infty(\Omega)\end{aligned}$$

for which the assumptions can be shown. We can apply Lemma 6.47 for $N = 1$ to show that $\tilde{\mathbb{X}}$ and $\tilde{\mathbb{D}}$ fulfill the assumptions on the spaces **(A1)**.

Lemma 6.49. *In addition to the standard assumptions in Section 6.1.1 let **(PGC1)** and **(PGC2)** hold. Then the translated j for the scalar valued problem* (121) *together with $\widetilde{\Phi_{ad}^{tan}}$ fulfills the assumptions on the problem **(A2)-(A7)**.*

Proof. The same as for Lemma 6.48. Note that $\delta\varphi \in \tilde{\mathbb{X}} \cap \tilde{\mathbb{D}}$ converges weakly in $\tilde{\mathbb{X}}$ and weakly-* in $\tilde{\mathbb{D}}$ if and only if $T'(\varphi)\delta\varphi = (\frac{1}{2}\delta\varphi, -\frac{1}{2}\delta\varphi)^T$ converges weakly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$. □

We don't want to do the translation to the problem (162) defined on Φ_{ad}^{tan} and back again every time. Thus we define the method on the original untranslated problem by first translating the problem to (162), then applying the method on this problem and finally translating the iterates back. This is no problem since the VMPT method is translation invariant, see Theorem 4.27. With this convention, we have that the iterates $\boldsymbol{\varphi}_k$ of the method are elements of $\Phi_{ad} \subset H^1(\Omega)^N$, whereas tangential vectors, such as the search directions $\boldsymbol{v}_k$, are elements of $H^1_{(0)}(\Omega)^N$.

We give examples of choices for the variable metric a_k in the VMPT method in the spaces $\mathbb{X}$ and $\mathbb{D}$ defined in (163), for which global convergence can be guaranteed. These inner products include:

$$a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2, \tag{165}$$

$$a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2, \tag{166}$$

$$a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2 + \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta u}_1) : \mathcal{E}(\boldsymbol{\delta u}_2) + \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta p}_1) : \mathcal{E}(\boldsymbol{\delta p}_2), \tag{167}$$

$$a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2 + \frac{\varepsilon}{\tau_k} \int_\Omega \boldsymbol{v}_1 \cdot \boldsymbol{v}_2, \tag{168}$$

$$a_k(v_1, v_2) = \gamma\varepsilon \int_\Omega \nabla v_1 \cdot \nabla v_2 + \left(\frac{\varepsilon}{\tau_k} - \frac{\gamma}{\varepsilon}\right) \int_\Omega v_1 v_2, \tag{169}$$

$$a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2 + \frac{1}{\tau_k}(\boldsymbol{v}_1, \boldsymbol{v}_2)_{H^{-1}} \tag{170}$$

The first choice is just the usual H^1 (semi) inner product. The second choice is endued with a $\gamma\varepsilon$-scaling, which will be motivated by interface thickness considerations in Section 6.12. In the third inner product $\boldsymbol{\delta u}_i$ are the linearized states and $\boldsymbol{\delta p}_i$ the linearized adjoint states corresponding to $\boldsymbol{\varphi}_k$ in direction $\boldsymbol{\delta\varphi} = \boldsymbol{v}_i$, $i = 1, 2$. The variable metric includes second order information, which can be motivated by the mean compliance problem as follows: For the mean compliance problem (112) the second order derivative of j is given as (cf. Theorem 6.44)

$$\begin{aligned} j''(\boldsymbol{\varphi})[\boldsymbol{\tau\varphi}, \boldsymbol{\delta\varphi}] = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{\tau\varphi} : \nabla \boldsymbol{\delta\varphi} + \frac{\gamma}{\varepsilon} \int_\Omega \psi_0''(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\boldsymbol{\tau\varphi} \\ - \int_\Omega (\boldsymbol{C}''(\boldsymbol{\varphi})\boldsymbol{\tau\varphi}\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{u})) \cdot \boldsymbol{\delta\varphi} - 2\int_\Omega (\nabla \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{\tau u}) : \mathcal{E}(\boldsymbol{u})) \cdot \boldsymbol{\delta\varphi} \\ + 2\int_\Omega \boldsymbol{f}_{\varphi,\varphi}(\boldsymbol{\varphi})[\boldsymbol{\tau\varphi}, \boldsymbol{\delta\varphi}] \cdot \boldsymbol{u} + 2\int_\Omega \boldsymbol{f}_\varphi(\boldsymbol{\varphi})^T \boldsymbol{\tau u} \cdot \boldsymbol{\delta\varphi} \\ + 2\int_{\Gamma_g} \boldsymbol{g}_{\varphi,\varphi}(\boldsymbol{\varphi})[\boldsymbol{\tau\varphi}, \boldsymbol{\delta\varphi}] \cdot \boldsymbol{u} + 2\int_{\Gamma_g} \boldsymbol{g}_\varphi(\boldsymbol{\varphi})^T \boldsymbol{\tau u} \cdot \boldsymbol{\delta\varphi}. \end{aligned}$$

On the other hand the inner product (167) can be reformulated to

$$\begin{aligned} a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2 - 2\int_\Omega (\nabla \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{u}_k) : \mathcal{E}(\boldsymbol{\delta u}_2)) \cdot \boldsymbol{v}_1 + 2\int_\Omega \boldsymbol{f}_\varphi(\boldsymbol{\varphi}_k)^T \boldsymbol{\delta u}_2 \cdot \boldsymbol{v}_1 \\ + 2\int_{\Gamma_g} \boldsymbol{g}_\varphi(\boldsymbol{\varphi}_k)^T \boldsymbol{\delta u}_2 \cdot \boldsymbol{v}_1 \end{aligned}$$

using the linearized state equation and the linearized adjoint equation. For the general calculation we refer to (235). Here we also used that it holds $\boldsymbol{u}_k = \boldsymbol{p}_k$ for the mean compliance problem. It can be observed that a_k coincides with certain terms in $j''(\boldsymbol{\varphi}_k)$. The terms are chosen such that a_k becomes positive definite. In fact, if $\boldsymbol{C}$, $\boldsymbol{g}$ and $\boldsymbol{f}$ depend linearly on $\boldsymbol{\varphi}$, then the corresponding second order derivatives in j'' vanish and we see that a_k coincides with $j''(\boldsymbol{\varphi}_k)$ up to the term $\frac{\gamma}{\varepsilon}\psi_0''(\boldsymbol{\varphi}_k)$, which is typically negative. Note that the second order inner product is in contrast to the first two inner products point based, i.e. it depends on the current iterate $\boldsymbol{\varphi}_k$. Moreover, it is only bounded in $H^1 \cap L^\infty$, whereas the first two inner products are bounded in H^1. Although the variable metric (167) is motivated by the mean compliance problem, we will see that the corresponding VMPT method converges for arbitrary cost functionals F.
The fourth and fifth choice of a_k stem from a pseudo time stepping of Allen-Cahn type with time step size τ_k, see Section 6.8. Note that we consider the fifth inner product only for scalar valued phase fields for simplicity. The last choice of a_k comes from a Cahn-Hilliard pseudo time stepping with time step size τ_k as will be also discussed in Section 6.8. For the definition of the H^{-1} inner product we refer to (161).

In the following we will show that each of the considered inner products fulfills the assumptions for global convergence of the VMPT method.

Lemma 6.50. *The metric*

$$a_k(\boldsymbol{p}, \boldsymbol{v}) = \int_\Omega \nabla \boldsymbol{p} : \nabla \boldsymbol{v}$$

*for all $\boldsymbol{p}$, $\boldsymbol{v} \in \mathbb{X} \cap \mathbb{D}$ and $k \in \mathbb{N}_0$ fulfills the assumptions **(A8)-(A12)** on the metric.*

Proof. Due to the Poincaré inequality for functions with vanishing mean value, there exists some $C > 0$, such that

$$C\|\boldsymbol{p}\|_{H^1}^2 \leq \int_\Omega |\nabla \boldsymbol{p}|^2 \quad \forall \boldsymbol{p} \in \mathbb{X} \cap \mathbb{D}$$

see e.g. [Alt12, p. 253]. Thus, **(A9)** and in particular **(A8)** is fulfilled. Because of

$$|a_k(\boldsymbol{p}, \boldsymbol{v})| \leq \|\boldsymbol{p}\|_{H^1} \|\boldsymbol{v}\|_{H^1} \leq \|\boldsymbol{p}\|_{\mathbb{X}\cap\mathbb{D}} \|\boldsymbol{v}\|_{\mathbb{X}\cap\mathbb{D}},$$

assumption **(A10)** is fulfilled. Since a_k is a continuous bilinear form on $\mathbb{X}$, which follows from the last estimate, we know that $a_k(\boldsymbol{\varphi}, \boldsymbol{p}_i) \to 0$ as $i \to \infty$ for all $\boldsymbol{\varphi} \in \Phi_{ad}^{tan}$ and each sequence $(\boldsymbol{p}_i)_i \subset \mathbb{X} \cap \mathbb{D}$ converging to $\mathbf{0}$ weakly in $\mathbb{X}$, resulting in **(A11)**. By the same argument we get that

$$a_{k_i}(\boldsymbol{p}_i, \boldsymbol{v}_i) = \int_\Omega \nabla \boldsymbol{p}_i : \nabla \boldsymbol{v}_i \to 0 \quad \text{as } i \to \infty$$

for each sequence $(\boldsymbol{p}_i)_i$ and $(\boldsymbol{v}_i)_i$ in $\mathbb{X}$, where $\boldsymbol{p}_i \to \boldsymbol{p}$ in H^1 for some $\boldsymbol{p} \in \mathbb{X} \cap \mathbb{D}$, and $\boldsymbol{v}_i \to \mathbf{0}$ in H^1, and for any subsequence a_{k_i}. Hence, also **(A12)** is fulfilled. □

Lemma 6.51. *The metric*

$$a_k(\boldsymbol{p}, \boldsymbol{v}) = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{p} : \nabla \boldsymbol{v}$$

*for all $\boldsymbol{p}$, $\boldsymbol{v} \in \mathbb{X} \cap \mathbb{D}$ and $k \in \mathbb{N}_0$ fulfills the assumptions **(A8)-(A12)** on the metric.*

Proof. The same as for Lemma 6.50. □

For the second order variable metric (167) we first of all we need some preparing lemmas.

Lemma 6.52. *Let **(PGC2)** hold. Let $\boldsymbol{\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$ and let $(\boldsymbol{\delta\varphi}_i)_i \subset H^1(\Omega)^N \cap L^\infty(\Omega)^N$ be a sequence with $\boldsymbol{\delta\varphi}_i \to \boldsymbol{\delta\varphi}$ weakly in H^1 and weakly-* in L^∞ for some $\boldsymbol{\delta\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$. Then it holds*

$$S'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}_i \to S'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi} \quad \text{weakly in } H^1_D.$$

Proof. First of all note that the statement is not trivial. From $S'(\boldsymbol{\varphi}) \in \mathcal{L}(H^1(\Omega)^N \cap L^\infty(\Omega)^N, H^1_D)$, we only get weak convergence of $S'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}_i$ if $\boldsymbol{\delta\varphi}_i \to \boldsymbol{\delta\varphi}$ weakly in $H^1(\Omega)^N \cap L^\infty(\Omega)^N$. We want to show that this also holds if $\boldsymbol{\delta\varphi}_i \to \boldsymbol{\delta\varphi}$ only weakly-* in $L^\infty(\Omega)^N$. From the linearity of $S'(\boldsymbol{\varphi})$ we can assume without loss of generality that $\boldsymbol{\delta\varphi} = 0$. Let $\boldsymbol{\delta u}_i := S'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}_i$. From the convergence of $(\boldsymbol{\delta\varphi}_i)_i$ we get that $(\boldsymbol{\delta\varphi}_i)_i$ is bounded in $H^1(\Omega)^N \cap L^\infty(\Omega)^N$ and thus, by

$$\|\boldsymbol{\delta u}_i\|_{H^1} = \|S'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}_i\|_{H^1} \le \|S'(\boldsymbol{\varphi})\|_{\mathcal{L}(H^1\cap L^\infty, H^1)} \|\boldsymbol{\delta\varphi}_i\|_{H^1\cap L^\infty}$$

we get also the boundedness of $(\boldsymbol{\delta u}_i)_i$ in H^1_D. Hence we can extract a subsequence, which we denote by $(\boldsymbol{\delta u}_i)_i$, for which it holds $\boldsymbol{\delta u}_i \to \boldsymbol{\delta u}$ weakly in H^1 for some $\boldsymbol{\delta u} \in H^1_D$. Now for all i it holds the linearized state equation (124)

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{\delta u}_i) : \mathcal{E}(\boldsymbol{\xi}) = -\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}_i\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) + \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}_i \cdot \boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}_i \cdot \boldsymbol{\xi}$$

for all $\boldsymbol{\xi} \in H^1_D$. We want to pass to the limit in this equation. Clearly, we have $\mathcal{E}(\boldsymbol{\delta u}_i) \to \mathcal{E}(\boldsymbol{\delta u})$ weakly in L^2 and thus

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{\delta u}_i) : \mathcal{E}(\boldsymbol{\xi}) \to \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{\delta u}) : \mathcal{E}(\boldsymbol{\xi}).$$

As in the proof of Lemma 6.48 we can conclude from **(PGC2)** that

$$-\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}_i\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) + \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}_i \cdot \boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}_i \cdot \boldsymbol{\xi} \to 0.$$

Thus we have $\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{\delta u}) : \mathcal{E}(\boldsymbol{\xi}) = 0$ for all $\boldsymbol{\xi} \in H^1_D$ and by the coercivity we get

$$\|\boldsymbol{\delta u}\|^2_{H^1} \le C \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{\delta u}) : \mathcal{E}(\boldsymbol{\delta u}) = 0.$$

Up to now we proved that $\boldsymbol{\delta u}_i \to \mathbf{0}$ weakly in H^1 for a subsequence. But since the same argument can be repeated for any subsequence, we get $\boldsymbol{\delta u}_i \to \mathbf{0}$ weakly in H^1 for the whole sequence, see Lemma 7.3. □

Lemma 6.53. *Let **(PGC1)** hold. Then for each $\boldsymbol{\varphi} \in \Phi_{ad}$, $\boldsymbol{u} \in H^1_D$ and $\boldsymbol{\delta u} \in H^1_D$ and for each sequence $(\boldsymbol{\varphi}_i)_i \subset H^1(\Omega)^N \cap L^\infty(\Omega)^N$ with $\boldsymbol{\varphi}_i \to \mathbf{0}$ weakly in $H^1(\Omega)^N$ and weakly-* in $L^\infty(\Omega)^N$ it holds*

$$F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}, \boldsymbol{u})[\boldsymbol{\varphi}_i, \boldsymbol{\delta u}] \to 0.$$

Proof. The proof is the same as for Lemma 4.37 by taking difference quotients with respect to $\boldsymbol{u}$ and exploiting that $Y \subset (H^1(\Omega)^N \cap L^\infty(\Omega)^N)^*$ (defined in Lemma 4.36) is closed and **(PGC1)**. Note that it holds $F_{\boldsymbol{u},\boldsymbol{\varphi}} = F_{\boldsymbol{\varphi},\boldsymbol{u}}$ due to **(AP8)**. □

The next lemma shows that the adjoint state depends continuously on the data.

Lemma 6.54. *Let $(\boldsymbol{\varphi}_i)_i \subset H^1(\Omega)^N \cap L^\infty(\Omega)^N$ be a sequence with $\boldsymbol{\varphi}_i \to \boldsymbol{\varphi}$ in $H^1 \cap L^\infty$ for some $\boldsymbol{\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$. Moreover, let $(\boldsymbol{u}_i)_i \subset H^1_D$ be a sequence with $\boldsymbol{u}_i \to \boldsymbol{u}$ in H^1 for some $\boldsymbol{u} \in H^1_D$. Denote the solution of the adjoint equation* (133) *with data $(\boldsymbol{\varphi}_i, \boldsymbol{u}_i)$ by $\boldsymbol{p}_i$ and the adjoint state for the data $(\boldsymbol{\varphi}, \boldsymbol{u})$ by $\boldsymbol{p}$. Then it holds*

$$\boldsymbol{p}_i \to \boldsymbol{p} \quad \text{in } H^1.$$

Proof. First of all note that $\|\boldsymbol{p}_i\|_{H^1}$ is uniformly bounded. This follows from the a priori estimate (134) and the continuity of $F_{\boldsymbol{u}}$. By using the H^1_D-coercivity of the bilinear form of the adjoint equation, we obtain

$$\begin{aligned}\|\boldsymbol{p}_i - \boldsymbol{p}\|^2_{H^1} &\le C \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{p}_i - \boldsymbol{p}) : \mathcal{E}(\boldsymbol{p}_i - \boldsymbol{p}) \\ &\le C\Big(\int_\Omega (\boldsymbol{C}(\boldsymbol{\varphi}_i)\mathcal{E}(\boldsymbol{p}_i) - \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{p})) : \mathcal{E}(\boldsymbol{p}_i - \boldsymbol{p}) \\ &\quad + \int_\Omega (\boldsymbol{C}(\boldsymbol{\varphi}) - \boldsymbol{C}(\boldsymbol{\varphi}_i))\mathcal{E}(\boldsymbol{p}_i) : \mathcal{E}(\boldsymbol{p}_i - \boldsymbol{p})\Big) \end{aligned} \tag{171}$$

We insert the adjoint equation for $\boldsymbol{p}_i$ and $\boldsymbol{p}$, each tested by $\boldsymbol{\xi} = \boldsymbol{p}_i - \boldsymbol{p}$ to get

$$\begin{aligned}\int_\Omega (\boldsymbol{C}(\boldsymbol{\varphi}_i)\mathcal{E}(\boldsymbol{p}_i) - \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{p})) : \mathcal{E}(\boldsymbol{p}_i - \boldsymbol{p}) &= \langle F_{\boldsymbol{u}}(\boldsymbol{\varphi}_i, \boldsymbol{u}_i) - F_{\boldsymbol{u}}(\boldsymbol{\varphi}, \boldsymbol{u}), \boldsymbol{p}_i - \boldsymbol{p}\rangle \\ &\le \|F_{\boldsymbol{u}}(\boldsymbol{\varphi}_i, \boldsymbol{u}_i) - F_{\boldsymbol{u}}(\boldsymbol{\varphi}, \boldsymbol{u})\|_{(H^1)^*} \|\boldsymbol{p}_i - \boldsymbol{p}\|_{H^1}.\end{aligned}$$

Furthermore, we use Hölder's inequality to obtain

$$\int_\Omega (\boldsymbol{C}(\boldsymbol{\varphi}) - \boldsymbol{C}(\boldsymbol{\varphi}_i))\mathcal{E}(\boldsymbol{p}_i) : \mathcal{E}(\boldsymbol{p}_i - \boldsymbol{p}) \le \|\boldsymbol{C}(\boldsymbol{\varphi}) - \boldsymbol{C}(\boldsymbol{\varphi}_i)\|_{L^\infty} \|\boldsymbol{p}_i\|_{H^1} \|\boldsymbol{p}_i - \boldsymbol{p}\|_{H^1}$$

We insert these estimates into (171) and divide by $\|\boldsymbol{p}_i - \boldsymbol{p}\|_{H^1}$. This yields

$$\|\boldsymbol{p}_i - \boldsymbol{p}\|_{H^1} \le C(\|F_{\boldsymbol{u}}(\boldsymbol{\varphi}_i, \boldsymbol{u}_i) - F_{\boldsymbol{u}}(\boldsymbol{\varphi}, \boldsymbol{u})\|_{(H^1)^*} + \|\boldsymbol{C}(\boldsymbol{\varphi}) - \boldsymbol{C}(\boldsymbol{\varphi}_i)\|_{L^\infty} \|\boldsymbol{p}_i\|_{H^1}) \to 0,$$

where we used the continuity of $F_{\boldsymbol{u}}$ and $\boldsymbol{C}$, as well as the boundedness of $\|\boldsymbol{p}_i\|_{H^1}$. □

We show that also the linearized adjoint state depends continuously on the data.

Lemma 6.55. *Let $(\boldsymbol{\varphi}_i, \boldsymbol{u}_i, \boldsymbol{p}_i, \boldsymbol{\delta u}_i, \boldsymbol{\delta\varphi}_i)_i \subset (H^1(\Omega)^N \cap L^\infty(\Omega)^N) \times H^1_D \times H^1_D \times H^1_D \times (H^1(\Omega)^N \cap L^\infty(\Omega)^N)$ be sequences with $(\boldsymbol{\varphi}_i, \boldsymbol{u}_i, \boldsymbol{p}_i, \boldsymbol{\delta u}_i, \boldsymbol{\delta\varphi}_i) \to (\boldsymbol{\varphi}, \boldsymbol{u}, \boldsymbol{p}, \boldsymbol{\delta u}, \boldsymbol{\delta\varphi})$ strongly in the corresponding space for some $(\boldsymbol{\varphi}, \boldsymbol{u}, \boldsymbol{p}, \boldsymbol{\delta u}, \boldsymbol{\delta\varphi})$. Denote the solution of the linearized adjoint equation* (155) *with data $(\boldsymbol{\varphi}_i, \boldsymbol{u}_i, \boldsymbol{p}_i, \boldsymbol{\delta u}_i, \boldsymbol{\delta\varphi}_i)$ by $\boldsymbol{\delta p}_i$ and the linearized adjoint state for the data $(\boldsymbol{\varphi}, \boldsymbol{u}, \boldsymbol{p}, \boldsymbol{\delta u}, \boldsymbol{\delta\varphi})$ by $\boldsymbol{\delta p}$. Then it holds*

$$\boldsymbol{\delta p}_i \to \boldsymbol{\delta p} \quad \text{in } H^1.$$

Proof. By the a priori estimate (156) for the linearized adjoint equation, we get

$$\begin{aligned}\|\boldsymbol{\delta p}_i\|_{H^1} \le C(&\|F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}_i, \boldsymbol{u}_i)\|_{\mathcal{L}(H^1 \cap L^\infty, (H^1)^*)} \|\boldsymbol{\delta\varphi}_i\|_{H^1 \cap L^\infty} + \|F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi}_i, \boldsymbol{u}_i)\|_{\mathcal{L}(H^1, (H^1)^*)} \|\boldsymbol{\delta u}_i\|_{H^1} \\ &+ \|\boldsymbol{C}'(\boldsymbol{\varphi}_i)\|_{L^\infty} \|\boldsymbol{\delta\varphi}_i\|_{L^\infty} \|\boldsymbol{p}_i\|_{H^1}) \le C,\end{aligned}$$

since the operators $F_{\boldsymbol{u},\boldsymbol{\varphi}}$, $F_{\boldsymbol{u},\boldsymbol{u}}$ and $\boldsymbol{C}'$ are continuous in the respective spaces. Analog to

the proof of Lemma 6.54, see (171), we use the coercivity to obtain

$$\begin{aligned}\|\boldsymbol{\delta p}_i - \boldsymbol{\delta p}\|_{H^1}^2 \le C\Big(&\int_\Omega (\boldsymbol{C}(\boldsymbol{\varphi}_i)\mathcal{E}(\boldsymbol{\delta p}_i) - \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{\delta p})) : \mathcal{E}(\boldsymbol{\delta p}_i - \boldsymbol{\delta p}) \\ &+ \int_\Omega (\boldsymbol{C}(\boldsymbol{\varphi}) - \boldsymbol{C}(\boldsymbol{\varphi}_i))\mathcal{E}(\boldsymbol{\delta p}_i) : \mathcal{E}(\boldsymbol{\delta p}_i - \boldsymbol{\delta p})\Big)\end{aligned} \tag{172}$$

We test the linearized adjoint equation for $\boldsymbol{\delta p}_i$ and $\boldsymbol{\delta p}$, respectively, by $\boldsymbol{\xi} = \boldsymbol{\delta p}_i - \boldsymbol{\delta p}$ to get for the first term

$$\begin{aligned}&\int_\Omega (\boldsymbol{C}(\boldsymbol{\varphi}_i)\mathcal{E}(\boldsymbol{\delta p}_i) - \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{\delta p})) : \mathcal{E}(\boldsymbol{\delta p}_i - \boldsymbol{\delta p}) \\ &\quad = F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}_i, \boldsymbol{u}_i)[\boldsymbol{\delta\varphi}_i, \boldsymbol{\delta p}_i - \boldsymbol{\delta p}] - F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}, \boldsymbol{u})[\boldsymbol{\delta\varphi}, \boldsymbol{\delta p}_i - \boldsymbol{\delta p}] \\ &\qquad + F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi}_i, \boldsymbol{u}_i)[\boldsymbol{\delta u}_i, \boldsymbol{\delta p}_i - \boldsymbol{\delta p}] - F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi}, \boldsymbol{u})[\boldsymbol{\delta u}, \boldsymbol{\delta p}_i - \boldsymbol{\delta p}] \\ &\qquad - \Big(\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}_i)\boldsymbol{\delta\varphi}_i\mathcal{E}(\boldsymbol{p}_i) : \mathcal{E}(\boldsymbol{\delta p}_i - \boldsymbol{\delta p}) - \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\mathcal{E}(\boldsymbol{p}) : \mathcal{E}(\boldsymbol{\delta p}_i - \boldsymbol{\delta p})\Big).\end{aligned}$$

For the first term therein we obtain

$$\begin{aligned}&F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}_i, \boldsymbol{u}_i)[\boldsymbol{\delta\varphi}_i, \boldsymbol{\delta p}_i - \boldsymbol{\delta p}] - F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}, \boldsymbol{u})[\boldsymbol{\delta\varphi}, \boldsymbol{\delta p}_i - \boldsymbol{\delta p}] \\ &\quad \le |(F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}_i, \boldsymbol{u}_i) - F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}, \boldsymbol{u}))[\boldsymbol{\delta\varphi}_i, \boldsymbol{\delta p}_i - \boldsymbol{\delta p}]| + |F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}, \boldsymbol{u})[\boldsymbol{\delta\varphi}_i - \boldsymbol{\delta\varphi}, \boldsymbol{\delta p}_i - \boldsymbol{\delta p}]| \\ &\quad \le (\|F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}_i, \boldsymbol{u}_i) - F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}, \boldsymbol{u})\|_{\mathcal{L}(H^1\cap L^\infty, (H^1_D)^*)}\|\boldsymbol{\delta\varphi}_i\|_{H^1\cap L^\infty} \\ &\qquad + \|F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}, \boldsymbol{u})\|_{\mathcal{L}(H^1\cap L^\infty, (H^1_D)^*)}\|\boldsymbol{\delta\varphi}_i - \boldsymbol{\delta\varphi}\|_{H^1\cap L^\infty})\|\boldsymbol{\delta p}_i - \boldsymbol{\delta p}\|_{H^1}.\end{aligned}$$

The same estimate can be derived for the second term

$$\begin{aligned}&F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi}_i, \boldsymbol{u}_i)[\boldsymbol{\delta u}_i, \boldsymbol{\delta p}_i - \boldsymbol{\delta p}] - F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi}, \boldsymbol{u})[\boldsymbol{\delta u}, \boldsymbol{\delta p}_i - \boldsymbol{\delta p}] \\ &\qquad \le (\|F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi}_i, \boldsymbol{u}_i) - F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi}, \boldsymbol{u})\|_{\mathcal{L}(H^1_D, (H^1_D)^*)}\|\boldsymbol{\delta u}_i\|_{H^1} \\ &\qquad\quad + \|F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi}, \boldsymbol{u})\|_{\mathcal{L}(H^1_D, (H^1_D)^*)}\|\boldsymbol{\delta u}_i - \boldsymbol{\delta u}\|_{H^1})\|\boldsymbol{\delta p}_i - \boldsymbol{\delta p}\|_{H^1}\end{aligned}$$

and similarly for the third term

$$\begin{aligned}&\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}_i)\boldsymbol{\delta\varphi}_i\mathcal{E}(\boldsymbol{p}_i) : \mathcal{E}(\boldsymbol{\delta p}_i - \boldsymbol{\delta p}) - \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\mathcal{E}(\boldsymbol{p}) : \mathcal{E}(\boldsymbol{\delta p}_i - \boldsymbol{\delta p}) \\ &\qquad \le (\|\boldsymbol{C}'(\boldsymbol{\varphi}_i) - \boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty}\|\boldsymbol{\delta\varphi}_i\|_{L^\infty}\|\boldsymbol{p}_i\|_{H^1} \\ &\qquad\quad + \|\boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty}\|\boldsymbol{\delta\varphi}_i - \boldsymbol{\delta\varphi}\|_{L^\infty}\|\boldsymbol{p}_i\|_{H^1} \\ &\qquad\quad + \|\boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty}\|\boldsymbol{\delta\varphi}\|_{L^\infty}\|\boldsymbol{p}_i - \boldsymbol{p}\|_{H^1})\|\boldsymbol{\delta p}_i - \boldsymbol{\delta p}\|_{H^1}.\end{aligned}$$

Now consider the second term in (172). By Hölder's inequality we get

$$\int_\Omega (\boldsymbol{C}(\boldsymbol{\varphi}) - \boldsymbol{C}(\boldsymbol{\varphi}_i))\mathcal{E}(\boldsymbol{\delta p}_i) : \mathcal{E}(\boldsymbol{\delta p}_i - \boldsymbol{\delta p})) \le \|\boldsymbol{C}(\boldsymbol{\varphi}) - \boldsymbol{C}(\boldsymbol{\varphi}_i)\|_{L^\infty}\|\boldsymbol{\delta p}_i\|_{H^1}\|\boldsymbol{\delta p}_i - \boldsymbol{\delta p}\|_{H^1}.$$

We insert all estimates in (172) and divide by $\|\boldsymbol{\delta p}_i - \boldsymbol{\delta p}\|_{H^1}$ and finally obtain

$$\begin{aligned}\|\boldsymbol{\delta p}_i - \boldsymbol{\delta p}\|_{H^1} \le C(&\|F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}_i,\boldsymbol{u}_i) - F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi},\boldsymbol{u})\|_{\mathcal{L}(H^1\cap L^\infty,(H^1_D)^*)}\|\boldsymbol{\delta\varphi}_i\|_{H^1\cap L^\infty}\\ &+ \|F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi},\boldsymbol{u})\|_{\mathcal{L}(H^1\cap L^\infty,(H^1_D)^*)}\|\boldsymbol{\delta\varphi}_i - \boldsymbol{\delta\varphi}\|_{H^1\cap L^\infty}\\ &+ \|F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi}_i,\boldsymbol{u}_i) - F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi},\boldsymbol{u})\|_{\mathcal{L}(H^1_D,(H^1_D)^*)}\|\boldsymbol{\delta u}_i\|_{H^1}\\ &+ \|F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi},\boldsymbol{u})\|_{\mathcal{L}(H^1_D,(H^1_D)^*)}\|\boldsymbol{\delta u}_i - \boldsymbol{\delta u}\|_{H^1}\\ &+ \|\boldsymbol{C}'(\boldsymbol{\varphi}_i) - \boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty}\|\boldsymbol{\delta\varphi}_i\|_{L^\infty}\|\boldsymbol{p}_i\|_{H^1}\\ &+ \|\boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty}\|\boldsymbol{\delta\varphi}_i - \boldsymbol{\delta\varphi}\|_{L^\infty}\|\boldsymbol{p}_i\|_{H^1}\\ &+ \|\boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty}\|\boldsymbol{\delta\varphi}\|_{L^\infty}\|\boldsymbol{p}_i - \boldsymbol{p}\|_{H^1}\\ &+ \|\boldsymbol{C}(\boldsymbol{\varphi}) - \boldsymbol{C}(\boldsymbol{\varphi}_i)\|_{L^\infty}\|\boldsymbol{\delta p}_i\|_{H^1}) \to 0.\end{aligned}$$

□

Now we are able to check the assumptions for the second order metric. Since the metric is point based we can either show the weak assumptions **(A8)-(A12)** directly or alternatively the sufficient conditions **(A8')-(A12')**. We decide to prove only the weaker assumptions since the estimates are shorter. However, also the stronger conditions **(A8')-(A12')** and in particular the continuity of $\boldsymbol{\varphi} \mapsto a_{\boldsymbol{\varphi}}$ can be shown combining the estimates in this section, which are uniform in the directions $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$.

Lemma 6.56. *Let **(PGC1)** and **(PGC2)** hold and let*

$$a_k(\boldsymbol{v}_1,\boldsymbol{v}_2) = \gamma\varepsilon\int_\Omega \nabla\boldsymbol{v}_1 : \nabla\boldsymbol{v}_2 + \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta u}_1) : \mathcal{E}(\boldsymbol{\delta u}_2) + \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta p}_1) : \mathcal{E}(\boldsymbol{\delta p}_2),$$

for all $\boldsymbol{v}_1$, $\boldsymbol{v}_2 \in \mathbb{X}\cap\mathbb{D}$ and $k \in \mathbb{N}_0$, where $\boldsymbol{\varphi}_k \in \Phi_{ad}$ is the iterate of the VMPT method on the untranslated problem (102) *in the kth step, $\boldsymbol{\delta u}_1 = S'(\boldsymbol{\varphi}_k)\boldsymbol{v}_1$, $\boldsymbol{\delta u}_2 = S'(\boldsymbol{\varphi}_k)\boldsymbol{v}_2$ and $\boldsymbol{\delta p}_1$ and $\boldsymbol{\delta p}_2$ are the solutions of the linearized adjoint equation* (155) *with data $(\boldsymbol{\varphi}_k, S(\boldsymbol{\varphi}_k), \boldsymbol{p}_k, \boldsymbol{\delta u}_1, \boldsymbol{v}_1)$ and $(\boldsymbol{\varphi}_k, S(\boldsymbol{\varphi}_k), \boldsymbol{p}_k, \boldsymbol{\delta u}_2, \boldsymbol{v}_2)$, respectively. Moreover, $\boldsymbol{p}_k$ is the solution of the adjoint equation* (133) *with data $(\boldsymbol{\varphi}_k, S(\boldsymbol{\varphi}_k))$.*
*Then a_k fulfills the assumptions **(A8)-(A12)** on the metric.*

Proof. One easily sees that a_k is symmetric and bilinear (note that $\boldsymbol{v}_i \mapsto \boldsymbol{\delta u}_i$ as well as $\boldsymbol{v}_i \mapsto \boldsymbol{\delta p}_i$ is linear). For **(A8)** and **(A9)**, consider

$$a_k(\boldsymbol{v}_1,\boldsymbol{v}_1) = \gamma\varepsilon\int_\Omega \nabla\boldsymbol{v}_1 : \nabla\boldsymbol{v}_1 + \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta u}_1) : \mathcal{E}(\boldsymbol{\delta u}_1) + \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta p}_1) : \mathcal{E}(\boldsymbol{\delta p}_1).$$

From Lemma 6.50 we know that $\int_\Omega \nabla\boldsymbol{v}_1 : \nabla\boldsymbol{v}_1 \ge C\|\boldsymbol{v}_1\|^2_{H^1}$ and from the coercivity of $\boldsymbol{C}(\boldsymbol{\varphi}_k)$ we get $\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta u}_1) : \mathcal{E}(\boldsymbol{\delta u}_1) \ge 0$ and $\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta p}_1) : \mathcal{E}(\boldsymbol{\delta p}_1) \ge 0$, thus

$$\|\boldsymbol{v}_1\|^2_{a_k} \ge C\|\boldsymbol{v}_1\|^2_{H^1}.$$

Next we turn to **(A10)**. Lemma 6.50 yields $\gamma\varepsilon\int_\Omega \nabla\boldsymbol{v}_1 : \nabla\boldsymbol{v}_2 \le C\|\boldsymbol{v}_1\|_{\mathbb{X}\cap\mathbb{D}}\|\boldsymbol{v}_2\|_{\mathbb{X}\cap\mathbb{D}}$. The estimate (104) gives us

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta u}_1) : \mathcal{E}(\boldsymbol{\delta u}_2) \le C\|\boldsymbol{\delta u}_1\|_{H^1}\|\boldsymbol{\delta u}_2\|_{H^1} \le C\|S'(\boldsymbol{\varphi}_k)\|^2_{\mathcal{L}(H^1\cap L^\infty,H^1)}\|\boldsymbol{v}_1\|_{\mathbb{X}\cap\mathbb{D}}\|\boldsymbol{v}_2\|_{\mathbb{X}\cap\mathbb{D}}$$

as well as

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta p}_1) : \mathcal{E}(\boldsymbol{\delta p}_2) \le C\|\boldsymbol{\delta p}_1\|_{H^1}\|\boldsymbol{\delta p}_2\|_{H^1}.$$

With the a priori estimate (156) we get for $i = 1, 2$

$$\|\boldsymbol{\delta p}_i\|_{H^1} \leq C(\|\boldsymbol{v}_i\|_{H^1 \cap L^\infty} + \|\boldsymbol{\delta u}_i\|_{H^1} + \|\boldsymbol{v}_i\|_{L^\infty}) \leq C\|\boldsymbol{v}_i\|_{\mathbb{X}\cap\mathbb{D}},$$

where $C > 0$ depends on $\boldsymbol{\varphi}_k$. We can conclude that there exists $C_k > 0$, such that

$$a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) \leq C_k\|\boldsymbol{v}_1\|_{\mathbb{X}\cap\mathbb{D}}\|\boldsymbol{v}_2\|_{\mathbb{X}\cap\mathbb{D}}.$$

To show **(A11)**, let $\boldsymbol{\varphi} \in \Phi_{ad}^{tan}$ and let $(\boldsymbol{v}_i)_i \subset \mathbb{X} \cap \mathbb{D}$ be a sequence with $\boldsymbol{v}_i \rightharpoonup \mathbf{0}$ weakly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$. Consider

$$a_k(\boldsymbol{\varphi}, \boldsymbol{v}_i) = \gamma\varepsilon \int_\Omega \nabla\boldsymbol{\varphi} : \nabla\boldsymbol{v}_i + \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta u}) : \mathcal{E}(\boldsymbol{\delta u}_i) + \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta p}) : \mathcal{E}(\boldsymbol{\delta p}_i),$$

where $\boldsymbol{\delta u}$ and $\boldsymbol{\delta p}$ belong to the direction $\boldsymbol{\varphi}$ and $\boldsymbol{\delta u}_i$ and $\boldsymbol{\delta p}_i$ belong to the direction $\boldsymbol{v}_i$. We show

$$a_k(\boldsymbol{\varphi}, \boldsymbol{v}_i) \to 0 \quad \text{as } i \to \infty.$$

For the first term we obviously have

$$\gamma\varepsilon \int_\Omega \nabla\boldsymbol{\varphi} : \nabla\boldsymbol{v}_i \to 0$$

because of the weak convergence in H^1. From Lemma 6.52 we get that $\boldsymbol{\delta u}_i = S'(\boldsymbol{\varphi}_k)\boldsymbol{v}_i \to \mathbf{0}$ weakly in H^1 as $i \to \infty$. This yields

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta u}) : \mathcal{E}(\boldsymbol{\delta u}_i) \to 0 \quad \text{as } i \to \infty.$$

For the last term we have to test the linearized adjoint equation (155) for $\boldsymbol{\delta p}_i$ by $\boldsymbol{\xi} = \boldsymbol{\delta p}$ to obtain

$$\begin{aligned}\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta p}_i) : \mathcal{E}(\boldsymbol{\delta p}) \\ = F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k, \boldsymbol{u}_k)[\boldsymbol{v}_i, \boldsymbol{\delta p}] + F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi}_k, \boldsymbol{u}_k)[\boldsymbol{\delta u}_i, \boldsymbol{\delta p}] - \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}_k)\boldsymbol{v}_i\mathcal{E}(\boldsymbol{p}_k) : \mathcal{E}(\boldsymbol{\delta p})\end{aligned}$$

with $\boldsymbol{u}_k := S(\boldsymbol{\varphi}_k)$. Due to Lemma 6.53 we have $F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k, \boldsymbol{u}_k)[\boldsymbol{v}_i, \boldsymbol{\delta p}] \to 0$ as $i \to \infty$. From $\boldsymbol{\delta u}_i \rightharpoonup \mathbf{0}$ weakly in H^1 and $F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi}_k, \boldsymbol{u}_k)[\,.\,, \boldsymbol{\delta p}] \in (H^1_D)^*$ we conclude

$$F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi}_k, \boldsymbol{u}_k)[\boldsymbol{\delta u}_i, \boldsymbol{\delta p}] \to 0 \quad \text{as } i \to \infty.$$

For the last term we note that $\nabla\boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{p}_k) : \mathcal{E}(\boldsymbol{\delta p}) \in L^1(\Omega)^N$, thus weak-* convergence of $\boldsymbol{v}_i$ in L^∞ leads to

$$\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}_k)\boldsymbol{v}_i\mathcal{E}(\boldsymbol{p}_k) : \mathcal{E}(\boldsymbol{\delta p}) \to 0 \quad \text{as } i \to \infty.$$

It remains to show the last assumption **(A12)**. Therefor, consider a subsequence $(\boldsymbol{\varphi}_{k_i})_i$ with $\boldsymbol{\varphi}_{k_i} \to \boldsymbol{\varphi}$ in $H^1 \cap L^\infty$ for some $\boldsymbol{\varphi} \in \Phi_{ad}$. Moreover, let $(\boldsymbol{v}_i^1)_i, (\boldsymbol{v}_i^2)_i \subset \mathbb{X} \cap \mathbb{D}$ be sequences with $\boldsymbol{v}_i^2 \rightharpoonup \mathbf{0}$ strongly in $\mathbb{X}$ and weakly-* in $\mathbb{D}$ and $\boldsymbol{v}_i^1 \to \boldsymbol{v}^1$ strongly in $\mathbb{X} \cap \mathbb{D}$ for some $\boldsymbol{v}^1 \in \mathbb{X} \cap \mathbb{D}$. We show

$$a_{k_i}(\boldsymbol{v}_i^1, \boldsymbol{v}_i^2) \to 0.$$

In Lemma 6.50 we already proved that

$$\gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_i^1 : \nabla \boldsymbol{v}_i^2 \to 0.$$

We define $\boldsymbol{\delta u}_i^j := S'(\boldsymbol{\varphi}_{k_i})\boldsymbol{v}_i^j$, $j = 1, 2$. We test the linearized state equation (124) for $\boldsymbol{\delta u}_i^2$ by $\boldsymbol{\xi} = \boldsymbol{\delta u}_i^1 \in H_D^1$ to obtain

$$\begin{aligned}\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_{k_i})\mathcal{E}(\boldsymbol{\delta u}_i^2) : \mathcal{E}(\boldsymbol{\delta u}_i^1) = &- \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}_{k_i})\boldsymbol{v}_i^2\mathcal{E}(\boldsymbol{u}_{k_i}) : \mathcal{E}(\boldsymbol{\delta u}_i^1) + \int_\Omega \boldsymbol{f}_\varphi(\boldsymbol{\varphi}_{k_i})\boldsymbol{v}_i^2 \cdot \boldsymbol{\delta u}_i^1 \\ &+ \int_{\Gamma_g} \boldsymbol{g}_\varphi(\boldsymbol{\varphi}_{k_i})\boldsymbol{v}_i^2 \cdot \boldsymbol{\delta u}_i^1\end{aligned} \tag{173}$$

We have that $\boldsymbol{\delta u}_i^1 \to \boldsymbol{\delta u}^1 := S'(\boldsymbol{\varphi})\boldsymbol{v}^1$ in H^1, since it holds the estimate

$$\begin{aligned}&\|\boldsymbol{\delta u}_i^1 - \boldsymbol{\delta u}^1\|_{H^1} \\ &\quad = \|S'(\boldsymbol{\varphi}_{k_i})\boldsymbol{v}_i^1 - S'(\boldsymbol{\varphi})\boldsymbol{v}^1\|_{H^1} \le \|(S'(\boldsymbol{\varphi}_{k_i}) - S'(\boldsymbol{\varphi}))\boldsymbol{v}_i^1\|_{H^1} + \|S'(\boldsymbol{\varphi})(\boldsymbol{v}_i^1 - \boldsymbol{v}^1)\|_{H^1} \\ &\quad \le \underbrace{\|S'(\boldsymbol{\varphi}_{k_i}) - S'(\boldsymbol{\varphi})\|_{\mathcal{L}(H^1\cap L^\infty, H^1)}}_{\to 0} \underbrace{\|\boldsymbol{v}_i^1\|_{H^1\cap L^\infty}}_{\le C} + \|S'(\boldsymbol{\varphi})\|_{\mathcal{L}(H^1\cap L^\infty, H^1)} \underbrace{\|\boldsymbol{v}_i^1 - \boldsymbol{v}^1\|_{H^1\cap L^\infty}}_{\to 0}.\end{aligned}$$

Moreover we have

$$\begin{aligned}\boldsymbol{C}'(\boldsymbol{\varphi}_{k_i}) &\to \boldsymbol{C}'(\boldsymbol{\varphi}) && \text{in } L^\infty \\ \boldsymbol{u}_{k_i} &\to \boldsymbol{u} := S(\boldsymbol{\varphi}) && \text{in } H^1 \\ \boldsymbol{f}_\varphi(\boldsymbol{\varphi}_{k_i}) &\to \boldsymbol{f}_\varphi(\boldsymbol{\varphi}) && \text{in } L^2 \\ \boldsymbol{g}_\varphi(\boldsymbol{\varphi}_{k_i}) &\to \boldsymbol{g}_\varphi(\boldsymbol{\varphi}) && \text{in } L^2\end{aligned}$$

due to continuity of the respective operators. We pass to the limit in (173). We apply the triangle inequality to the first term to obtain

$$\begin{aligned}&\left|\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}_{k_i})\boldsymbol{v}_i^2\mathcal{E}(\boldsymbol{u}_{k_i}) : \mathcal{E}(\boldsymbol{\delta u}_i^1)\right| \\ &\le \left|\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{v}_i^2\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\delta u}^1)\right| + \left|\int_\Omega (\boldsymbol{C}'(\boldsymbol{\varphi}_{k_i}) - \boldsymbol{C}'(\boldsymbol{\varphi}))\boldsymbol{v}_i^2\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\delta u}^1)\right| \\ &\quad + \left|\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}_{k_i})\boldsymbol{v}_i^2\mathcal{E}(\boldsymbol{u}_{k_i} - \boldsymbol{u}) : \mathcal{E}(\boldsymbol{\delta u}^1)\right| + \left|\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}_{k_i})\boldsymbol{v}_i^2\mathcal{E}(\boldsymbol{u}_{k_i}) : \mathcal{E}(\boldsymbol{\delta u}_i^1 - \boldsymbol{\delta u}^1)\right| \\ &\le \underbrace{\left|\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{v}_i^2\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\delta u}^1)\right|}_{\to 0} + \underbrace{\|\boldsymbol{C}'(\boldsymbol{\varphi}_{k_i}) - \boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty}}_{\to 0} \|\boldsymbol{v}_i^2\|_{L^\infty}\|\boldsymbol{u}\|_{H^1}\|\boldsymbol{\delta u}^1\|_{H^1} \\ &\quad + \|\boldsymbol{C}'(\boldsymbol{\varphi}_{k_i})\|_{L^\infty}\|\boldsymbol{v}_i^2\|_{L^\infty} \underbrace{\|\boldsymbol{u}_{k_i} - \boldsymbol{u}\|_{H^1}}_{\to 0} \|\boldsymbol{\delta u}^1\|_{H^1} + \|\boldsymbol{C}'(\boldsymbol{\varphi}_{k_i})\|_{L^\infty}\|\boldsymbol{v}_i^2\|_{L^\infty}\|\boldsymbol{u}_{k_i}\|_{H^1} \underbrace{\|\boldsymbol{\delta u}_i^1 - \boldsymbol{\delta u}^1\|_{H^1}}_{\to 0},\end{aligned}$$

where the first term converges because of the weak-* convergence of v_i^2 in L^∞. The other norms in the estimate stay bounded. Consider the second term in (173). Again by triangle

inequality we get

$$\begin{aligned}
&\left|\int_\Omega \boldsymbol{f}_\varphi(\boldsymbol{\varphi}_{k_i})\boldsymbol{v}_i^2\cdot\boldsymbol{\delta u}_i^1\right| \\
&\le \left|\int_\Omega \boldsymbol{f}_\varphi(\boldsymbol{\varphi})\boldsymbol{v}_i^2\cdot\boldsymbol{\delta u}^1\right| + \left|\int_\Omega (\boldsymbol{f}_\varphi(\boldsymbol{\varphi}_{k_i})-\boldsymbol{f}_\varphi(\boldsymbol{\varphi}))\boldsymbol{v}_i^2\cdot\boldsymbol{\delta u}^1\right| + \left|\int_\Omega \boldsymbol{f}_\varphi(\boldsymbol{\varphi}_{k_i})\boldsymbol{v}_i^2\cdot(\boldsymbol{\delta u}_i^1-\boldsymbol{\delta u}^1)\right| \\
&\le \underbrace{\left|\int_\Omega \boldsymbol{f}_\varphi(\boldsymbol{\varphi})\boldsymbol{v}_i^2\cdot\boldsymbol{\delta u}^1\right|}_{\to 0} + \underbrace{\|\boldsymbol{f}_\varphi(\boldsymbol{\varphi}_{k_i})-\boldsymbol{f}_\varphi(\boldsymbol{\varphi})\|_{L^2}}_{\to 0}\|\boldsymbol{v}_i^2\|_{L^\infty}\|\boldsymbol{\delta u}^1\|_{L^2} \\
&\quad + \|\boldsymbol{f}_\varphi(\boldsymbol{\varphi}_{k_i})\|_{L^2}\|\boldsymbol{v}_i^2\|_{L^\infty}\underbrace{\|\boldsymbol{\delta u}_i^1-\boldsymbol{\delta u}^1\|_{L^2}}_{\to 0}
\end{aligned}$$

The third term in (173) can be handled differently because of the higher regularity of $\boldsymbol{g}_\varphi$ and we use the trace $H^1(\Omega)\to L^2(\partial\Omega)$.

$$\left|\int_{\Gamma_g}\boldsymbol{g}_\varphi(\boldsymbol{\varphi}_{k_i})\boldsymbol{v}_i^2\cdot\boldsymbol{\delta u}_i^1\right| \le \|\boldsymbol{g}_\varphi(\boldsymbol{\varphi}_{k_i})\|_{L^\infty(\Gamma_g)}\underbrace{\|\boldsymbol{v}_i^2\|_{H^1(\Omega)}}_{\to 0}\|\boldsymbol{\delta u}_i^1\|_{H^1(\Omega)}$$

It should be noted that $(\boldsymbol{g}_\varphi(\boldsymbol{\varphi}_{k_i}))_i$ is uniformly bounded in $L^\infty(\Gamma_g)$, since $\boldsymbol{\varphi}_{k_i}\in\Phi_{ad}$ for all i and **(PGC2)** holds. Summarizing, we proved

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_{k_i})\mathcal{E}(\boldsymbol{\delta u}_i^2):\mathcal{E}(\boldsymbol{\delta u}_i^1)\to 0,$$

which is the second term in the definition of a_k. It remains to show

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_{k_i})\mathcal{E}(\boldsymbol{\delta p}_i^1):\mathcal{E}(\boldsymbol{\delta p}_i^2)\to 0,$$

where $\boldsymbol{\delta p}_i^j$ is the solution of the linearized adjoint equation (155) with data $(\boldsymbol{\varphi}_{k_i},\boldsymbol{u}_{k_i},\boldsymbol{p}_{k_i},\boldsymbol{\delta u}_i^j,\boldsymbol{v}_i^j)$ for $j=1,2$. By Lemma 6.54 we get that $\boldsymbol{p}_{k_i}\to\boldsymbol{p}$ in H^1, where $\boldsymbol{p}$ is the adjoint state for $(\boldsymbol{\varphi},\boldsymbol{u})$. Using this, we obtain by Lemma 6.55 that $\boldsymbol{\delta p}_i^1\to\boldsymbol{\delta p}^1$ in H^1, where $\boldsymbol{\delta p}^1$ is the linearized adjoint state for $(\boldsymbol{\varphi},\boldsymbol{u},\boldsymbol{p},\boldsymbol{\delta u}^1,\boldsymbol{v}^1)$.
We prove $\boldsymbol{\delta u}_i^2\to 0$ weakly in H^1. We cannot apply Lemma 6.52 directly, since not only $\boldsymbol{v}_i^2$ varies, but also $\boldsymbol{\varphi}_{k_i}$. Let $l\in(H^1_D)^*$ be an arbitrary functional. Then

$$\begin{aligned}
\left\langle l,\boldsymbol{\delta u}_i^2\right\rangle &= \left\langle l,S'(\boldsymbol{\varphi}_{k_i})\boldsymbol{v}_i^2\right\rangle = \left\langle l,S'(\boldsymbol{\varphi})\boldsymbol{v}_i^2\right\rangle + \left\langle l,(S'(\boldsymbol{\varphi}_{k_i})-S'(\boldsymbol{\varphi}))\boldsymbol{v}_i^2\right\rangle \\
&\le \underbrace{\left\langle l,S'(\boldsymbol{\varphi})\boldsymbol{v}_i^2\right\rangle}_{\to 0} + \|l\|_{(H^1_D)^*}\underbrace{\|S'(\boldsymbol{\varphi}_{k_i})-S'(\boldsymbol{\varphi})\|_{\mathcal{L}(H^1\cap L^\infty,H^1)}}_{\to 0}\|\boldsymbol{v}_i^2\|_{H^1\cap L^\infty},
\end{aligned}$$

where the first term converges to zero due to Lemma 6.52.
We test the linearized adjoint equation (155) for $\boldsymbol{\delta p}_i^2$ by $\boldsymbol{\xi}=\boldsymbol{\delta p}_i^1$ to obtain

$$\begin{aligned}
\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_{k_i})\mathcal{E}(\boldsymbol{\delta p}_i^2):\mathcal{E}(\boldsymbol{\delta p}_i^1) &= F_{\boldsymbol{u},\varphi}(\boldsymbol{\varphi}_{k_i},\boldsymbol{u}_{k_i})[\boldsymbol{v}_i^2,\boldsymbol{\delta p}_i^1] + F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi}_{k_i},\boldsymbol{u}_{k_i})[\boldsymbol{\delta u}_i^2,\boldsymbol{\delta p}_i^1] \\
&\quad - \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}_{k_i})\boldsymbol{v}_i^2\mathcal{E}(\boldsymbol{p}_{k_i}):\mathcal{E}(\boldsymbol{\delta p}_i^1)
\end{aligned} \tag{174}$$

We show that each term on the right hand side vanishes as $i\to\infty$. As already done often in this proof, we use triangle inequality, Hölder's inequality and operator norm estimates,

thus we skip the intermediate steps and state only the results. For the first term we get

$$\begin{aligned}
&F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}_{k_i},\boldsymbol{u}_{k_i})[\boldsymbol{v}_i^2,\boldsymbol{\delta p}_i^1]\\
&\quad\le \|F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}_{k_i},\boldsymbol{u}_{k_i})-F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi},\boldsymbol{u})\|_{\mathcal{L}(H^1\cap L^\infty,(H^1_D)^*)}\|\boldsymbol{v}_i^2\|_{H^1\cap L^\infty}\|\boldsymbol{\delta p}_i^1\|_{H^1}\\
&\qquad+\|F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi},\boldsymbol{u})\|_{\mathcal{L}(H^1\cap L^\infty,(H^1_D)^*)}\|\boldsymbol{v}_i^2\|_{H^1\cap L^\infty}\|\boldsymbol{\delta p}_i^1-\boldsymbol{\delta p}^1\|_{H^1}+F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi},\boldsymbol{u})[\boldsymbol{v}_i^2,\boldsymbol{\delta p}^1]\to 0,
\end{aligned}$$

where we can pass to the limit in the last term due to Lemma 6.53. The second term in (174) yields

$$\begin{aligned}
&F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi}_{k_i},\boldsymbol{u}_{k_i})[\boldsymbol{\delta u}_i^2,\boldsymbol{\delta p}_i^1]\\
&\qquad\le \|F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi}_{k_i},\boldsymbol{u}_{k_i})-F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi},\boldsymbol{u})\|_{\mathcal{L}(H^1_D,(H^1_D)^*)}\|\boldsymbol{\delta u}_i^2\|_{H^1}\|\boldsymbol{\delta p}_i^1\|_{H^1}\\
&\qquad\quad+\|F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi},\boldsymbol{u})\|_{\mathcal{L}(H^1_D,(H^1_D)^*)}\|\boldsymbol{\delta u}_i^2\|_{H^1}\|\boldsymbol{\delta p}_i^1-\boldsymbol{\delta p}^1\|_{H^1}+F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi},\boldsymbol{u})[\boldsymbol{\delta u}_i^2,\boldsymbol{\delta p}^1]\to 0,
\end{aligned}$$

using the weak convergence of $\boldsymbol{\delta u}_i^2$ in H^1 for the last term. Finally, for the last term in (174) we obtain

$$\begin{aligned}
&\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}_{k_i})\boldsymbol{v}_i^2\mathcal{E}(\boldsymbol{p}_{k_i}):\mathcal{E}(\boldsymbol{\delta p}_i^1)\\
&\le \|\boldsymbol{C}'(\boldsymbol{\varphi}_{k_i})-\boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty}\|\boldsymbol{v}_i^2\|_{L^\infty}\|\boldsymbol{p}_{k_i}\|_{H^1}\|\boldsymbol{\delta p}_i^1\|_{H^1}+\|\boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty}\|\boldsymbol{v}_i^2\|_{L^\infty}\|\boldsymbol{p}_{k_i}-\boldsymbol{p}\|_{H^1}\|\boldsymbol{\delta p}_i^1\|_{H^1}\\
&\quad+\|\boldsymbol{C}'(\boldsymbol{\varphi})\|_{L^\infty}\|\boldsymbol{v}_i^2\|_{L^\infty}\|\boldsymbol{p}\|_{H^1}\|\boldsymbol{\delta p}_i^1-\boldsymbol{\delta p}^1\|_{H^1}+\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{v}_i^2\mathcal{E}(\boldsymbol{p}):\mathcal{E}(\boldsymbol{\delta p}^1)\to 0,
\end{aligned}$$

where we utilize the weak-* convergence of $\boldsymbol{v}_i^2$ in L^∞ for the last term. □

Now we turn to the fourth choice of inner product.

Lemma 6.57. *Let*

$$a_k(\boldsymbol{v}_1,\boldsymbol{v}_2)=\gamma\varepsilon\int_\Omega \nabla\boldsymbol{v}_1:\nabla\boldsymbol{v}_2+\frac{\varepsilon}{\tau_k}\int_\Omega \boldsymbol{v}_1\cdot\boldsymbol{v}_2$$

for all $\boldsymbol{v}_1,\ \boldsymbol{v}_2\in\mathbb{X}\cap\mathbb{D}$ and $k\in\mathbb{N}_0$ where $\tau_k\in[\tau_{min},\infty)$ for some $\tau_{min}>0$.
*Then a_k fulfills the assumptions **(A8)-(A12)** on the metric.*

Proof. Due to the Poincaré inequality for functions with vanishing mean value, there exists some $C>0$, such that

$$C\|\boldsymbol{v}\|_{H^1}^2\le\gamma\varepsilon\int_\Omega|\nabla\boldsymbol{v}|^2\le\|\boldsymbol{v}\|_{a_k}^2\le\max\left\{\gamma\varepsilon,\frac{\varepsilon}{\tau_{min}}\right\}\|\boldsymbol{v}\|_{H^1}^2\quad\forall\boldsymbol{v}\in\mathbb{X}\cap\mathbb{D},\ k\in\mathbb{N}_0.$$

Lemma 4.19 then proves the statement. □

Recall that we consider the next inner product only for scalar valued phase fields.

Lemma 6.58. *Let*

$$a_k(v_1,v_2)=\gamma\varepsilon\int_\Omega\nabla v_1\cdot\nabla v_2+\left(\frac{\varepsilon}{\tau_k}-\frac{\gamma}{\varepsilon}\right)\int_\Omega v_1v_2$$

for all $v_1,\ v_2\in\tilde{\mathbb{X}}\cap\tilde{\mathbb{D}}$ and $k\in\mathbb{N}_0$ where $\tau_k\in[\tau_{min},\varepsilon^2/\gamma]$, $0<\tau_{min}\le\varepsilon^2/\gamma$.
*Then a_k fulfills the assumptions **(A8)-(A12)** on the metric.*

Proof. As in the proof of Lemma 6.57 we get

$$C\|v\|_{H^1}^2 \leq \gamma\varepsilon \int_\Omega |\nabla v|^2 \leq \|v\|_{a_k}^2 \leq \max\left\{\gamma\varepsilon, \frac{\varepsilon}{\tau_{min}} - \frac{\gamma}{\varepsilon}\right\} \|v\|_{H^1}^2 \quad \forall v \in \tilde{\mathbb{X}} \cap \tilde{\mathbb{D}},\ k \in \mathbb{N}_0.$$

Note that the assumption $\tau_k \leq \varepsilon^2/\gamma$ is equivalent to $\left(\frac{\varepsilon}{\tau_k} - \frac{\gamma}{\varepsilon}\right) \geq 0$. □

The following lemma treats the Cahn-Hilliard type inner product.

Lemma 6.59. *Let*

$$a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2 + \frac{1}{\tau_k}(\boldsymbol{v}_1, \boldsymbol{v}_2)_{H^{-1}}$$

for all $\boldsymbol{v}_1$, $\boldsymbol{v}_2 \in \mathbb{X} \cap \mathbb{D}$ and $k \in \mathbb{N}_0$ where $\tau_k \in [\tau_{min}, \infty]$ for some $\tau_{min} > 0$.
*Then a_k fulfills the assumptions **(A8)-(A12)** on the metric.*

Proof. As in the proof of Lemma 6.57 we get

$$C\|\boldsymbol{v}\|_{H^1}^2 \leq \gamma\varepsilon \int_\Omega |\nabla \boldsymbol{v}|^2 \leq \|\boldsymbol{v}\|_{a_k}^2 \leq \max\left\{\gamma\varepsilon, \frac{1}{\tau_{min}}\right\} \|\boldsymbol{v}\|_{H^1}^2 \quad \forall \boldsymbol{v} \in \mathbb{X} \cap \mathbb{D},\ k \in \mathbb{N}_0,$$

where we use that the embedding $H^1_{(0)}(\Omega)^N \hookrightarrow (H^1_{(0)}(\Omega)^N)^*$, $\boldsymbol{\varphi} \mapsto \int_\Omega \boldsymbol{\varphi} \,.\, \mathrm{d}x$ is continuous. □

From Theorem 4.14 we get global convergence of the VMPT method applied to the topology optimization problem for any discussed choice of inner product a_k.

Corollary 6.60. *In addition to the standard assumptions in Section 6.1.1 let **(PGC1)** and **(PGC2)** hold. Let a_k be one of the inner products defined in* (165), (166), (167), (168), (169) *and* (170)*, where τ_k fulfills the assumptions in Lemma 6.57, Lemma 6.58 and Lemma 6.59, respectively. Moreover, let for the scaling parameter within the VMPT method hold $\lambda_{min} \leq \lambda_k \leq \lambda_{max}$ with $\lambda_{min} > 0$.*
Then all assumptions for global convergence are fulfilled and the statements in Theorem 4.14 apply. □

Finally we give some counterexamples for choices of a_k, which do not fulfill the assumptions. First of all, the choice

$$a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = j''(\boldsymbol{\varphi}_k)[\boldsymbol{v}_1, \boldsymbol{v}_2]$$

does not fit in the framework of VMPT methods, since the topology optimization problem is not convex and thus the second order derivative j'' is not necessarily positive definite. However, this choice fits in the context of Josephy-Newton methods, which will be discussed in Section 6.9. Only local convergence can be expected for this choice of a_k.
The second counterexample for a_k is the L^2 inner product

$$a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \int_\Omega \boldsymbol{v}_1 \cdot \boldsymbol{v}_2,$$

which is used often in the literature for the numerical treatment of optimal control problems, see [Trö09, KS92]. For the topology optimization problem, this choice of a_k will not be possible, since there are no spaces $\mathbb{X}$ and $\mathbb{D}$ such that the assumptions are fulfilled. Assume that such spaces exist. Then from **(A9)** it would follow that $L^2(\Omega)^N \hookrightarrow \mathbb{X}$, hence $L^2(\Omega)^N \cap \mathbb{D} \hookrightarrow \mathbb{X} \cap \mathbb{D}$ and thus j has to be differentiable in $L^2(\Omega)^N \cap \mathbb{D}$ due to **(A5)**.

If we assume that j is not differentiable in a better space than $H^1(\Omega)^N \cap L^\infty(\Omega)^N$, it is necessary that $L^2(\Omega)^N \cap \mathbb{D} \hookrightarrow H^1(\Omega)^N \cap L^\infty(\Omega)^N$. Since Φ_{ad} is not bounded in H^1, but bounded in L^2, we get that Φ_{ad} is unbounded in $\mathbb{D}$, hence **(A4)** is not fulfilled. This is an example for a metric, for which the discretized method is well defined, but not the method in the continuous setting. In Section 6.13.11 we will show numerically that the L^2 inner product will not give rise to a mesh independent method.

6.8 Global convergence of certain pseudo time stepping methods with adaptive time step sizes

In Section 5 we pointed out that most of the literature about phase field methods in structural topology optimization use a pseudo time stepping approach as a numerical solver. However, no convergence results are yet available for these methods. In this section we show global convergence for the pseudo time stepping approaches used in [BGS+12, BFGS14] with minor changes, which stem from a gradient flow dynamic of Allen-Cahn and Cahn-Hilliard type. We show that the pseudo time stepping schemes are equivalent to a VMPT method for certain choices of inner products a_k and apply the developed convergence theory of the VMPT method. Moreover, using this equivalence, we are able to suggest a rigorous method for choosing the pseudo time step size τ_k, which gives rapid evolution in time but still preserves global convergence. In fact, numerical experiments in Section 6.13.7 show that τ_k can tend to infinity without destroying global convergence.
Moreover we introduce a rigorous stopping criterion, which also did not exist before.

Consider the pseudo time stepping method used in [BFGS14] for the minimization of the mean compliance or for the compliant mechanism problem. For simplicity we will focus here only on the mean compliance problem with $\boldsymbol{f} \equiv 0$ and $\boldsymbol{g}$ independent of $\boldsymbol{\varphi}$. Literally the same calculation can be carried out for the compliant mechanism problem.
In [BFGS14] an L^2 gradient flow is used, which fulfills for any $t > 0$

$$\boldsymbol{\varphi} \in \Phi_{ad}, \quad \varepsilon \int_\Omega \frac{\partial \boldsymbol{\varphi}}{\partial t}(\boldsymbol{\eta} - \boldsymbol{\varphi}) + \gamma\varepsilon \int_\Omega \nabla\boldsymbol{\varphi} : \nabla(\boldsymbol{\eta} - \boldsymbol{\varphi}) + \frac{\gamma}{\varepsilon} \int_\Omega \psi_0'(\boldsymbol{\varphi})(\boldsymbol{\eta} - \boldsymbol{\varphi}) \\ - \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})(\boldsymbol{\eta} - \boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{u}) \geq 0 \quad \forall \boldsymbol{\eta} \in \Phi_{ad}$$

As usual for Allen-Cahn type evolutions, time is scaled by the factor ε. This variational inequality is discretized semi-implicitly in time, where the gradient term is taken implicitly and the remaining terms explicitly. In [BFGS14] a fixed time step size τ is used, but we will allow varying time steps τ_k here. Let $\boldsymbol{\varphi}_k$ be the solution of the kth time step. The solution $\boldsymbol{\varphi}$ of the $(k+1)$st time step is then given by the variational inequality

$$\boldsymbol{\varphi} \in \Phi_{ad}, \quad \frac{\varepsilon}{\tau_k} \int_\Omega (\boldsymbol{\varphi} - \boldsymbol{\varphi}_k) \cdot (\boldsymbol{\eta} - \boldsymbol{\varphi}) + \gamma\varepsilon \int_\Omega \nabla\boldsymbol{\varphi} : \nabla(\boldsymbol{\eta} - \boldsymbol{\varphi}) + \frac{\gamma}{\varepsilon} \int_\Omega \psi_0'(\boldsymbol{\varphi}_k)(\boldsymbol{\eta} - \boldsymbol{\varphi}) \\ - \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}_k)(\boldsymbol{\eta} - \boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u}_k) : \mathcal{E}(\boldsymbol{u}_k) \geq 0 \quad \forall \boldsymbol{\eta} \in \Phi_{ad}, \quad (175)$$

where $\boldsymbol{u}_k := S(\boldsymbol{\varphi}_k)$. Using the metric a_k defined in (168), this VI is equivalent to

$$\boldsymbol{\varphi} \in \Phi_{ad}, \quad a_k(\boldsymbol{\varphi} - \boldsymbol{\varphi}_k, \boldsymbol{\eta} - \boldsymbol{\varphi}) + \gamma\varepsilon \int_\Omega \nabla\boldsymbol{\varphi}_k : \nabla(\boldsymbol{\eta} - \boldsymbol{\varphi}) + \frac{\gamma}{\varepsilon} \int_\Omega \psi_0'(\boldsymbol{\varphi}_k)(\boldsymbol{\eta} - \boldsymbol{\varphi}) \\ - \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}_k)(\boldsymbol{\eta} - \boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u}_k) : \mathcal{E}(\boldsymbol{u}_k) \geq 0 \quad \forall \boldsymbol{\eta} \in \Phi_{ad}$$

or

$$\boldsymbol{\varphi} \in \Phi_{ad}, \quad a_k(\boldsymbol{\varphi} - \boldsymbol{\varphi}_k, \boldsymbol{\eta} - \boldsymbol{\varphi}) + \left\langle j'(\boldsymbol{\varphi}_k), \boldsymbol{\eta} - \boldsymbol{\varphi} \right\rangle \geq 0 \quad \forall \boldsymbol{\eta} \in \Phi_{ad}$$

Comparing this to the variational inequality (26) of the projection type subproblem, we see that the time steps of this flow coincide with the iterates of the VMPT method using this specific inner product a_k and $\lambda_k = \alpha_k = 1$.

We propose Algorithm 6.1 for choosing the time step size τ_k.

Algorithm 6.1 Adaptive pseudo time stepping

1: Choose $\boldsymbol{\varphi}_0 \in \Phi_{ad}$, $0 < \beta < 1$, $0 < \sigma < 1$, $\tau_{min} > 0$ and $\tau_{-1} \geq \tau_{min}$.
2: $k := 0$.
3: **while** $k \leq k_{\max}$ **do**
4: Calculate the time step size τ_k by Armijo type backtracking, i.e. set $\tau_k := \beta^{m_k - 1}\tau_{k-1}$ where $m_k \in \mathbb{N}_0$ is the minimal power such that it holds

$$j(\boldsymbol{\varphi}_k + \boldsymbol{v}(\tau_k)) \leq j(\boldsymbol{\varphi}_k) + \sigma \left\langle j'(\boldsymbol{\varphi}_k), \boldsymbol{v}(\tau_k) \right\rangle,$$

where $\boldsymbol{v}(\tau_k) := \boldsymbol{\varphi}^*(\tau_k) - \boldsymbol{\varphi}_k$ and $\boldsymbol{\varphi}^*(\tau_k)$ is the solution of the VI (175), which depends on τ_k.
5: **if** such an m_k does not exist or $\tau_k < \tau_{min}$ **then**
6: $\tau_k := \tau_{min}$.
7: Calculate the step length $0 < \alpha_k \leq 1$ by Armijo backtracking in direction $\boldsymbol{v}(\tau_k)$, i.e. find the minimal power $m_k \in \mathbb{N}_0$ such that $\alpha_k := \beta^{m_k}$ fulfills

$$j(\boldsymbol{\varphi}_k + \alpha_k \boldsymbol{v}(\tau_k)) \leq j(\boldsymbol{\varphi}_k) + \alpha_k \sigma \left\langle j'(\boldsymbol{\varphi}_k), \boldsymbol{v}(\tau_k) \right\rangle.$$

8: **else**
9: $\alpha_k := 1$
10: **end if**
11: $\boldsymbol{v}_k := \boldsymbol{v}(\tau_k)$.
12: Update $\boldsymbol{\varphi}_{k+1} := \boldsymbol{\varphi}_k + \alpha_k \boldsymbol{v}_k$
13: **if** $\sqrt{\gamma\varepsilon}\|\nabla \boldsymbol{v}_k\|_{L^2} < tol$ **then**
14: **return**
15: **end if**
16: $k := k + 1$
17: **end while**

The idea is to start with a larger trial time step size $\tau_k = \beta^{-1}\tau_{k-1} > \tau_{k-1}$ than in the previous time step and then successively decrease it until the Armijo condition is fulfilled. We note that a similar update for τ_k is employed in [BC03]. However, since such a step size may not exist, we also include Armijo backtracking in α as backup, similar to the hybrid method in Section 4.6.2. Note that if τ_k already fulfills the Armijo condition, then $\alpha = 1$ automatically fulfills the Armijo condition, thus Algorithm 6.1 is a special instance of the

abstract VMPT method in Algorithm 4.1. We note that in all numerical experiments in Section 6.13.7 it is possible to find a time step size $\tau_k > \tau_{min}$ fulfilling the Armijo condition, thus the backtracking in α is never used.
Since the proposed method assures that $\tau_k \geq \tau_{min}$ we get together with Lemma 6.57 and Theorem 4.14 and Corollary 6.60, respectively, the following global convergence result.

Corollary 6.61. *Let $(\boldsymbol{\varphi}_k)_k$ denote the sequence of iterates generated by the pseudo time stepping algorithm 6.1, which stems from a time discrete L^2-gradient flow, where the gradient term is taken implicitly and the remaining terms explicitly in time. Let the initial guess $\boldsymbol{\varphi}_0 \in \Phi_{ad}$ be arbitrary. Then every accumulation point of $(\boldsymbol{\varphi}_k)_k$ in $H^1(\Omega)^N \cap L^\infty(\Omega)^N$ is a stationary point of j.* □

As already mentioned, in the numerical experiments in Section 6.13.7 it turns out that $\tau_k \to \infty$ if the proposed update scheme for τ_k is used. We show that in this case the iterates of the pseudo time stepping algorithm 6.1 approach the iterates of the VMPT method with $a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \varepsilon\gamma \int \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2$ and $\lambda_k = 1$, i.e. the scaled projected H^1-gradient method.

Lemma 6.62. *Let $\boldsymbol{\varphi}_k \in \Phi_{ad}$ be arbitrary. Denote by $\boldsymbol{\varphi}_\tau$ the solution of the projection type subproblem* (26) *for*

$$a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2 + \frac{\varepsilon}{\tau} \int_\Omega \boldsymbol{v}_1 \cdot \boldsymbol{v}_2,$$

and $\lambda_k = 1$ and denote by $\boldsymbol{\varphi}_\infty$ the solution of the projection type subproblem (26) *for*

$$a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2,$$

and $\lambda_k = 1$. Then $\boldsymbol{\varphi}_\tau \to \boldsymbol{\varphi}_\infty$ in H^1 as $\tau \to \infty$.

Proof. We test the VI (26) for $\boldsymbol{\varphi}_\tau$ by $\boldsymbol{\eta} = \boldsymbol{\varphi}_\infty$ and vice versa and add them up. The j' term drops out and we obtain

$$\gamma\varepsilon \int_\Omega \nabla(\boldsymbol{\varphi}_\infty - \boldsymbol{\varphi}_k) : \nabla(\boldsymbol{\varphi}_\tau - \boldsymbol{\varphi}_\infty) + \gamma\varepsilon \int_\Omega \nabla(\boldsymbol{\varphi}_\tau - \boldsymbol{\varphi}_k) : \nabla(\boldsymbol{\varphi}_\infty - \boldsymbol{\varphi}_\tau) + \frac{\varepsilon}{\tau} \int_\Omega (\boldsymbol{\varphi}_\tau - \boldsymbol{\varphi}_k) \cdot (\boldsymbol{\varphi}_\infty - \boldsymbol{\varphi}_\tau) \geq 0$$

Rearranging the terms and applying the Poincaré and Hölder inequality leads to

$$\|\boldsymbol{\varphi}_\infty - \boldsymbol{\varphi}_\tau\|_{H^1}^2 \leq \frac{C}{\tau} \int_\Omega (\boldsymbol{\varphi}_\tau - \boldsymbol{\varphi}_k) \cdot (\boldsymbol{\varphi}_\infty - \boldsymbol{\varphi}_\tau) \leq \frac{C}{\tau} \|\boldsymbol{\varphi}_\tau - \boldsymbol{\varphi}_k\|_{L^\infty} \|\boldsymbol{\varphi}_\infty - \boldsymbol{\varphi}_\tau\|_{H^1}$$

From $\|\boldsymbol{\varphi}_\tau - \boldsymbol{\varphi}_k\|_{L^\infty} \leq C$ we get

$$\|\boldsymbol{\varphi}_\infty - \boldsymbol{\varphi}_\tau\|_{H^1} \leq \frac{C}{\tau}$$

and thus the statement. □

In the presented time discretization, only the gradient term is taken implicitly. Other time discretizations are also possible. For instance in [BGS+12] in addition the potential is taken implicitly in time. As in [BGS+12] we restrict ourselves to the mean compliance problem and the scalar valued case with potential $\psi_0(\varphi) = \frac{1}{2}(1 - \varphi^2)$. Taking the potential

term $\psi_0'(\varphi) = -\varphi$ implicitly in time leads to the time step

$$\varphi \in \Phi_{ad}, \quad \frac{\varepsilon}{\tau_k}\int_\Omega (\varphi-\varphi_k)(\eta-\varphi) + \gamma\varepsilon\int_\Omega \nabla\varphi\cdot\nabla(\eta-\varphi) - \frac{\gamma}{\varepsilon}\int_\Omega \varphi(\eta-\varphi) \\ - \int_\Omega \boldsymbol{C}'(\varphi_k)(\eta-\varphi)\mathcal{E}(\boldsymbol{u}_k):\mathcal{E}(\boldsymbol{u}_k) \geq 0 \quad \forall \eta\in\Phi_{ad}, \quad (176)$$

which is equivalent to

$$\varphi \in \Phi_{ad}, \quad a_k(\varphi-\varphi_k,\eta-\varphi) + \gamma\varepsilon\int_\Omega \nabla\varphi_k\cdot\nabla(\eta-\varphi) - \frac{\gamma}{\varepsilon}\int_\Omega \varphi_k(\eta-\varphi) \\ - \int_\Omega \boldsymbol{C}'(\varphi_k)(\eta-\varphi)\mathcal{E}(\boldsymbol{u}_k):\mathcal{E}(\boldsymbol{u}_k) \geq 0 \quad \forall \eta\in\Phi_{ad},$$

with a_k defined as in (169), i.e.

$$a_k(v_1,v_2) := \gamma\varepsilon\int_\Omega \nabla v_1\cdot\nabla v_2 + \left(\frac{\varepsilon}{\tau_k}-\frac{\gamma}{\varepsilon}\right)\int_\Omega v_1 v_2.$$

Again this is the same as

$$a_k(\varphi-\varphi_k,\eta-\varphi) + \left\langle j'(\varphi_k),\eta-\varphi\right\rangle \geq 0 \quad \forall\eta\in\Phi_{ad}.$$

Thus, taking the potential term implicitly in time leads to a different scaling a_k. Equivalently, this can also be seen as a rescaling in time, since there is a one-to-one correspondence between the range of $\left(\frac{\varepsilon}{\tau_k}-\frac{\gamma}{\varepsilon}\right)$ and $\frac{\varepsilon}{\tilde\tau_k}$ in the respective intervals $\tau_k\in[\tau_{min},\varepsilon^2/\gamma)$ and $\tilde\tau_k\in[\tilde\tau_{min},\infty)$ for suitable $\tilde\tau_{min}$. Nevertheless one can apply Algorithm 6.1 with an additional modification such that $\tau_k < \varepsilon^2/\gamma$ is ensured and one obtains global convergence.

Proposition 6.63. *Let $(\boldsymbol{\varphi}_k)_k$ denote the sequence of iterates generated by the pseudo time stepping algorithm 6.1, where the VI* (175) *is replaced by the VI* (176)*. This method stems from a time discrete L^2-gradient flow, where the gradient term and the potential term are taken implicitly and the remaining terms explicitly in time. Let the initial guess $\boldsymbol{\varphi}_0\in\Phi_{ad}$ be arbitrary and replace the statement $\tau_k := \beta^{m_k-1}\tau_{k-1}$ in line 4 of algorithm 6.1 by e.g. $\tau_k := \min\{\beta^{m_k-1}\tau_{k-1}, 0.99\varepsilon^2/\gamma\}$. Then every accumulation point of $(\boldsymbol{\varphi}_k)_k$ in $H^1(\Omega)^N\cap L^\infty(\Omega)^N$ is a stationary point of j.* □

At last we want to consider a H^{-1} gradient flow, which is of Cahn-Hilliard type, for the mean compliance problem in the case of scalar valued phase fields. This is also considered in [BGS+12], where they use the degenerate mobility $B(\varphi) = \frac{9}{4}(1-\varphi^2)^2$. We emphasize that using this $B(\varphi)$ is not possible in our framework, since for the following calculation we need that the operator $w\mapsto\int_\Omega B(\varphi)\nabla w\cdot\nabla(.)$ is invertible, which is not the case for the degenerate mobility. Thus we restrict our calculation to the case $B(\varphi)=1$. Another difference to the presented approach and that in [BGS+12] is that we will use a double obstacle potential and will take the potential term explicitly in time, whereas in [BGS+12] they use a smooth potential and take a linear Taylor expansion of the potential implicitly in time.

As usual for Cahn-Hilliard equations, one introduces the chemical potential w. A single

time step is then given as (cf. [BGS+12, BBG11])

$$\begin{aligned} |\varphi| &\le 1 \\ \frac{1}{\tau_k}\int_\Omega (\varphi-\varphi_k)\eta + \int_\Omega \nabla w\cdot\nabla\eta = 0 \quad &\forall \eta \\ -\int_\Omega w(\eta-\varphi) + \gamma\varepsilon\int_\Omega \nabla\varphi\cdot\nabla(\eta-\varphi) + \frac{\gamma}{\varepsilon}\int_\Omega \psi_0'(\varphi_k)(\eta-\varphi) & \\ -\int_\Omega \boldsymbol{C}'(\varphi_k)(\eta-\varphi)\mathcal{E}(\boldsymbol{u}_k):\mathcal{E}(\boldsymbol{u}_k) \ge 0 \quad &\forall -1\le\eta\le 1 \end{aligned} \tag{177}$$

With the definition $v := w - \fint_\Omega w$ we get from the second equation that $v = -\frac{1}{\tau_k}(-\Delta_N)^{-1}(\varphi-\varphi_k)$, which is the unique weak solution of the pure Neumann problem

$$-\Delta v = -\frac{1}{\tau_k}(\varphi-\varphi_k) \text{ in } \Omega, \quad \fint_\Omega v = 0, \quad \partial_n v = 0 \text{ on } \partial\Omega,$$

as well as $\fint_\Omega \varphi = \fint_\Omega \varphi_k = \mathfrak{m}$. As in [BE91b, BBG11] we get that the whole system is then equivalent to

$$\begin{aligned} \varphi\in\Phi_{ad}, \quad \frac{1}{\tau_k}\int_\Omega (-\Delta_N)^{-1}(\varphi-\varphi_k)(\eta-\varphi) + \gamma\varepsilon\int_\Omega \nabla\varphi\cdot\nabla(\eta-\varphi) & \\ +\frac{\gamma}{\varepsilon}\int_\Omega \psi_0'(\varphi_k)(\eta-\varphi) - \int_\Omega \boldsymbol{C}'(\varphi_k)(\eta-\varphi)\mathcal{E}(\boldsymbol{u}_k):\mathcal{E}(\boldsymbol{u}_k)\ge 0 \quad &\forall \eta\in\Phi_{ad}. \end{aligned} \tag{178}$$

From the identity $\int_\Omega (-\Delta_N)^{-1}(\varphi-\varphi_k)(\eta-\varphi) = (\varphi-\varphi_k,\eta-\varphi)_{H^{-1}}$ we get the equivalence to

$$\begin{aligned} \varphi\in\Phi_{ad}, \quad a_k(\varphi-\varphi_k,\eta-\varphi) + \gamma\varepsilon\int_\Omega \nabla\varphi_k\cdot\nabla(\eta-\varphi) + \frac{\gamma}{\varepsilon}\int_\Omega \psi_0'(\varphi_k)(\eta-\varphi) & \\ -\int_\Omega \boldsymbol{C}'(\varphi_k)(\eta-\varphi)\mathcal{E}(\boldsymbol{u}_k):\mathcal{E}(\boldsymbol{u}_k)\ge 0 \quad &\forall \eta\in\Phi_{ad}, \end{aligned}$$

with a_k as in (170), i.e.

$$a_k(v_1,v_2) := \gamma\varepsilon\int_\Omega \nabla v_1\cdot\nabla v_2 + \frac{1}{\tau_k}(v_1,v_2)_{H^{-1}}.$$

Finally we obtain

$$\varphi\in\Phi_{ad}, \quad a_k(\varphi-\varphi_k,\eta-\varphi) + \left\langle j'(\varphi_k),\eta-\varphi\right\rangle \ge 0 \quad \forall\eta\in\Phi_{ad}.$$

Thus, the time steps of the discrete H^{-1} flow coincide with the iterates of the VMPT method with scaling a_k in (170) and $\lambda_k = \alpha_k = 1$. Algorithm 6.1, with the variational inequality (175) replaced by the VI (178), or equivalently by the system (177), can be used to choose the time step sizes τ_k. Again we get global convergence from Lemma 6.59 and Theorem 4.14. Note that by Lemma 4.6 we also get the unique solvability of the VI (178).

Proposition 6.64. *Let $(\boldsymbol{\varphi}_k)_k$ denote the sequence of iterates generated by the pseudo time stepping algorithm 6.1, where the VI* (175) *is replaced by the system* (177)*. This method stems from a time discrete H^{-1}-gradient flow, where the gradient term is taken implicitly in time and the remaining terms explicitly. Let the initial guess $\boldsymbol{\varphi}_0\in\Phi_{ad}$ be arbitrary. Then every accumulation point of $(\boldsymbol{\varphi}_k)_k$ in $H^1(\Omega)^N\cap L^\infty(\Omega)^N$ is a stationary point of j.* □

Literally as in Lemma 6.62 one shows that $\varphi_\tau \to \varphi_\infty$ in H^1 as $\tau \to \infty$, where one uses that $\|\varphi\|_{H^{-1}} \le C\|\varphi\|_{L^2}$. Thus, also the pseudo time stepping scheme of Cahn-Hilliard type approximates the scaled projected H^1-gradient method as $\tau_k \to \infty$.

6.9 SQP method on the reduced problem, Josephy-Newton method

In this section we derive an SQP method for the topology optimization problem (102). The goal is to compare the VMPT method with a state-of-the-art SQP method, which will be done numerically in Section 6.13.8.
Here, we apply the SQP method on the reduced problem, i.e. we eliminate the state equation first and then apply the SQP method. Thus the state equation is not linearized and $\boldsymbol{u}_k = S(\boldsymbol{\varphi}_k)$ holds in every iteration. The idea of the SQP method is to subsequently solve a subproblem, where the cost functional is approximated by its second order Taylor polynomial with j'' replaced by $\mathcal{L}_{\varphi,\varphi}$, where $\mathcal{L}$ is the Lagrange functional. In addition, the constraints are linearized in the subproblem, see (9)-(11). Since for the considered topology optimization problem the constraints are already linear, no linearization is necessary. For the Lagrange functional we get

$$\mathcal{L}(\boldsymbol{\varphi}, \boldsymbol{\lambda}, \Lambda, \boldsymbol{\mu}) = j(\boldsymbol{\varphi}) - \sum_i \lambda_i \left(\int_\Omega \varphi_i - \mathfrak{m}_i |\Omega| \right) - \left\langle \Lambda, \sum_{i=1}^N \varphi_i - 1 \right\rangle - \langle \boldsymbol{\mu}, \boldsymbol{\varphi} \rangle_{(L^\infty)^*, L^\infty} .$$

Recall that we proved the existence and uniqueness of the Lagrange multipliers $\boldsymbol{\lambda}$, Λ and $\boldsymbol{\mu}$ in Section 6.5.
By the linearity of the constraints we see that $\mathcal{L}_{\varphi,\varphi}(\boldsymbol{\varphi}, \boldsymbol{\lambda}, \Lambda, \boldsymbol{\mu}) = j''(\boldsymbol{\varphi})$. Thus, the SQP subproblem (9)-(11) reads

$$\min \frac{1}{2} j''(\boldsymbol{\varphi}_k)[\boldsymbol{y} - \boldsymbol{\varphi}_k, \boldsymbol{y} - \boldsymbol{\varphi}_k] + \left\langle j'(\boldsymbol{\varphi}_k), \boldsymbol{y} - \boldsymbol{\varphi}_k \right\rangle \tag{179}$$

$$\boldsymbol{y} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N \tag{180}$$

$$\boldsymbol{y} \ge 0, \quad \sum_{i=1}^N y_i = 1, \quad \fint \boldsymbol{y} = \mathfrak{m}. \tag{181}$$

Note that this problem is not necessarily uniquely solvable. Moreover, in contrast to (9)-(11) it does not depend on the Lagrange multipliers of the previous SQP step due to the linearity of the constraints. Thus the Lagrange multipliers don't need to be updated. Starting with an initial guess $\boldsymbol{\varphi}_0$, the SQP subproblem is solved for $\boldsymbol{y}$, where the solution $\boldsymbol{y}$ closest to $\boldsymbol{\varphi}_k$ is chosen if the subproblem attains multiple solutions. This solution becomes the new iterate, i.e. $\boldsymbol{\varphi}_{k+1} = \boldsymbol{y}$.

The SQP method given here formally coincides with the VMPT method without line search and with the choices $a_k = j''(\boldsymbol{\varphi}_k)$ and $\lambda_k = 1$. In [Ber99] this is also called constrained Newton's method. However, since j is not convex its second order derivative j'' is not necessarily positive definite and thus does not define an inner product. Hence only local convergence can be expected from the SQP method without any further globalization strategy.

As mentioned in Section 3, the SQP method can also be derived as Josephy-Newton method applied to the KKT system, which can be seen as follows, cf. also [HPUU08]. Consider the KKT system given in Theorem 6.33. The contained complementarity condi-

tion

$$\boldsymbol{\varphi} \geq 0, \quad \langle \boldsymbol{\mu}, \boldsymbol{\varphi} \rangle = 0, \quad \langle \boldsymbol{\mu}, \boldsymbol{\eta} \rangle \geq 0 \quad \forall \boldsymbol{\eta} \geq 0$$

is equivalent to the variational inequality

$$\boldsymbol{\varphi} \geq 0 : \quad \langle \boldsymbol{\mu}, \boldsymbol{\eta} - \boldsymbol{\varphi} \rangle \geq 0 \quad \forall \boldsymbol{\eta} \geq 0.$$

For simplicity we leave away the function space, which can always be chosen as $H^1(\Omega)^N \cap L^\infty(\Omega)^N$. This is no restriction as already discussed after Theorem 6.37. The first direction of the equivalence can be seen by subtracting the latter two (in-)equalities, the other direction is obtained by using the test functions $\boldsymbol{\eta} = 2\boldsymbol{\varphi}$, $\boldsymbol{\eta} = 0$ and $\boldsymbol{\eta} = \tilde{\boldsymbol{\eta}} + \boldsymbol{\varphi}$ for arbitrary $\tilde{\boldsymbol{\eta}} \geq 0$. The variational inequality in turn can be written as generalized equation

$$\boldsymbol{\mu} \in N_{\geq}(\boldsymbol{\varphi}),$$

with the normal cone mapping

$$N_{\geq}(\boldsymbol{\varphi}) = \begin{cases} \{\boldsymbol{\mu} \mid \langle \boldsymbol{\mu}, \boldsymbol{\eta} - \boldsymbol{\varphi} \rangle \geq 0 \ \forall \boldsymbol{\eta} \geq 0\} & \boldsymbol{\varphi} \geq 0 \\ \varnothing & \text{else.} \end{cases}$$

Thus, the KKT system in Theorem 6.33 is equivalent to the generalized equation

$$0 \in G(\boldsymbol{\varphi}, \boldsymbol{\lambda}, \Lambda, \boldsymbol{\mu}) + N(\boldsymbol{\varphi}, \boldsymbol{\lambda}, \Lambda, \boldsymbol{\mu})$$

with the differentiable function

$$G(\boldsymbol{\varphi}, \boldsymbol{\lambda}, \Lambda, \boldsymbol{\mu}) := \begin{pmatrix} \int_\Omega \varphi_1 - \mathfrak{m}_1 |\Omega| \\ \vdots \\ \int_\Omega \varphi_{N-1} - \mathfrak{m}_{N-1} |\Omega| \\ \sum_{i=1}^N \varphi_i - 1 \\ -\boldsymbol{\mu} \\ H(\boldsymbol{\varphi}, \boldsymbol{\lambda}, \Lambda, \boldsymbol{\mu}) \end{pmatrix},$$

where

$$\langle H(\boldsymbol{\varphi}, \boldsymbol{\lambda}, \Lambda, \boldsymbol{\mu}), \boldsymbol{\eta} \rangle := \langle j'(\boldsymbol{\varphi}), \boldsymbol{\eta} \rangle - \sum_{i=1}^{N-1} \int_\Omega \eta_i \lambda_i - \left\langle \Lambda, \sum_{i=1}^N \eta_i \right\rangle - \langle \boldsymbol{\mu}, \boldsymbol{\eta} \rangle \quad \forall \boldsymbol{\eta}$$

and the set-valued mapping

$$N(\boldsymbol{\varphi}, \boldsymbol{\lambda}, \Lambda, \boldsymbol{\mu}) := \{0\}^{N-1} \times \{0\} \times N_{\geq}(\boldsymbol{\varphi}) \times \{\mathbf{0}\}.$$

Recall that the Josephy-Newton method applied to the generalized equation linearizes G and evaluates N at the new iterate, see (7). Since G is affine except for the nonlinearity $j'(\boldsymbol{\varphi})$, the linearization process only affects the latter. Thus, the $(k+1)$th iterate of the

Josephy-Newton method is given by the solution $(\boldsymbol{\varphi}, \boldsymbol{\lambda}, \Lambda, \boldsymbol{\mu})$ of the linear system

$$\int_\Omega \boldsymbol{\varphi} = \mathfrak{m}|\Omega|, \tag{182}$$

$$\sum_{i=1}^{N} \varphi_i = 1, \tag{183}$$

$$\boldsymbol{\varphi} \geq 0, \tag{184}$$

$$\left\langle j'(\boldsymbol{\varphi}_k), \boldsymbol{\eta} \right\rangle + j''(\boldsymbol{\varphi}_k)[\boldsymbol{\varphi} - \boldsymbol{\varphi}_k, \boldsymbol{\eta}] \tag{185}$$

$$- \sum_{i=1}^{N-1} \int_\Omega \eta_i \lambda_i - \left\langle \Lambda, \sum_{i=1}^{N} \eta_i \right\rangle - \langle \boldsymbol{\mu}, \boldsymbol{\eta} \rangle = 0 \quad \forall \boldsymbol{\eta}, \tag{186}$$

$$\langle \boldsymbol{\mu}, \boldsymbol{\eta} \rangle \geq 0 \quad \forall \boldsymbol{\eta} \geq 0, \tag{187}$$

$$\langle \boldsymbol{\mu}, \boldsymbol{\varphi} \rangle = 0. \tag{188}$$

which is according to Theorem 6.33 exactly the KKT system of the SQP subproblem (179)-(181). Note that the Lagrange multipliers are unique for a given solution $\boldsymbol{\varphi}$. However, there may be multiple solutions $\boldsymbol{\varphi}$ of the SQP subproblem.
We note that due to the linearity of the constraints the reduced SQP method presented here also coincides with the Josephy-Newton method applied to the variational inequality

$$\boldsymbol{\varphi} \in \Phi_{ad}, \quad \left\langle j'(\boldsymbol{\varphi}), \boldsymbol{\eta} - \boldsymbol{\varphi} \right\rangle \geq 0 \quad \forall \boldsymbol{\eta} \in \Phi_{ad},$$

which is a first order optimality condition of the topology optimization problem. The subproblem in the corresponding Josephy-Newton method is the linearized variational inequality (cf. (8))

$$\boldsymbol{\varphi} \in \Phi_{ad}, \quad \left\langle j'(\boldsymbol{\varphi}_k), \boldsymbol{\eta} - \boldsymbol{\varphi} \right\rangle + j''(\boldsymbol{\varphi}_k)[\boldsymbol{\varphi} - \boldsymbol{\varphi}_k, \boldsymbol{\eta} - \boldsymbol{\varphi}] \geq 0 \quad \forall \boldsymbol{\eta} \in \Phi_{ad},$$

which is a first order condition of the SQP subproblem (179)-(181).
The results for the Josephy-Newton method can thus be used to prove local convergence of the SQP method. Therefor, strong regularity (in the sense of Robinson) of the solution has to be shown, which we will not prove here. The numerical results in Section 6.13.8 indicate that the SQP method converges locally with at least q-superlinear rate. Moreover, the method is mesh independent.

We solve the SQP subproblem and its KKT system (182)-(188), respectively, by a semismooth Newton method which will be described in Section 6.10.

Inserting the adjoint representation for j' and j'' as derived in Proposition 6.30 and

Theorem 6.44 yields the SQP subproblem

$$\begin{aligned}
\min_{y\in\Phi_{ad}} \frac{1}{2}\Bigg[& \gamma\varepsilon \int_\Omega \nabla(\boldsymbol{y}-\boldsymbol{\varphi}_k) : \nabla(\boldsymbol{y}-\boldsymbol{\varphi}_k) + \frac{\gamma}{\varepsilon}\int_\Omega \psi_0''(\boldsymbol{\varphi}_k)(\boldsymbol{y}-\boldsymbol{\varphi}_k)(\boldsymbol{y}-\boldsymbol{\varphi}_k) \\
& + F_{\boldsymbol{\varphi},\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k,\boldsymbol{u}_k)[\boldsymbol{y}-\boldsymbol{\varphi}_k,\boldsymbol{y}-\boldsymbol{\varphi}_k] + F_{\boldsymbol{\varphi},\boldsymbol{u}}(\boldsymbol{\varphi}_k,\boldsymbol{u}_k)[\boldsymbol{\delta u},\boldsymbol{y}-\boldsymbol{\varphi}_k] \\
& - \int_\Omega (\boldsymbol{C}''(\boldsymbol{\varphi}_k)(\boldsymbol{y}-\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{u}_k) : \mathcal{E}(\boldsymbol{p}_k))\cdot(\boldsymbol{y}-\boldsymbol{\varphi}_k) \\
& - \int_\Omega (\nabla \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta u}) : \mathcal{E}(\boldsymbol{p}_k))\cdot(\boldsymbol{y}-\boldsymbol{\varphi}_k) \\
& - \int_\Omega (\nabla \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{u}_k) : \mathcal{E}(\boldsymbol{\delta p}))\cdot(\boldsymbol{y}-\boldsymbol{\varphi}_k) + \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi},\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k)[\boldsymbol{y}-\boldsymbol{\varphi}_k,\boldsymbol{y}-\boldsymbol{\varphi}_k]\cdot\boldsymbol{p}_k \\
& + \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k)^T\boldsymbol{\delta p}\cdot(\boldsymbol{y}-\boldsymbol{\varphi}_k) + \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi},\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k)[\boldsymbol{y}-\boldsymbol{\varphi}_k,\boldsymbol{y}-\boldsymbol{\varphi}_k]\cdot\boldsymbol{p}_k \\
& + \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k)^T\boldsymbol{\delta p}\cdot(\boldsymbol{y}-\boldsymbol{\varphi}_k)\Bigg] + \\
& + \gamma\varepsilon\int_\Omega \nabla\boldsymbol{\varphi}_k\cdot\nabla(\boldsymbol{y}-\boldsymbol{\varphi}_k) + \frac{\gamma}{\varepsilon}\int_\Omega \psi_0'(\boldsymbol{\varphi}_k)(\boldsymbol{y}-\boldsymbol{\varphi}_k) + \langle F_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k,\boldsymbol{u}_k),\boldsymbol{y}-\boldsymbol{\varphi}_k\rangle \\
& - \int_\Omega (\nabla \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{u}_k) : \mathcal{E}(\boldsymbol{p}_k))\cdot(\boldsymbol{y}-\boldsymbol{\varphi}_k) + \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k)^T\boldsymbol{p}_k\cdot(\boldsymbol{y}-\boldsymbol{\varphi}_k) \\
& + \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k)^T\boldsymbol{p}_k\cdot(\boldsymbol{y}-\boldsymbol{\varphi}_k).
\end{aligned}$$

Here, $\boldsymbol{u}_k$ is the weak solution of the state equation

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{u}_k) : \mathcal{E}(\boldsymbol{\xi}) = \int_\Omega \boldsymbol{f}(\boldsymbol{\varphi}_k)\cdot\boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g}(\boldsymbol{\varphi}_k)\cdot\boldsymbol{\xi} \quad \forall \boldsymbol{\xi}\in H_D^1$$

and $\boldsymbol{p}_k$ is the weak solution of the adjoint equation

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{p}_k) : \mathcal{E}(\boldsymbol{\xi}) = \langle F_{\boldsymbol{u}}(\boldsymbol{\varphi}_k,\boldsymbol{u}_k),\boldsymbol{\xi}\rangle_{(H_D^1)^*,H_D^1} \quad \forall \boldsymbol{\xi}\in H_D^1.$$

Moreover, $\boldsymbol{\delta u}$ is the weak solution of the linearized state equation

$$\begin{aligned}
\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta u}) : \mathcal{E}(\boldsymbol{\xi}) = & -\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}_k)(\boldsymbol{y}-\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{u}_k) : \mathcal{E}(\boldsymbol{\xi}) + \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k)(\boldsymbol{y}-\boldsymbol{\varphi}_k)\cdot\boldsymbol{\xi} \\
& + \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k)(\boldsymbol{y}-\boldsymbol{\varphi}_k)\cdot\boldsymbol{\xi} \quad \forall \boldsymbol{\xi}\in H_D^1
\end{aligned}$$

and $\boldsymbol{\delta p}$ is the weak solution of the linearized adjoint equation

$$\begin{aligned}
\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta p}) : \mathcal{E}(\boldsymbol{\xi}) = & F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k,\boldsymbol{u}_k)[\boldsymbol{y}-\boldsymbol{\varphi}_k,\boldsymbol{\xi}] + F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi}_k,\boldsymbol{u}_k)[\boldsymbol{\delta u},\boldsymbol{\xi}] \\
& - \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}_k)(\boldsymbol{y}-\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{p}_k) : \mathcal{E}(\boldsymbol{\xi}) \quad \forall \boldsymbol{\xi}\in H_D^1.
\end{aligned}$$

Note that $\boldsymbol{u}_k$ and $\boldsymbol{p}_k$ only depend on $\boldsymbol{\varphi}_k$ but not on $\boldsymbol{y}$ and thus both have to be computed only once in each SQP step. On the other hand, $\boldsymbol{\delta u}$ and $\boldsymbol{\delta p}$ also depend on $\boldsymbol{y}$ and hence have to be recomputed for each evaluation of $j''(\boldsymbol{\varphi}_k)$ in a certain direction. We also refer to the discussion in Section 6.10.4. The numerical results for the SQP method can be found in Section 6.13.8.

6.10 General discrete PDAS method as a semismooth Newton method

In this section we mostly consider a general optimization problem, which is not necessarily the topology optimization problem. We derive a primal-dual active set (PDAS) method as a semismooth Newton (SSN) method applied to a nonsmooth system of equations, which is equivalent to the KKT system of the general optimization problem. The analysis of the method will only be performed on the discrete level, since for the considered optimization problems convergence of the method cannot be shown on the continuous level due to the lack of semismoothness. This is also reflected in the numerical results, where a mild mesh dependency can be observed, see Section 6.13.3 and Section 6.13.9.
The derived PDAS method will be used as a solver for the subproblems arising in the VMPT method and the SQP method. Moreover, we will apply the PDAS/SSN method to the topology optimization problem itself and we will compare it to the VMPT method in Section 6.13.9.
We show local superlinear convergence for the PDAS method applied to the projection type subproblem in the VMPT method under mild assumptions. Moreover, we show local superlinear convergence for a general objective functional under the additional assumption of a second order sufficient condition and strict complementarity. We also derive the PDAS method for the scalar valued case involving two phases and show that the PDAS methods for the scalar and vector problem are equivalent in a certain sense. The same local convergence properties are shown for the scalar PDAS method. Finally, we discuss some details about the implementation and the computational cost.
We note that the PDAS method is also used as a solver for the time steps in the Allen-Cahn and Cahn-Hilliard variational inequalities, see [BGSS13b, BSS12, BGSS13a, Sar10, But12, BBG11]. In fact, the local convergence results here generalize the results in [BGSS13a], since the cost functional can be arbitrary.

We consider a general objective $f : H^1(\Omega)^N \cap L^\infty(\Omega)^N$, which is two times continuously Fréchet differentiable. In our applications, f will either be the reduced cost functional j itself, the functional $g(\boldsymbol{\varphi}) = \frac{1}{2}a_k(\boldsymbol{\varphi}-\boldsymbol{\varphi}_k, \boldsymbol{\varphi}-\boldsymbol{\varphi}_k)+\lambda_k \langle j'(\boldsymbol{\varphi}_k), \boldsymbol{\varphi}-\boldsymbol{\varphi}_k\rangle$ of the projection-type subproblem, or the functional of the SQP subproblem. We consider the problem

$$\begin{aligned} \min f(\boldsymbol{\varphi}) & \\ \boldsymbol{\varphi} &\geq 0 \quad \text{a.e. in } \Omega \\ \sum_{i=1}^{N} \varphi_i &= 1 \quad \text{a.e. in } \Omega \\ \fint \boldsymbol{\varphi} &= \mathfrak{m}, \end{aligned} \tag{189}$$

where $\boldsymbol{\varphi}(x) \geq 0$ is to be understood component-wise as usual. In the following we assume $\mathfrak{m} > 0$. Let $\overline{\boldsymbol{\varphi}}$ be a local minimizer of problem (189). From Theorem 6.33 and Theorem 6.37 we get the existence and uniqueness of Lagrange multipliers $\boldsymbol{\lambda} \in \mathbb{R}^{N-1}$, $\Lambda \in (H^1(\Omega) \cap$

$L^\infty(\Omega))^*$ and $\boldsymbol{\mu} \in (L^\infty(\Omega)^N)^*$, such that the KKT system

$$\left\langle f'(\overline{\boldsymbol{\varphi}}), \boldsymbol{\eta} \right\rangle - \sum_{i=1}^{N-1} \int_\Omega \eta_i \lambda_i - \left\langle \Lambda, \sum_{i=1}^{N} \eta_i \right\rangle - \langle \boldsymbol{\mu}, \boldsymbol{\eta} \rangle_{(L^\infty)^*, L^\infty} = 0 \quad \forall \boldsymbol{\eta} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N, \tag{190}$$

$$\langle \boldsymbol{\mu}, \boldsymbol{\eta} \rangle_{(L^\infty)^*, L^\infty} \geq 0 \quad \forall \boldsymbol{\eta} \in L^\infty(\Omega)^N, \boldsymbol{\eta} \geq 0, \tag{191}$$

$$\overline{\boldsymbol{\varphi}} \geq 0, \tag{192}$$

$$\langle \boldsymbol{\mu}, \overline{\boldsymbol{\varphi}} \rangle_{(L^\infty)^*, L^\infty} = 0, \tag{193}$$

$$\int_\Omega \overline{\varphi}_i = \mathfrak{m}_i |\Omega| \quad i = 1, \dots, N-1, \tag{194}$$

$$\sum_{i=1}^{N} \overline{\varphi}_i = 1. \tag{195}$$

holds. As in Section 6.5 we drop the redundant constraint $\int_\Omega \varphi_N = \mathfrak{m}_N$.
The KKT system includes inequalities and thus the semismooth Newton method cannot be applied directly. First one has to reformulate the KKT system to a system of (non-smooth) equations whereon the semismooth Newton method can be applied. There are multiple possibilities to reformulate the system. We give some examples.
Suppose that we consider the topology optimization problem, i.e. $f = j$. Then there is a state variable $\boldsymbol{u}$ and an adjoint variable $\boldsymbol{p}$ present. It is possible to eliminate these variables by means of the state equation and the adjoint equation, respectively. Then $\boldsymbol{u}$ and $\boldsymbol{p}$ don't appear as unknowns in the Newton system. On the other hand it is possible to keep $\boldsymbol{u}$ and $\boldsymbol{p}$ as independent variables. In this case the unknowns of the Newton system are $\boldsymbol{\varphi}, \boldsymbol{u}, \boldsymbol{p}$ and the dual variables connected to the other constraints. Moreover, since $\boldsymbol{u}$ and $\boldsymbol{p}$ are independent of $\boldsymbol{\varphi}$, they do not solve the state equation and the adjoint equation during the Newton iteration, but only in the limit. An advantage of the latter approach is that the linearized state and adjoint equations don't have to be solved exactly in each Newton iteration to gain local convergence of the method. This is known as inexact Newton method. However, by eliminating $\boldsymbol{u}$ and $\boldsymbol{p}$ one can reduce the number of unknowns in the Newton system drastically. For example consider the binary case ($N = 2$) in 3-D. The vector of unknowns $(\varphi, \boldsymbol{u}, \boldsymbol{p}, \mu, \boldsymbol{\lambda})$ consists of 8 scalar functions (neglecting $\boldsymbol{\lambda} \in \mathbb{R}^{N-1}$), whereas the reduced vector of unknowns $(\varphi, \mu, \boldsymbol{\lambda})$ only consists of 2 functions. For multiple phases, this reduction is less drastic, but still present. We therefore decide to eliminate $\boldsymbol{u}$ and $\boldsymbol{p}$.
In case that f is the functional in the SQP subproblem or the functional in the projection type subproblem with a_k as in (167), the linearized state $\boldsymbol{\delta u}$ and the linearized adjoint state $\boldsymbol{\delta p}$ are present. For the same reason we decide to eliminate $\boldsymbol{\delta u}$ and $\boldsymbol{\delta p}$.
In case that f is the functional in the projection type subproblem with a_k being e.g. the H^1 inner product, there is no state variable present and $\boldsymbol{\varphi}$ is the only primal unknown.
In all cases it is additionally possible to eliminate the Lagrange multiplier $\boldsymbol{\mu}$ by means of the gradient equation (190) in the KKT system. Since we want to end up with the primal dual active set method, we have to treat $\boldsymbol{\mu}$ as an independent variable here.
For more information on these different reformulations of the KKT system we refer to [HPUU08].

6.10.1 Derivation of the PDAS method as a semismooth Newton method

To get a better idea how the inequalities in the KKT system are reformulated to a nonsmooth equation, assume that Λ and $\boldsymbol{\mu}$ are functions. The complementarity condition in the KKT system thus holds pointwise almost everywhere, i.e.

$$\boldsymbol{\mu} \geq 0, \quad \overline{\boldsymbol{\varphi}} \geq 0, \quad \boldsymbol{\mu} \cdot \overline{\boldsymbol{\varphi}} = 0.$$

By an elementary proof one shows that for any $c > 0$ this is equivalent to

$$\boldsymbol{\mu}(x) - P_{[0,\infty)}(\boldsymbol{\mu}(x) - c\overline{\boldsymbol{\varphi}}(x)) = 0 \quad \text{a.e.}, \tag{196}$$

where $P_{[0,\infty)} : \mathbb{R} \to \mathbb{R}$ denotes the projection on $[0, \infty)$, which is applied componentwise and pointwise. On the other hand, since the complementarity condition is symmetric in $\overline{\boldsymbol{\varphi}}$ and $\boldsymbol{\mu}$, another equivalent equation is

$$\overline{\boldsymbol{\varphi}}(x) - P_{[0,\infty)}(\overline{\boldsymbol{\varphi}}(x) - c\boldsymbol{\mu}(x)) = 0 \quad \text{a.e.} \tag{197}$$

Here we choose the reformulation (196) in order to end up with the primal dual active set strategy. Note that in Section 6.10 we reformulated the complementarity condition to a smooth generalized equation, whereas here we reformulate it to a nonsmooth equation.
Thus, we have reformulated the KKT system to the following system of nonsmooth equations, where we assume for simplicity that $\nabla f(\overline{\boldsymbol{\varphi}})$ exists as a function.

$$\begin{aligned}
\nabla f(\overline{\boldsymbol{\varphi}}) - \boldsymbol{\lambda} - \Lambda \boldsymbol{e} - \boldsymbol{\mu} &= 0 \\
\boldsymbol{\mu} - P_{[0,\infty)}(\boldsymbol{\mu} - c\overline{\boldsymbol{\varphi}}) &= 0 \\
\int_\Omega \overline{\varphi}_i &= \mathfrak{m}_i |\Omega| \quad i = 1, \dots, N-1, \\
\sum_{i=1}^{N} \overline{\varphi}_i &= 1.
\end{aligned}$$

On this system, which we write as $G(\overline{\boldsymbol{\varphi}}, \boldsymbol{\lambda}, \Lambda, \boldsymbol{\mu}) = \mathbf{0}$, the semismooth Newton method is applied.
In order to get local convergence in the continuous function space setting one needs two properties. The operator G has to be semismooth at the solution and the linearized operators have to be invertible near the solution with uniformly bounded inverse, see Theorem 3.4. We want to point out that this is not given here. Even in the case that the Lagrange multipliers and $f'(\varphi)$ are functions these assumptions would not be fulfilled. The reason is as follows. Assume $\boldsymbol{\mu} \in L^2(\Omega)^N$. To get invertibility of the linearized operators, one has to choose a subspace $L^2(\Omega)^N$ for the projection equation, i.e. the operator

$$L^2(\Omega)^N \times (H^1(\Omega)^N \cap L^\infty(\Omega)^N) \ni (\boldsymbol{\mu}, \boldsymbol{\varphi}) \mapsto \boldsymbol{\mu} - P_{[0,\infty)}(\boldsymbol{\mu} - c\boldsymbol{\varphi}) \in L^2(\Omega)^N$$

has to be semismooth at the solution. In general this requires that the Nemytskii operator $P_{[0,\infty)} : L^2(\Omega) \to L^2(\Omega)$ is semismooth. It is well known that a Nemytskii operator $P : L^p(\Omega) \to L^p(\Omega)$ is Fréchet differentiable for $1 \leq p < \infty$ if and only if P is affine linear [KZPP76]. The same result holds for semismoothness of Nemytskii operators [HPUU08]. Since $P_{[0,\infty)}$ is not affine-linear, it thus cannot be semismooth in $L^2(\Omega)$. One though can show that $P_{[0,\infty)} : L^p(\Omega) \to L^q(\Omega)$ is semismooth for $1 \leq q < p \leq \infty$ [IK08]. One thus has to get a smoothing operator inside the projection, which maps into L^q for some $q > 2$. For some optimal control problems this is possible when using the projection equation (197)

with $\boldsymbol{\mu}$ eliminated, see e.g. [HPUU08] and the examples in [Trö09]. For a special choice of c, equation (197) there becomes the usual projection formula

$$\overline{\varphi} - P_{[0,\infty)}\left(-\frac{1}{\lambda}p\right) = 0,$$

where $p \in H^1(\Omega)$ is the adjoint state and $\lambda > 0$ is a constant, cf. also the optimal control problem in Section 4.12. Because of $p \in H^1(\Omega) \hookrightarrow L^p(\Omega)$ for some $p > 2$, semismoothness in L^2 can be shown.
However, in our setting convergence of the semismooth Newton method cannot be shown in the continuous setting. This is also reflected by the numerical results, which show a moderate mesh dependent behavior of the method.
A typical approach to overcome this mesh dependency is to regularize the problem in some way. As an example, the Yosida-Moreau approximation will be discussed in Remark 6.67 below.

Because the method will not converge in the continuous setting, we first discretize the nonlinear system and then apply the semismooth Newton method on the discrete system. Therefor, let $\mathcal{T}_h$ be a triangulation of Ω and let $S_h \subset H^1(\Omega) \cap L^\infty(\Omega)$ be the standard P1 finite element space, i.e.

$$S_h = \left\{\varphi \in C(\overline{\Omega}) \mid \varphi|_T \in P_1(T) \ \forall T \in \mathcal{T}_h\right\},$$

where $P_1(T)$ denotes the space of all affine linear functions on the triangle T and h denotes the mesh parameter, i.e. the largest diameter of the triangles (resp. tetrahedra in 3D). See also Section 6.11 for more information about the used discretization. Let $\{p_i\}_{i=1}^J$ denote the set of nodes of the triangulation. The standard nodal basis functions of S_h are defined by

$$\chi_i(p_j) = \delta_{ij} \quad i,j = 1,\dots,J,$$

with $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. We discretize $H^1(\Omega)^N \cap L^\infty(\Omega)^N$ by S_h^N with the basis given by $\chi_i \boldsymbol{e}_j$ with $i = 1,\dots,J$ and $j = 1,\dots,N$ and $\boldsymbol{e}_j$ being the standard basis vectors of $\mathbb{R}^N$. For $\boldsymbol{\varphi}_h \in S_h$ we write $\boldsymbol{\varphi}_h = \sum_{ij} \varphi_i^j \chi_i \boldsymbol{e}_j$ with coordinates $\varphi_i^j \in \mathbb{R}$. We discretize the gradient equation (190) by the finite element method based on the weak formulation, since boundary integrals and the Laplacian coming from the Ginzburg-Landau energy can be handled this way. This leads to the discretized gradient equation

$$\frac{\langle f'(\boldsymbol{\varphi}), \chi_i \boldsymbol{e}_j\rangle}{m_i} - \lambda_j - \frac{\langle \Lambda, \chi_i\rangle}{m_i} - \frac{\langle \boldsymbol{\mu}, \chi_i \boldsymbol{e}_j\rangle_{(L^\infty)^*, L^\infty}}{m_i} = 0 \quad \forall i = 1,\dots,J,\ j = 1,\dots,N,$$

where we additionally divided each equation by $m_i := \int_\Omega \chi_i$. Since $f'(\boldsymbol{\varphi})$, Λ and $\boldsymbol{\mu}$ are in general only functionals, we don't discretize them as functions. We rather formulate the system in the unknowns φ_i^j, as well as in the coordinates

$$\Lambda_i := \frac{\langle \Lambda, \chi_i\rangle}{m_i},$$
$$\mu_i^j := \frac{\langle \boldsymbol{\mu}, \chi_i \boldsymbol{e}_j\rangle_{(L^\infty)^*, L^\infty}}{m_i},$$

which somehow corresponds to a discretization using the dual basis of $(\chi_i/m_i)_i$. We divide

by m_i to get that the coordinates Λ_i and μ_i^j are independent of the mesh parameter h in case that Λ and $\boldsymbol{\mu}$ are functions. This is the case in all considered experiments where the boundary traction $\boldsymbol{g}$ is independent of $\boldsymbol{\varphi}$. For example let Λ be a constant function, then $\Lambda_i = \frac{\langle \Lambda, \chi_i \rangle}{m_i} = \Lambda \frac{\int_\Omega \chi_i}{m_i} = \Lambda$ independent of h. However, when the Lagrange are no functions, these values can depend on h. For example let Λ be a Dirac measure, i.e. $\langle \Lambda, \eta \rangle = \eta(p_j)$ for some mesh point $p_j \in \Omega$ and for all continuous functions η. Then we have $\Lambda_j = \frac{\langle \Lambda, \chi_j \rangle}{m_j} = \frac{1}{m_j} = \mathcal{O}(h^{-d})$. If Λ includes measures concentrated on the boundary $\partial\Omega$, it holds $\Lambda_i = \mathcal{O}(h^{-1})$, which is also observed in the numerical experiments in Section 6.13.10. With the additional definition

$$D_i^j(\boldsymbol{\varphi}) := \frac{\left\langle f'(\sum_{k,l} \varphi_k^l \chi_k \boldsymbol{e}_l), \chi_i \boldsymbol{e}_j \right\rangle}{m_i} \quad \forall \boldsymbol{\varphi} = (\varphi_k^l)_{kl}$$

we finally get the discretized gradient equation

$$D_i^j(\boldsymbol{\varphi}) - \lambda_j - \Lambda_i - \mu_i^j = 0 \quad \forall i = 1, \dots, J,\ j = 1, \dots, N.$$

We discretize equation (191) also weakly by

$$\langle \boldsymbol{\mu}, \boldsymbol{\eta} \rangle_{(L^\infty)^*, L^\infty} \geq 0 \quad \forall \boldsymbol{\eta} \in S_h^N, \boldsymbol{\eta} \geq 0,$$

which is equivalent to

$$\mu_i^j \geq 0 \quad \forall i = 1, \dots, J,\ j = 1, \dots, N.$$

Equation (192) is for $\boldsymbol{\varphi} \in S_h^N$ equivalent to

$$\varphi_i^j \geq 0 \quad \forall i = 1, \dots, J,\ j = 1, \dots, N.$$

For equation (193) we get

$$0 = \langle \boldsymbol{\mu}, \boldsymbol{\varphi} \rangle_{(L^\infty)^*, L^\infty} = \sum_{i,j} \varphi_i^j \langle \boldsymbol{\mu}, \chi_i \boldsymbol{e}_j \rangle_{(L^\infty)^*, L^\infty} = \sum_{i,j} \varphi_i^j \mu_j^i m_i.$$

Using that it holds $\varphi_i^j \geq 0$, $\mu_i^j \geq 0$ and $m_i > 0$, this is equivalent to

$$\varphi_i^j \mu_j^i = 0 \quad \forall i = 1, \dots, J,\ j = 1, \dots, N.$$

As above, the resulting complementarity condition

$$\mu_i^j \geq 0, \qquad \varphi_i^j \geq 0, \qquad \varphi_i^j \mu_j^i = 0$$

is for any $c > 0$ equivalent to

$$\mu_i^j - P_{[0,\infty)}(\mu_i^j - c\varphi_i^j) = 0.$$

The discretization of the equality constraints is straight forward. The mass constraint (194) is equivalent to

$$\sum_i m_i \varphi_i^j = \mathfrak{m}_j |\Omega| \quad \forall j = 1, \dots, N-1$$

and the sum constraint (195) is equivalent to

$$\sum_j \varphi_i^j = 1 \quad \forall i = 1, \dots, J.$$

The final discretized KKT system thus reads

$$D_i^j(\boldsymbol{\varphi}) - \lambda_j - \Lambda_i - \mu_i^j = 0 \quad \forall i = 1, \dots, J,\ j = 1, \dots, N \tag{198}$$

$$\sum_i m_i \varphi_i^j - \mathfrak{m}_j |\Omega| = 0 \quad \forall j = 1, \dots, N-1 \tag{199}$$

$$\sum_j \varphi_i^j - 1 = 0 \quad \forall i = 1, \dots, J \tag{200}$$

$$\mu_i^j - P_{[0,\infty)}(\mu_i^j - c\varphi_i^j) = 0 \quad \forall i = 1, \dots, J,\ j = 1, \dots, N. \tag{201}$$

To show the existence of solutions of this system, we remark that this system is equivalent to the KKT system of the discretized optimization problem

$$\begin{aligned} &\min f(\boldsymbol{\varphi}_h) \\ &\boldsymbol{\varphi}_h \in S_h^N \cap \Phi_{ad} \end{aligned} \tag{202}$$

with rescaled Lagrange multipliers, which can be seen easily. Thus the choice of the discretization is done such that the approaches discretize-then-optimize and optimize-then-discretize are equivalent. From this it immediately follows that for each minimizer of (202) there exist Lagrange multipliers, since the constraints are affine and thus the Abadie constraint qualification is met [NW06]. On the other hand, $S_h^N \cap \Phi_{ad}$ is compact and f is continuous, thus we also get the existence of a minimizer of (202). Hence, the discretized KKT system and its equivalent formulation always have a solution.
Note that although in the continuous setting the uniqueness of Lagrange multipliers can be proved, this is not the case for the discrete KKT system. The reason is that in the continuous setting one can show that the inactive phases are connected (in the sense of graphs, see Theorem 6.35), which cannot be shown in the discrete case. However, we will show below that the LICQ constraint qualifications hold if the discrete inactive sets are connected and thus uniqueness of Lagrange multipliers is obtained in this case, see Corollary 6.70. In Lemma 6.72 we will also show that the connectedness is fulfilled in practice due to some compatibility condition.

The derived discrete system (198)-(201) can be written as $G(\boldsymbol{\varphi}, \boldsymbol{\lambda}, \boldsymbol{\Lambda}, \boldsymbol{\mu}) = 0$ with $G : \mathbb{R}^{J\times N} \times \mathbb{R}^{N-1} \times \mathbb{R}^J \times \mathbb{R}^{J\times N} \to \mathbb{R}^{J\times N} \times \mathbb{R}^{N-1} \times \mathbb{R}^J \times \mathbb{R}^{J\times N}$. As opposed to the infinite dimensional equations, semismoothness is given for G, which we show in the following. By the calculation rules for semismooth functions (see [HPUU08]), it is sufficient to show that D_i^j is continuously differentiable and that $P_{[0,\infty)}$ is semismooth. The former follows from the assumption $f \in C^2$. By chain rule, it holds for the derivative of D_i^j

$$(D')_i^j(\boldsymbol{\varphi})\boldsymbol{\eta} = \frac{f''(\sum_{k,l} \varphi_k^l \chi_k \boldsymbol{e}_l)[\chi_i \boldsymbol{e}_j, \sum_{k,l} \eta_k^l \chi_k \boldsymbol{e}_l]}{m_i} \quad \forall \boldsymbol{\varphi} = (\varphi_k^l)_{kl}, \boldsymbol{\eta} = (\eta_k^l)_{kl}.$$

The projection $P_{[0,\infty)}$ is ∂P-semismooth in every $x \in \mathbb{R}$, see [HPUU08], with generalized

differential

$$\partial P : \mathbb{R} \rightrightarrows \mathbb{R}$$
$$\partial P(x) = \begin{cases} \{0\} & x < 0 \\ [0,1] & x = 0 \\ \{1\} & x > 0 \end{cases}$$

In the following we will always choose the element $N(x) \in \partial P(x)$ of the subdifferential with $N(0) = 0$. Thus G fulfills the semismoothness assumption.
Let $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\Lambda}}, \overline{\boldsymbol{\mu}})$ be given. A semismooth Newton step to compute the new iterate $(\boldsymbol{\varphi}, \boldsymbol{\lambda}, \boldsymbol{\Lambda}, \boldsymbol{\mu})$ is given by the linear system

$$M(\boldsymbol{\varphi} - \overline{\boldsymbol{\varphi}}, \boldsymbol{\lambda} - \overline{\boldsymbol{\lambda}}, \boldsymbol{\Lambda} - \overline{\boldsymbol{\Lambda}}, \boldsymbol{\mu} - \overline{\boldsymbol{\mu}}) = -G(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\Lambda}}, \overline{\boldsymbol{\mu}}),$$

with $M \in \partial G(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\Lambda}}, \overline{\boldsymbol{\mu}})$, which is uniquely given by the choice of $N(x) \in \partial P(x)$ above. Defining the update $\boldsymbol{\delta\varphi} := \boldsymbol{\varphi} - \overline{\boldsymbol{\varphi}}$, the following linear system has to be solved for the unknowns $(\boldsymbol{\delta\varphi}, \boldsymbol{\lambda}, \boldsymbol{\Lambda}, \boldsymbol{\mu})$.

$$(D')_i^j(\overline{\boldsymbol{\varphi}})\boldsymbol{\delta\varphi} - \lambda_j - \Lambda_i - \mu_i^j = -D_i^j(\overline{\boldsymbol{\varphi}}) \quad \forall i = 1, \dots, J,\ j = 1, \dots, N \tag{203}$$
$$\sum_i m_i \delta\varphi_i^j = -\sum_i m_i \overline{\varphi}_i^j + \mathfrak{m}_j |\Omega| \quad \forall j = 1, \dots, N-1$$
$$\sum_j \delta\varphi_i^j = -\sum_j \overline{\varphi}_i^j + 1 \quad \forall i = 1, \dots, J \tag{204}$$
$$\mu_i^j - N_i^j(\mu_i^j - \overline{\mu}_i^j - c\delta\varphi_i^j) = P_{[0,\infty)}(\overline{\mu}_i^j - c\overline{\varphi}_i^j) \quad \forall i = 1, \dots, J,\ j = 1, \dots, N. \tag{205}$$

with $N_i^j \in \partial P(\overline{\mu}_i^j - c\overline{\varphi}_i^j)$ chosen as above.
We show that this system is equivalent to a step in a primal dual active set method in the case that the cost functional f is quadratic. Additionally we will see that the system can be decoupled such that only a linear system of lower dimension has to be solved. This we show similarly to [BGSS13a]. We therefor introduce the active and inactive sets

$$\mathcal{A}^j := \{i \in \{1, \dots, J\} \mid \overline{\mu}_i^j - c\overline{\varphi}_i^j > 0\}, \quad j = 1, \dots, N$$
$$\mathcal{I}^j := (\mathcal{A}^j)^c, \quad j = 1, \dots, N.$$

Therewith we get $N_i^j = 1$ for all $i \in \mathcal{A}^j$. Moreover, by the special choice of N we get $N_i^j = 0$ for all $i \in \mathcal{I}^j$ as noted above. Equation (205) can now be rewritten to

$$\delta\varphi_i^j = -\overline{\varphi}_i^j \quad \forall i \in \mathcal{A}^j,\ j = 1, \dots, N \tag{206}$$
$$\mu_i^j = 0 \quad \forall i \in \mathcal{I}^j,\ j = 1, \dots, N. \tag{207}$$

To get a better understanding of the resulting system, we note that for quadratic f the

system roughly speaking corresponds to

$$\begin{aligned} \nabla f(\boldsymbol{\varphi}) - \boldsymbol{\lambda} - \Lambda \boldsymbol{e} - \boldsymbol{\mu} &= 0 \\ \int_\Omega \boldsymbol{\varphi} &= \mathfrak{m}|\Omega|, \\ \textstyle\sum_j \varphi^j &= 1, \\ \mu^j &= 0 \quad \text{in } \mathcal{I}^j, \\ \varphi^j &= 0 \quad \text{in } \mathcal{A}^j, \end{aligned}$$

which can be seen as first order condition of the equality constrained problem

$$\begin{aligned} &\min f(\boldsymbol{\varphi}) \\ &\int_\Omega \boldsymbol{\varphi} = \mathfrak{m}|\Omega|, \\ &\textstyle\sum_j \varphi^j = 1, \\ &\varphi^j = 0 \quad \text{in } \mathcal{A}^j. \end{aligned}$$

This has the typical form of an active set method, see [IK08].

As in [BGSS13a] we introduce the index sets

$$\begin{aligned} \mathcal{D}^j &:= \mathcal{I}^j \cap \left(\bigcup_{k \neq j} \mathcal{I}^k \right) \quad j = 1, \dots, N, \\ \mathcal{D} &:= \bigcup_j \mathcal{D}^j, \end{aligned}$$

where $\mathcal{D}^j$ is the set where phase j and at least an additional phase is inactive. If only a single phase is inactive, i.e. for $i \in \mathcal{I}^j \setminus \mathcal{D}^j = \mathcal{I}^j \cap \left(\bigcap_{k \neq j} \mathcal{A}^k \right)$, we get from (206) that $\delta\varphi_i^k = -\overline{\varphi}_i^k$ holds for all $k \neq j$ and thus by the sum constraint

$$\delta\varphi_i^j = -\overline{\varphi}_i^j + 1 \quad \forall i \in \mathcal{I}^j \setminus \mathcal{D}^j. \tag{208}$$

From the partition $\{1, \dots, J\} = \mathcal{A}^j \sqcup \mathcal{D}^j \sqcup (\mathcal{I}^j \setminus \mathcal{D}^j)$, we see that φ_i^j is unknown only for $i \in \mathcal{D}^j$.
To further reduce the linear system, we observe that the gradient equation (203) for index (i, j) only depends on μ_i^j and Λ_i, but not on other components of $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$. Thus, by (207), we can eliminate $\boldsymbol{\mu}$ in the gradient equations for $i \in \mathcal{D}^j \subset \mathcal{I}^j$. Moreover, it is sufficient to consider the sum constraint only for indices $i \in \mathcal{D}$, since $\delta\varphi_i^j$ is already known otherwise. We thus obtain the reduced system

$$(D')_i^j(\overline{\boldsymbol{\varphi}})\boldsymbol{\delta\varphi} - \lambda_j - \Lambda_i = -D_i^j(\overline{\boldsymbol{\varphi}}) \quad \forall i \in \mathcal{D}^j \;\; j = 1, \dots, N, \tag{209}$$

$$\sum_i m_i \delta\varphi_i^j = -\sum_i m_i \overline{\varphi}_i^j + \mathfrak{m}_j|\Omega| \quad \forall j = 1, \dots, N-1, \tag{210}$$

$$\sum_j \delta\varphi_i^j = -\sum_j \overline{\varphi}_i^j + 1 \quad \forall i \in \mathcal{D}. \tag{211}$$

After eliminating $\delta\varphi_i^j$ for $i \in (\mathcal{D}^j)^c$ by equations (206) and (208), we end up with a linear system of dimension $\sum_{j=1}^N |\mathcal{D}^j| + N - 1 + |\mathcal{D}|$ with unknowns $\delta\varphi_i^j$ for $i \in \mathcal{D}^j$, $j = 1, \dots, N$, Λ_i for $i \in \mathcal{D}$ and $\lambda_1, \dots, \lambda_{N-1}$. After this linear reduced system is solved, $\boldsymbol{\delta\varphi}$ and $\boldsymbol{\lambda}$ are fully

determined, as well as Λ_i for $i \in \mathcal{D}$ and μ_i^j for $i \in \mathcal{I}^j$. To compute the remaining values of Λ_i for $i \in \mathcal{D}^c$, we have to assume that for any $i \in \{1, \dots, J\}$ there exists a $j \in \{1, \dots, N\}$ such that $i \in \mathcal{I}^j$, i.e. in each point at least one phase is inactive. If this would not be the case then from (206) we would get $\delta\varphi_i^j = -\overline{\varphi}_i^j$ for all j and thus $\sum_j \delta\varphi_i^j = -\sum_j \overline{\varphi}_i^j$, which contradicts the sum constraint (211). Thus, the Newton system would not be solvable. We note that this assumption is fulfilled for all $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\Lambda}}, \overline{\boldsymbol{\mu}})$ in a neighborhood of a solution of $G(\boldsymbol{\varphi}, \boldsymbol{\lambda}, \boldsymbol{\Lambda}, \boldsymbol{\mu}) = 0$, see the proof of Theorem 6.69 below. With this assumption fulfilled we get from $i \in \mathcal{D}^c = \bigcap_k (\mathcal{D}^k)^c = \bigcap_k (\mathcal{A}^k \cup (\bigcap_{l \neq k} \mathcal{A}^l))$, that there exists a unique phase $k(i)$, such that $i \in \mathcal{I}^{k(i)} \cap (\bigcap_{l \neq k(i)} \mathcal{A}^l)$, i.e. all phases are active except phase $k(i)$. We use the gradient equation and $\mu_i^{k(i)} = 0$ to compute the value of Λ_i by

$$\Lambda_i = (D')_i^{k(i)}(\overline{\boldsymbol{\varphi}})\boldsymbol{\delta\varphi} - \lambda_{k(i)} + D_i^{k(i)}(\overline{\boldsymbol{\varphi}}) \quad \forall i \in \mathcal{D}^c. \tag{212}$$

It remains to compute μ_i^j for $i \in \mathcal{A}^j$. Again from the gradient equation (203) we get

$$\mu_i^j = (D')_i^j(\overline{\boldsymbol{\varphi}})\boldsymbol{\delta\varphi} - \lambda_j - \Lambda_i + D_i^j(\overline{\boldsymbol{\varphi}}) \quad \forall i \in \mathcal{A}^j,\ j = 1, \dots, N. \tag{213}$$

We thus are able to reduce the Newton system to a linear system in the variables $(\delta\varphi^1_{\mathcal{D}^1}, \dots, \delta\varphi^N_{\mathcal{D}^N}, \lambda, \Lambda_{\mathcal{D}})$. The remaining unknowns can be computed explicitly without solving a linear system. This leads to Algorithm 6.2, where the iterates and sets in the k-th step are denoted by an additional index k. Note that the variables $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\Lambda}}, \overline{\boldsymbol{\mu}})$ in the equations above have to be replaced by $(\boldsymbol{\varphi}_k, \boldsymbol{\lambda}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\mu}_k)$ in the algorithm. In general this method is a semismooth Newton (SSN) method applied to the system (198)-(201). However, if the cost functional f is quadratic we refer to the method as primal dual active set (PDAS) method, which we motivated above. The name 'primal dual' comes from the fact that primal and dual information is used to define the active set, i.e. $\mathcal{A}$ depends on $\boldsymbol{\varphi}$ and $\boldsymbol{\mu}$.

Algorithm 6.2 PDAS/SSN method for the discretized general problem (189)

1: Choose $\boldsymbol{\varphi}_0$, $\boldsymbol{\lambda}_0$, $\boldsymbol{\Lambda}_0$, $\boldsymbol{\mu}_0$, $c > 0$.
2: $k := 0$.
3: **while** $k \leq k_{\max}$ **do**
4: Define the active sets

$$\mathcal{A}_k^j := \{i \in \{1, \dots, J\} \mid (\mu_k)_i^j - c(\varphi_k)_i^j > 0\}, \quad j = 1, \dots, N$$
$$\mathcal{I}_k^j := (\mathcal{A}_k^j)^c, \quad j = 1, \dots, N.$$

5: Set $(\boldsymbol{\delta\varphi}_k)^j_{(\mathcal{D}^j)^c}$ by equations (206) and (208) for $j = 1, \dots, N$.
6: Compute $((\boldsymbol{\delta\varphi}_k)^1_{\mathcal{D}^1}, \dots, (\boldsymbol{\delta\varphi}_k)^N_{\mathcal{D}^N}, \boldsymbol{\lambda}_{k+1}, (\boldsymbol{\Lambda}_{k+1})_{\mathcal{D}})$ by solving the linear system (209)-(211).
7: Set $\boldsymbol{\varphi}_{k+1} := \boldsymbol{\varphi}_k + \boldsymbol{\delta\varphi}_k$.
8: Set $(\boldsymbol{\Lambda}_{k+1})_{\mathcal{D}^c}$ by (212).
9: Set $\boldsymbol{\mu}_{k+1}$ by (207) and (213).
10: **end while**

Remark 6.65. The iterate $(\boldsymbol{\varphi}_{k+1}, \boldsymbol{\lambda}_{k+1}, \boldsymbol{\Lambda}_{k+1}, \boldsymbol{\mu}_{k+1})$ for $k \geq 0$ only depends on $\mathcal{A}_k^j$, $j = 1, \dots, N$ and $\boldsymbol{\varphi}_k$, but not on $\boldsymbol{\lambda}_k$, $\boldsymbol{\Lambda}_k$ or $\boldsymbol{\mu}_k$. Moreover, if f is quadratic then the iterate

even doesn't depend on $\boldsymbol{\varphi}_k$, since then D_i^j is linear and it holds

$$(D')_i^j(\boldsymbol{\varphi}_k)\boldsymbol{\delta\varphi}_k + D_i^j(\boldsymbol{\varphi}_k) = D_i^j(\boldsymbol{\varphi}_{k+1})$$

and the linear system (209)-(211) is equivalent to

$$\begin{aligned} D_i^j(\boldsymbol{\varphi}_{k+1}) - (\lambda_{k+1})_j - (\Lambda_{k+1})_i = 0 \quad & \forall i \in \mathcal{D}^j \ j = 1, \dots, N, \\ \sum_i m_i (\varphi_{k+1})_i^j = \mathfrak{m}_j |\Omega| \quad & \forall j = 1, \dots, N-1, \\ \sum_j (\varphi_{k+1})_i^j = 1 \quad & \forall i \in \mathcal{D}, \end{aligned}$$

which does not depend on $\boldsymbol{\varphi}_k$ anymore. It is a typical feature of active set methods that the iterates depend only on the active sets.

Remark 6.66. We note that the iterates of Algorithm 6.2 are (almost) independent of the choice of the constant c. This is also the case for other unilaterally constrained problems, see [HIK02]. The only influence of c is the determination of the initial active sets $\mathcal{A}_0^j$. This can be seen as follows. Let $k \geq 1$. The active set depends on the value of $(\mu_k)_i^j - c(\varphi_k)_i^j$. If $(\mu_k)_i^j \neq 0$ then $i \in \mathcal{A}_{k-1}^j$ due to (207) and thus $(\varphi_k)_i^j = 0$ by (206). Hence, it holds

$$A_k^j = \{(\mu_k)_i^j - c(\varphi_k)_i^j > 0\} = \{(\mu_k)_i^j > 0\} \cup \{(\varphi_k)_i^j < 0\}, \quad k \geq 1$$

and the set on the right hand side is independent of c. Since c only occurs in the definition of the active sets, but not in the remaining equations, the statement follows. Moreover, if the initial guess is chosen such that the complementarity $(\mu_0)_i^j(\varphi_0)_i^j = 0$ holds for all i, j, then by the arguments above the whole method is independent of c.
The iteration stays independent of c even if the reduced Newton system (209)-(211) is only solved inexactly (e.g. by an iterative solver), since $(\mu_k)_i^j = 0$ for $i \in \mathcal{I}_{k-1}^j$ and $(\varphi_k)_i^j = 0$ for $i \in \mathcal{A}_{k-1}^j$ is set explicitly in the algorithm and the above consideration remains true.

Remark 6.67. As already discussed, local convergence of the presented SSN method can only be shown for the discrete problem, but not in the continuous setting due to the lack of semismoothness. This leads to a mild mesh dependency of the method. To overcome this, one could regularize the optimization problem by e.g. using a Moreau-Yosida relaxation of the constraint $\boldsymbol{\varphi} \geq 0$ as discussed in [IK08]. This approximation replaces the hard inequality constraint $\boldsymbol{\varphi} \geq 0$ by adding the nonsmooth penalty term $\frac{1}{2d} \| \min(0, \tilde{\boldsymbol{\mu}} + d\boldsymbol{\varphi}) \|_{L^2}^2$ to the cost functional, where $\tilde{\boldsymbol{\mu}}$ is a shift parameter and $d > 0$ is a penalty parameter. The penalty term is Lipschitz continuously differentiable and thus a SSN method can be used to solve the corresponding KKT system. In this case the SSN method is well defined on the continuous level and thus mesh independence is obtained. However, this approximation leads to an important disadvantage. It does not hold

$$\begin{aligned} (\boldsymbol{\varphi}_{k+1})^j_{\mathcal{A}_k^j} &= 0 \\ (\boldsymbol{\varphi}_{k+1})^j_{\mathcal{I}_k^j \setminus \mathcal{D}_k^j} &= 1 \end{aligned}$$

anymore in each Newton iteration, cf. (206) and (208), and thus the Newton system cannot be reduced to a smaller system with degrees of freedom only on the interfacial sets $\mathcal{D}^j$. This leads to a Newton system with a much higher number of unknowns, which is discussed below. Moreover, one has to compute solutions for a sequence of regularized problems while letting $d \to \infty$. We experienced that for this procedure one needs in fact

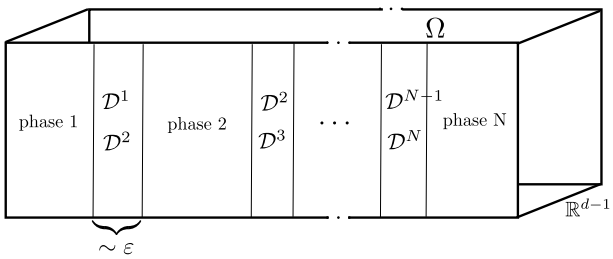


Figure 5: A typical phase distribution in $\Omega \subset \mathbb{R}^d$

more Newton iterations than for the original unregularized problem itself, even on the finest meshes we used.
Because of these reasons it is in our case more efficient to use the mesh dependent PDAS method on the unregularized problem, since the mesh dependency is very mild, see Section 6.13.3.

We give a short comparison of the degrees of freedom (DOFs) one can save by reducing the Newton system to DOFs on the interface (i.e. on the index sets $\mathcal{D}^j$). Assume that the phase field $\boldsymbol{\varphi}$ looks like in Figure 5, which shows a typical structure of the solution, where bulk regions of pure phases are separated by thin interfaces of order ε. In this example the phases are ordered from left to right, i.e. phase 1 is present on the left hand side, which is followed by an interface between phase 1 and phase 2 ($\mathcal{D}^1$ in the figure) and so on. Here we assume that on the interface separating phase i and phase j, the phases other than i and j vanish. This holds true if the potential ψ_0 is chosen in a symmetric way, but in general also other phases are present, see Section 6.12. Further assume that $\Omega \subset \mathbb{R}^d$ is a cuboid as shown in Figure 5.
As a first case we assume that Ω is discretized by an equidistant mesh with M mesh points in each direction. Let k be the number of mesh points across the interface. A typical value would be $k = 10$. In this case we get NM^d DOFs for $\boldsymbol{\varphi}$ in the full system. In the reduced Newton system (209)-(211) only the values of $\boldsymbol{\varphi}^j$ in $\mathcal{D}^j$ appear. In this example it holds $|\mathcal{D}^1| = |\mathcal{D}^N| = kM^{d-1}$ and $|\mathcal{D}^j| = 2kM^{d-1}$, $1 < j < N$, since the phases in the middle have interfaces on the left and on the right hand side. This gives a total of $2kM^{d-1}(N-1)$ unknowns for $\boldsymbol{\varphi}$ in the reduced Newton system. For a concrete comparison consider the realistic values $N = 3$, $M = 512$, $k = 10$ and $d = 2$. In the full Newton system we get about 786000 DOFs for $\boldsymbol{\varphi}$, whereas in the reduced Newton system we get only 20000, which is about 2.5% of the original number.
Now assume that Ω is discretized by a locally refined mesh, which is fine on $\mathcal{D}^j$, $j = 1, \dots, N$ and coarse in the bulk regions. We assume that the same mesh is used for the different components φ^j of $\boldsymbol{\varphi}$. Such a mesh is typically used for phase field evolutions, e.g. in [BNS04, BGSS13a, BFGS14]. We can neglect the DOFs in the bulk region and get $NkM^{d-1}(N-1)$ DOFs for $\boldsymbol{\varphi}$ in the unreduced Newton system and $2kM^{d-1}(N-1)$ DOFs in the reduced system as above. This means that for $N = 2$ the unreduced system has the same number of DOFs, for $N = 3$ we get 50% more DOFs and for $N = 4$ we get 100% more DOFs.
It is also possible to use different meshes for each component φ^j. If the mesh Ω_h^j for φ^j is chosen fine on $\mathcal{D}^j$ and coarse elsewhere, then the number of DOFs for the unreduced system would be the same as for the reduced system. However, the discretization and implementation of the Newton method would be more involved in this case. In particular it is not obvious how to discretize $\boldsymbol{\Lambda}$, which appears in the gradient equation for every

phase.
Note that when considering the topology optimization problem it is not reasonable to take a very coarse mesh in the bulk region, since this would lead to very high discretization errors in the elasticity equation. Unless one uses different meshes for the control and the state variable, one has to consider also mesh points in the bulk region. In this case the savings of DOFs lies somewhere between the equidistant mesh and the adaptive mesh discussed above. An adaptive mesh which can resolve the control and the state simultaneously can be found in Figure 7.
Table 1 summarizes the results.
In the case that an unsymmetric potential ψ_0 is used, each phase can be inactive on each

mesh	DOFs for full system	DOFs for reduced system
equidistant	NM^d	$2kM^{d-1}(N-1)$
coarse in bulk	$NkM^{d-1}(N-1)$	$2kM^{d-1}(N-1)$

Table 1: Degrees of freedom for $\boldsymbol{\varphi}$ in the Newton system on a mesh with M^d mesh points, k points across the interface and N phases. The situation in Figure 5 is considered.

interface, i.e. each set $\mathcal{D}^j$, $j = 1, \dots, N$, can have contributions on the region between phase 1 and phase 2 in Figure 5. In the worst case we have thus $NkM^{d-1}(N-1)$ DOFs for $\boldsymbol{\varphi}$ in the reduced Newton system.

6.10.2 Stopping criterion, initial guess and damping strategy

In the case that f is quadratic, we stop the method if the active sets do not change anymore. In this case we get from the definition of the active sets that

$$\begin{aligned} (\boldsymbol{\varphi}_k)^j = 0, \quad (\boldsymbol{\mu}_k)^j > 0 \quad \text{on } \mathcal{A}_k^j \\ (\boldsymbol{\varphi}_k)^j \geq 0, \quad (\boldsymbol{\mu}_k)^j = 0 \quad \text{on } \mathcal{I}_k^j \end{aligned}$$

and thus the complementarity condition is fulfilled. Note that in the used definition of the active sets an index i, which lacks strict complementarity, i.e. with $\varphi_i^j = \mu_i^j = 0$, counts as inactive. For quadratic f the remaining equations in the KKT system are linear and thus are fulfilled in each Newton step if exact arithmetic is assumed. We conclude that a solution of the KKT system is found if the active sets do not change anymore.
In the case that f is nonlinear, e.g. if $f = j$, the gradient equation in the KKT system is in general a nonlinear equation. Thus the KKT system may not be fulfilled even if the active set do not change anymore. We therefore stop the method if $\|G(\boldsymbol{\varphi}_k, \boldsymbol{\lambda}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\mu}_k)\|$ is below a given tolerance. This stopping criterion is usually not used for Newton type methods, since it depends on the scaling of G. However, for our purpose this gives satisfactory results.

As initial guess one needs values for $(\boldsymbol{\varphi}_0, \boldsymbol{\lambda}_0, \boldsymbol{\Lambda}_0, \boldsymbol{\mu}_0)$. If only an initial guess for $\boldsymbol{\varphi}$ is given, e.g. as an approximation of the minimizer computed by the VMPT method, then one can utilize the discrete versions of equations (145), (146) and (135), which were used to show uniqueness of Lagrange multipliers, to compute initial guesses for $\boldsymbol{\lambda}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\mu}$. This is done if $f = j$, which works very well in practice. In case that f is the functional in a projection type subproblem, we use the solution of the previous projection as initial guess. Note that in this case the method is active set driven, as already discussed, and thus it is sufficient to provide an initial guess for the active sets A_0^j instead of $(\boldsymbol{\varphi}_0, \boldsymbol{\lambda}_0, \boldsymbol{\Lambda}_0, \boldsymbol{\mu}_0)$. This

warm start technique normally leads to very good performance, which justifies the PDAS method to be an adequate solver for the projection type subproblem, being advantageous over e.g. interior point methods.

Note that there are also other kinds of PDAS methods to solve a nonlinear complementarity problem. For instance in [IK08] a PDAS method is given, where in each iteration a nonlinear equation has to be solved. However, such methods don't stem from a Newton method. The equivalence of our method to a semismooth Newton method enables us to develop a globalization strategy. It is known that Newton-type methods only converge locally, i.e. if the initial guess is near the solution, cf. Theorem 6.69. For certain classes of problems one can even show global convergence of the PDAS method [IK08], i.e. convergence independent of the initial guess. However, this does not apply to our applications. Thus we use a damping strategy which is inspired by ideas in [IK08]. It will not result in global convergence, but will enhance the convergence radius considerably. We use the following crude damping: Let $(\boldsymbol{\delta\varphi}_k, \boldsymbol{\delta\lambda}_k, \boldsymbol{\delta\Lambda}_k, \boldsymbol{\delta\mu}_k)$ be the solution of the Newton system

$$M(\boldsymbol{\delta\varphi}_k, \boldsymbol{\delta\lambda}_k, \boldsymbol{\delta\Lambda}_k, \boldsymbol{\delta\mu}_k) = -G(\boldsymbol{\varphi}_k, \boldsymbol{\lambda}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\mu}_k),$$

which can be calculated as in Algorithm 6.2. Then, find an $\alpha \in (0, 1]$, such that

$$\|G(\boldsymbol{\varphi}_k + \alpha\boldsymbol{\delta\varphi}_k, \boldsymbol{\lambda}_k + \alpha\boldsymbol{\delta\lambda}_k, \boldsymbol{\Lambda}_k + \alpha\boldsymbol{\delta\Lambda}_k, \boldsymbol{\mu}_k + \alpha\boldsymbol{\delta\mu}_k)\| < \|G(\boldsymbol{\varphi}_k, \boldsymbol{\lambda}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\mu}_k)\| \tag{214}$$

is fulfilled. Finally we update the iterate by

$$(\boldsymbol{\varphi}_{k+1}, \boldsymbol{\lambda}_{k+1}, \boldsymbol{\Lambda}_{k+1}, \boldsymbol{\mu}_{k+1}) = (\boldsymbol{\varphi}_k + \alpha\boldsymbol{\delta\varphi}_k, \boldsymbol{\lambda}_k + \alpha\boldsymbol{\delta\lambda}_k, \boldsymbol{\Lambda}_k + \alpha\boldsymbol{\delta\Lambda}_k, \boldsymbol{\mu}_k + \alpha\boldsymbol{\delta\mu}_k).$$

At least this strategy prevents cycling in the active sets for quadratic f, since the norm of G decreases monotonically and thus an active set cannot occur twice during the iteration. Note that an α fulfilling (214) may not exist. Also note that although the iterates of the undamped PDAS method do not depend on the constant c, cf. Remark 6.66, the norm of G does and thus c influences the damping method. We emphasize that the local convergence analysis of the PDAS method below is for the undamped method.
For the PDAS method applied to the projection type subproblem this damping works in almost all cases. Only at the first few iterations of the VMPT method it can happen that the PDAS method does not converge since the initial guess can be far away from the solution of the projection. In this case we reduce the weight $\tilde{\lambda}$ of the derivative j' in the projection type subproblem and restart the PDAS iteration. This works since we proved that for $\tilde{\lambda} \to 0$ the solution of the projection type subproblem converges to the previous iterate of the VMPT method, which is explicitly known and can be used as a good initial guess for the PDAS method, see Corollary 4.33.

6.10.3 Local convergence theory

In the following we show local superlinear convergence of the discrete PDAS method. In the case that f is the functional in the projection type subproblem it will only be needed that the inactive sets are connected. For general functionals f a second order sufficient condition and strict complementarity is assumed additionally. We will also see that the inactive sets will be connected in practice and that the LICQ constraint qualification then holds.

We first show local convergence in the case that f is the functional of the projection type subproblem. In this case we have

$$f(\boldsymbol{\varphi}) = \frac{1}{2}a(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}},\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}) + \tilde{\lambda}\left\langle j'(\tilde{\boldsymbol{\varphi}}),\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}\right\rangle \quad \forall \boldsymbol{\varphi}\in H^1(\Omega)^N\cap L^\infty(\Omega)^N$$
$$D_i^j(\boldsymbol{\varphi}) = \frac{a(\sum_{k,l}\varphi_k^l\chi_k\boldsymbol{e}_l-\tilde{\boldsymbol{\varphi}},\chi_i\boldsymbol{e}_j)+\tilde{\lambda}\left\langle j'(\tilde{\boldsymbol{\varphi}}),\chi_i\boldsymbol{e}_j\right\rangle}{m_i} \quad \forall \boldsymbol{\varphi}=(\varphi_k^l)_{kl}$$
$$(D')_i^j(\boldsymbol{\varphi})\boldsymbol{\eta} = \frac{a(\sum_{k,l}\eta_k^l\chi_k\boldsymbol{e}_l,\chi_i\boldsymbol{e}_j)}{m_i} \quad \forall \boldsymbol{\varphi}=(\varphi_k^l)_{kl},\ \boldsymbol{\eta}=(\eta_k^l)_{kl}.$$

for some $\tilde{\boldsymbol{\varphi}}\in H^1(\Omega)^N\cap L^\infty(\Omega)^N$ (which is the current iterate in the outer VMPT method). We put a tilde above the scaling parameter λ to distinguish it from the Lagrange multiplier for the mass constraint. For the convergence theory of the VMPT method we assumed that the inner products fulfill **(A8)-(A12)**. However, here we will only need that a is an inner product i.e. only **(A8)**. Thus the theory also applies if the metric a comes from a BFGS update.
Recall that the discrete projection problem is strictly convex and thus exhibits a unique solution. If the inactive sets are connected, there also exist unique Lagrange multipliers (see Corollary 6.70 below), and the KKT system is equivalent to the projection type subproblem.
As in [BGSS13a] we have to impose a condition on the solution of the projection type subproblem in order to get local convergence. This condition involves connectivity of the *s*trict inactive sets, which we define as

$$\mathcal{B}^j := \{i\in\{1,\dots,J\}\mid \mu_i^j - c\varphi_i^j < 0\}\subset\mathcal{I}^j,\quad j=1,\dots,N.$$

The reason why we use the strict inactive set rather than $\mathcal{I}^j$ is that $i\in\mathcal{B}^j$ is stable under small perturbations in μ_i^j and φ_i^j, which is not given for $\mathcal{I}^j$. We also note that in the solution of the KKT system it holds

$$\mathcal{B}^j = \{i\in\{1,\dots,J\}\mid \overline{\varphi}_i^j > 0\},$$
$$\mathcal{I}^j = \{i\in\{1,\dots,J\}\mid \overline{\mu}_i^j = 0\}.$$

Thus, if strict complementarity is given, we have $\mathcal{B}^j=\mathcal{I}^j$.
For arbitrary $\boldsymbol{\varphi}$ and $\boldsymbol{\mu}$ we define the graph $\mathcal{G}(\boldsymbol{\varphi},\boldsymbol{\mu})$ consisting of the nodes $\{1,\dots,N\}$ with an edge between k and j if and only if $\mathcal{B}^k\cap\mathcal{B}^j\neq\emptyset$.
Let $M\in\partial G$ be given by the choice $N\in\partial P_{[0,\infty)}$ with $N(0)=0$ as noted above. To apply the abstract convergence result for semismooth Newton methods, see Theorem 3.4, one has to show $\{M\}$-semismoothness of G, invertibility of M near the solution and uniform boundedness of M^{-1} near the solution.
We start by showing invertibility of M under certain conditions.

Lemma 6.68. *Let f be the functional in the projection type subproblem and let $\boldsymbol{\varphi}$, $\boldsymbol{\lambda}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\mu}$ be arbitrary. Let $M\in\partial G(\boldsymbol{\varphi},\boldsymbol{\lambda},\boldsymbol{\Lambda},\boldsymbol{\mu})$ as above, assume $\bigcup_j\mathcal{I}^j=\{1,\dots,J\}$ and that $\mathcal{G}(\boldsymbol{\varphi},\boldsymbol{\mu})$ is connected.*
Then M is invertible.

Proof. The proof is similar to [BGSS13a]. We consider the discrete case, thus M is in-

vertible if and only if its nullspace is trivial. Let $M(\boldsymbol{\delta\varphi}, \boldsymbol{\delta\lambda}, \boldsymbol{\delta\Lambda}, \boldsymbol{\delta\mu}) = 0$, i.e.

$$(D')_i^j(\boldsymbol{\varphi})\boldsymbol{\delta\varphi} - \delta\lambda_j - \delta\Lambda_i - \delta\mu_i^j = 0 \quad \forall i = 1,\dots,J,\ j = 1,\dots,N \tag{215}$$

$$\sum_i m_i \delta\varphi_i^j = 0 \quad \forall j = 1,\dots,N-1 \tag{216}$$

$$\sum_j \delta\varphi_i^j = 0 \quad \forall i = 1,\dots,J \tag{217}$$

$$\delta\mu_i^j - N_i^j(\delta\mu_i^j - c\delta\varphi_i^j) = 0 \quad \forall i = 1,\dots,J,\ j = 1,\dots,N \tag{218}$$

with

$$N_i^j = \begin{cases} 1 & \mu_i^j - c\varphi_i^j > 0 \\ 0 & \mu_i^j - c\varphi_i^j \le 0 \end{cases} \tag{219}$$

and $\delta\lambda_N := 0$. We multiply (215) by $m_i\delta\varphi_i^j$ and sum up over $i = 1,\dots,J$ and $j = 1,\dots,N$. We obtain

$$\underbrace{\sum_{i,j} m_i\delta\varphi_i^j (D')_i^j(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}}_{a(\sum_{k,l}\delta\varphi_k^l\chi_k\boldsymbol{e}_l, \sum_{k,l}\delta\varphi_k^l\chi_k\boldsymbol{e}_l)} - \sum_j \underbrace{\delta\lambda_j \sum_i m_i\delta\varphi_i^j}_{=0} - \sum_i \delta\Lambda_i m_i \underbrace{\sum_j \delta\varphi_i^j}_{=0} - \sum_{i,j} m_i \underbrace{\delta\varphi_i^j\delta\mu_i^j}_{=0} = 0, \tag{220}$$

where we used (216)-(218) and $\delta\lambda_N = 0$. Note that from (218) we get $\delta\varphi_i^j\delta\mu_i^j = 0$ for all i and j. Using that a is positive definite on $\mathbb{X}\cap\mathbb{D} = H^1_{(0)}(\Omega)^N \cap L^\infty(\Omega)^N$ we get $\boldsymbol{\delta\varphi} = 0$. Note that $\sum_{k,l}\delta\varphi_k^l\chi_k\boldsymbol{e}_l \in \mathbb{X}\cap\mathbb{D}$ holds, since

$$\int_\Omega \sum_{k,l}\delta\varphi_k^l\chi_k\boldsymbol{e}_l = \sum_l \boldsymbol{e}_l \sum_k \delta\varphi_k^l \int_\Omega \chi_k = \sum_l \boldsymbol{e}_l \sum_k \delta\varphi_k^l m_k = 0.$$

From the remaining equation

$$\delta\lambda_j + \delta\Lambda_i + \delta\mu_i^j = 0 \quad \forall i = 1,\dots,J,\ j = 1,\dots,N$$

we get that the Lagrange multipliers vanish by the same arguments we used to prove uniqueness of Lagrange multipliers: Subtracting the equations for $j = k$ and $j = l$ we get

$$\delta\lambda_k - \delta\lambda_l + \delta\mu_i^k - \delta\mu_i^l = 0 \quad \forall i = 1,\dots,J.$$

From (218) we get $\delta\mu_i^j = 0$ for all $i \in \mathcal{B}^j$. Thus, if $\mathcal{B}^k \cap \mathcal{B}^l \neq \varnothing$ we can choose $i \in \mathcal{B}^k \cap \mathcal{B}^l$ and arrive at

$$\delta\lambda_k - \delta\lambda_l = 0 \quad \text{if } \mathcal{B}^k \cap \mathcal{B}^l \neq \varnothing.$$

From the connectedness of $\mathcal{G}(\boldsymbol{\varphi}, \boldsymbol{\mu})$ and from $\delta\lambda_N = 0$ we get $\boldsymbol{\delta\lambda} = 0$. From the assumption $\bigcup_j \mathcal{I}^j = \{1,\dots,J\}$ we find for any i some index k such that $i \in \mathcal{I}^k$. For this k we get, using (218),

$$0 = \delta\Lambda_i + \delta\mu_i^k = \delta\Lambda_i,$$

thus $\boldsymbol{\delta\Lambda} = 0$ and finally $\boldsymbol{\delta\mu} = 0$ by (215). □

Theorem 6.69. *Let f be the functional in the projection type subproblem and let $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\Lambda}},$*

$\overline{\boldsymbol{\mu}}$) *be a solution of the discrete KKT system and assume that* $\mathcal{G}(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$ *is connected. Then the iterates of the PDAS Algorithm 6.2 converge superlinearly to the solution provided that* $\|\boldsymbol{\varphi}_0 - \overline{\boldsymbol{\varphi}}\| + \|\boldsymbol{\mu}_0 - \overline{\boldsymbol{\mu}}\|$ *is sufficiently small.*

Proof. First of all we note that the iterates are independent of the choice of $\boldsymbol{\lambda}_0$ and $\boldsymbol{\Lambda}_0$. Thus it is not necessary that $(\boldsymbol{\lambda}_0, \boldsymbol{\Lambda}_0)$ is close to $(\overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\Lambda}})$. In the following we check the assumptions of the abstract Theorem 3.4. First of all we have to show $\bigcup_j \mathcal{I}^j = \{1, \dots, J\}$ to get the invertibility of M due to the previous lemma. Let $i \in \{1, \dots, J\}$. From the sum constraint $\sum_{j=1}^N \overline{\varphi}_i^j = 1$ we get $\overline{\varphi}_i^k \geq \frac{1}{N}$ for some k. From complementarity we get $\overline{\mu}_i^k = 0$. Thus we have $\mu_i^k - c\varphi_i^k \leq 0$ for all $(\boldsymbol{\varphi}, \boldsymbol{\mu})$ close to $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$, i.e. the k-th phase is inactive at node i. Since i was arbitrary we find for any i an inactive phase and thus $\bigcup_j \mathcal{I}^j = \{1, \dots, J\}$ holds in a neighborhood of $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$. This holds since the set $\{1, \dots, J\}$ is finite.
The next thing we have to show is that $\mathcal{G}(\boldsymbol{\varphi}, \boldsymbol{\mu})$ is connected near the solution. Let (j, k) be an edge in the graph $\mathcal{G}(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$. Then we find some node i such that $i \in \mathcal{B}^k \cap \mathcal{B}^j$, i.e. it holds $\overline{\mu}_i^j - c\overline{\varphi}_i^j < 0$ and $\overline{\mu}_i^k - c\overline{\varphi}_i^k < 0$. Since this also holds in a neighborhood of $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$ we conclude that (j, k) is an edge in any graph $\mathcal{G}(\boldsymbol{\varphi}, \boldsymbol{\mu})$ in that neighborhood and thus $\mathcal{G}(\boldsymbol{\varphi}, \boldsymbol{\mu})$ is connected near the solution. Lemma 6.68 then guarantees the invertibility of M near $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$.
To see that M^{-1} is uniformly bounded we note that M only depends on the active sets rather than on $(\boldsymbol{\varphi}, \boldsymbol{\lambda}, \boldsymbol{\Lambda}, \boldsymbol{\mu})$ directly. This is true since $(D')_i^j(\boldsymbol{\varphi})$ does not depend on $\boldsymbol{\varphi}$. In the discrete case there are only finitely many choices of active sets and thus we conclude that M^{-1} has to be uniformly bounded.
The semismoothness of G has already been discussed above. The main argument is that D_i^j is continuously differentiable and that the projection $P_{[0,\infty)}$ is semismooth in finite dimension. By a chain rule for semismooth functions [HPUU08, Theorem 2.10], also $\mu_i^j - P_{[0,\infty)}(\mu_i^j - c\varphi_i^j)$ is semismooth.
Thus all assumptions are fulfilled and Theorem 3.4 proves the statement. □

Corollary 6.70. *Let* f *be arbitrary and let* $\overline{\boldsymbol{\varphi}}$ *be a solution of the discretized problem* (202)*, where the redundant constraint* $\boldsymbol{m}^T \boldsymbol{\varphi}^N = \mathfrak{m}_N |\Omega|$ *is dropped. If* $\mathcal{G}(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$ *is connected then the LICQ constraint qualification holds at* $\overline{\boldsymbol{\varphi}}$*. In particular the Lagrange multipliers are unique.*

Proof. First of all note that $\mathcal{G}(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$ is independent of $\overline{\boldsymbol{\mu}}$ since $\mathcal{B}^j = \{i \mid \overline{\varphi}_i^j > 0\}$.
We have to show that the gradients of the active constraints are linearly independent, i.e. from

$$a_j \boldsymbol{m} + \boldsymbol{b} + \boldsymbol{c}^j = 0, \quad \forall j = 1, \dots, N$$

with $\boldsymbol{a} \in \mathbb{R}^N$, $\boldsymbol{b} \in \mathbb{R}^J$, $\boldsymbol{c} \in \mathbb{R}^{J \times N}$, $a_N := 0$ and $c_i^j := 0$ if $i \in \mathcal{B}^j$ it follows $\boldsymbol{a} = 0$, $\boldsymbol{b} = 0$ and $\boldsymbol{c} = 0$. Here, $\boldsymbol{m} \in \mathbb{R}^J$ is the finite element mass vector, $\boldsymbol{a}$ corresponds to the mass constraint, $\boldsymbol{b}$ to the sum constraint and $\boldsymbol{c}$ to the active inequality constraints.
As in the proof of Theorem 6.69 we get $\bigcup_j \mathcal{B}^j = \{1, \dots, J\}$ and then the assumption follows as in the second part of the proof of Theorem 6.68. □

The next lemma yields convergence in finitely many steps, which is typically given for PDAS methods.

Lemma 6.71. *Let* f *be quadratic, e.g. the functional in the projection type subproblem or in the SQP subproblem. Assume the unique solvability of the Newton system in each iteration. If the iterates of the PDAS algorithm 6.2 converge to the solution then they converge in finitely many steps.*

Proof. If f is quadratic then the solution of the Newton system only depends on the active sets but not on $\boldsymbol{\varphi}_k$, cf. Remark 6.65. A combination of active sets cannot occur twice, otherwise the iterates would cycle and not converge to the solution. Since there are only finitely many combinations of active sets possible, the statement follows. □

In [BGSS13a] it is argued that $\mathcal{G}(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$ is connected if the mesh size h is small enough, since the graph in the continuous setting is connected. Here we want to use the PDAS method also for coarse meshes. We thus show that $\mathcal{G}(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$ is disconnected only if some compatibility condition is fulfilled involving the mesh and the mass $\mathfrak{m}$, which are independent data. In practice, $\mathcal{G}(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$ is always connected for any h.

Lemma 6.72. *Let f be arbitrary and let $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\Lambda}}, \overline{\boldsymbol{\mu}})$ be a solution of $G(\boldsymbol{\varphi}, \boldsymbol{\lambda}, \boldsymbol{\Lambda}, \boldsymbol{\mu}) = 0$. If $\mathcal{G}(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$ is not connected then the following compatibility condition has to be fulfilled: There exist proper subsets $\mathcal{C} \subsetneq \{1, \dots, J\}$ and $\mathcal{L} \subsetneq \{1, \dots, N\}$, $\mathcal{L} \neq \emptyset$, such that*

$$\frac{1}{|\Omega|} \sum_{i \in \mathcal{C}} m_i = \sum_{j \in \mathcal{L}} \mathfrak{m}_j \in (0, 1). \tag{221}$$

Recall that $m_i = \int \chi_i$ are the finite element masses and $\mathfrak{m}_j$ are the masses from the mass constraint.

Proof. The first part of the proof is the same argumentation as in [BGSS13a]. Since $\mathcal{G}(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$ is not connected there exists a partition of the nodes $\mathcal{M} \sqcup \mathcal{L} = \{1, \dots, N\}$, such that $\mathcal{M}$ and $\mathcal{L}$ are not connected and $\mathcal{L} \neq \emptyset$, $\mathcal{M} \neq \emptyset$. We define

$$v_i := \sum_{j \in \mathcal{M}} \overline{\varphi}_i^j, \quad w_i := \sum_{j \in \mathcal{L}} \overline{\varphi}_i^j \quad i = 1, \dots, J.$$

From $\overline{\boldsymbol{\varphi}} \geq 0$ and $\sum_j \overline{\boldsymbol{\varphi}}^j = 1$ we get

$$\boldsymbol{v} \geq 0, \quad \boldsymbol{w} \geq 0, \quad \boldsymbol{v} + \boldsymbol{w} = \mathbf{1}.$$

Let $v_i > 0$ then we find $k \in \mathcal{M}$, such that $\overline{\varphi}_i^k > 0$, thus $\overline{\mu}_i^k = 0$ and $i \in \mathcal{B}^k$. Since $\mathcal{M}$ and $\mathcal{L}$ are not connected we have $\mathcal{B}^k \cap \mathcal{B}^j = \emptyset$ for all $j \in \mathcal{L}$ and thus $\overline{\varphi}_i^j = 0$ for all $j \in \mathcal{L}$ and $w_i = 0$. Hence, $v_i = 1$. In the case $v_i = 1$ we get $w_i = 0$. Thus v_i and w_i can only have values 0 and 1.
From the mass constraint we get

$$\frac{1}{|\Omega|} \sum_i w_i m_i = \sum_{j \in \mathcal{L}} \frac{1}{|\Omega|} \sum_i \overline{\varphi}_i^j m_i = \sum_{j \in \mathcal{L}} \mathfrak{m}_j \in (0, 1). \tag{222}$$

Recall that it holds $\mathfrak{m} > 0$, $\sum_{j=1}^N \mathfrak{m}_j = 1$ and $\sum_{i=1}^J m_i = \int_\Omega 1 = |\Omega|$. Let $\mathcal{C} := \{i \mid w_i = 1\}$. By (222) we get $\emptyset \subsetneq \mathcal{C} \subsetneq \{1, \dots, J\}$ and the statement follows. □

The left hand side of the compatibility condition (221) only depends on the given mesh, whereas the right hand side is given by the prescribed material masses $\mathfrak{m}$, which are independent of the mesh. Thus, (221) holds only in special cases and local convergence of the PDAS method is in most cases given, also for coarse meshes.
The second condition for the solvability of the Newton system, namely $\bigcup_j \mathcal{I}^j = \{1, \dots, J\}$, is fulfilled in most cases, even away from the minimizer. Assume that the initial guess fulfills $\sum_j \varphi_i^j = 1$ for all i, then this holds also for every iterate of the PDAS method, which follows from (211). This holds even if the described damping strategy is used, since we

have $\sum_j \delta\varphi_i^j = 0$ for all i. From the sum constraint we then get $\bigcup_j \mathcal{I}^j = \{1, \dots, J\}$ as in the proof of Theorem 6.69.

For general functional f, we assume in addition the following second order sufficient condition in order to show local convergence of the PDAS method. For $F(\boldsymbol{\varphi}) := f(\sum \varphi_i^j \chi_i \boldsymbol{e}_j)$ let the following second order sufficient condition hold.

$$F''(\overline{\boldsymbol{\varphi}})[\boldsymbol{\delta\varphi}, \boldsymbol{\delta\varphi}] > 0 \quad \forall \boldsymbol{\delta\varphi} \neq 0,\ \boldsymbol{\delta\varphi} \in \mathcal{C}(\overline{\boldsymbol{\varphi}}) \tag{223}$$

with the critical cone

$$\mathcal{C}(\overline{\boldsymbol{\varphi}}) := \{\beta(\boldsymbol{\varphi} - \overline{\boldsymbol{\varphi}}) \mid \boldsymbol{m}^T\boldsymbol{\varphi} = \mathfrak{m}|\Omega|,\ \sum_j \boldsymbol{\varphi}^j = 1,\ \boldsymbol{\varphi} \geq 0,\ F'(\overline{\boldsymbol{\varphi}})(\boldsymbol{\varphi} - \overline{\boldsymbol{\varphi}}) = 0,\ \beta \geq 0\}.$$

One can check that this definition of the critical cone coincides with the one given in [NW06]. In particular, one can show for $\boldsymbol{w} \in \mathcal{C}(\overline{\boldsymbol{\varphi}})$ that $w_i^j = 0$ if $\mu_j^i > 0$ by using the gradient equation in the KKT system. We note that $F''(\overline{\boldsymbol{\varphi}})$ coincides with the second order derivative of the Lagrange functional $\mathcal{L}_{\varphi,\varphi}$, since the control constraints are linear. It is well known that if (223) holds for a solution $\overline{\boldsymbol{\varphi}}$ of the KKT system, then $\overline{\boldsymbol{\varphi}}$ is a strict local minimizer of F [NW06].

Theorem 6.73. *Let f be arbitrary and let $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\Lambda}}, \overline{\boldsymbol{\mu}})$ be a solution of the discrete KKT system where the second order sufficient condition* (223) *holds and assume that $\mathcal{G}(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$ is connected. Moreover, let strict complementarity hold, i.e. $\overline{\mu}_i^j + \overline{\varphi}_i^j > 0$ for all i and j. Then the iterates of the PDAS/SSN algorithm 6.2 converge superlinearly to the solution provided that $\|\boldsymbol{\varphi}_0 - \overline{\boldsymbol{\varphi}}\| + \|\boldsymbol{\mu}_0 - \overline{\boldsymbol{\mu}}\|$ is sufficiently small.*

Proof. We have to show that M is invertible and that M^{-1} is uniformly bounded in a neighborhood of $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$. The rest of the proof is the same as in Theorem 6.69.
Other than in the case that f is quadratic, M now also depends on $\boldsymbol{\varphi}$ and not only on the active sets. Thus the uniform boundedness of M^{-1} cannot be deduced from the finite number of possible active sets. However, since strict complementarity is given, yielding $\mathcal{I}^j = \{\overline{\mu}_i^j - c\overline{\varphi}_i^j < 0\}$, the active sets do not change near $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$ and thus M depends continuously on $(\boldsymbol{\varphi}, \boldsymbol{\mu})$. Hence, we get invertibility of M near $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$ and the uniform boundedness of M^{-1} by continuity arguments if we prove that M is invertible in $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$.
Without strict complementarity given, the active sets are not stable under small perturbations in $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$ and thus, M does not depend continuously on $(\boldsymbol{\varphi}, \boldsymbol{\mu})$ since N_i^j in (227) can jump from 0 to 1.
Let $(\boldsymbol{\delta\varphi}, \boldsymbol{\delta\lambda}, \boldsymbol{\delta\Lambda}, \boldsymbol{\delta\mu})$ be a solution of

$$(D')_i^j(\overline{\boldsymbol{\varphi}})\boldsymbol{\delta\varphi} - \delta\lambda_j - \delta\Lambda_i - \delta\mu_i^j = 0 \quad \forall i = 1, \dots, J,\ j = 1, \dots, N \tag{224}$$

$$\sum_i m_i \delta\varphi_i^j = 0 \quad \forall j = 1, \dots, N-1 \tag{225}$$

$$\sum_j \delta\varphi_i^j = 0 \quad \forall i = 1, \dots, J \tag{226}$$

$$\delta\mu_i^j - N_i^j(\delta\mu_i^j - c\delta\varphi_i^j) = 0 \quad \forall i = 1, \dots, J,\ j = 1, \dots, N \tag{227}$$

with

$$N_i^j = \begin{cases} 1 & \overline{\mu}_i^j - c\overline{\varphi}_i^j > 0 \\ 0 & \overline{\mu}_i^j - c\overline{\varphi}_i^j \leq 0. \end{cases} \tag{228}$$

We show $(\boldsymbol{\delta\varphi}, \boldsymbol{\delta\lambda}, \boldsymbol{\delta\Lambda}, \boldsymbol{\delta\mu}) = 0$.
The proof is similar to the proof of Lemma 6.68. We multiply (224) by $\delta\varphi_i^j m_i$ and sum up over i and j to obtain

$$\underbrace{\sum_{i,j} \delta\varphi_i^j m_i (D')_i^j(\overline{\boldsymbol{\varphi}})\boldsymbol{\delta\varphi}}_{=F''(\overline{\boldsymbol{\varphi}})[\boldsymbol{\delta\varphi},\boldsymbol{\delta\varphi}]} - \sum_j \delta\lambda_j \underbrace{\sum_i m_i \delta\varphi_i^j}_{=0} - \sum_i \delta\Lambda_i m_i \underbrace{\sum_j \delta\varphi_i^j}_{=0} - \sum_{i,j} m_i \underbrace{\delta\mu_i^j \delta\varphi_i^j}_{=0} = 0 \tag{229}$$

and thus $F''(\overline{\boldsymbol{\varphi}})[\boldsymbol{\delta\varphi}, \boldsymbol{\delta\varphi}] = 0$. We show $\boldsymbol{\delta\varphi} \in \mathcal{C}(\overline{\boldsymbol{\varphi}})$. Then we get $\boldsymbol{\delta\varphi} = 0$ by (223). Let $\boldsymbol{\varphi} := \overline{\boldsymbol{\varphi}} + \frac{1}{\beta}\boldsymbol{\delta\varphi}$ for some $\beta > 0$, cf. the definition of $\mathcal{C}(\overline{\boldsymbol{\varphi}})$. Obviously it holds $\boldsymbol{m}^T\boldsymbol{\varphi} = \mathfrak{m}|\Omega|$ and $\sum_j \boldsymbol{\varphi}^j = 1$ for all $\beta > 0$. To prove $\boldsymbol{\varphi} \geq 0$ we use strict complementarity. Let $\overline{\varphi}_i^j = 0$ for some i and j. Then we have $\overline{\mu}_i^j > 0$, thus $i \in \mathcal{A}^j$ and from (227) we get $\delta\varphi_i^j = 0$. We conclude $\varphi_i^j = 0$. There are only finitely many indices for which $\overline{\varphi}_i^j > 0$ holds, hence we can choose $\beta > 0$ so large that $\boldsymbol{\varphi} \geq 0$ holds. It remains to prove $F'(\overline{\boldsymbol{\varphi}})(\boldsymbol{\varphi} - \overline{\boldsymbol{\varphi}}) = 0$, being equivalent to $F'(\overline{\boldsymbol{\varphi}})\boldsymbol{\delta\varphi} = 0$. We exploit that $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\Lambda}}, \overline{\boldsymbol{\mu}})$ is a solution of the gradient equation (198) in the discrete KKT system. We multiply (198) by $\delta\varphi_i^j m_i$ and sum up over i and j to obtain

$$\underbrace{\sum_{i,j} \delta\varphi_i^j m_i D_i^j(\overline{\boldsymbol{\varphi}})}_{=F'(\overline{\boldsymbol{\varphi}})\boldsymbol{\delta\varphi}} - \sum_j \overline{\lambda}_j \underbrace{\sum_i m_i \delta\varphi_i^j}_{=0} - \sum_i \overline{\Lambda}_i m_i \underbrace{\sum_j \delta\varphi_i^j}_{=0} - \sum_{i,j} m_i \underbrace{\overline{\mu}_i^j \delta\varphi_i^j}_{=0} = 0,$$

where $\overline{\mu}_i^j \delta\varphi_i^j = 0$ since $\overline{\mu}_i^j > 0$ implies $i \in \mathcal{A}^j$ and thus $\delta\varphi_i^j = 0$.
Thus, it follows $\boldsymbol{\delta\varphi} = 0$ by (223). To get $(\boldsymbol{\delta\lambda}, \boldsymbol{\delta\Lambda}, \boldsymbol{\delta\mu}) = 0$ we use the same arguments as in Lemma 6.68. Note that it holds $\bigcup_j \mathcal{I}^j = \{1, \ldots, J\}$, cf. the proof of Theorem 6.69. □

Note that strict complementarity is not needed in case of the projection type subproblem (see Lemma 6.68), since in this case the condition (223) holds for all $\boldsymbol{\delta\varphi}$ with $\boldsymbol{m}^T\boldsymbol{\delta\varphi} = 0$ and not only for critical directions.
If strict complementarity is not given for general f, then one cannot show $\boldsymbol{\delta\varphi} \in \mathcal{C}(\overline{\boldsymbol{\varphi}})$ anymore in the proof of Theorem 6.73. However, if the second order sufficient condition (223) is tightened by dropping the condition $\boldsymbol{\varphi} \geq 0$ in the definition of the critical cone, then one can show the invertibility of M in the solution $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$ without strict complementarity. This yields the invertibility and uniform boundedness of M near the solution by the perturbation arguments in [IK08, Thm. 8.3].

6.10.4 Numerical solution of the reduced Newton system

Next we briefly discuss how we solve the reduced Newton system (209)-(211) numerically. To obtain a symmetric linear system we first of all multiply equation (209) by m_i, (210) by -1 and equation (211) by $-m_i$, leading to

$$m_i(D')_i^j(\overline{\boldsymbol{\varphi}})\boldsymbol{\delta\varphi} - m_i\lambda_j - m_i\Lambda_i = -m_i D_i^j(\overline{\boldsymbol{\varphi}}) \quad \forall i \in \mathcal{D}^j \;\; j = 1, \ldots, N, \tag{230}$$

$$-\sum_i m_i \delta\varphi_i^j = \sum_i m_i \overline{\varphi}_i^j - \mathfrak{m}_j|\Omega| \quad \forall j = 1, \ldots, N-1, \tag{231}$$

$$-\sum_j m_i \delta\varphi_i^j = \sum_j m_i \overline{\varphi}_i^j - m_i \quad \forall i \in \mathcal{D}. \tag{232}$$

We want to write this system in a more compact form. Therefore we introduce matrices $Q^{j,k} \in \mathbb{R}^{J\times J}$, $j,k = 1,\dots,N$ with

$$Q^{j,k}\boldsymbol{\delta\varphi} := (m_i(D')_i^j(\overline{\boldsymbol{\varphi}})\widetilde{\boldsymbol{\delta\varphi}})_{i=1}^J = (f''(\sum_{m,n}\overline{\varphi}_n^m\chi_n\boldsymbol{e}_m)[\chi_i\boldsymbol{e}_j,\sum_n\delta\varphi_n\chi_n\boldsymbol{e}_k])_{i=1}^J,$$

where $\widetilde{\delta\varphi}_i^j := \delta\varphi_i\delta_{kj}$. Note that $Q^{j,k}$ can depend on the current iterate $\overline{\boldsymbol{\varphi}}$. Moreover, we introduce the mass vector $\boldsymbol{m} := (m_i)_{i=1}^J$ and the lumped mass matrix $M := diag(m_1,\dots,m_J)$. We also group the index i in $\boldsymbol{\delta\varphi}^j := (\delta\varphi_i^j)_{i=1}^J$. For an index set $\mathcal{A} \subset \{1,\dots,J\}$ we introduce the restriction operator $R_\mathcal{A} : \mathbb{R}^J \to \mathbb{R}^{|\mathcal{A}|}$. Then $R_\mathcal{A}^T : \mathbb{R}^{|\mathcal{A}|} \to \mathbb{R}^J$ is the extension by zero operator. For two index sets $\mathcal{I},\mathcal{A} \subset \{1,\dots,J\}$ and a matrix $B \in \mathbb{R}^{J\times J}$, we define $B_{\mathcal{I},\mathcal{A}} := R_\mathcal{I} B R_\mathcal{A}^T \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{A}|}$. For a vector $\boldsymbol{\delta\varphi} \in \mathbb{R}^J$ we similarly define $\boldsymbol{\delta\varphi}_\mathcal{A} := R_\mathcal{A}\boldsymbol{\delta\varphi}$. We also abbreviate $\boldsymbol{g}^j := (m_i D_i^j(\overline{\boldsymbol{\varphi}}))_{i=1}^J$.
With this notation, equation (230) can be written as

$$\sum_{k=1}^N Q^{j,k}_{\mathcal{D}^j,\mathcal{D}^k}\boldsymbol{\delta\varphi}^k_{\mathcal{D}^k} - \boldsymbol{m}_{\mathcal{D}^j}\lambda_j - M_{\mathcal{D}^j,\mathcal{D}}\boldsymbol{\Lambda}_\mathcal{D} = -\boldsymbol{g}^j_{\mathcal{D}^j} - \sum_{k=1}^N Q^{j,k}_{\mathcal{D}^j,(\mathcal{D}^k)^c}\boldsymbol{\delta\varphi}^k_{(\mathcal{D}^k)^c} \quad \forall j = 1,\dots,N \tag{233}$$

Since $\boldsymbol{\delta\varphi}^k_{(\mathcal{D}^k)^c}$ is given by (206) and (208), we put it on the right hand side. Equation (231) becomes

$$-\boldsymbol{m}^T_{\mathcal{D}^j}\boldsymbol{\delta\varphi}^j_{\mathcal{D}^j} = \boldsymbol{m}^T\overline{\boldsymbol{\varphi}}^j + \boldsymbol{m}^T_{(\mathcal{D}^j)^c}\boldsymbol{\delta\varphi}^j_{(\mathcal{D}^j)^c} - \mathfrak{m}_j|\Omega| \quad \forall j = 1,\dots,N-1$$

and equation (232) transforms to

$$-\sum_j M_{\mathcal{D},\mathcal{D}^j}\boldsymbol{\delta\varphi}^j_{\mathcal{D}^j} = \sum_j M_{\mathcal{D},}\overline{\boldsymbol{\varphi}}^j + \sum_j M_{\mathcal{D},(\mathcal{D}^j)^c}\boldsymbol{\delta\varphi}^j_{(\mathcal{D}^j)^c} - \boldsymbol{m}_\mathcal{D}.$$

For $N = 3$, the resulting saddle point system is thus

$$\begin{pmatrix} Q^{1,1}_{\mathcal{D}^1,\mathcal{D}^1} & Q^{1,2}_{\mathcal{D}^1,\mathcal{D}^2} & Q^{1,3}_{\mathcal{D}^1,\mathcal{D}^3} & -\boldsymbol{m}_{\mathcal{D}^1} & 0 & -M_{\mathcal{D}^1,\mathcal{D}} \\ Q^{2,1}_{\mathcal{D}^2,\mathcal{D}^1} & Q^{2,2}_{\mathcal{D}^2,\mathcal{D}^2} & Q^{2,3}_{\mathcal{D}^2,\mathcal{D}^3} & 0 & -\boldsymbol{m}_{\mathcal{D}^2} & -M_{\mathcal{D}^2,\mathcal{D}} \\ Q^{3,1}_{\mathcal{D}^3,\mathcal{D}^1} & Q^{3,2}_{\mathcal{D}^3,\mathcal{D}^2} & Q^{3,3}_{\mathcal{D}^3,\mathcal{D}^3} & 0 & 0 & -M_{\mathcal{D}^3,\mathcal{D}} \\ -\boldsymbol{m}^T_{\mathcal{D}^1} & 0 & 0 & 0 & 0 & 0 \\ 0 & -\boldsymbol{m}^T_{\mathcal{D}^2} & 0 & 0 & 0 & 0 \\ -M_{\mathcal{D},\mathcal{D}^1} & -M_{\mathcal{D},\mathcal{D}^2} & -M_{\mathcal{D},\mathcal{D}^3} & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\delta\varphi}^1_{\mathcal{D}^1} \\ \boldsymbol{\delta\varphi}^2_{\mathcal{D}^2} \\ \boldsymbol{\delta\varphi}^3_{\mathcal{D}^3} \\ \lambda_1 \\ \lambda_2 \\ \boldsymbol{\Lambda}_\mathcal{D} \end{pmatrix} = rhs. \tag{234}$$

$$rhs := \begin{pmatrix} -\boldsymbol{g}^1_{\mathcal{D}^1} - \sum_{k=1}^3 Q^{1,k}_{\mathcal{D}^1,(\mathcal{D}^k)^c}\boldsymbol{\delta\varphi}^k_{(\mathcal{D}^k)^c} \\ -\boldsymbol{g}^2_{\mathcal{D}^2} - \sum_{k=1}^3 Q^{2,k}_{\mathcal{D}^2,(\mathcal{D}^k)^c}\boldsymbol{\delta\varphi}^k_{(\mathcal{D}^k)^c} \\ -\boldsymbol{g}^3_{\mathcal{D}^3} - \sum_{k=1}^3 Q^{3,k}_{\mathcal{D}^3,(\mathcal{D}^k)^c}\boldsymbol{\delta\varphi}^k_{(\mathcal{D}^k)^c} \\ \boldsymbol{m}^T\overline{\boldsymbol{\varphi}}^1 + \boldsymbol{m}^T_{(\mathcal{D}^1)^c}\boldsymbol{\delta\varphi}^1_{(\mathcal{D}^1)^c} - \mathfrak{m}_1|\Omega| \\ \boldsymbol{m}^T\overline{\boldsymbol{\varphi}}^2 + \boldsymbol{m}^T_{(\mathcal{D}^2)^c}\boldsymbol{\delta\varphi}^2_{(\mathcal{D}^2)^c} - \mathfrak{m}_2|\Omega| \\ \sum_j M_{\mathcal{D},}\overline{\boldsymbol{\varphi}}^j + \sum_j M_{\mathcal{D},(\mathcal{D}^j)^c}\boldsymbol{\delta\varphi}^j_{(\mathcal{D}^j)^c} - \boldsymbol{m}_\mathcal{D} \end{pmatrix}$$

Since f'' is symmetric, we get that $(Q^{j,k})^T = Q^{k,j}$ from the following equation.

$$\begin{aligned}(\boldsymbol{\eta}, Q^{j,k}\boldsymbol{\delta\varphi}) &= \sum_i f''(\sum_{n,m} \overline{\varphi}^n_m \chi_m \boldsymbol{e}_n)[\chi_i \boldsymbol{e}_j, \sum_l \delta\varphi_l \chi_l \boldsymbol{e}_k]\eta_i \\ &= \sum_l f''(\sum_{n,m} \overline{\varphi}^n_m \chi_m \boldsymbol{e}_n)[\chi_l \boldsymbol{e}_k, \sum_i \eta_i \chi_i \boldsymbol{e}_j]\delta\varphi_l \\ &= (Q^{k,j}\boldsymbol{\eta}, \boldsymbol{\delta\varphi}).\end{aligned}$$

Moreover, it follows that

$$(Q^{j,k}_{\mathcal{D}^j,\mathcal{D}^k})^T = (R_{\mathcal{D}^j} Q^{j,k} R^T_{\mathcal{D}^k})^T = R_{\mathcal{D}^k} Q^{k,j} R^T_{\mathcal{D}^j} = Q^{k,j}_{\mathcal{D}^k,\mathcal{D}^j}.$$

Similarly, $(M_{\mathcal{D},\mathcal{D}^j})^T = M_{\mathcal{D}^j,\mathcal{D}}$. Thus, the matrix in (234) is symmetric. We therefore use a MINRES solver [PS75] for the reduced Newton system (234). If the entries of the matrices $Q^{j,k}$ can be calculated explicitly and the dimension J is not too high, i.e. if the mesh parameter h is not too small, we use the direct solver UMFPACK [Dav07] to solve the saddle point system. This is usually much faster than the MINRES solver. The entries of $Q^{j,k}$ are explicitly known if e.g. the H^1 projection is considered. In this case we have

$$(Q^{j,k})_{o,p} = a(\chi_o \boldsymbol{e}_j, \chi_p \boldsymbol{e}_k) = \delta_{k,j} \int_\Omega \nabla\chi_o \cdot \nabla\chi_p.$$

Thus, $Q^{j,k} = \delta_{k,j} S$, where S is the stiffness matrix with $S_{i,j} = \int_\Omega \nabla\chi_i \cdot \nabla\chi_j$ and the matrix in (234) becomes

$$\begin{pmatrix} S_{\mathcal{D}^1,\mathcal{D}^1} & 0 & 0 & -\boldsymbol{m}_{\mathcal{D}^1} & 0 & -M_{\mathcal{D}^1,\mathcal{D}} \\ 0 & S_{\mathcal{D}^2,\mathcal{D}^2} & 0 & 0 & -\boldsymbol{m}_{\mathcal{D}^2} & -M_{\mathcal{D}^2,\mathcal{D}} \\ 0 & 0 & S_{\mathcal{D}^3,\mathcal{D}^3} & 0 & 0 & -M_{\mathcal{D}^3,\mathcal{D}} \\ -\boldsymbol{m}^T_{\mathcal{D}^1} & 0 & 0 & 0 & 0 & 0 \\ 0 & -\boldsymbol{m}^T_{\mathcal{D}^2} & 0 & 0 & 0 & 0 \\ -M_{\mathcal{D},\mathcal{D}^1} & -M_{\mathcal{D},\mathcal{D}^2} & -M_{\mathcal{D},\mathcal{D}^3} & 0 & 0 & 0 \end{pmatrix}$$

Note that this matrix is very similar to the matrix obtained for one time step of the non-local vector-valued Allen-Cahn variational inequality, cf. [BGSS13a, Sar10]. However, our matrix is more sparse, since we decided to drop the Lagrange multiplier for the redundant mass constraint $\boldsymbol{m}^T\boldsymbol{\varphi}^N = \mathfrak{m}_N|\Omega|$, whereas in [BGSS13a, Sar10], the additional constraint $\sum_{i=1}^N \lambda_i = 0$ is introduced. With this additional constraint, the system matrix would be

$$\begin{pmatrix} S_{\mathcal{D}^1,\mathcal{D}^1} & 0 & 0 & -\boldsymbol{m}_{\mathcal{D}^1} & 0 & -M_{\mathcal{D}^1,\mathcal{D}} \\ 0 & S_{\mathcal{D}^2,\mathcal{D}^2} & 0 & 0 & -\boldsymbol{m}_{\mathcal{D}^2} & -M_{\mathcal{D}^2,\mathcal{D}} \\ 0 & 0 & S_{\mathcal{D}^3,\mathcal{D}^3} & \boldsymbol{m}_{\mathcal{D}^3} & \boldsymbol{m}_{\mathcal{D}^3} & -M_{\mathcal{D}^3,\mathcal{D}} \\ -\boldsymbol{m}^T_{\mathcal{D}^1} & 0 & \boldsymbol{m}^T_{\mathcal{D}^3} & 0 & 0 & 0 \\ 0 & -\boldsymbol{m}^T_{\mathcal{D}^2} & \boldsymbol{m}^T_{\mathcal{D}^3} & 0 & 0 & 0 \\ -M_{\mathcal{D},\mathcal{D}^1} & -M_{\mathcal{D},\mathcal{D}^2} & -M_{\mathcal{D},\mathcal{D}^3} & 0 & 0 & 0 \end{pmatrix}$$

where in the third row λ_3 is eliminated using $-\boldsymbol{m}_{\mathcal{D}^3}\lambda_3 = \boldsymbol{m}_{\mathcal{D}^3}\lambda_1 + \boldsymbol{m}_{\mathcal{D}^3}\lambda_2$. Also the mass constraints have to be amended in order to get a symmetric matrix.

On the other hand, the entries of $Q^{j,k}$ are not explicitly known in the case $f = j$, or if the SQP subproblem is considered, or the projection type subproblem with the second order metric a_k as in (167), since evaluation of $Q^{j,k}$ involves solving PDEs. Also if the inner product in the projection type subproblem is updated by a BFGS formula, then the

entries of $Q^{j,k}$ are not known, since evaluation is done by a recursion. In these cases the MINRES algorithm has to be used for solving the Newton system (234).

In the case that we use the PDAS method within the VMPT method to solve the projection type subproblem and we use MINRES for the Newton system, we have three nested loops:
VMPT —> PDAS —> MINRES.

When applying the iterative solver MINRES, the usage of a preconditioner is reasonable. We refer to [BSS12], where the preconditioning of the Allen-Cahn system is discussed, which has the same saddle point structure as the system here. The preconditioner is based upon a good preconditioner for the Q-block. In case of the H^1-projection one would use a multigrid method as a preconditioner for the Q-block, since it has a block diagonal structure consisting of discrete Laplacians. However, since in most cases we use UMFPACK for the H^1-projection, and also the MINRES solver works quite well in most cases without preconditioning, we didn't implement this.

The computational cost of one PDAS step is concentrated in the solution of the reduced Newton system, which is the saddle point system (209)-(211). The cost thereof highly depends on the choice of f. For the H^1-projection this is rather cheap, since UMFPACK can be used as solver. Contrary to this, in the case $f = j$, the solution of the reduced Newton system is far more expensive. The reason is that the system has to be solved iteratively, since the entries of the matrix are unknown. For each step of the iterative linear solver (inner iteration), the Q-block has to be evaluated. If $f = j$, this involves evaluation of the second order derivative of j in a certain direction $\boldsymbol{\tau\varphi}$. As we have seen in Theorem 6.44, this means computation of the state $\boldsymbol{u}$, the adjoint state $\boldsymbol{p}$, the linearized state $\boldsymbol{\tau u}$ and the linearized adjoint state $\boldsymbol{\tau p}$. The state $\boldsymbol{u}$ and adjoint state $\boldsymbol{p}$ do not change during the inner iteration and thus have to be computed only once in each PDAS iteration. On the other hand the linearized state $\boldsymbol{\tau u}$ and the linearized adjoint state $\boldsymbol{\tau p}$ have to be recomputed in each inner iteration, which is expensive. Since the linearized state and linearized adjoint equations have to be solved for varying right hand sides, it is advantageous to use a direct solver for these discrete PDEs and compute a factorization of the respective matrix at the beginning of each PDAS step. In each inner iteration it then remains to perform a forward and backward substitution. But this is of course not possible for very fine meshes.
As already seen, it holds $\boldsymbol{\tau u} = \boldsymbol{\tau p}$ for the mean compliance problem, thus the computational cost can be halved in this case.

Finally, we want to discuss how the calculation and evaluation of the Q-block is done. To evaluate the left hand side of (233) we have to calculate

$$\begin{aligned}
\sum_{k=1}^{N} Q^{j,k}_{\mathcal{D}^j,\mathcal{D}^k} \boldsymbol{\delta\varphi}^k_{\mathcal{D}^k} &= R_{\mathcal{D}^j} \sum_{k=1}^{N} Q^{j,k} R^T_{\mathcal{D}^k} \boldsymbol{\delta\varphi}^k_{\mathcal{D}^k} \\
&= R_{\mathcal{D}^j} \sum_{k=1}^{N} \left(f''\Big(\sum_{m,n} \overline{\varphi}^m_n \chi_n \boldsymbol{e}_m\Big)[\chi_i \boldsymbol{e}_j, \sum_n (R^T_{\mathcal{D}^k} \boldsymbol{\delta\varphi}^k_{\mathcal{D}^k})_n \chi_n \boldsymbol{e}_k] \right)_{i=1}^{J} \\
&= \left(f''\Big(\sum_{m,n} \overline{\varphi}^m_n \chi_n \boldsymbol{e}_m\Big)[\chi_i \boldsymbol{e}_j, \sum_{k=1}^{N} \sum_{n \in \mathcal{D}^k} \boldsymbol{\delta\varphi}^k_n \chi_n \boldsymbol{e}_k] \right)_{i \in \mathcal{D}^j}
\end{aligned}$$

for $j = 1, \dots, N$. Thus to evaluate the whole Q-block we have to calculate $f''(\overline{\boldsymbol{\varphi}}_h)[\chi_i \boldsymbol{e}_j, \boldsymbol{\delta\varphi}_h]$

for all i and j and for some given $\overline{\boldsymbol{\varphi}}_h \in S_h^N$ and $\boldsymbol{\delta\varphi}_h \in S_h^N$. In case of the projection type subproblem with inner product a it holds

$$f''(\overline{\boldsymbol{\varphi}}_h)[\chi_i \boldsymbol{e}_j, \boldsymbol{\delta\varphi}_h] = a(\chi_i \boldsymbol{e}_j, \boldsymbol{\delta\varphi}_h).$$

Assume that the inner product is given as one of the following

$$\begin{aligned}
a(\boldsymbol{v}_1, \boldsymbol{v}_2) &= \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2, \\
a(\boldsymbol{v}_1, \boldsymbol{v}_2) &= \gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2, \\
a(\boldsymbol{v}_1, \boldsymbol{v}_2) &= \gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2 + \frac{\varepsilon}{\tau_k} \int_\Omega \boldsymbol{v}_1 \cdot \boldsymbol{v}_2, \\
a(v_1, v_2) &= \gamma\varepsilon \int_\Omega \nabla v_1 \cdot \nabla v_2 + \left(\frac{\varepsilon}{\tau_k} - \frac{\gamma}{\varepsilon} \right) \int_\Omega v_1 v_2,
\end{aligned}$$

which correspond to the H^1-projection, the scaled H^1-projection and the pseudo time stepping of Allen-Cahn type with potential term taken explicitly and implicitly, respectively, as discussed in Section 6.7. Then the Q-block coincides with the tensor $(a(\chi_i \boldsymbol{e}_j, \chi_k \boldsymbol{e}_l))_{ijkl}$, which can be assembled easily, since it consists of a linear combination of mass matrix and stiffness matrix. For example for third inner product it holds

$$a(\chi_i \boldsymbol{e}_j, \chi_k \boldsymbol{e}_l) = \delta_{jl} \left(\gamma\varepsilon \int_\Omega \nabla \chi_i \cdot \nabla \chi_k + \frac{\varepsilon}{\tau_k} \int_\Omega \chi_i \chi_k \right).$$

By this way the entries of the Q-block can be calculated explicitly.

For the Cahn-Hilliard inner product

$$a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2 + \frac{1}{\tau_k} (\boldsymbol{v}_1, \boldsymbol{v}_2)_{H^{-1}}$$

the entries of Q cannot be calculated explicitly because of the H^{-1} inner product. Recall that the H^{-1} inner product can be calculated by (see (161))

$$(\boldsymbol{v}_1, \boldsymbol{v}_2)_{H^{-1}} = \int_\Omega \boldsymbol{w}_1 : \boldsymbol{v}_2$$

where $\boldsymbol{w}_1 := (-\Delta_N)^{-1}(\boldsymbol{v}_1 - \fint \boldsymbol{v}_1)$ is the weak solution of the pure Neumann problem

$$\begin{aligned}
-\Delta \boldsymbol{w}_1 &= \boldsymbol{v}_1 - \fint \boldsymbol{v}_1 \text{ in } \Omega \\
\fint \boldsymbol{w}_1 &= \boldsymbol{0} \\
\partial_\nu \boldsymbol{w}_1 &= \boldsymbol{0} \text{ on } \partial\Omega.
\end{aligned}$$

This system decouples into N scalar equations. For the numerical solution of the pure Neumann problem we refer to [BL05]. Thus, for each evaluation of the Q-block we have to solve N scalar Laplace equations. We choose the described implementation since it fits in the abstract framework of this section and we can reuse the existing code without changes. A better implementation for the H^{-1} inner product would introduce the slack variable $\boldsymbol{w}_1$ as an independent variable and append the equations of the pure Neumann problem to the Newton system. This would blow up the Newton system, but no Laplace equation has to be solved when evaluating the Q-block. The variable $\boldsymbol{w}_1$ then coincides

with the chemical potential up to the factor $-\tau_k$ and an additive constant. For details we refer to [BBG11]. However, since we don't want to examine the Cahn-Hilliard pseudo time stepping in detail, we use the inefficient implementation which eliminates $\boldsymbol{w}_1$.

If a BFGS update is considered the inner product a_k is defined recursively by

$$a_{k+1}(\boldsymbol{v}_1, \boldsymbol{v}_2) = \rho_k \left(a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) - \frac{a_k(\boldsymbol{v}_1, \boldsymbol{p}_k) a_k(\boldsymbol{p}_k, \boldsymbol{v}_2)}{a_k(\boldsymbol{p}_k, \boldsymbol{p}_k)} \right) + \frac{\langle \boldsymbol{y}_k, \boldsymbol{v}_1 \rangle \langle \boldsymbol{y}_k, \boldsymbol{v}_2 \rangle}{\langle \boldsymbol{y}_k, \boldsymbol{p}_k \rangle},$$

see (69). When assembling $(a_{k+1}(\boldsymbol{\delta\varphi}_h, \chi_i \boldsymbol{e}_j))_{ij}$, one has to be careful not to evaluate a_k four times as suggested by the recursion formula, since this would lead to very high computational cost. In fact it suffices to assemble only $(a_k(\boldsymbol{\delta\varphi}_h, \chi_i \boldsymbol{e}_j))_{ij}$. Then, $a_k(\boldsymbol{\delta\varphi}_h, \boldsymbol{p}_k)$ can be computed as a linear combination of the values $(a_k(\boldsymbol{\delta\varphi}_h, \chi_i \boldsymbol{e}_j))_{ij}$. Moreover, $(a_k(\boldsymbol{p}_k, \chi_i \boldsymbol{e}_j))_{ij}$ can be assembled in advance and stored for later use. Similarly, $a_k(\boldsymbol{p}_k, \boldsymbol{p}_k)$, $(\langle \boldsymbol{y}_k, \chi_i \boldsymbol{e}_j \rangle)_{ij}$ and $\langle \boldsymbol{y}_k, \boldsymbol{p}_k \rangle$ can be computed in advance. This procedure is similar to the (unconstrained) L-BFGS method [NW06]. However, we evaluate the BFGS-matrix and not its inverse.
Thus, $(a_{k+1}(\boldsymbol{\delta\varphi}_h, \chi_i \boldsymbol{e}_j))_{ij}$ is assembled recursively, where in each recursion one has to assemble $(a_k(\boldsymbol{\delta\varphi}_h, \chi_i \boldsymbol{e}_j))_{ij}$, compute two inner products in $\mathbb{R}^{J \times N}$ for $a_k(\boldsymbol{\delta\varphi}_h, \boldsymbol{p}_k)$ and $\langle \boldsymbol{y}_k, \boldsymbol{\delta\varphi}_h \rangle$ and put all together, which costs $5NJ$ flops. At the innermost level of the recursion, $(a_0(\boldsymbol{\delta\varphi}_h, \chi_i \boldsymbol{e}_j))_{ij}$ has to be assembled, which accounts for N matrix-vector multiplications in $\mathbb{R}^J$ if a_0 is the H^1 inner product.
The computational cost of evaluating the Q-block grows linearly with the number k of the current VMPT iteration, since k recursions have to be evaluated. To prevent that the iteration becomes too expensive we allow as in the L-BFGS method a maximum of n recursions for some $n \in \mathbb{N}$, i.e. we set $a_{k-n} = a_0$ and thus drop secant information from iterations prior to $(k-n)$. Hence, the cost grows linearly in the first n VMPT iterations and then stays constant.
As a preconditioner, one can use a preconditioner for the initialization a_0, e.g. a multigrid method if a_0 is the H^1 inner product, together with the Sherman-Morrison formula to approximate the inverse of the BFGS matrix.

Evaluating the Q-block is more involved when using the second order inner product (167), i.e.

$$a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2 + \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k) \mathcal{E}(\boldsymbol{\delta u}_1) : \mathcal{E}(\boldsymbol{\delta u}_2) + \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k) \mathcal{E}(\boldsymbol{\delta p}_1) : \mathcal{E}(\boldsymbol{\delta p}_2).$$

Recall that $\boldsymbol{\delta u}_i$ and $\boldsymbol{\delta p}_i$ are solutions of the linearized state equation and the linearized adjoint equation, respectively, which depend on $\boldsymbol{v}_i$. The inner product a_k in this form is not computable, since assembling $(a_k(\boldsymbol{\delta\varphi}_h, \chi_i \boldsymbol{e}_j))_{ij}$ would need the solution of $(2NJ+2)$ PDEs. To make a_k computable, we reformulate it using an adjoint approach. Testing the linearized state equation for $\boldsymbol{\delta u}_1$ by $\boldsymbol{\xi} = \boldsymbol{\delta u}_2$ and the linearized adjoint equation for $\boldsymbol{\delta p}_1$ by $\boldsymbol{\xi} = \boldsymbol{\delta p}_2$, we end up with

$$\begin{aligned} \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k) \mathcal{E}(\boldsymbol{\delta u}_1) : \mathcal{E}(\boldsymbol{\delta u}_2) = &- \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}_k) \boldsymbol{v}_1 \mathcal{E}(\boldsymbol{u}_k) : \mathcal{E}(\boldsymbol{\delta u}_2) + \int_\Omega \boldsymbol{f}_\varphi(\boldsymbol{\varphi}_k) \boldsymbol{v}_1 \cdot \boldsymbol{\delta u}_2 \\ &+ \int_{\Gamma_g} \boldsymbol{g}_\varphi(\boldsymbol{\varphi}_k) \boldsymbol{v}_1 \cdot \boldsymbol{\delta u}_2 \end{aligned}$$

and

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta p}_1) : \mathcal{E}(\boldsymbol{\delta p}_2) = F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k, \boldsymbol{u}_k)[\boldsymbol{v}_1, \boldsymbol{\delta p}_2] + F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi}_k, \boldsymbol{u}_k)[\boldsymbol{\delta u}_1, \boldsymbol{\delta p}_2] - \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}_k)\boldsymbol{v}_1\mathcal{E}(\boldsymbol{p}_k) : \mathcal{E}(\boldsymbol{\delta p}_2).$$

This yields

$$\begin{aligned} a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2 - \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}_k)\boldsymbol{v}_1\mathcal{E}(\boldsymbol{u}_k) : \mathcal{E}(\boldsymbol{\delta u}_2) + \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k)\boldsymbol{v}_1 \cdot \boldsymbol{\delta u}_2 \\ + \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k)\boldsymbol{v}_1 \cdot \boldsymbol{\delta u}_2 + F_{\boldsymbol{u},\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k, \boldsymbol{u}_k)[\boldsymbol{v}_1, \boldsymbol{\delta p}_2] + F_{\boldsymbol{u},\boldsymbol{u}}(\boldsymbol{\varphi}_k, \boldsymbol{u}_k)[\boldsymbol{\delta u}_1, \boldsymbol{\delta p}_2] \\ - \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}_k)\boldsymbol{v}_1\mathcal{E}(\boldsymbol{p}_k) : \mathcal{E}(\boldsymbol{\delta p}_2). \end{aligned} \tag{235}$$

For the mean compliance problem (112) and the compliant mechanism problem (116) the term $F_{\boldsymbol{u},\boldsymbol{u}}$ vanishes. For these problems we can assemble $(a_k(\chi_i \boldsymbol{e}_j, \boldsymbol{\delta\varphi}_h))_{ij}$ by first solving the linearized state equation for $\boldsymbol{\delta u}_2$, then solving the linearized adjoint equation for $\boldsymbol{\delta p}_2$ and then computing the integrals in (235) by standard finite element methods. Note that the state $\boldsymbol{u}_k$ and the adjoint $\boldsymbol{p}_k$ are already known, since they are needed to compute j and j'. Moreover, $\boldsymbol{u}_k$ and $\boldsymbol{p}_k$ don't change during the PDAS iteration and thus have to be computed only once in advance. If a direct solver is used for the linearized equations, only two forward and backward substitutions are needed per evaluation of a_k, since the left hand side of the linearized PDEs does not change during the PDAS iteration.
It is also possible to introduce $\boldsymbol{u}_2$ and $\boldsymbol{p}_2$ as independent variables in the KKT system and append the linearized state and linearized adjoint equations as additional equality constraints. Numerical experiments show that in this case the MINRES solver for the Newton system does not converge without preconditioner. The reason is probably that the linearized equations are ill conditioned if an ersatz material is used, cf. Section 5. When taking a LU decomposition as a preconditioner for the linearized equations the MINRES method converges, but it is still slower than the original MINRES method applied to the system where $\boldsymbol{u}_2$ and $\boldsymbol{p}_2$ are eliminated. We therefore decide to eliminate $\boldsymbol{u}_2$ and $\boldsymbol{p}_2$ in the KKT system.

In the case that f is the functional in the SQP subproblem, or j itself, the situation is similar to the case where the preceding a_k is used. This means the solution of the linearized state equation and the linearized adjoint equation is necessary in each MINRES step, as already mentioned above.

We note that it may happen that the graph $\mathcal{G}(\boldsymbol{\varphi}_k, \boldsymbol{\mu}_k)$ becomes disconnected during the PDAS iteration. In particular this happens at the beginning of the outer VMPT iteration if a coarse mesh is used. In this case the reduced Newton system is not solvable. To overcome this issue, we reduce the scaling parameter $\tilde{\lambda}$ in the outer VMPT iteration and restart the PDAS iteration as described in Section 6.10.2, which works very well in practice.

6.10.5 Special treatment in case of two phases

Literally to Section 6.1.3 we can reduce the general problem (189) for $N = 2$ to the following problem involving only a scalar valued phase field variable φ.

$$\begin{aligned} &\min f(\varphi) \\ &\quad -1 \leq \varphi \leq 1 \quad \text{a.e. in } \Omega \\ &\quad \fint \varphi = \mathfrak{m}, \end{aligned} \tag{236}$$

where the cost functional f has to be amended appropriately. In the following we will derive a semi-smooth Newton method applied to the KKT system similar to above. Again, this will be equivalent to a primal dual active set method if f is quadratic, and the Newton system can be reduced to a smaller linear system. Since many steps in the derivation are analogous to the vector valued case we describe it only briefly.
From Theorem 6.38 we get the existence and uniqueness of Lagrange multipliers λ and μ_1, μ_2 such that the KKT system

$$-1 \leq \overline{\varphi} \leq 1 \quad \text{a.e. in } \Omega$$

$$\fint \overline{\varphi} = \mathfrak{m}$$

$$\left\langle f'(\overline{\varphi}), \eta \right\rangle - \lambda \int_\Omega \eta - \langle \mu_1 - \mu_2, \eta \rangle_{(L^\infty)^*, L^\infty} = 0 \quad \forall \eta \in H^1(\Omega) \cap L^\infty(\Omega) \tag{237}$$

$$\langle \mu_1, \eta \rangle_{(L^\infty)^*, L^\infty} \geq 0 \quad \forall \eta \in L^\infty(\Omega), \eta \geq 0 \tag{238}$$

$$\langle \mu_2, \eta \rangle_{(L^\infty)^*, L^\infty} \geq 0 \quad \forall \eta \in L^\infty(\Omega), \eta \geq 0 \tag{239}$$

$$\langle \mu_1, 1 + \overline{\varphi} \rangle_{(L^\infty)^*, L^\infty} = 0 \tag{240}$$

$$\langle \mu_2, 1 - \overline{\varphi} \rangle_{(L^\infty)^*, L^\infty} = 0 \tag{241}$$

is satisfied. Note that the KKT system is equivalent to the vector-valued KKT system for $N = 2$ as shown in Theorem 6.38. As in the vector valued case the semismooth Newton method is not well defined in the continuous setting, thus we discretize the KKT system first. We use the same discretization for $H^1(\Omega) \cap L^\infty(\Omega)$ with P1 finite elements and the same nodal basis functions $(\chi_i)_{i=1}^J$ as above. With the coordinate vectors

$$\mu_i := \frac{\langle \mu_2 - \mu_1, \chi_i \rangle_{(L^\infty)^*, L^\infty}}{m_i}$$

$$D_i(\boldsymbol{\varphi}) := \frac{\langle f'(\sum_k \varphi_k \chi_k), \chi_i \rangle}{m_i} \quad \forall \boldsymbol{\varphi} = (\varphi_k)_k$$

we get the following discretized KKT system.

$$D(\overline{\boldsymbol{\varphi}}) - \lambda \boldsymbol{e} + \boldsymbol{\mu} = \mathbf{0} \tag{242}$$

$$\boldsymbol{m}^T \overline{\boldsymbol{\varphi}} - \mathfrak{m}|\Omega| = 0 \tag{243}$$

$$\boldsymbol{\mu} - \max(\mathbf{0}, \boldsymbol{\mu} + c(\overline{\boldsymbol{\varphi}} - \mathbf{1})) - \min(\mathbf{0}, \boldsymbol{\mu} + c(\overline{\boldsymbol{\varphi}} + \mathbf{1})) = \mathbf{0} \tag{244}$$

We don't discretize μ_1 and μ_2 separately, but only the difference $\mu_2 - \mu_1$, since μ_2 and μ_1 can be recovered from the difference as the positive and negative part, respectively. The equivalence of the min-max-equation to the complementarity condition is shown in [IK08]. The operators $\max(\mathbf{0}, .)$ and $\min(\mathbf{0}, .)$ are applied coordinatewise. Since $\max(\mathbf{0}, .)$ and $\min(\mathbf{0}, .)$ are semismooth in finite dimension [IK08], we can apply a semismooth Newton

method to solve the KKT system. For given $(\overline{\boldsymbol{\varphi}}, \overline{\lambda}, \overline{\boldsymbol{\mu}})$, the next iterate of the semismooth Newton method is $(\overline{\boldsymbol{\varphi}} + \boldsymbol{\delta\varphi}, \lambda, \boldsymbol{\mu})$, where $(\boldsymbol{\delta\varphi}, \lambda, \boldsymbol{\mu})$ is the solution of the Newton system

$$D'(\overline{\boldsymbol{\varphi}})\boldsymbol{\delta\varphi} - \lambda \boldsymbol{e} + \boldsymbol{\mu} = -D(\overline{\boldsymbol{\varphi}}) \tag{245}$$

$$\boldsymbol{m}^T \boldsymbol{\delta\varphi} = -\boldsymbol{m}^T \overline{\boldsymbol{\varphi}} + \mathfrak{m}|\Omega| \tag{246}$$

$$\boldsymbol{\mu} - N_1(\boldsymbol{\mu} - \overline{\boldsymbol{\mu}} + c\boldsymbol{\delta\varphi}) - N_2(\boldsymbol{\mu} - \overline{\boldsymbol{\mu}} + c\boldsymbol{\delta\varphi}) = \max(\boldsymbol{0}, \overline{\boldsymbol{\mu}} + c(\overline{\boldsymbol{\varphi}} - \boldsymbol{1})) + \min(\boldsymbol{0}, \overline{\boldsymbol{\mu}} + c(\overline{\boldsymbol{\varphi}} + \boldsymbol{1})). \tag{247}$$

Here, $N_1 = diag((N_1)_1, \dots, (N_1)_J) \in \mathbb{R}^{J\times J}$ and we choose

$$(N_1)_i = \begin{cases} 0 & \overline{\mu}_i + c(\overline{\varphi}_i - 1) \le 0 \\ 1 & \overline{\mu}_i + c(\overline{\varphi}_i - 1) > 0 \end{cases}.$$

Similarly, $N_2 = diag((N_2)_1, \dots, (N_2)_J)$ and we choose

$$(N_2)_i \begin{cases} 0 & \overline{\mu}_i + c(\overline{\varphi}_i + 1) \ge 0 \\ 1 & \overline{\mu}_i + c(\overline{\varphi}_i + 1) < 0 \end{cases}.$$

We introduce the active sets for the lower and upper bound,

$$\begin{aligned} \mathcal{A}^- &:= \{i \mid \overline{\mu}_i + c(\overline{\varphi}_i + 1) < 0\} \\ \mathcal{A}^+ &:= \{i \mid \overline{\mu}_i + c(\overline{\varphi}_i - 1) > 0\} \\ \mathcal{I} &:= (\mathcal{A}^- \cup \mathcal{A}^+)^c \end{aligned}$$

and observe that the min-max-equation in the Newton system is equivalent to

$$\delta\varphi_i = -\overline{\varphi}_i - 1 \quad \forall i \in \mathcal{A}^- \tag{248}$$

$$\delta\varphi_i = -\overline{\varphi}_i + 1 \quad \forall i \in \mathcal{A}^+ \tag{249}$$

$$\mu_i = 0 \quad \forall i \in \mathcal{I}. \tag{250}$$

We restrict (245) to $\mathcal{I}$ in order to compute the remaining unknowns $\boldsymbol{\delta\varphi}_{\mathcal{I}}$ and λ. Together with the mass constraint (246) we get

$$\begin{aligned} D_i'(\overline{\boldsymbol{\varphi}})\boldsymbol{\delta\varphi} - \lambda &= -D_i(\overline{\boldsymbol{\varphi}}) \quad \forall i \in \mathcal{I} \\ \boldsymbol{m}^T \boldsymbol{\delta\varphi} &= -\boldsymbol{m}^T \overline{\boldsymbol{\varphi}} + \mathfrak{m}|\Omega|, \end{aligned}$$

where the already known variables $\boldsymbol{\delta\varphi}_{\mathcal{I}^c}$ can be eliminated. As above, this can be reformulated to the symmetric linear system

$$\begin{pmatrix} Q_{\mathcal{I},\mathcal{I}} & -\boldsymbol{m}_{\mathcal{I}} \\ -\boldsymbol{m}_{\mathcal{I}}^T & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\delta\varphi}_{\mathcal{I}} \\ \lambda \end{pmatrix} = \begin{pmatrix} -\boldsymbol{g}_{\mathcal{I}} - Q_{\mathcal{I},(\mathcal{I})^c} \boldsymbol{\delta\varphi}_{(\mathcal{I})^c} \\ \boldsymbol{m}^T \overline{\boldsymbol{\varphi}} + \boldsymbol{m}_{(\mathcal{I})^c}^T \boldsymbol{\delta\varphi}_{(\mathcal{I})^c} - \mathfrak{m}|\Omega| \end{pmatrix} \tag{251}$$

with $Q := (m_i D_i'(\overline{\boldsymbol{\varphi}}))_i \in \mathbb{R}^{J\times J}$ and $\boldsymbol{g} := (m_i D_i(\overline{\boldsymbol{\varphi}}))_i \in \mathbb{R}^J$. This system is very similar to the system solved in each time step of the scalar Allen-Cahn variational inequality with nonlocal constraints [BGSS13b]. The Lagrange multiplier $\boldsymbol{\mu}$ can finally be computed by using (245),

$$\boldsymbol{\mu}_{\mathcal{I}^c} = -D_{\mathcal{I}^c}(\overline{\boldsymbol{\varphi}}) - D_{\mathcal{I}^c}'(\overline{\boldsymbol{\varphi}})\boldsymbol{\delta\varphi} + \lambda \boldsymbol{e}_{\mathcal{I}^c}. \tag{252}$$

We end up with Algorithm 6.3.

Algorithm 6.3 PDAS/SSN method for the discretized general problem (236) involving two phases

Choose $\boldsymbol{\varphi}_0$, λ_0, $\boldsymbol{\mu}_0$, $c > 0$.
$k := 0$.
while $k \leq k_{\max}$ **do**
  Define the active sets

$$\begin{aligned}\mathcal{A}_k^- &:= \{i \in \{1,\dots,J\} \mid (\mu_k)_i + c((\varphi_k)_i + 1) < 0\}\\ \mathcal{A}_k^+ &:= \{i \in \{1,\dots,J\} \mid (\mu_k)_i + c((\varphi_k)_i - 1) > 0\}\\ \mathcal{I}_k &:= (\mathcal{A}_k^- \cup \mathcal{A}_k^+)^c.\end{aligned}$$

  Set $(\boldsymbol{\delta\varphi}_k)_{\mathcal{I}^c}$ by equations (248) and (249).
  Compute $((\boldsymbol{\delta\varphi}_k)_{\mathcal{I}}, \lambda_{k+1})$ by solving the linear system (251).
  Set $\boldsymbol{\varphi}_{k+1} := \boldsymbol{\varphi}_k + \boldsymbol{\delta\varphi}_k$.
  Set $\boldsymbol{\mu}_{k+1}$ by (250) and (252).
end while

We remark that the iterates of Algorithm 6.3 depend on the choice of the constant c as opposed to the vector valued Algorithm 6.2. For bilateral constraints the constant c usually influences the likelihood that a node can switch from one active set to the other [IK08], e.g. if $i \in \mathcal{A}_{k-1}^-$, then $(\varphi_k)_i = -1$. It then holds $i \in \mathcal{A}_k^+$ if and only if $(\mu_k)_i - 2c > 0$, which is more probable for smaller c.

We prove local superlinear convergence of the method. In the vector valued case it is required that the graph $\mathcal{G}(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$ is connected. For $N = 2$ this is equivalent to the existence of a point on the interface, i.e. that it holds $0 < \varphi_{i_0}^1 < 1$ for some i_0. This in turn corresponds to $-1 < \varphi_{i_0} < 1$ in the scalar valued case, which is again needed to show unique solvability of the Newton system. For the invertibility of the vector valued system also the assumption $\bigcup_j \mathcal{I}^j = \{1,\dots,J\}$ is needed. We don't need a corresponding assumption in the scalar valued case since this is somehow fulfilled automatically, cf. Lemma 6.78, 1) and (268).

Theorem 6.74. *Let f be the functional in the projection type subproblem. Let $(\overline{\boldsymbol{\varphi}}, \overline{\lambda}, \overline{\boldsymbol{\mu}})$ be a solution of the discrete KKT system* (242)-(244)*. Let there exist $i_0 \in \{1,\dots,J\}$, such that $-1 < \overline{\varphi}_{i_0} < 1$.*
Then the iterates of the PDAS algorithm 6.3 converge superlinearly to the solution provided that $\|\boldsymbol{\varphi}_0 - \overline{\boldsymbol{\varphi}}\| + \|\boldsymbol{\mu}_0 - \overline{\boldsymbol{\mu}}\|$ is sufficiently small.

Proof. From $-1 < \overline{\varphi}_{i_0} < 1$ we get from the complementarity that $\overline{\mu}_{i_0} = 0$, thus $\mu_{i_0} + c(\varphi_{i_0} - 1) < 0$ and $\mu_{i_0} + c(\varphi_{i_0} + 1) > 0$ holds for all $(\boldsymbol{\varphi}, \boldsymbol{\mu})$ in a neighborhood of $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$. Hence, in this neighborhood it holds $(N_1)_{i_0} = (N_2)_{i_0} = 0$. We show that

$$D'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi} - \delta\lambda \boldsymbol{e} + \boldsymbol{\delta\mu} = 0 \tag{253}$$

$$\boldsymbol{m}^T \boldsymbol{\delta\varphi} = 0 \tag{254}$$

$$\boldsymbol{\delta\mu} - N_1(\boldsymbol{\delta\mu} + c\boldsymbol{\delta\varphi}) - N_2(\boldsymbol{\delta\mu} + c\boldsymbol{\delta\varphi}) = 0 \tag{255}$$

implies $(\boldsymbol{\delta\varphi}, \delta\lambda, \boldsymbol{\delta\mu}) = \mathbf{0}$ for all $(\boldsymbol{\varphi}, \boldsymbol{\mu})$ in this neighborhood, which is equivalent to the invertibility of the linearized operator. From (255) we get $\delta\varphi_i \delta\mu_i = 0$ for all i. Multiply (253) by $(m_i \delta\varphi_i)_i$ to get from the positive definiteness of the inner product a that $\boldsymbol{\delta\varphi} = 0$.

From (253) we then have $\delta\lambda = \delta\mu_i$ for all i. Equation (255) leads to $\delta\mu_{i_0} = 0$ and thus $\delta\lambda = 0$ and $\boldsymbol{\delta\mu} = 0$. The inverse is uniformly bounded since it does not depend on $\boldsymbol{\varphi}$ and there are only finitely many choices for the active sets. The statement follows from the abstract convergence theory in Theorem 3.4 as in the vector valued case. □

Note that if $\overline{\varphi}_i \in \{\pm 1\}$ and $\mu_i \neq 0$ for all i holds in the solution of the KKT system, then $\{(\mathbf{0}, a, a\boldsymbol{e}) \mid a \in \mathbb{R}\}$ is the kernel of the linearized operator given in (253)-(255), since (255) is then equivalent to $\boldsymbol{\delta\varphi} = 0$. In this case also the discrete Lagrange multipliers are not unique, since one can show that $\lambda \pm a$ and $\boldsymbol{\mu} \pm a\boldsymbol{e}$ are also solutions of the KKT system for a small enough.
The assumption $-1 < \overline{\varphi}_{i_0} < 1$ for some i_0 should be fulfilled in most cases, see Lemma 6.72.

As in the vector valued case we get two corollaries.

Corollary 6.75. *Let f be quadratic, e.g. the functional in the projection type subproblem or in the SQP subproblem. Assume the unique solvability of the Newton system in each iteration. If the iterates of the scalar PDAS algorithm 6.3 converge to the solution then they converge in finitely many steps.* □

Corollary 6.76. *Let f be arbitrary and let $\overline{\boldsymbol{\varphi}}$ be a solution of the discretized problem. If there exists $i_0 \in \{1, \dots, J\}$, such that $-1 < \overline{\varphi}_{i_0} < 1$, then the LICQ constraint qualification holds at $\overline{\boldsymbol{\varphi}}$.* □

For a general functional f, one needs as in the vector valued case the following second order sufficient condition in order to show local convergence of the PDAS method. For $F(\boldsymbol{\varphi}) := f(\sum \varphi_i \chi_i)$ let the following second order sufficient condition hold.

$$F''(\overline{\boldsymbol{\varphi}})[\boldsymbol{\delta\varphi}, \boldsymbol{\delta\varphi}] > 0 \quad \forall \boldsymbol{\delta\varphi} \neq 0,\ \boldsymbol{\delta\varphi} \in \mathcal{C}(\overline{\boldsymbol{\varphi}}) \tag{256}$$

with the critical cone

$$\mathcal{C}(\overline{\boldsymbol{\varphi}}) := \left\{ \beta(\boldsymbol{\varphi} - \overline{\boldsymbol{\varphi}}) \;\middle|\; \boldsymbol{m}^T\boldsymbol{\varphi} = \mathfrak{m}|\Omega|,\ -1 \le \boldsymbol{\varphi} \le 1,\ F'(\overline{\boldsymbol{\varphi}})(\boldsymbol{\varphi} - \overline{\boldsymbol{\varphi}}) = 0,\ \beta \ge 0 \right\}$$

Theorem 6.77. *Let $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\mu}})$ be a solution of the discrete KKT system (242)-(244) for general cost functional f, where the second order sufficient condition (256) holds and assume that there exists $i_0 \in \{1, \dots, J\}$, such that $-1 < \overline{\varphi}_{i_0} < 1$. Moreover, let strict complementarity hold, i.e. from $\overline{\mu}_i = 0$ it follows $-1 < \overline{\varphi}_i < 1$ for all i.*
Then the iterates of the PDAS algorithm 6.3 converge superlinearly to the solution provided that $\|\boldsymbol{\varphi}_0 - \overline{\boldsymbol{\varphi}}\| + \|\boldsymbol{\mu}_0 - \overline{\boldsymbol{\mu}}\|$ is sufficiently small.

Proof. The proof is very similar to the vector valued case. From strict complementarity we get that $\overline{\mu}_i + c(\overline{\varphi}_i + 1) \neq 0$ and $\overline{\mu}_i + c(\overline{\varphi}_i - 1) \neq 0$ for all i. Thus, the active sets do not change in a neighborhood of $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$. By the same arguments as in the vector valued case it remains to show that M is invertible in $(\overline{\boldsymbol{\varphi}}, \overline{\boldsymbol{\mu}})$. Therefor, assume

$$D'(\overline{\boldsymbol{\varphi}})\boldsymbol{\delta\varphi} - \delta\lambda\boldsymbol{e} + \boldsymbol{\delta\mu} = 0 \tag{257}$$

$$\boldsymbol{m}^T\boldsymbol{\delta\varphi} = 0 \tag{258}$$

$$\boldsymbol{\delta\mu} - N_1(\boldsymbol{\delta\mu} + c\boldsymbol{\delta\varphi}) - N_2(\boldsymbol{\delta\mu} + c\boldsymbol{\delta\varphi}) = 0 \tag{259}$$

Multiply (257) by $m_i\delta\varphi_i$ and sum up over i to get $F''(\overline{\boldsymbol{\varphi}})[\boldsymbol{\delta\varphi}, \boldsymbol{\delta\varphi}] = 0$. As in the vector valued case one shows that $\boldsymbol{\delta\varphi} \in \mathcal{C}(\overline{\boldsymbol{\varphi}})$. In particular one can show by strict complementarity that $\delta\varphi_i = 0$ holds if $\overline{\varphi}_i \in \{\pm 1\}$. From the second order condition one gets $\boldsymbol{\delta\varphi} = 0$ and as in Theorem 6.74 also $\delta\lambda = 0$ and $\boldsymbol{\delta\mu} = 0$ follows. □

For two phases we now have two different PDAS methods, namely the vector valued PDAS Algorithm 6.2 for $N = 2$ and the scalar valued PDAS Algorithm 6.3, where one of the phase field variables is eliminated. It often can be disadvantageous to eliminate variables and apply a numerical method on the reduced system. We show that this is not the case for the PDAS method. More precisely we show that the PDAS method on the vector valued problem for $N = 2$ is equivalent to the PDAS method on the scalar valued problem provided that c is large enough. Note that although the equations on which the semismooth Newton methods are applied are equivalent, this does not imply that the Newton iterations are equivalent. We first show some properties of the iterates of the vector valued PDAS method.

Lemma 6.78. *Consider the vector valued PDAS method (Algorithm 6.2) for $N = 2$. Let for the initial guess hold*

$$\{1, \dots, J\} = \mathcal{I}_0^1 \cup \mathcal{I}_0^2$$

Then we get for all $k \geq 1$:

1. *It holds*

$$\{1, \dots, J\} = \mathcal{I}_k^1 \cup \mathcal{I}_k^2$$

2. *For $i \in \mathcal{A}_k^1$ it holds*

$$\begin{aligned} &\text{either } (\mu_k)_i^1 = 0, \quad (\mu_k)_i^2 = 0, \quad (\varphi_k)_i^1 < 0, \quad (\varphi_k)_i^2 > 1 \\ &\text{or } (\mu_k)_i^1 > 0, \quad (\mu_k)_i^2 = 0, \quad (\varphi_k)_i^1 = 0, \quad (\varphi_k)_i^2 = 1 \end{aligned}$$

3. *For $i \in \mathcal{A}_k^2$ it holds*

$$\begin{aligned} &\text{either } (\mu_k)_i^1 = 0, \quad (\mu_k)_i^2 = 0, \quad (\varphi_k)_i^1 > 1, \quad (\varphi_k)_i^2 < 0 \\ &\text{or } (\mu_k)_i^1 = 0, \quad (\mu_k)_i^2 > 0, \quad (\varphi_k)_i^1 = 1, \quad (\varphi_k)_i^2 = 0 \end{aligned}$$

4. *It holds*

$$(\mu_k)_i^1 (\mu_k)_i^2 = 0 \quad \forall i$$

5. *It holds*

$$(\mu_k)_i^1 (\varphi_k)_i^1 = (\mu_k)_i^2 (\varphi_k)_i^2 = 0 \quad \forall i$$

Here we assume that there exists a unique solution for each Newton step. Sufficient conditions for solvability in case of the projection type subproblem are given in Lemma 6.68.

Proof. 1) The proof is by induction. For $k = 0$ the statement holds by assumption. Let $k \geq 1$ and let the statement hold for $k - 1$. Let $i \in \mathcal{A}_k^1$, i.e.

$$(\mu_k)_i^1 - c(\varphi_k)_i^1 > 0. \tag{260}$$

We show $i \in \mathcal{I}_k^2$. It holds $\{0, \dots, J\} = \mathcal{A}_{k-1}^1 \cup \mathcal{A}_{k-1}^2 \cup (\mathcal{I}_{k-1}^1 \cap \mathcal{I}_{k-1}^2)$.
In the first case, $i \in \mathcal{A}_{k-1}^1$, we get from the Newton system $(\varphi_k)_i^1 = 0$ and $(\varphi_k)_i^2 = 1$. From induction hypothesis we get $i \in \mathcal{I}_{k-1}^2$, thus from the Newton system we conclude $(\mu_k)_i^2 = 0$

and from (260) we get $(\mu_k)_i^1 > 0$. Thus it holds $(\mu_k)_i^2 - c(\varphi_k)_i^2 = -c < 0$ and hence $i \in \mathcal{I}_k^2$.
In the second case, $i \in \mathcal{A}_{k-1}^2$ we get from the Newton system $(\varphi_k)_i^2 = 0$ and $(\varphi_k)_i^1 = 1$. From induction hypothesis we get $i \in \mathcal{I}_{k-1}^1$, thus from the Newton system we conclude $(\mu_k)_i^1 = 0$. This is a contradiction to (260).
In the third case, $i \in \mathcal{I}_{k-1}^1 \cap \mathcal{I}_{k-1}^2$, we get from the Newton system $(\mu_k)_i^1 = (\mu_k)_i^2 = 0$. From (260) we conclude $(\varphi_k)_i^1 < 0$ and thus $(\varphi_k)_i^2 = 1 - (\varphi_k)_i^1 > 1$. Thus it holds $(\mu_k)_i^2 - c(\varphi_k)_i^2 < -c < 0$ and hence $i \in \mathcal{I}_k^2$.
2) This follows from the proof of 1). The first case of the statement corresponds to $i \in \mathcal{I}_{k-1}^1 \cap \mathcal{I}_{k-1}^2$ and the second case to $i \in \mathcal{A}_{k-1}^1$.
3) The proof is analog to 2). Let $i \in \mathcal{A}_k^2$, i.e.

$$(\mu_k)_i^2 - c(\varphi_k)_i^2 > 0. \tag{261}$$

In the first case, $i \in \mathcal{A}_{k-1}^1$, we get from the Newton system $(\varphi_k)_i^1 = 0$ and $(\varphi_k)_i^2 = 1$. Hence, $i \in \mathcal{I}_{k-1}^2$ and $(\mu_k)_i^2 = 0$. This is a contradiction to (261).
In the second case, $i \in \mathcal{A}_{k-1}^2$ we get from the Newton system $(\varphi_k)_i^2 = 0$ and $(\varphi_k)_i^1 = 1$. From (261) we get $(\mu_k)_i^2 > 0$. From 1) we get $i \in \mathcal{I}_{k-1}^1$ and thus $(\mu_k)_i^1 = 0$.
In the third case, $i \in \mathcal{I}_{k-1}^1 \cap \mathcal{I}_{k-1}^2$, we get from the Newton system $(\mu_k)_i^1 = (\mu_k)_i^2 = 0$. From (261) we conclude $(\varphi_k)_i^2 < 0$ and thus $(\varphi_k)_i^1 = 1 - (\varphi_k)_i^2 > 1$.
4) From 1) it follows that for every i it holds either $i \in \mathcal{I}_{k-1}^1$ and thus $(\mu_k)_i^1 = 0$, or $i \in \mathcal{I}_{k-1}^2$ and thus $(\mu_k)_i^2 = 0$.
5) Let $(\mu_k)_i^1 \neq 0$. Then $i \in \mathcal{A}_{k-1}^1$ and thus $(\varphi_k)_i^1 = 0$. The same holds for $(\mu_k)_i^2$ and $(\varphi_k)_i^2$. □

Note that the assumption $\{1, \dots, J\} = \mathcal{I}_0^1 \cup \mathcal{I}_0^2$ of Lemma 6.78 is necessary for the solvability of the Newton system.
In Lemma 6.78 we derived possible ranges for $\boldsymbol{\mu}_k$ and $\boldsymbol{\varphi}_k$ if $i \in \mathcal{A}_k^1$ or $i \in \mathcal{A}_k^2$. We note that in the remaining case, $i \in \mathcal{I}_k^1 \cap \mathcal{I}_k^2$, no conclusions about $\boldsymbol{\mu}$ and $\boldsymbol{\varphi}$ are possible.

We now show the equivalence of the PDAS methods. As before we denote variables that are connected to the scalar valued PDAS with a tilde. Also recall the transformation $T(\tilde{\varphi}) = (\frac{1+\tilde{\varphi}}{2}, \frac{1-\tilde{\varphi}}{2})^T$. In Theorem 6.38 we showed that the KKT systems of the continuous vector valued and scalar problems are equivalent and derived the relation

$$\begin{aligned}
\boldsymbol{\varphi} &= T(\tilde{\varphi}) \\
\lambda_1 &= 2\tilde{\lambda} \\
\langle \boldsymbol{\mu}, \boldsymbol{\eta} \rangle_{(L^\infty)^*, L^\infty} &= 2(\langle \tilde{\mu}_1, \eta_1 \rangle_{(L^\infty)^*, L^\infty} + \langle \tilde{\mu}_2, \eta_2 \rangle_{(L^\infty)^*, L^\infty}) \quad \forall \boldsymbol{\eta} \in L^\infty(\Omega)^2
\end{aligned}$$

between the solutions of the respective continuous KKT system. By the same consideration we get that for the solutions of the discrete KKT systems it holds

$$\begin{aligned}
\varphi_i^1 &= \frac{1 + \tilde{\varphi}_i}{2} \\
\varphi_i^2 &= \frac{1 - \tilde{\varphi}_i}{2} \\
\lambda_1 &= 2\tilde{\lambda} \\
\mu_i^1 &= -2\min(\tilde{\mu}_i, 0) \\
\mu_i^2 &= 2\max(\tilde{\mu}_i, 0)
\end{aligned}$$

and it holds $\tilde{\mu}_i = \frac{1}{2}(\mu_i^2 - \mu_i^1)$ for all i and j. These relations also hold partially for the

iterates of the Newton method as we show now. In the following, $\frac{1+\tilde{\varphi}}{2}$ for $\tilde{\varphi} \in \mathbb{R}^J$ has to be understood componentwise. This defines the transformation $T : \mathbb{R}^J \to \mathbb{R}^{J\times 2}$.

Theorem 6.79. *The vector valued PDAS Algorithm 6.2 for $N = 2$ and the scalar valued PDAS Algorithm 6.3 are equivalent in the following sense.*
Let the initial guess $(\tilde{\varphi}_0, \tilde{\lambda}_0, \tilde{\mu}_0)$ for the scalar PDAS method be given. Take

$$\left(\left(\frac{1+\tilde{\varphi}_0}{2}, \frac{1-\tilde{\varphi}_0}{2}\right)^T, 2\tilde{\lambda}_0, 0, \left(-\frac{c}{2\tilde{c}}\tilde{\mu}_0, \frac{c}{2\tilde{c}}\tilde{\mu}_0\right)^T\right)$$

as initial guess for the vector PDAS method. Denote by $(\tilde{\varphi}_k, \tilde{\lambda}_k, \tilde{\mu}_k)$ the iterates of the scalar PDAS method and by $(\varphi_k, \lambda_k, \Lambda_k, \mu_k)$ the iterates of the vector PDAS method and assume that each Newton step is uniquely solvable. Let the constant $\tilde{c}$ of the scalar PDAS method be chosen so large that $\tilde{c} \geq -\frac{1}{4}\mu_k$ for all $k \geq 1$.
Then it holds for all $k \geq 0$

$$\mathcal{A}_k^- = \mathcal{A}_k^1 \tag{262}$$
$$\mathcal{A}_k^+ = \mathcal{A}_k^2 \tag{263}$$
$$(\varphi_k)^1 = \frac{1+\tilde{\varphi}_k}{2} \tag{264}$$
$$(\varphi_k)^2 = \frac{1-\tilde{\varphi}_k}{2} \tag{265}$$
$$\tilde{\lambda}_k = \frac{1}{2}\lambda_k, \tag{266}$$

and for all $k \geq 1$

$$\tilde{\mu}_k = \frac{1}{2}((\mu_k)^2 - (\mu_k)^1). \tag{267}$$

Proof. We show the statement by induction. From the choice of the initial guess we get

$$\begin{aligned}\mathcal{A}_0^1 &= \left\{(\mu_0)_i^1 - c(\varphi_0)_i^1 > 0\right\} = \left\{-\frac{c}{2\tilde{c}}(\tilde{\mu}_0)_i - c\frac{1+(\tilde{\varphi}_0)_i}{2} > 0\right\} = \{(\tilde{\mu}_0)_i + \tilde{c}((\tilde{\varphi}_0)_i + 1) < 0\} \\ &= \mathcal{A}_0^- \\ \mathcal{A}_0^2 &= \left\{(\mu_0)_i^2 - c(\varphi_0)_i^2 > 0\right\} = \left\{\frac{c}{2\tilde{c}}(\tilde{\mu}_0)_i - c\frac{1-(\tilde{\varphi}_0)_i}{2} > 0\right\} = \{(\tilde{\mu}_0)_i + \tilde{c}((\tilde{\varphi}_0)_i - 1) > 0\} \\ &= \mathcal{A}_0^+\end{aligned}$$

and also (264)-(266) holds for $k = 0$, hence the base case is shown. For the inductive step we assume that (262)-(266) holds for $(k-1)$ and we show (262)-(267) for k. We start with (264)-(267) by showing that $(2(\delta\varphi_{k-1})^1, \frac{1}{2}\lambda_k, \frac{1}{2}((\mu_k)^2 - (\mu_k)^1))$ is the solution of the scalar Newton system (245)-(247).
First of all note that we have two different transformations T, one acting on functions and the other acting on coordinate vectors. However, it holds

$$\frac{1+\sum_j \tilde{\varphi}_j\chi_j}{2} = \frac{\sum_j \chi_j + \sum_j \tilde{\varphi}_j\chi_j}{2} = \sum_j \left(\frac{1+\tilde{\varphi}}{2}\right)_j \chi_j$$

and thus $T(\sum_j \tilde{\varphi}_j\chi_j) = \sum_j T(\tilde{\varphi})_j\chi_j$. From $\tilde{f} = f \circ T$ we get

$$\begin{aligned}\tilde{D}_i(\tilde{\boldsymbol{\varphi}}_{k-1}) &= \frac{\left\langle \tilde{f}'(\sum_j(\tilde{\varphi}_{k-1})_j\chi_j), \chi_i\right\rangle}{m_i} = \frac{\left\langle f'(T(\sum_j(\tilde{\varphi}_{k-1})_j\chi_j)), T'(\sum_j(\tilde{\varphi}_{k-1})_j\chi_j)\chi_i\right\rangle}{m_i} \\ &= \frac{1}{2}\left(\frac{\left\langle f'(\sum_{m,l}((\varphi_{k-1})^l_m\chi_m\boldsymbol{e}_l), \chi_i\boldsymbol{e}_1\right\rangle}{m_i} - \frac{\left\langle f'(\sum_{m,l}((\varphi_{k-1})^l_m\chi_m\boldsymbol{e}_l), \chi_i\boldsymbol{e}_2\right\rangle}{m_i}\right) \\ &= \frac{1}{2}(D^1_i(\boldsymbol{\varphi}_{k-1}) - D^2_i(\boldsymbol{\varphi}_{k-1})),\end{aligned}$$

where we used (264)-(265). We get by the chain rule

$$\begin{aligned}\tilde{D}'_i(\tilde{\boldsymbol{\varphi}}_{k-1})\widetilde{\boldsymbol{\delta\varphi}} &= \frac{1}{2}((D')^1_i(\boldsymbol{\varphi}_{k-1}) - (D')^2_i(\boldsymbol{\varphi}_{k-1}))T'(\tilde{\boldsymbol{\varphi}}_{k-1})(\widetilde{\boldsymbol{\delta\varphi}}) \\ &= \frac{1}{2}((D')^1_i(\boldsymbol{\varphi}_{k-1}) - (D')^2_i(\boldsymbol{\varphi}_{k-1}))\frac{1}{2}(\widetilde{\boldsymbol{\delta\varphi}}, -\widetilde{\boldsymbol{\delta\varphi}})^T.\end{aligned}$$

From the vector SSN system, we get by the sum constraint (204)

$$(\boldsymbol{\delta\varphi}_{k-1})^1 + (\boldsymbol{\delta\varphi}_{k-1})^2 = -((\boldsymbol{\varphi}_{k-1})^1 + (\boldsymbol{\varphi}_{k-1})^2) + 1 = \mathbf{0}.$$

Thus it holds

$$\boldsymbol{\delta\varphi}_{k-1} = ((\boldsymbol{\delta\varphi}_{k-1})^1, -(\boldsymbol{\delta\varphi}_{k-1})^1)^T = T'(\tilde{\boldsymbol{\varphi}}_{k-1})(2(\boldsymbol{\delta\varphi}_{k-1})^1).$$

We subtract the two gradient equations (203) in the vector SSN system to obtain

$$((D')^1(\boldsymbol{\varphi}_{k-1}) - (D')^2(\boldsymbol{\varphi}_{k-1}))\boldsymbol{\delta\varphi}_{k-1} - \lambda_k\boldsymbol{e} + ((\boldsymbol{\mu}_k)^2 - (\boldsymbol{\mu}_k)^1) = -(D^1(\boldsymbol{\varphi}_{k-1}) - D^2(\boldsymbol{\varphi}_{k-1})).$$

We use the transformations for D from above to get

$$2\tilde{D}'(\tilde{\boldsymbol{\varphi}}_{k-1})(2(\boldsymbol{\delta\varphi}_{k-1})^1) - 2(\frac{1}{2}\lambda_k\boldsymbol{e}) + 2(\frac{1}{2}((\boldsymbol{\mu}_k)^2 - (\boldsymbol{\mu}_k)^1)) = -2\tilde{D}(\tilde{\boldsymbol{\varphi}}_{k-1}).$$

Hence, $(2(\boldsymbol{\delta\varphi}_{k-1})^1, \frac{1}{2}\lambda_k, \frac{1}{2}((\boldsymbol{\mu}_k)^2 - (\boldsymbol{\mu}_k)^1))$ is a solution of the gradient equation (245) in the scalar SSN system. To see that this is also a solution to the mass equation (246) in the scalar SSN system we calculate

$$\begin{aligned}\boldsymbol{m}^T(2(\boldsymbol{\delta\varphi}_{k-1})^1) &= -2(\boldsymbol{m}^T(\boldsymbol{\varphi}_{k-1})^1) + 2\mathfrak{m}_1|\Omega| \\ &= -2\left(\frac{1}{2}|\Omega| + \frac{1}{2}\boldsymbol{m}^T\tilde{\boldsymbol{\varphi}}_{k-1}\right) + 2\mathfrak{m}_1|\Omega| = -\boldsymbol{m}^T\tilde{\boldsymbol{\varphi}}_{k-1} + \tilde{\mathfrak{m}}|\Omega|\end{aligned}$$

where we used that $\boldsymbol{m}^T\boldsymbol{e} = |\Omega|$, that $\boldsymbol{\delta\varphi}_{k-1}$ solves the mass equation in the vector SSN system, the induction hypothesis for (264) and $2\mathfrak{m}_1 - 1 = \tilde{\mathfrak{m}}$. It remains to show that $(2(\boldsymbol{\delta\varphi}_{k-1})^1, \frac{1}{2}\lambda_k, \frac{1}{2}((\boldsymbol{\mu}_k)^2 - (\boldsymbol{\mu}_k)^1))$ solves the projection equation in the scalar PDAS system. Therefor we use the equality of the active sets (262)-(263) for $(k-1)$. We have to show (247), which can be written as

$$\begin{aligned}-2(\delta\varphi_{k-1})^1_i &= (\tilde{\varphi}_{k-1})_i + 1 && \forall i \in \mathcal{A}^-_{k-1} \\ -2(\delta\varphi_{k-1})^1_i &= (\tilde{\varphi}_{k-1})_i - 1 && \forall i \in \mathcal{A}^+_{k-1} \\ \frac{1}{2}((\mu_k)^2_i - (\mu_k)^1_i)) &= 0 && \forall i \in \mathcal{I}_{k-1}.\end{aligned}$$

In the first case, $i \in \mathcal{A}_{k-1}^- = \mathcal{A}_{k-1}^1$, we get from the projection equation in the vector SSN system that

$$(\delta\varphi_{k-1})_i^1 = -(\varphi_{k-1})_i^1 = -\frac{1}{2}(1 + (\tilde{\varphi}_{k-1})_i).$$

In the second case, $i \in \mathcal{A}_{k-1}^+ = \mathcal{A}_{k-1}^2$, we similarly get

$$-(\delta\varphi_{k-1})_i^1 = (\delta\varphi_{k-1})_i^2 = -(\varphi_{k-1})_i^2 = -\frac{1}{2}(1 - (\tilde{\varphi}_{k-1})_i).$$

In the third case, $i \in \mathcal{I}_{k-1} = (\mathcal{A}_{k-1}^- \cup \mathcal{A}_{k-1}^+)^c = \mathcal{I}_{k-1}^1 \cap \mathcal{I}_{k-1}^2$, we get

$$(\mu_k)_i^1 = (\mu_k)_i^2 = 0.$$

Thus, we showed that $(2(\boldsymbol{\delta\varphi}_{k-1})^1, \frac{1}{2}\lambda_k, \frac{1}{2}((\boldsymbol{\mu}_k)^2 - (\boldsymbol{\mu}_k)^1))$ is a solution of the scalar SSN system. From the unique solvability we get

$$(\widetilde{\boldsymbol{\delta\varphi}}_{k-1}, \tilde{\lambda}_k, \tilde{\boldsymbol{\mu}}_k) = \left(2(\boldsymbol{\delta\varphi}_{k-1})^1, \frac{1}{2}\lambda_k, \frac{1}{2}((\boldsymbol{\mu}_k)^2 - (\boldsymbol{\mu}_k)^1)\right).$$

This leads to

$$(\boldsymbol{\varphi}_k)^1 = (\boldsymbol{\varphi}_{k-1})^1 + (\boldsymbol{\delta\varphi}_{k-1})^1 = \frac{1 + \tilde{\boldsymbol{\varphi}}_{k-1}}{2} + \frac{1}{2}\widetilde{\boldsymbol{\delta\varphi}}_{k-1} = \frac{1 + \tilde{\boldsymbol{\varphi}}_k}{2}$$
$$(\boldsymbol{\varphi}_k)^2 = (\boldsymbol{\varphi}_{k-1})^2 + (\boldsymbol{\delta\varphi}_{k-1})^2 = \frac{1 - \tilde{\boldsymbol{\varphi}}_{k-1}}{2} - \frac{1}{2}\widetilde{\boldsymbol{\delta\varphi}}_{k-1} = \frac{1 - \tilde{\boldsymbol{\varphi}}_k}{2}.$$

Up to now we showed (264)-(267), thus it remains to show the equality of the active sets (262)-(263). We want to use the results of Lemma 6.78, hence it has to hold the assumption $\{1, \dots, J\} = \mathcal{I}_0^1 \cup \mathcal{I}_0^2$, which follows from

$$\mathcal{I}_0^1 \cup \mathcal{I}_0^2 = (\mathcal{A}_0^1 \cap \mathcal{A}_0^2)^c = (\mathcal{A}_0^- \cap \mathcal{A}_0^+)^c = \varnothing^c = \{1, \dots, J\}. \tag{268}$$

In the following we show the inclusions $\mathcal{A}_k^1 \subset\supset \mathcal{A}_k^-$ and $\mathcal{A}_k^2 \subset\supset \mathcal{A}_k^+$.
'$\mathcal{A}_k^1 \subset \mathcal{A}_k^-$': Let $i \in \mathcal{A}_k^1$. From Lemma 6.78, 2) we get

$$(\tilde{\mu}_k)_i + \tilde{c}((\tilde{\varphi}_k)_i + 1) = \frac{1}{2}((\mu_k)_i^2 - (\mu_k)_i^1) + 2\tilde{c}(\varphi_k)_i^1 < 0,$$

thus $i \in \mathcal{A}_k^-$.
'$\mathcal{A}_k^- \subset \mathcal{A}_k^1$': Let $i \in \mathcal{A}_k^-$, i.e.

$$\frac{1}{2}((\mu_k)_i^2 - (\mu_k)_i^1) + 2\tilde{c}(\varphi_k)_i^1 < 0. \tag{269}$$

Assume $(\mu_k)_i^2 \neq 0$. From Lemma 6.78, 4) and 5) we get $(\mu_k)_i^1 = (\varphi_k)_i^2 = 0$, thus $(\varphi_k)_i^1 = 1$. From (269) we then get a contradiction to the assumption $\tilde{c} \geq -\frac{1}{4}(\mu_k)_i^2$. Thus $(\mu_k)_i^2 = 0$. Assume now $(\mu_k)_i^1 \neq 0$ From Lemma 6.78, 5), $(\varphi_k)_i^1 = 0$ and from (269) we get $(\mu_k)_i^1 > 0$. Hence, $(\mu_k)_i^1 - c(\varphi_k)_i^1 > 0$ and $i \in \mathcal{A}_k^1$. In case $(\mu_k)_i^1 = 0$ we get from (269) that $(\varphi_k)_i^1 < 0$ and again it holds $(\mu_k)_i^1 - c(\varphi_k)_i^1 > 0$ and $i \in \mathcal{A}_k^1$.
'$\mathcal{A}_k^2 \subset \mathcal{A}_k^+$': Let $i \in \mathcal{A}_k^2$. From Lemma 6.78, 3) we get

$$(\tilde{\mu}_k)_i + \tilde{c}((\tilde{\varphi}_k)_i - 1) = \frac{1}{2}((\mu_k)_i^2 - (\mu_k)_i^1) - 2\tilde{c}(\varphi_k)_i^2 > 0,$$

thus $i \in \mathcal{A}_k^+$.
'$\mathcal{A}_k^+ \subset \mathcal{A}_k^2$': Let $i \in \mathcal{A}_k^+$, i.e.

$$\frac{1}{2}((\mu_k)_i^2 - (\mu_k)_i^1) - 2\tilde{c}(\varphi_k)_i^2 > 0. \tag{270}$$

Assume $(\mu_k)_i^1 \neq 0$. From Lemma 6.78, 4) and 5) we get $(\mu_k)_i^2 = (\varphi_k)_i^1 = 0$, thus $(\varphi_k)_i^2 = 1$. From (270) we then get a contradiction to the assumption $\tilde{c} \geq -\frac{1}{4}(\mu_k)_i^1$. Thus $(\mu_k)_i^1 = 0$. Assume now $(\mu_k)_i^2 \neq 0$ From Lemma 6.78, 5), $(\varphi_k)_i^2 = 0$ and from (270) we get $(\mu_k)_i^2 > 0$. Hence, $(\mu_k)_i^2 - c(\varphi_k)_i^2 > 0$ and $i \in \mathcal{A}_k^2$. In case $(\mu_k)_i^2 = 0$ we get from (270) that $(\varphi_k)_i^2 < 0$ and again it holds $(\mu_k)_i^2 - c(\varphi_k)_i^2 > 0$ and $i \in \mathcal{A}_k^2$. □

We remark that in Theorem 6.79 it does not matter which initial guess for $\boldsymbol{\mu}$, λ and $\boldsymbol{\Lambda}$ is used as long as (262)-(263) holds for $k = 0$.

We also note that numerical experiments show that the scalar PDAS and vector PDAS method can generate nonequivalent iterates if the assumption $\tilde{c} \geq -\frac{1}{4}\boldsymbol{\mu}_k$ is not fulfilled.

6.11 Discretization and adaptive mesh

In this section we discuss the discretization of the optimization problem as well as of the VMPT method. For simplicity we assume that the cost functional F has the form

$$F(\boldsymbol{\varphi}, \boldsymbol{u}) = \int_\Omega \alpha(x, \boldsymbol{\varphi}(x), \boldsymbol{u}(x))\,\mathrm{d}x + \int_{\partial\Omega} \beta(x, \boldsymbol{\varphi}(x), \boldsymbol{u}(x))\,\mathrm{d}x,$$

which holds for all numerical examples. We discretize the control and the state equation by standard piecewise linear finite elements. For convenience of the reader we recall this type of discretization briefly. For a detailed introduction we refer to [Bra01, Cia02].
Let $\mathcal{T}_h$ be a decomposition of Ω into triangles in 2D and into tetrahedra in 3D, respectively. For simplicity we assume that Ω has a polygonal boundary, such that no further discretization errors appear due to the triangulation. Let $S_h \subset H^1(\Omega)$ be the P1 finite element space, i.e.

$$S_h = \left\{\varphi \in C(\overline{\Omega}) \mid \varphi|_T \in P_1(T)\ \forall T \in \mathcal{T}_h\right\},$$

where $P_1(T)$ denotes the space of all affine linear functions on T. For the discretization of the control and state variables we define the vector valued finite element spaces $S_h^n := (S_h)^n$ for $n \in \mathbb{N}$. The Dirichlet boundary condition in the state equation is incorporated into the finite element space

$$S_{h,D}^d = \left\{\boldsymbol{u} \in S_h^d \mid \boldsymbol{u}(p) = \mathbf{0} \text{ for each mesh node } p \in \Gamma_D\right\}.$$

To avoid discretization errors due to the discretization of Γ_D, we choose the meshes in the numerical examples such that Γ_D is a union of triangle edges (in 2D) or a union of tetrahedron faces (in 3D). Let $\{p_i\}_{i=1}^J$ denote the set of nodes of the triangulation. The standard nodal basis functions of S_h are defined by

$$\chi_i(p_j) = \delta_{ij},$$

with $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. This induces the basis $\chi_i \boldsymbol{e}_j$, $i = 1, \ldots, J$, $j = 1, \ldots, d$ of S_h^d, where $\boldsymbol{e}_j$ denotes the jth unit vector in $\mathbb{R}^d$. Any function $\boldsymbol{u}_h \in S_h^d$ can be written in the basis representation $\boldsymbol{u}_h = \sum_{ij} u_i^j \chi_i \boldsymbol{e}_j$ with coordinates u_i^j.

The discretization of the state equation now reads as follows. Find $\boldsymbol{u} \in S^d_{h,D}$, such that

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) = \int_\Omega \boldsymbol{f}(x,\boldsymbol{\varphi}) \cdot \boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g}(x,\boldsymbol{\varphi}) \cdot \boldsymbol{\xi} \quad \forall \boldsymbol{\xi} \in S^d_{h,D}.$$

Here we again assume that Γ_g is a union of edges or faces. This linear equation can be written in coordinates as

$$\sum_{kl} a_{ijkl} u_l^k = b_{ij} \quad \forall i = 1,\dots,d, j = 1,\dots,J \text{ s.t. } p_j \notin \Gamma_D \tag{271}$$

$$u_l^k = 0 \quad \forall k = 1,\dots,d, l = 1,\dots,J \text{ s.t. } p_l \in \Gamma_D \tag{272}$$

with

$$a_{ijkl} \coloneqq \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\chi_l \boldsymbol{e}_k) : \mathcal{E}(\chi_j \boldsymbol{e}_i) \tag{273}$$

$$b_{ij} \coloneqq \int_\Omega \boldsymbol{f}(x,\boldsymbol{\varphi}) \cdot \boldsymbol{e}_i \chi_j + \int_{\Gamma_g} \boldsymbol{g}(x,\boldsymbol{\varphi}) \cdot \boldsymbol{e}_i \chi_j. \tag{274}$$

The solution operator $\boldsymbol{\varphi} \mapsto \boldsymbol{u}$ of the discretized state equation (271)-(272) is denoted by S^h. If it holds $\boldsymbol{\varphi} \in S_h^N$ and if $\boldsymbol{C}$, $\boldsymbol{f}$ and $\boldsymbol{g}$ are polynomial in $\boldsymbol{\varphi}$ and x, then also $\boldsymbol{C}(\boldsymbol{\varphi}(x))$, $\boldsymbol{f}(x,\boldsymbol{\varphi}(x))$ and $\boldsymbol{g}(x,\boldsymbol{\varphi}(x))$ are piecewise polynomial functions. In this case the integrals in (273) and (274) can be computed by exact quadrature rules, which is done in the numerical experiments. As solver for the discrete state equation (271)-(272) we use the direct solver UMFPACK [Dav07] in case of d=2 and if the number of unknowns is not too large. Otherwise we use the CG method. Note that the state equation is ill conditioned if there is a huge difference between the stiffness tensors of the distinct phases, i.e. if a hard material and a very weak material is present. In particular this holds for the ersatz material approach, see Section 5.

The linearized state equation (124), the adjoint equation (133) and the linearized adjoint equation (155) are discretized by the same way. We note that it is very common in the literature to use Q1-elements for solving the elasticity equation, i.e. quadrilaterals are used instead of triangles and the finite element functions are bilinear on each quadrilateral. However, we use P1-elements here.

The control variable is also discretized by linear finite elements. Note that $S_h^N \subset H^1(\Omega)^N \cap L^\infty(\Omega)^N$. The constraints can be adopted without further approximation. Thus the reduced discrete optimization problem reads

$$\begin{gathered} \min j_h(\boldsymbol{\varphi}) = \gamma E(\boldsymbol{\varphi}) + F_h(\boldsymbol{\varphi}, S^h(\boldsymbol{\varphi})) \\ \boldsymbol{\varphi} \in S_h^N \\ \boldsymbol{\varphi} \geq 0, \quad \sum_{i=1}^N \varphi_i = 1, \quad \fint_\Omega \boldsymbol{\varphi} = \mathfrak{m}. \end{gathered} \tag{275}$$

Note that the constraints can be written in coordinates as

$$\begin{aligned} \varphi_i^j &\geq 0 \quad \forall j = 1, \dots, N, i = 1, \dots, J \\ \sum_{j=1}^{N} \varphi_i^j &= 1 \quad \forall i = 1, \dots, J \\ \sum_{i=1}^{J} m_i \varphi_i^j &= \mathfrak{m}_j |\Omega| \quad \forall j = 1, \dots, N, \end{aligned}$$

where $m_i := \int_\Omega \chi_i$ is the mass of the ith basis vector. We assume that the potential ψ_0 is polynomial in $\boldsymbol{\varphi}$ and thus the Ginzburg-Landau energy $E(\boldsymbol{\varphi})$ can be computed exactly for $\boldsymbol{\varphi} \in S_h^N$. In certain cases the energy F has to be approximated by some F_h e.g. the tracking functional $\int_{\Omega_{obs}} |\boldsymbol{u} - \boldsymbol{u}_\Omega|^2$ is approximated by $\int_{\tilde{\Omega}_{obs}} |\boldsymbol{u} - I_h \boldsymbol{u}_\Omega|^2$, where $\tilde{\Omega}_{obs}$ is a union of elements approximating Ω_{obs} and $I_h : C(\overline{\Omega})^d \to S_h^d$ is the interpolation operator which interpolates each component of the function piecewise linearly in the mesh nodes. The function F on the right hand side of the adjoint equation then also has to be exchanged by F_h. Note that the compliance can be calculated exactly, thus we take $F_h = F$ in this case.

The discretization of the VMPT method basically boils down to the discretization of the projection type subproblem. If a discrete initial guess $\boldsymbol{\varphi}_0 \in S_h^N$ is chosen and the discrete solution operator $\mathcal{P}_{k,h}$ of the projection type subproblem maps into S_h^N, then automatically all iterates $\boldsymbol{\varphi}_k$ of the VMPT method are elements of S_h^N. Recall the projection type subproblem

$$\min_{\boldsymbol{y} \in \Phi_{ad}} \quad \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{\varphi}_k\|_{a_k}^2 + \lambda_k \left\langle j'(\boldsymbol{\varphi}_k), \boldsymbol{y} - \boldsymbol{\varphi}_k \right\rangle. \tag{276}$$

We discretize the VMPT method in a consistent way such that the approaches discretize-then-optimize and optimize-then-discretize are the same. Thereby we mean that the discretized VMPT method coincides with the VMPT method applied to the discretized problem (275). This has the advantage that global convergence of the discretized method is given, i.e. on a fixed mesh the iterates of the method converge in the sense of Theorem 4.14. Note that $\Phi_{ad} \cap S_h^N$ is compact and thus there always exists a convergent subsequence for the discrete method. Without a consistent discretization one would have to refine the mesh during the optimization procedure to obtain convergence. Moreover the search direction may not be a descent direction for the discrete cost functional if the mesh is too coarse. The discretized subproblem thus reads

$$\min_{\boldsymbol{y} \in \Phi_{ad} \cap S_h^N} \quad \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{\varphi}_k\|_{a_k}^2 + \lambda_k \left\langle j_h'(\boldsymbol{\varphi}_k), \boldsymbol{y} - \boldsymbol{\varphi}_k \right\rangle. \tag{277}$$

Note that for the inner products a_k used here, the norm $\|\boldsymbol{y} - \boldsymbol{\varphi}_k\|_{a_k}^2$ can be calculated exactly for $\boldsymbol{y}, \boldsymbol{\varphi}_k \in S_h^N$. An exception is the Cahn-Hilliard inner product (170), where the H^{-1} inner product has to be approximated, but we will not discuss this here.
To compute $j_h'(\boldsymbol{\varphi}_k)$, we first of all observe by the proof of Theorem 6.9 that the discretized linearized state equation coincides with the linearization of the discretized state equation. Moreover, since we use the same finite element discretization for the linearized state equation and for the adjoint equation, i.e. we use $S_{h,D}^d$ as test and trial space for both equations, we get by the proof of Lemma 6.29 that the discretized adjoint equation coincides with the adjoint equation of the discretized problem. By Proposition 6.30 we

then get the formula for $j_h'(\boldsymbol{\varphi}_k)$,

$$\begin{aligned}\left\langle j_h'(\boldsymbol{\varphi}_k), \boldsymbol{\delta\varphi}\right\rangle = \gamma\varepsilon \int_\Omega \nabla\boldsymbol{\varphi}_k : \nabla\boldsymbol{\delta\varphi} + \frac{\gamma}{\varepsilon}\int_\Omega \psi_0'(\boldsymbol{\varphi}_k)\boldsymbol{\delta\varphi} \\ + \left\langle (F_h)_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k, \boldsymbol{u}_k), \boldsymbol{\delta\varphi}\right\rangle - \int_\Omega (\nabla \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{u}_k) : \mathcal{E}(\boldsymbol{p}_k)) \cdot \boldsymbol{\delta\varphi} \\ + \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k)^T \boldsymbol{p}_k \cdot \boldsymbol{\delta\varphi} + \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k)^T \boldsymbol{p}_k \cdot \boldsymbol{\delta\varphi} \quad \forall \boldsymbol{\delta\varphi} \in S_h^N,\end{aligned}$$

where $\boldsymbol{u}_k$ and $\boldsymbol{p}_k$ are the solutions of the discrete state and adjoint equation, respectively. Since it holds $\boldsymbol{y} - \boldsymbol{\varphi}_k \in S_h^N$ in (277), we can write

$$\left\langle j_h'(\boldsymbol{\varphi}_k), \boldsymbol{y} - \boldsymbol{\varphi}_k\right\rangle = \sum_{ij} \left\langle j_h'(\boldsymbol{\varphi}_k), \boldsymbol{e}_j\chi_i\right\rangle (y_i^j - (\varphi_k)_i^j).$$

Thus it suffices to compute $\left\langle j_h'(\boldsymbol{\varphi}_k), \boldsymbol{e}_j\chi_i\right\rangle$ for all basis functions, which is given by

$$\begin{aligned}\left\langle j_h'(\boldsymbol{\varphi}_k), \boldsymbol{e}_j\chi_i\right\rangle = \gamma\varepsilon \int_\Omega \nabla(\boldsymbol{\varphi}_k)_j \cdot \nabla\chi_i + \frac{\gamma}{\varepsilon}\int_\Omega \partial_j\psi_0(\boldsymbol{\varphi}_k)\chi_i + \int_\Omega \partial_{\boldsymbol{\varphi}_j}\alpha_h(x, \boldsymbol{\varphi}_k, \boldsymbol{u}_k)\chi_i \\ + \int_{\partial\Omega} \partial_{\boldsymbol{\varphi}_j}\beta_h(x, \boldsymbol{\varphi}_k, \boldsymbol{u}_k)\chi_i + - \int_\Omega (\partial_j \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{u}_k) : \mathcal{E}(\boldsymbol{p}_k))\chi_i \\ + \int_\Omega \partial_j \boldsymbol{f}(\boldsymbol{\varphi}_k)^T \boldsymbol{p}_k\chi_i + \int_{\Gamma_g} \partial_j \boldsymbol{g}(\boldsymbol{\varphi}_k)^T \boldsymbol{p}_k\chi_i\end{aligned}$$

and can be assembled by standard finite element techniques. The discrete projection type subproblem is solved by the PDAS method as described in Section 6.10. The Armijo rule (20) in the VMPT method is discretized by replacing the continuous function j by its discretization j_h.

We emphasize that convergence of the VMPT method is only given for exact arithmetic. In practice rounding errors are always present. Moreover, when an iterative solver is used for the state equation, then there are in addition approximation errors present when evaluating the operator S^h. Usually these errors play a minor role when $\boldsymbol{\varphi}_k$ is far away from a minimizer. However, when $\boldsymbol{\varphi}_k$ is close to a minimizer these errors may dominate. As an example, consider the plot of j_h along the search direction $\alpha \mapsto j_h(\boldsymbol{\varphi}_k + \alpha\boldsymbol{v}_k)$ in Figure 6, where the evaluation of j_h at the depicted points involves approximation errors. Since j_h is smooth, its graph should also be smooth. However, it can be seen that the values oscillate due to errors. In this example $\boldsymbol{\varphi}_k$ is chosen near a minimum and thus $\boldsymbol{v}_k$ is small. Since j is almost constant along the line, the approximation errors can be observed, which are at a magnitude of 10^{-11}. In this case the backtracking algorithm may fail to compute a step length which fulfills the Armijo condition and the VMPT method breaks down. Also the solution of the projection type subproblem involves approximation and rounding errors, which may dominate if $\boldsymbol{\varphi}_k$ is near a local minimum. If the errors are too high it may happen that the computed search direction is not a descent direction for j_h and the VMPT method breaks down. In practice this occurs around a residual of $\sqrt{\gamma\varepsilon}\|\nabla\boldsymbol{v}_k\|_{L^2} \approx 10^{-6}$. If a high number of degrees of freedom is involved, this break down can also occur earlier, see e.g. Figure 29b. Note that these errors are independent of the discretization errors. Refining the mesh does not reduce these errors.

An important task is the choice of an adequate mesh. When discretizing a phase field variable, typically a locally refined mesh is used, which is fine on the interface and coarse in the bulk region [WR12, BBG11, BGSS13a]. This is plausible, since only the interfacial

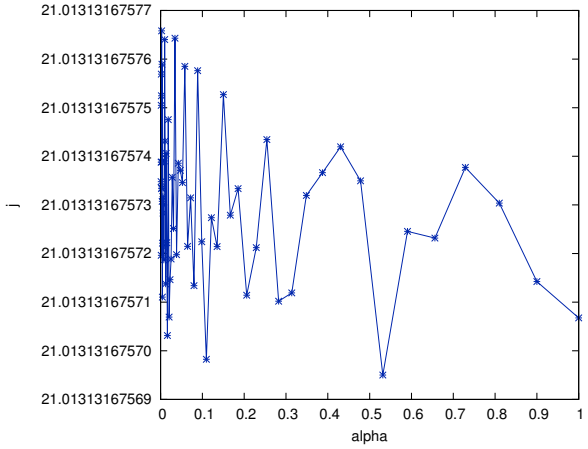


Figure 6: Plot of $\alpha \mapsto j_h(\boldsymbol{\varphi}_k + \alpha \boldsymbol{v}_k)$ reveals approximation errors.

region contributes to the integrals in the Ginzburg-Landau energy. Such a mesh refinement strategy for two phases is given in [BNS04], in which the mesh for the current time step depends on the phase field at the previous time step. For given mesh parameters h_{min} and h_{max} and for given phase field φ of the previous time step a triangle T is refined, if it or one of its neighbors satisfies

$$|\min_{x \in \overline{T}} |\varphi(x)| - 1| > 0.1 \cdot tol,$$

until the diameter h_{min} is reached. On the other hand, if T satisfies

$$|\min_{x \in \overline{T}} |\varphi(x)| - 1| < 0.001 \cdot tol,$$

the triangle is coarsened up to a maximal diameter of h_{max}. By this way an adaptive mesh is created with mesh size h_{max} in the bulk and h_{min} around the interface. We use a similar adaptive strategy here. However, since we don't solve an evolution equation but an optimization problem, our strategy is rather
SOLVE — REFINE — SOLVE — REFINE ...,
i.e. we compute an approximate minimizer of the optimization problem on a coarse equidistant mesh and based on this solution we adaptively refine the mesh. This procedure is repeated until the desired refinement level is reached. Thus, given an approximate vector-valued solution $\boldsymbol{\varphi}$, we start with an equidistant mesh of size h_{max} and recursively refine each triangle T satisfying

$$tol \leq \varphi_i(x) \leq 1 - tol \tag{278}$$

for some $x \in \overline{T}$ and some $i = 1, \ldots, N$, until the minimal diameter h_{min} is reached. In the experiments we set $tol = 10^{-3}$. Note that the implementation also has to pay attention to the case that a very coarse mesh is present such that $|\boldsymbol{\varphi}(x)| = 1$ holds in the vertices of the triangle and condition (278) is only fulfilled in the interior or along an edge. It may also happen that the interfacial region of $\boldsymbol{\varphi}$ leaves the refined region of the mesh during the optimization process. In this case we update the adaptive mesh every 20 optimization steps.
The Ginzburg-Landau energy can be resolved by a mesh which is only fine on the interface. However, also the state equation has to be solved on the mesh and thus also mesh points

in the bulk region are necessary. In principle it is possible to use different meshes for the control and the state variable. In this case the integrals in (273) and (274) cannot be computed exactly without further implementation effort, and one would interpolate the control variable onto the state mesh before solving the state equation. However, then the approaches optimize-then-discretize and discretize-then optimize would not be the same anymore. For this reason we decide to use the same mesh for all functions. Thus the mesh has to be chosen such that not only the Ginzburg-Landau energy, but also the other part F of the cost functional can be resolved. To give an impression how these meshes should look like, we give an example in the case of the mean compliance problem. We use the cantilever beam and bridge experiments described in Example 6.83 and Example 6.84 later on. First we generate an adaptive mesh by the strategy described above with $h_{max} = 1/8$ and $h_{min} = 1/256$. In the second step we further refine the mesh based on the goal-oriented dual weighted residual (DWR) error estimator [BR01], which measures the discretization error of the state in the compliance functional (i.e. the error $|F(u)-F(u_h)|$), and which is shipped with the finite element toolbox FEniCS [LMW+12]. Figure 7 shows the final adaptive meshes together with the optimal phase fields and the deformed shapes. Here, the blue region corresponds to hard material and the red region corresponds to void. We observe in both examples that almost no mesh points are needed in the void, whereas in the hard material the mesh is chosen finer, especially for the cantilever beam. We also observe that there are certain regions where the mesh is particularly fine. In the cantilever experiment these points are both left corners as well as the point on the bottom line next to the right bottom corner. The former are the points where the Neumann boundary touches the Dirichlet boundary and the latter point is the point where the boundary traction starts to act, i.e. there is a discontinuity in the Neumann boundary data. In the bridge experiment also the points between Dirichlet and Neumann boundary are particularly refined and also the traction boundary Γ_g is very fine.
In most experiments we will however use an equidistant mesh in the bulk for simplicity instead of using the DWR error estimator, i.e. we use the described refinement strategy with a rather low value for h_{max}. Also for the phase field it is often necessary to have mesh points in the bulk region such that nucleation of new phases can occur. We note that in the topology optimization literature, especially in the engineering community, the elasticity equation is very often solved on an equidistant mesh.
We finally note that only few mesh points across the interface are necessary in order for the VMPT method to behave the same as on a very fine mesh, see the experiments corresponding to Figure 26. However, discretization errors may be very high if the mesh is chosen too coarse.

6.12 Choice of parameters

In this section we summarize which parameters we use for the VMPT method and for the topology optimization problem in the numerical experiments.

First we give a motivation for the variable metrics (165)-(166), in particular we will justify the $\gamma\varepsilon$-scaling. The analysis of the VMPT method for the topology optimization problem is performed in the space $H^1_{(0)}(\Omega)^N \cap L^\infty(\Omega)^N$, where the space $H^1_{(0)}(\Omega)^N$ is equipped with the H^1 inner product or equivalently with the semi inner product $\int \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2$. Thus our first choice of the inner product in the VMPT method is $a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \int \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2$.
To motivate a better scaling for a_k, we perform the following consideration. Assume that

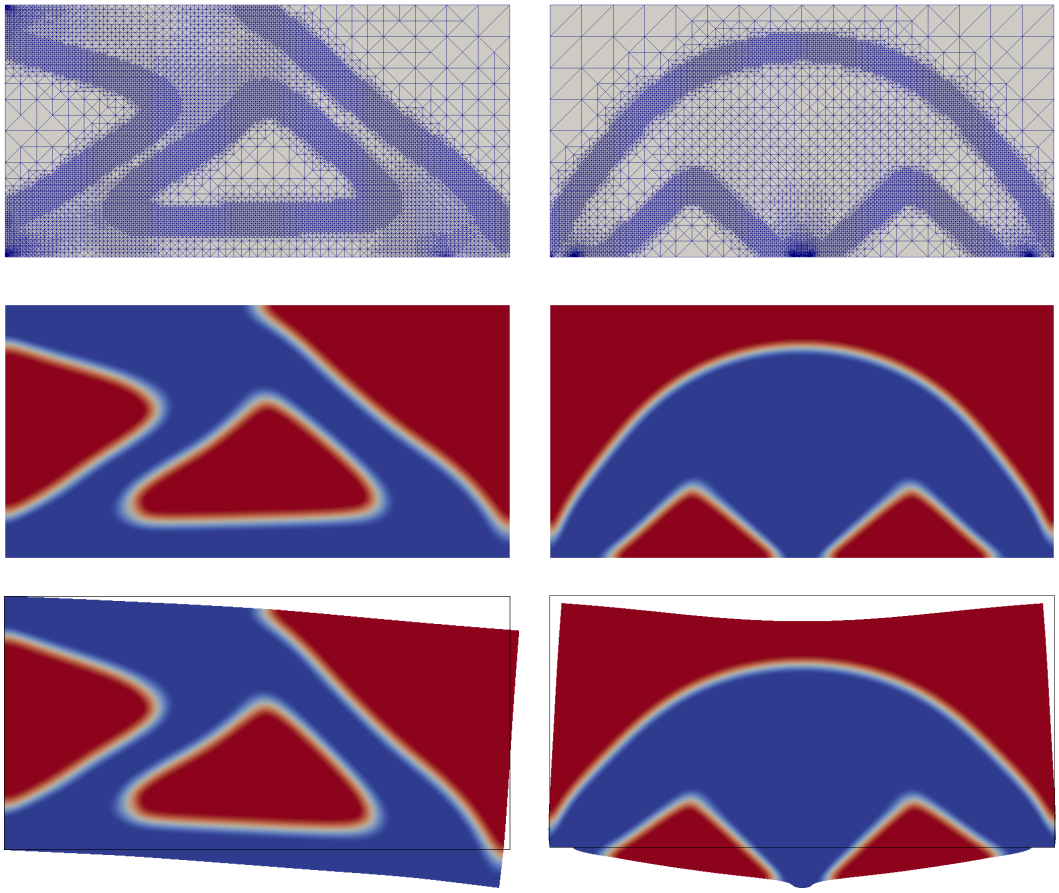


Figure 7: Adaptively refined meshes which are fine along the interface and are in addition locally refined based on the DWR error estimator. The middle row shows the corresponding controls and the deformed shapes are depicted in the bottom row.

φ_ε, $0 < \varepsilon < 1$, are global minimizers of the optimization problem for two phases and varying ε. Let $\varphi_\varepsilon \to \varphi_0$ in $L^1(\Omega)$. From the Γ-convergence result we then get that $j_\varepsilon(\varphi_\varepsilon)$ converges as $\varepsilon \to 0$, cf. Theorem 6.20. This implies the boundedness of the term $\int_\Omega \frac{\varepsilon}{2}|\nabla\varphi_\varepsilon|^2$ in the cost functional. Thus we have

$$\|\nabla\varphi_\varepsilon\|_{L^2} \leq \frac{C}{\sqrt{\varepsilon}}. \tag{279}$$

In fact, numerical experiments reveal that equality holds, $\|\nabla\varphi_\varepsilon\|_{L^2} = \frac{C}{\sqrt{\varepsilon}}$, see Figure 8, where Example 6.83 is considered with $\gamma = 0.5$. Thus, to have a norm which does not depend on ε, one should better make use of the scaled norm $\sqrt{\varepsilon}\|\nabla\varphi_\varepsilon\|_{L^2}$. The preceding consideration holds for $\varepsilon \to 0$, i.e. for ε sufficiently small. For the case when ε is large we perform the following more general calculation. Assume for simplicity that the interface is located in the strip $\Gamma \times (-\frac{d}{2}, \frac{d}{2}) \subset \Omega$ with interface thickness d. Further assume that φ has the typical form

$$\varphi(x,y) = \begin{cases} -1 & y < -\frac{d}{2} \\ \sin(\frac{y\pi}{d}) & -\frac{d}{2} \leq y \leq \frac{d}{2} \\ 1 & y > \frac{d}{2} \end{cases}$$

and thus $\nabla\varphi = (0, \frac{\pi}{d}\cos(\frac{y\pi}{d}))^T$ for $-\frac{d}{2} \leq y \leq \frac{d}{2}$. By transformation rule we calculate

$$\|\nabla\varphi\|_{L^2}^2 = \int_\Gamma \int_{-\frac{d}{2}}^{\frac{d}{2}} |\nabla\varphi(x,y)|^2 = |\Gamma| \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{\pi^2}{d^2}\cos^2(y)\frac{d}{\pi} = |\Gamma|\frac{\pi^2}{2d}.$$

Thus, $\sqrt{d}\|\nabla\varphi_\varepsilon\|_{L^2}$ should be the norm of choice, which is independent of the interface width. All parameters influencing the interface width should enter the constant d. Under certain conditions the interface thickness decreases with the parameter γ (see Section 6.13.2), thus we choose $d = \gamma\varepsilon$. Note that the linear dependency on γ is not mandatory. Based on the numerical results of Section 6.13.2 one could also take $d = \gamma^{0.4}\varepsilon$. Nevertheless, we take $d = \gamma\varepsilon$, since this gives satisfactory numerical results. Thus we consider the rescaled inner product $a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \gamma\varepsilon \int \nabla\boldsymbol{v}_1 : \nabla\boldsymbol{v}_2$. Moreover, when looking at the second order derivative of the Ginzburg-Landau energy

$$E''(\boldsymbol{\varphi})[\boldsymbol{v}_1, \boldsymbol{v}_2] = \gamma\varepsilon \int_\Omega \nabla\boldsymbol{v}_1 : \nabla\boldsymbol{v}_2 + \frac{\gamma}{\varepsilon}\int_\Omega \psi_0''(\boldsymbol{\varphi})[\boldsymbol{v}_1, \boldsymbol{v}_2],$$

we observe that the first term coincides with the used rescaled inner product a_k. Thus, the scaling $d = \gamma\varepsilon$ can also be seen as second order information from the Ginzburg-Landau energy.

Note that the choice $\lambda_k = 1$ and $a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \gamma\varepsilon \int \nabla\boldsymbol{v}_1 : \nabla\boldsymbol{v}_2$ is equivalent to the choice $\lambda_k = (\gamma\varepsilon)^{-1}$ and $a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \int \nabla\boldsymbol{v}_1 : \nabla\boldsymbol{v}_2$. In the numerical experiments we use the latter formulation, thus we consider the two choices $\lambda_k = 1$ and $\lambda_k = (\gamma\varepsilon)^{-1}$. As a third choice for λ_k we look at the following update rule:

$$\begin{aligned} &\text{Choose } \lambda_0 > 0,\ 0 < c < 1 \text{ and } 0 < \lambda_{min} < \lambda_{max}. \\ &\text{If } \alpha_{k-1} = 1 \text{ then set } \tilde{\lambda}_k = \lambda_{k-1}/c, \text{ else set } \tilde{\lambda}_k = c\lambda_{k-1}. \\ &\text{Finally set } \lambda_k = \max\{\lambda_{min}, \min\{\lambda_{max}, \tilde{\lambda}_k\}\}. \end{aligned} \tag{280}$$

Recall that $\alpha_k \leq 1$ is the step size generated by Armijo backtracking. The last adjustment

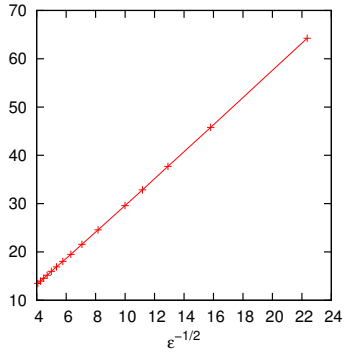


Figure 8: The growth of $\|\nabla\varphi_\varepsilon\|_{L^2}$ is $\mathcal{O}(\varepsilon^{-1/2})$.

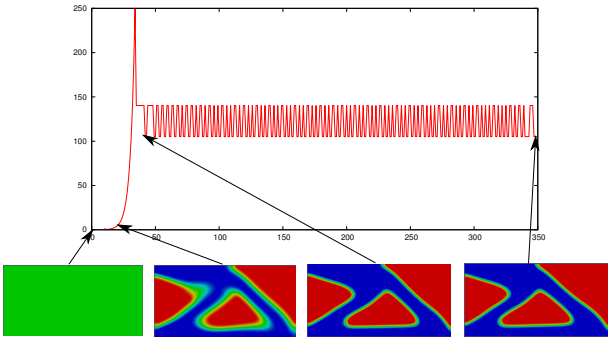
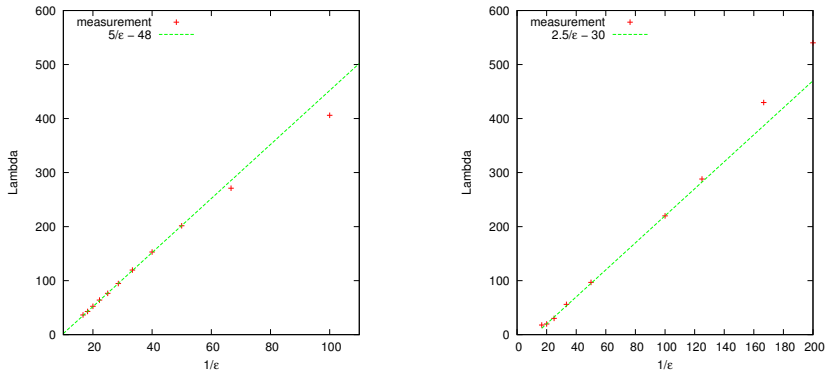


Figure 9: Parameters λ_k generated by the update scheme (280).

ensures that $\lambda_{min} \leq \lambda_k \leq \lambda_{max}$ holds for all k and thus λ_k fulfills the assumptions of the VMPT method. The idea of this update scheme is to increase λ_k as soon as the full step $\alpha_{k-1} = 1$ is taken in the previous step to have λ_k as large as possible, which can improve the efficiency of the method. For the numerical experiments we choose $c = 0.75$, $\lambda_{min} = 10^{-10}$ and $\lambda_{max} = 10^{10}$, where both limits are never reached. The choice of λ_0 depends on the respective experiment and ranges from $\lambda_0 = 10^{-3}$ to $\lambda_0 = 1$. Another motivation for the update scheme is the following. We often start the VMPT iteration with $\boldsymbol{\varphi}_0$ being a homogeneous mixture of the distinct phases. In this stage no interface of width d is yet present. Experience shows that in this case a smaller value for λ_k is advantageous. Thus we start with a smaller value λ_0 and the update scheme ensures that λ_k is increased whenever possible until the final desired scaling $\lambda_k = \gamma\varepsilon$ is reached as soon as the current phase field $\boldsymbol{\varphi}_k$ has an interface of thickness d. A typical sequence of scalings λ_k generated by the update scheme is depicted in Figure 9, where in the bottom row also the corresponding phase fields $\boldsymbol{\varphi}_k$ are shown. Starting with a small value, λ_k is gradually increased. Eventually, λ_k oscillates between two values. Figure 10 shows the mean value of these two final values of λ_k for varying ε and fixed $\gamma = 0.5$ in two different numerical experiments. For comparison, also the reference lines $5/\varepsilon - 48$ and $2.5/\varepsilon - 30$, respectively, are included in the plot. It can be observed that the final scaling generated by the update grows approximately like $\mathcal{O}(\varepsilon^{-1})$ as it was motivated above. This experiment confirms that our considerations about the ε-scaling are correct, as well as that the update scheme (280) eventually generates these optimal scalings.

Based on the well scaled inner product introduced above we also consider a BFGS



(a) Cantilever beam experiment (b) Bridge experiment

Figure 10: The final parameter λ_k generated by the update (280) scales like $\mathcal{O}(\varepsilon^{-1})$

update, i.e. we set

$$a_0(\boldsymbol{v}_1, \boldsymbol{v}_2) = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2 \tag{281}$$

and define recursively for $k = 0, 1, \ldots$

$$a_{k+1}(\boldsymbol{v}_1, \boldsymbol{v}_2) = \rho_k \left(a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) - \frac{a_k(\boldsymbol{v}_1, \boldsymbol{p}_k) a_k(\boldsymbol{p}_k, \boldsymbol{v}_2)}{a_k(\boldsymbol{p}_k, \boldsymbol{p}_k)} \right) + \frac{\langle \boldsymbol{y}_k, \boldsymbol{v}_1 \rangle \langle \boldsymbol{y}_k, \boldsymbol{v}_2 \rangle}{\langle \boldsymbol{y}_k, \boldsymbol{p}_k \rangle}. \tag{282}$$

where $\boldsymbol{p}_k = \boldsymbol{\varphi}_{k+1} - \boldsymbol{\varphi}_k$ and $\boldsymbol{y}_k = j'(\boldsymbol{\varphi}_{k+1}) - j'(\boldsymbol{\varphi}_k)$, see (69). Since the cost of evaluating a_k grows with k due to the growing recursion, it is common to use an L-BFGS variant, where the recursion depth is limited to some number L and only the vectors $\boldsymbol{y}_k$, $\boldsymbol{p}_k$ are kept in memory rather than a_k. Thus, at the $(k+1)$th iterate we determine a_{k+1} by the above recursion where we set $a_{(k+1)-L} = a_0$. For the numerical experiments we choose $\rho_k = 1$ and $L = 10$. Moreover, we set $\lambda_{max} = 1$ in the update scheme (280). Thereby we avoid the oscillation of λ_k as in Figure 10 and save computational effort in the backtracking step, since usually $\alpha_k = 1$ is accepted for $\lambda_k = 1$ if the BFGS update is used.
As a third choice for a_k we consider the inner product (167) including second order information, i.e.

$$a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2 + \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta u}_1) : \mathcal{E}(\boldsymbol{\delta u}_2) + \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta p}_1) : \mathcal{E}(\boldsymbol{\delta p}_2), \tag{283}$$

where $\boldsymbol{\delta u}_i$ and $\boldsymbol{\delta p}_i$ are the solutions of the linearized state and linearized adjoint equations, respectively. Also for this method we set $\lambda_{max} = 1$.
For the parameters in the Armijo backtracking scheme (20) within the VMPT method we choose $\sigma = 10^{-4}$ and $\beta = 0.75$ or sometimes $\beta = 0.6$. For the stopping criterion we choose according to the above scaling considerations $\sqrt{\gamma\varepsilon}\|\nabla \boldsymbol{v}_k\|_{L^2} \leq tol$, which is independent of the interface width. Therefore we will later on refer to the term $\sqrt{\gamma\varepsilon}\|\nabla \boldsymbol{v}_k\|_{L^2}$ as residual. In fact this is the well scaled norm of the residual of the first order condition $\boldsymbol{\varphi} = \mathcal{P}_k(\boldsymbol{\varphi})$, see Lemma 4.8. Moreover, recall that $\boldsymbol{v}_k \to 0$ in $H^1(\Omega)^N$, see Theorem 4.14.

Summarizing, we use the following parameters for the VMPT method:

1. $a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2$ with $\lambda_k \in \{1, (\gamma\varepsilon)^{-1}, \text{update (280)}\}$

2. a_k = L-BFGS update (281)-(282) and λ_k as in (280) with $\lambda_{max} = 1$.
3. a_k = inner product (283) and λ_k as in (280) with $\lambda_{max} = 1$.

To the first method we will refer as (scaled) projected H^1-gradient method or simply H^1-gradient method, the second method we call H^1-BFGS method and the last we call second order VMPT method. Note that the derived $\gamma\varepsilon$-scaling also enters the H^1-BFGS method by its initialization a_0.

In the following we describe the parameter we choose for the topology optimization problem. We consider homogeneous isotropic materials, thus the stiffness tensor of a single material can be written as [EGK08, Cia93]

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

with Lamé constants $\lambda > 0$ and $\mu > 0$. Note that this tensor fulfills the assumptions **(AP3)** and **(AP4)** [Cia93]. In all numerical experiments a phase representing void will be present. We use here the commonly used ersatz material approach, where the void is approximated by a very weak elastic material as described in Section 5. Thus, we use $\delta\boldsymbol{C}$ as stiffness tensor for the void phase, where $\delta > 0$ is a small parameter (usually 10^{-3}) and $\boldsymbol{C}$ is the stiffness tensor of some other material. The advantage of the ersatz material approach is that the state equation can be solved on the whole domain Ω and not just on the subset where material is present. Also the displacement field $\boldsymbol{u}$ is then defined on whole Ω.
On the interface the stiffness tensors $\boldsymbol{C}_1, \dots, \boldsymbol{C}_N$ of the distinct materials are interpolated. In the numerical experiments we consider the linear interpolation (109) and the quadratic interpolation (110), i.e.

$$\boldsymbol{C}(\boldsymbol{\varphi}) = \sum_{i=1}^{N} \varphi_i \boldsymbol{C}_i$$
$$\boldsymbol{C}(\boldsymbol{\varphi}) = \sum_{i,j=1}^{N} \varphi_i\varphi_j \boldsymbol{C}_{max\{i,j\}}$$

with a suitable extension outside the Gibbs simplex, see Section 6.1.2. For the quadratic interpolation we assume that the tensors $\boldsymbol{C}_i$ are ordered from stiff to elastic.

In case of two phases, i.e. for the scalar-valued phase field we take the standard potential [BE93]

$$\psi_0(\varphi) = \frac{1}{2}(1 - \varphi^2).$$

For multiple phases we consider the potential

$$\psi_0(\boldsymbol{\varphi}) = \frac{1}{2}\boldsymbol{\varphi}^T A\boldsymbol{\varphi} \tag{284}$$

with

$$A = \begin{pmatrix} 0 & 0.1 & \cdots & 0.1 & 1 \\ 0.1 & 0 & \cdots & 0.1 & 1 \\ \vdots & & \ddots & & \vdots \\ 0.1 & 0.1 & \cdots & 0 & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \in \mathbb{R}^{N\times N}. \tag{285}$$

The reason for this choice is that we want to have a certain angle condition in the limit $\varepsilon \to 0$ (cf. Section 6.4). For this choice of the potential the angles in a triple junction involving the Nth phase is such that the angle for the Nth phase is larger than the angles of the two other phases. Since we choose the Nth phase always to be the void phase this means that the boundary between material and void is more like a straight line and doesn't have a 120° angle. The 120° angle condition would be obtained by the choice $A = \mathbf{1} \otimes \mathbf{1} - Id$. We refer to [BFGS14] for a numerical comparison of these potentials. We also refer to [WZ07], where the effect of a 120° angle condition can be observed very well. We now compute the surface tensions and thereby the angle condition in the triple junctions for our choice of the potential. Therefor we proceed as described in [GNS99, Gar00], i.e. we solve the optimization problem

$$\begin{aligned} &\min \int_{-\infty}^{\infty} \left\{ \frac{1}{2}|\gamma'|^2 + \psi_0(\gamma) \right\} \\ &\text{s.t. } \gamma : \mathbb{R} \to \mathbb{R}^N \text{ is Lipschitz continuous, } \gamma(-\infty) = \boldsymbol{e}_i,\ \gamma(\infty) = \boldsymbol{e}_j, \\ &\quad \sum_i \gamma_i = 1,\ \gamma_i \geq 0 \end{aligned}$$

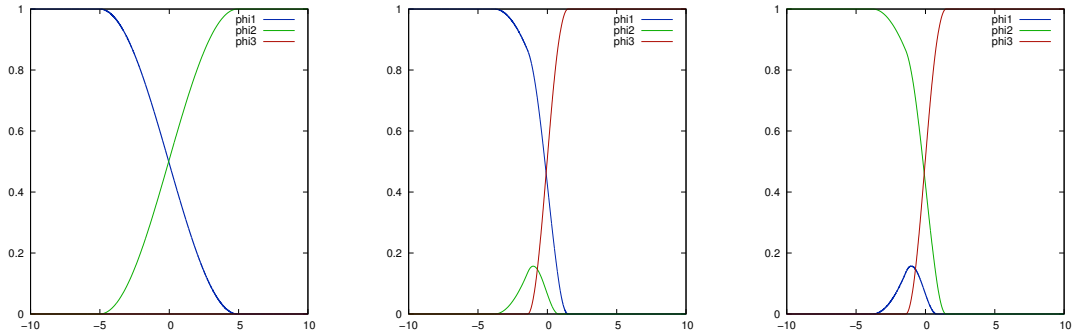
numerically (using the VMPT method), which gives as optimal value the surface tension σ_{ij} of the sharp interface between the ith and the jth phase (see also [Ste91]). Since for obstacle potentials the diffuse interface has finite thickness we can replace the value ∞ by the value 10 in the above optimization problem, which is sufficient in this case. For $N = 3$ we obtain the surface tensions

$$\sigma_{12} \approx 0.248 \quad \sigma_{13} \approx 0.765 \quad \sigma_{23} \approx 0.765.$$

The corresponding optimal phase transitions are depicted in Figure 11 (up to reparametrization) and the optimal curves γ connecting the corners $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ within the Gibbs simplex are shown in Figure 12. By Young's law (see Section 6.4) we get the following angles in a triple junction:

$$\theta_3 \approx 162^\circ \quad \theta_2 \approx 99^\circ \quad \theta_1 \approx 99^\circ, \tag{286}$$

which are shown on the right hand side of Figure 12. The angle for the void phase θ_3 is larger than the other angles, which gives rise to more robust optimal shapes. In Figure 11 it can be observed that on the interface between phase 1 and phase 3 also phase 2 is present (and similarly for the 2–3 interface). For the PDAS method this means that in the reduced Newton system φ_2 has also degrees of freedom on the 1–3 interface, resulting in a larger set $\mathcal{D}^2$ (see Section 6.10). As a consequence, the linear systems in the PDAS method get larger. From a computational point of view it would thus be better to take a perturbation of the potential (284)-(285) such that the geodesics in Figure 12 are straight lines between the corners of the Gibbs simplex. In this case only phases i and j would be present on the i–j transition and thus the PDAS systems would be smaller. However,



(a) 1–2 phase transition. (b) 1–3 phase transition. (c) 2–3 phase transition.

Figure 11: Optimal phase transitions for potential (284)-(285).

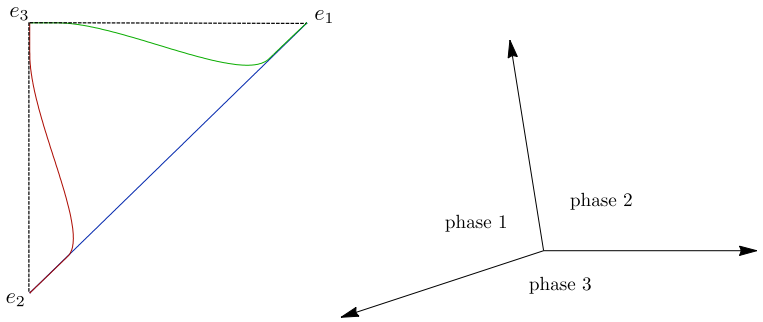


Figure 12: Geodesics and angle condition for potential (284)-(285).

it is not an easy task to find such a potential, since in general the geodesics don't move along the edges of the Gibbs simplex. In [GNS99] this is achieved by using higher order potentials.

We finally explain how we visualize a phase field $\boldsymbol{\varphi}$. A plot of the distinct components φ_i is often not very insightful. A better visualization is to plot the level sets $\{\varphi_i \geq 0.5\}$ in a single graphic. However, in this case the interface cannot be seen, which is of particular interest for phase field methods. Therefore we decided to plot the function

$$\zeta := \sum_{i=1}^{N} (i-1)\varphi_i.$$

This function takes the value $(i-1)$ in the ith phase and has a smooth transition in between. Therefore, also the position and width of the interface can be seen. However, the colors in the corresponding plot can be ambiguous. For example if 3 phases are present, then the second phase as well as the interface is shown in green, see Figure 49. On the other hand the area including the second phase is extensive, whereas the interface is typically a thin strip. Thus the ambiguous coloring should not lead to confusion. In the case of two phases we plot the scalar-valued phase field φ itself. It is common practice in the topology optimization literature to present post-processed designs. However, the plots presented here are all unprocessed.

All computations are done using the finite element toolbox FEniCS [LMW+12] and its C++ interface DOLFIN [LWH12]. As direct solver for the linear systems we use

UMFPACK [Dav07]. A personal computer with 3GHz and 4GB RAM running Debian 7 is used for the computations.

6.13 Numerical results for the mean compliance problem

In this section we present all numerical results for the mean compliance problem. On the one hand we use the mean compliance problem to extensively study the VMPT method for various parameters. On the other hand we use the VMPT method to study the mean compliance problem, e.g. the dependency of the optimal design on the various given model parameters. We also study the influence of the model parameters on the VMPT method. We embed the VMPT method in the existing literature in the context of topology optimization and compare the optimal designs obtained by the phase field model to the optimal designs obtained in literature using other models and numerical methods. Finally we compare the VMPT method to other state-of-the-art optimization methods such as pseudo time stepping based on gradient flows, the SQP method and the semismooth Newton method.

For convenience of the reader we recall at this point the optimization problem in the special case of the compliance minimization and give the derivatives of the reduced cost functional, which are used by the VMPT method. The mean compliance minimization problem reads

$$\begin{aligned}
\min \gamma \int_\Omega \left\{ \frac{\varepsilon}{2} |\nabla \boldsymbol{\varphi}|^2 + \frac{1}{\varepsilon} \psi_0(\boldsymbol{\varphi}) \right\} + \int_\Omega \boldsymbol{f}(x, \boldsymbol{\varphi}) \cdot \boldsymbol{u} + \int_{\Gamma_g} \boldsymbol{g}(x, \boldsymbol{\varphi}) \cdot \boldsymbol{u} \\
\boldsymbol{\varphi} \in H^1(\Omega)^N, \quad \boldsymbol{u} \in H^1_D \\
\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}) \mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) = \int_\Omega \boldsymbol{f}(x, \boldsymbol{\varphi}) \cdot \boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g}(x, \boldsymbol{\varphi}) \cdot \boldsymbol{\xi} \quad \forall \boldsymbol{\xi} \in H^1_D \\
\boldsymbol{\varphi} \geq 0, \quad \sum_{i=1}^N \varphi_i = 1, \quad \fint \boldsymbol{\varphi} = \mathfrak{m}.
\end{aligned}$$

The Fréchet derivative of the reduced cost functional is given as (cf. Proposition 6.30)

$$\begin{aligned}
\left\langle j'(\boldsymbol{\varphi}), \boldsymbol{\delta\varphi} \right\rangle = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{\varphi} : \nabla \boldsymbol{\delta\varphi} + \frac{\gamma}{\varepsilon} \int_\Omega \psi_0'(\boldsymbol{\varphi}) \boldsymbol{\delta\varphi} - \int_\Omega (\nabla \boldsymbol{C}(\boldsymbol{\varphi}) \mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{u})) \cdot \boldsymbol{\delta\varphi} \\
+ 2 \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})^T \boldsymbol{u} \cdot \boldsymbol{\delta\varphi} + 2 \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})^T \boldsymbol{u} \cdot \boldsymbol{\delta\varphi}.
\end{aligned}$$

Recall that in case of the mean compliance problem the adjoint state $\boldsymbol{p}$ coincides with state $\boldsymbol{u}$ and thus no adjoint equation has to be solved in order to compute j'. Also the linearized adjoint state $\boldsymbol{\delta p}$ coincides with the linearized state $\boldsymbol{\delta u}$. The second order Fréchet derivative is given as (cf. Theorem 6.44)

$$\begin{aligned}
j''(\boldsymbol{\varphi})[\boldsymbol{\tau\varphi}, \boldsymbol{\delta\varphi}] = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{\tau\varphi} : \nabla \boldsymbol{\delta\varphi} + \frac{\gamma}{\varepsilon} \int_\Omega \psi_0''(\boldsymbol{\varphi}) \boldsymbol{\delta\varphi\tau\varphi} \\
- \int_\Omega (\boldsymbol{C}''(\boldsymbol{\varphi}) \boldsymbol{\tau\varphi} \mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{u})) \cdot \boldsymbol{\delta\varphi} - 2 \int_\Omega (\nabla \boldsymbol{C}(\boldsymbol{\varphi}) \mathcal{E}(\boldsymbol{\tau u}) : \mathcal{E}(\boldsymbol{u})) \cdot \boldsymbol{\delta\varphi} \\
+ 2 \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi},\boldsymbol{\varphi}}(\boldsymbol{\varphi})[\boldsymbol{\tau\varphi}, \boldsymbol{\delta\varphi}] \cdot \boldsymbol{u} + 2 \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})^T \boldsymbol{\tau u} \cdot \boldsymbol{\delta\varphi} \\
+ 2 \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi},\boldsymbol{\varphi}}(\boldsymbol{\varphi})[\boldsymbol{\tau\varphi}, \boldsymbol{\delta\varphi}] \cdot \boldsymbol{u} + 2 \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})^T \boldsymbol{\tau u} \cdot \boldsymbol{\delta\varphi},
\end{aligned}$$

where $\boldsymbol{\tau u}$ is the solution of the linearized state equation (124) in direction $\boldsymbol{\tau\varphi}$, i.e.

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{\tau u}) : \mathcal{E}(\boldsymbol{\xi}) = -\int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\tau\varphi}\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{\xi}) + \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\boldsymbol{\tau\varphi}\cdot\boldsymbol{\xi} + \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\boldsymbol{\tau\varphi}\cdot\boldsymbol{\xi}$$

for all $\forall \boldsymbol{\xi} \in H^1_D$. Recall that the mean compliance functional fulfills all assumptions which are needed for the global convergence of the VMPT method (see Section 6.7). Also all assumptions for the Γ-convergence as $\varepsilon \to 0$ are fulfilled (see Section 6.4). We can also prove that j' is Lipschitz continuous by the following

Lemma 6.80. *For the mean compliance problem it holds*

$$\|j''(\boldsymbol{\varphi})\|_{\mathcal{L}(H^1\cap L^\infty,(H^1\cap L^\infty)^*)} \le C \quad \forall \boldsymbol{\varphi} \in \Phi_{ad}$$

for some $C > 0$. In particular we get $j \in C^{1,1}(\Phi_{ad})$.

Proof. From **(AP2)** and **(AP11)** we get $|\psi_0''(\boldsymbol{\varphi})| \le C$, $|\boldsymbol{C}'(\boldsymbol{\varphi})| \le C$ and $|\boldsymbol{C}''(\boldsymbol{\varphi})| \le C$ for all $\boldsymbol{\varphi} \in \Delta^{N-1}$. From (106) and (107) we get $\|\boldsymbol{f}(\boldsymbol{\varphi})\|_{L^2} \le C$, $\|\boldsymbol{f}_{\varphi}(\boldsymbol{\varphi})\|_{L^2} \le C$, $\|\boldsymbol{f}_{\varphi,\varphi}(\boldsymbol{\varphi})\|_{L^2} \le C$, $\|\boldsymbol{g}(\boldsymbol{\varphi})\|_{L^2} \le C$, $\|\boldsymbol{g}_{\varphi}(\boldsymbol{\varphi})\|_{L^2} \le C$, $\|\boldsymbol{g}_{\varphi,\varphi}(\boldsymbol{\varphi})\|_{L^2} \le C$ for all $\boldsymbol{\varphi} \in \Phi_{ad}$. From the a priori estimate (123) we get $\|S(\boldsymbol{\varphi})\|_{H^1_D} \le C$ for all $\boldsymbol{\varphi} \in \Phi_{ad}$ and the a priori estimate (125) together with the preceding estimates implies $\|S'(\boldsymbol{\varphi})\boldsymbol{\tau\varphi}\|_{H^1_D} \le C\|\boldsymbol{\tau\varphi}\|_{H^1\cap L^\infty}$ for all $\boldsymbol{\varphi} \in \Phi_{ad}$. The statement then follows by Hölder and trace estimates. □

Recall that the regularity $j \in C^{1,1}(\Phi_{ad})$ is used in Theorem 4.14 to show that $\langle j'(\boldsymbol{\varphi}_k)\boldsymbol{v}_k\rangle \to 0$ and $\boldsymbol{v}_k \to 0$ in H^1, where $\boldsymbol{\varphi}_k$ are the iterates of the VMPT method and $\boldsymbol{v}_k$ the respective search directions.

For the SIMP method it is known that the mean compliance problem is convex if the penalization parameter $p = 1$ is chosen. Moreover, the problem is also well posed without regularization [Ben83]. A similar result can be shown for the phase field model:

Lemma 6.81. *Assume that $\boldsymbol{f}$ and $\boldsymbol{g}$ are independent of $\boldsymbol{\varphi}$, $\|\boldsymbol{f}\|_{L^2} + \|\boldsymbol{g}\|_{L^2} > 0$ and that $\boldsymbol{C}(\boldsymbol{\varphi})$ is linear. Then the mean compliance problem with $\gamma = 0$ is strictly convex.*

Proof. Let $\boldsymbol{\varphi}, \boldsymbol{\delta\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$ with $\boldsymbol{\delta\varphi} \neq 0$. Under the assumptions it holds

$$\begin{aligned} j''(\boldsymbol{\varphi})[\boldsymbol{\delta\varphi},\boldsymbol{\delta\varphi}] &= -2\int_\Omega (\nabla \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{\delta u}) : \mathcal{E}(\boldsymbol{u}))\cdot\boldsymbol{\delta\varphi} \\ &= 2\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{\delta u}) : \mathcal{E}(\boldsymbol{\delta u}), \end{aligned}$$

where we inserted the linearized state equation tested with $\boldsymbol{\xi} = \boldsymbol{\delta u}$. It remains to show $\boldsymbol{\delta u} \neq 0$. Since $\boldsymbol{f}$ or $\boldsymbol{g}$ does not vanish we get from the state equation $\boldsymbol{u} \neq 0$. If we assume $\boldsymbol{\delta u} = 0$ we get from the linearized state equation (tested with $\boldsymbol{\xi} = \boldsymbol{u}$) that $0 = \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi})\boldsymbol{\delta\varphi}\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{u}) = \int_\Omega \boldsymbol{C}(\boldsymbol{\delta\varphi})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{u}) > 0$, which is a contradiction. □

From that we can conclude that in case of linear stiffness interpolation also the unregularized problem is well posed.

Lemma 6.82. *Let the assumptions of Lemma 6.81 hold. Then there exists a unique local (and therefore global) minimizer of the mean compliance problem with $\gamma = 0$.*

Proof. The mean compliance functional is continuous in $L^1(\Omega)^N$ (see e.g. [BGHR15] or Section 6.4) and thus also in $L^2(\Omega)^N$. From convexity we get the L^2 weak lower

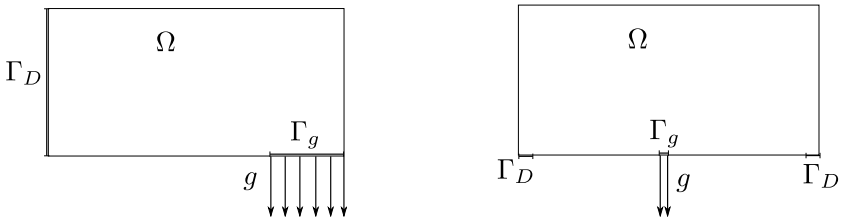


Figure 13: The cantilever beam and bridge setup.

semicontinuity of j. Since Φ_{ad} is convex, closed and bounded in L^2 we get the existence of a minimizer by the direct method in the calculus of variations. The uniqueness follows from strict convexity. □

Since the linearized state and linearized adjoint equations coincide, the second order metric (283) simplifies to

$$a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2 + 2 \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta u}_1) : \mathcal{E}(\boldsymbol{\delta u}_2),$$

where $\boldsymbol{\delta u}_i$ is the solution of the linearized state equation in direction $\boldsymbol{v}_i$, $i = 1, 2$. Equivalently, this inner product can be written as (see (235))

$$\begin{aligned} a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2 - 2 \int_\Omega \boldsymbol{C}'(\boldsymbol{\varphi}_k)\boldsymbol{v}_1\mathcal{E}(\boldsymbol{u}_k) : \mathcal{E}(\boldsymbol{\delta u}_2) + 2 \int_\Omega \boldsymbol{f}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k)\boldsymbol{v}_1 \cdot \boldsymbol{\delta u}_2 \\ + 2 \int_{\Gamma_g} \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_k)\boldsymbol{v}_1 \cdot \boldsymbol{\delta u}_2. \end{aligned}$$

To investigate the VMPT method we use mainly two different setups, namely a cantilever beam and a bridge setup. These are considered e.g. in [BFGS14, WR12]. Especially the cantilever beam problem is used as a benchmark problem throughout the literature. The reason to use these experiments here is that the VMPT method works very well on them and thus we are able to examine the method also on fine meshes and for various parameters ε and γ.

Example 6.83. The first experiment is the cantilever beam setting as shown on the left hand side of Figure 13. The design domain is $\Omega = (-1, 1) \times (0, 1)$. The boundary force $\boldsymbol{g} \equiv (0, -250)^T$ is acting on $\Gamma_g = (0.75, 1) \times \{0\}$ and we set the volume force to $\boldsymbol{f} \equiv \boldsymbol{0}$. The structure is supported at the left hand side, i.e. $\Gamma_D = \{-1\} \times (0, 1)$. If not stated otherwise then the Lamé constants $\mu = 5000$ and $\lambda = 5000$ are taken for the hard material phase and $\mu = 10$ and $\lambda = 10$ for the weak material phase, which approximates void. A typical solution for this problem is given in Figure 7 on the left hand side.

Example 6.84. The setup of the bridge experiment is depicted on the right hand side of Figure 13. The design domain is again $\Omega = (-1, 1) \times (0, 1)$ with fixed boundary $\Gamma_D = \{1 - 3/32 \le |x_1| \le 1\} \cap \{x_2 = 0\}$. The boundary traction $\boldsymbol{g} \equiv (0, -5000)^T$ acts on $\Gamma_g = \{|x_1| \le 1/32\} \cap \{x_2 = 0\}$ and no body force is present. The same Lamé constants are used as in Example 6.83. A typical solution for this problem is given in Figure 7 on the right hand side.

We emphasize that the obtained optimal shapes are only optimal for infinitesimal small displacements, since we use the equations of linearized elasticity in the model. However, even if the displacement $\boldsymbol{u}$ of the optimal design is large, the solution is still reasonable in the sense of the following rescaling consideration. Let an optimal design $\boldsymbol{\varphi}$ be given with

corresponding displacement $\boldsymbol{u}$. We rescale $\boldsymbol{u}$ by some constant c, such that $c\boldsymbol{u}$ stays in the regime of linearized elasticity. The rescaled deformation solves the state equation for the rescaled forces $c\boldsymbol{f}$ and $c\boldsymbol{g}$ because of linearity. The corresponding compliance is then scaled by the factor c^2. To balance the weights in the cost functional one has to use $c^2\gamma$ as scaling for the Ginzburg-Landau energy. Thus, if the pair $(\boldsymbol{\varphi}, \boldsymbol{u})$ is a local optimum, then also $(\boldsymbol{\varphi}, c\boldsymbol{u})$ is a local optimum for the rescaled topology optimization problem with data $c\boldsymbol{f}$, $c\boldsymbol{g}$ and $c^2\gamma$.

The numerical results are structured as follows:
In Section 6.13.1 we will compare the usage of an obstacle potential to a smooth potential. It turns out that the mass constraint $\fint \boldsymbol{\varphi} = \mathfrak{m}$ is incompatible with a smooth potential in the sense that numerical solutions for large ε are not physical, which is on the other hand not a problem for an obstacle potential.
Section 6.13.2 deals with the comparison of linear and quadratic interpolation of the stiffness tensors. Choosing a reasonable scaling, the VMPT method is much more efficient for quadratic interpolation. Also the optimal designs are better for quadratic interpolation since the phases are well separated even for large ε.
In Section 6.13.3 mesh independency of the VMPT method for various setups, number of phases, stiffness interpolations, scalings and inner products is shown numerically. A nested iteration in the mesh parameter h is proposed to enhance the efficiency of the VMPT method.
In Section 6.13.4 various inner products used in the VMPT are compared among each other in terms of efficiency in iteration numbers and CPU times and in terms of the quality of the obtained optimal shape. The performance can be increased considerably by the $\gamma\varepsilon$-scaling and the BFGS update. The second order metric gives rise to lower local minima.
In Section 6.13.5 it is shown that the complexity of the optimal design increases as the parameter γ is decreased. Moreover, convergence of the optimal phase fields to characteristic functions is shown as $\varepsilon \to 0$ and the dependence of the VMPT method on ε is studied.
The optimal shapes obtained by the VMPT method are compared to results in the literature in Section 6.13.6. Additionally, we show that the numerical method used in [TP13] is for certain parameters an instance of the VMPT method and that the method in [Tav14] is quite different from the VMPT method, although it looks similar at first glance.
In Section 6.13.7 we present numerical results for the time adaptivity proposed for pseudo time stepping in Section 6.8. We show that the usage of adaptive time step sizes increases the efficiency of the time stepping considerably. However, the H^1-gradient method is still more efficient.
Section 6.13.8 is devoted to the comparison of the VMPT method with the SQP method described in Section 6.9. It turns out that the efficiency of the VMPT method is better in terms of computation time. Also the radius of local convergence is very tiny for the SQP method, whereas the VMPT method convergences for any initial guess.
The comparison to the semismooth Newton (SSN) method is performed in Section 6.13.9. In contrast to the VMPT method, the SSN method is mesh dependent. Moreover, just as for the SQP method, the SSN method needs more computation time than the VMPT method and the radius of local convergence is very small.
In Section 6.13.10 we show that the Lagrange multipliers are in general no functions as indicated by the theoretical results in Section 6.5. Moreover, we show that the Lagrange multiplier for the constraint $\varphi \leq 1$ is in certain cases a scaled characteristic function.
We know that the L^2 inner product does not satisfy the assumptions for global conver-

gence of the VMPT method. In Section 6.13.11 we show the numerical consequences of this result, namely the mesh dependency of the projected L^2-gradient method. The ill-posedness of the projection type subproblem is shown. The same holds for the L^2-BFGS method.

A final remark on nesting in γ: It will turn out that it is very efficient to nest the iteration in the mesh parameter h and in the phase field parameter ε, i.e. the iteration is started using a large h (coarse mesh) and large ε (broad interface) and h as well as ε are decreased during the iteration. This is reasonable since the limit $h \to 0$ exists because everything is well posed in the function space setting. Also the limit $\varepsilon \to 0$ exists due to the Γ-convergence result. However, a nested iteration in γ is not reasonable, although the VMPT method works better for larger γ. Since the optimization problem does not converge as $\gamma \to 0$ (except for special cases, see Lemma 6.82), an optimal design for large γ is not a good approximation for an optimal design for smaller γ. This can be also observed in the numerics, see Section 6.13.5.

6.13.1 Difficulties arising for smooth potential

First of all we want to give a motivation why we use an obstacle potential rather than a smooth potential. It will turn out that a smooth potential causes certain difficulties with the mass constraint $⨍ \varphi = \mathfrak{m}$ for large values of ε. For simplicity we consider here only the case that two phases are present. The considerations are also valid for multiple phases.

In the presented model (121) for topology optimization an obstacle potential is used, i.e. the Ginzburg-Landau energy

$$E(\varphi) = \int_\Omega \left\{ \frac{\varepsilon}{2} |\nabla \varphi|^2 + \frac{1}{\varepsilon} \psi(\varphi) \right\}$$

is used with an obstacle potential

$$\psi(\varphi) = \begin{cases} \psi_0(\varphi) & -1 \le \varphi \le 1 \\ \infty & \text{else} \end{cases}.$$

To get a smooth cost functional the potential ψ in the energy is replaced by ψ_0 and $-1 \le \varphi \le 1$ is introduced as a hard constraint. In the context of phase field models it is also common to take a smooth potential like the standard double well potential $\psi(\varphi) = \frac{1}{4}(1-\varphi^2)^2$, cf. Figure 14. In this case there are no box constraints for φ present and after elimination of the state variable the optimization problem

$$\begin{aligned} &\min j(\varphi) \\ &⨍_\Omega \varphi = \mathfrak{m} \end{aligned}$$

is obtained. Usually the smooth potential has global minima in $\varphi = -1$ and $\varphi = 1$, such that these values are favored by the Ginzburg-Landau energy. However, in contrast to an obstacle potential φ can obtain values outside the interval $[-1, 1]$ for positive ε.
In the limit $\varepsilon \to 0$ one formally gets $\varphi = \pm 1$ a.e. in Ω. In the following we present a numerical result for the case that a smooth potential is used. For the stiffness tensor we

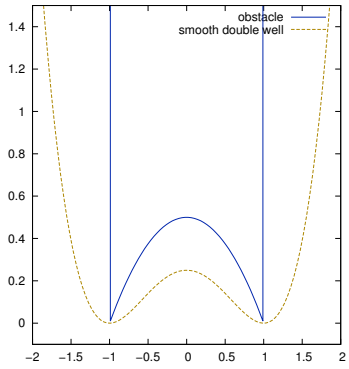


Figure 14: Smooth double well and nonsmooth double obstacle potential.

take the interpolation

$$\boldsymbol{C}(\varphi) = \begin{cases} \boldsymbol{C}_1 & \varphi <= -1 \\ \boldsymbol{C}_1 + (1 + \frac{1}{8}\varphi(15 - 10\varphi^2 + 3\varphi^4))\frac{\boldsymbol{C}_2 - \boldsymbol{C}_1}{2} & -1 < \varphi < 1 \\ \boldsymbol{C}_2 & \varphi >= 1. \end{cases} \tag{287}$$

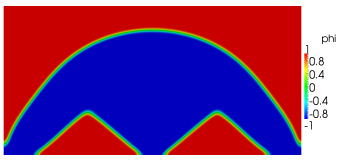
Note that the optimal design now also depends on $\boldsymbol{C}(\varphi)$ for $\varphi \in \mathbb{R} \setminus [-1, 1]$. The used interpolation has C^2-regularity, which is important since we want to apply a second order method.

We reformulate the optimization problem in the following way. The feasible set is an affine hyperplane. As described in Section 6.7 we translate the problem by a constant vector such that the feasible set becomes $\{\varphi \in H^1(\Omega) \cap L^\infty(\Omega) \mid \int \varphi = 0\}$, which is a linear space. In this sense we get an unconstrained optimization problem, for which efficient solvers are available. We use a Trust-Region-Newton-Steihaug-cg solver, which is e.g. described in [NW06, CGT00].

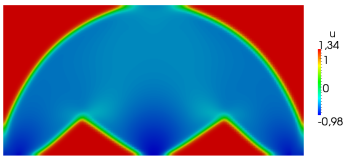
As numerical experiment we take the bridge problem (Example 6.84). We choose $\boldsymbol{C}_2 = \varepsilon^2 \boldsymbol{C}_1$, thus $\varphi = -1$ corresponds to material and $\varphi = 1$ corresponds to void. The final designs for varying ε are depicted in Figure 15. We also include a reference solution in Figure 15a, which is computed using an obstacle potential. It can be observed that the value of φ in the red region is not near $\varphi = 1$ as it should be, but ranges from $\varphi = 1.34$ for $\varepsilon = 0.02$ to $\varphi = 1.06$ for $\varepsilon = 0.001$. We also see that in the bluish region the value of φ is not constant, but e.g. for $\varepsilon = 0.02$ takes values in $[-0.98, -0.5]$. Table 2 lists the mean value of φ in the bluish region and in the red region of the structure. Moreover the area occupied by the bluish region also depends on ε. The larger ε is the larger is the bluish region. This seems to be contradictory to the mass constraint $\fint_\Omega \varphi = \mathfrak{m}$. The behavior can be explained by the following consideration. The structure with the lowest mean compliance surely consists of material everywhere in Ω. This trivial solution is prevented by the mass constraint. However, since φ can have values in whole $\mathbb{R}$ one can construct the phase field

$$\varphi(x) = \begin{cases} \frac{|\Omega|}{|B_\delta(x_0)|}(\mathfrak{m} + 1) - 1 & x \in B_\delta(x_0) \\ -1 & \text{else} \end{cases}$$

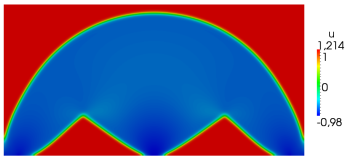
for some $x_0 \in \Omega$ and small $\delta > 0$. By construction φ fulfills the mass constraint $\fint_\Omega \varphi = \mathfrak{m}$ and it holds $\boldsymbol{C}(\varphi(x)) = \varepsilon^2 \boldsymbol{C}_1$ in $B_\delta(x_0)$ and $\boldsymbol{C}(\varphi(x)) = \boldsymbol{C}_1$ in $\Omega \setminus B_\delta(x_0)$ if δ is small enough. Thus, material is put everywhere except for the δ-ball and therefore this phase



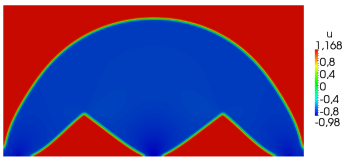
(a) $\varepsilon = 0.02$ (obstacle potential)



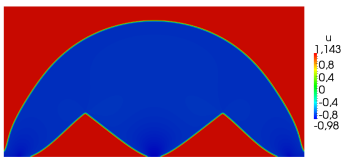
(b) $\varepsilon = 0.02$



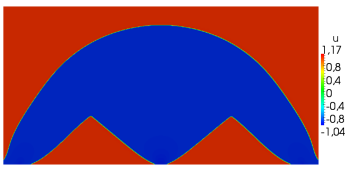
(c) $\varepsilon = 0.01$



(d) $\varepsilon = 0.005$



(e) $\varepsilon = 0.0025$



(f) $\varepsilon = 0.001$

Figure 15: Optimal bridge designs for different ε using a smooth double well potential. The pure phases take values of $\pm 1 + \mathcal{O}(\varepsilon)$.

field describes a structure with optimal stiffness for $\delta \to 0$. However, the Ginzburg-Landau energy is very high for this φ and hence we don't see this solution. But the problem stays the same. Since φ can attain values outside of $[-1, 1]$, the mean value $⨍ \varphi$ does not model the volume of the material phase anymore. By choosing large values for φ in the void, one obtains additional mass for the material phase, which can be observed in Figure 15. Note that this difficulty arises for positive ε. For $\varepsilon \to 0$ the values of φ are forced to the minima ± 1 of the potential. On the other hand one wants to choose ε as large as possible in the numerics. This shows that using a smooth potential is disadvantageous. For the obstacle potential the discussed difficulty does not arise since φ only has values within $[-1, 1]$. Figure 15a shows that even for large ε a reasonable structure can be obtained, which has the correct mass and which is a much better approximation of the sharp interface solution than the corresponding solution with smooth potential in Figure 15b. Therefore we don't consider a smooth potential in this work.

We note that in the work of Penzler, Rumpf and Wirth [PRW12] also a phase field model with smooth double well potential for compliance minimization is used. However, they don't model the mass by the integral $⨍_\Omega \varphi$, but by the nonlinear term

$$\frac{1}{4} \int_\Omega (\varphi + 1)^2.$$

Using this model the phenomenon described above cannot occur since the values below the integral are always nonnegative. On the other hand, the constraint $\frac{1}{4} \int_\Omega (\varphi + 1)^2 = \mathfrak{m}$ gives rise to nonconvex control constraints and thus cannot be handled by the VMPT method. However, it is also possible to add the mass as a penalization term in the cost functional

instead of using a mass constraint, which is done in [PRW12].

ε	φ in blue region	φ in red region
0.02	-0.6014	1.339
0.01	-0.6953	1.213
0.005	-0.8073	1.136
0.0025	-0.8437	1.114
0.001	-0.9308	1.058

Table 2: The mean value of φ in the red and blue region for different ε. The data corresponds to the pictures in Figure 15.

Another difficulty arising for smooth potential is that in the numerics the stiffness tensor $\boldsymbol{C}(\varphi)$ has to be defined for all $\varphi \in \mathbb{R}$. In the obstacle case only the values of $\boldsymbol{C}(\varphi)$ for $\varphi \in [-1,1]$ are used by the VMPT method, since all iterates are feasible. It is not easy to construct a simple extension of $\boldsymbol{C}(\varphi)$ on the whole real line. For the extension (287) used here a higher order interpolating polynomial has to be used to gain the C^2-regularity of $\boldsymbol{C}(\varphi)$. An extension of a linear or quadratic interpolation on $[-1,1]$ can theoretically be constructed as e.g. in (111). However, to compute the needed parameter δ one has to know estimates θ, Θ for the lowest and highest eigenvalues of the stiffness tensors $\boldsymbol{C}_i$ of the distinct materials. When using an obstacle potential the implementation of the interpolation $\boldsymbol{C}(\varphi)$ is much simpler.
Another aspect that may cause problems is that the stiffness of the material corresponding to $\varphi \notin [-1,1]$ can be higher than the stiffness of the pure phases. As an example consider the quadratic interpolation (110) for two phases, which can be written as $\boldsymbol{C}(\varphi) = k(\varphi)\boldsymbol{C}_1$ for some quadratic function $k : \mathbb{R} \to \mathbb{R}$ and $\varphi \in [-1,1]$ (see also (288) below). If $\varphi = -1$ corresponds to material and $\varphi = 1$ corresponds to void, then it holds $k'(-1) < 0$. If $k(\varphi)$ is extended smoothly on $\mathbb{R}$ it follows that $k(-1-\delta) > 1$ for all δ small enough. Thus, the stiffness of the material corresponding to $\varphi = -1-\delta$ is higher than the stiffness for the pure phase $\varphi = -1$. As a consequence, the compliance term in the cost functional will favor phase fields with $\varphi < -1$.
In the case of multiple phases it is even more difficult to construct a suitable extension of the stiffness tensor outside of the Gibbs simplex. Also the smooth potential ψ has to be defined adequately on $\mathbb{R}^N$.

6.13.2 Influence of the stiffness interpolation scheme

In this section we have a closer look at the choice of the stiffness tensor $\boldsymbol{C}(\boldsymbol{\varphi})$. We restrict ourselves to the case of two phases, namely material and void. We will compare the two different choices of linear and quadratic interpolation as given by (109) and (110), i.e.

$$\boldsymbol{C}(\boldsymbol{\varphi}) = \sum_{i=1}^{N} \varphi_i \boldsymbol{C}_i$$

$$\boldsymbol{C}(\boldsymbol{\varphi}) = \sum_{i,j=1}^{N} \varphi_i \varphi_j \boldsymbol{C}_{max\{i,j\}}$$

for all $\boldsymbol{\varphi}$ in the Gibbs simplex with a suitable extension to $\mathbb{R}^N$. As already mentioned, the choice of the extension does not influence the VMPT method, since the values of the iterates $\boldsymbol{\varphi}_i$ are always within the Gibbs simplex. Since we consider only two phases we

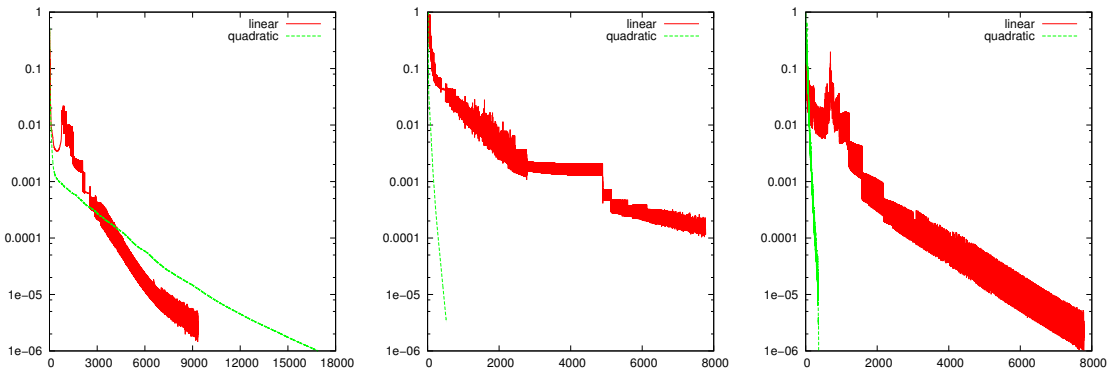
can use a scalar-valued phase field. Assume that the stiffness of void is given as $\boldsymbol{C}_2 = \delta \boldsymbol{C}_1$ where $\boldsymbol{C}_1$ is the stiffness tensor of the material and δ is a small constant. The stiffness interpolations corresponding to the scalar valued phase field are then

$$\begin{aligned} \boldsymbol{C}(\varphi) &= \frac{1}{2}(1+\varphi+(1-\varphi)\delta)\boldsymbol{C}_1 \\ \boldsymbol{C}(\varphi) &= \left(\frac{1}{4}(1-\delta)\varphi^2 + \frac{1}{2}(1-\delta)\varphi + \frac{1}{4}(1-\delta) + \delta\right)\boldsymbol{C}_1. \end{aligned} \tag{288}$$

Here, $\varphi = 1$ corresponds to material, whereas $\varphi = -1$ corresponds to void.
For these two different interpolations we compare on the one hand the shape of the local minimizers, i.e. the model, which is independent of the numerical method, and on the other hand the performance of the projected H^1-gradient method, i.e. the numerical method to compute a local minimizer of the model. We emphasize that the choice of $\boldsymbol{C}(\varphi)$ is part of the model and not of the numerical method. Moreover, in the limit problem for $\varepsilon \to 0$, the function $\varphi \in BV(\Omega, \{\pm 1\})$ only attains the function values +1 and −1 (cf. Section 6.4). Thus the choice of the stiffness tensor interpolation does not influence the Γ-limit problem, but only the regularized problem for positive ε.
As experiment we choose the cantilever beam (Example 6.83) with $\varepsilon = 0.04$ and $\gamma = 0.5$. The mesh is equidistant with $h = 2^{-6}$. No stopping criterion is used, but the method is carried out until it breaks down. This happens when the iterate is very close to the minimum such that the computed search direction is no descent direction anymore because of high approximation errors as already discussed.
We compare the VMPT method using the inner product $a_k = (\nabla ., \nabla .)_{L^2}$ and different choices for λ_k, namely

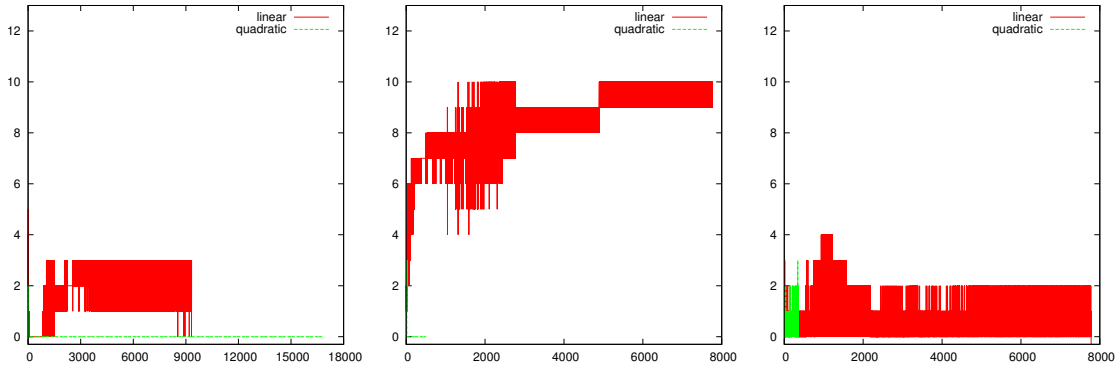
1. $\lambda_k = 1$
2. $\lambda_k = (\varepsilon\gamma)^{-1}$
3. λ_k updated by the method described in (280)

Note that the choice of $\lambda_k = (\varepsilon\gamma)^{-1}$ is equivalent to the choice of $a_k = \varepsilon\gamma(\nabla ., \nabla .)_{L^2}$ together with $\lambda_k = 1$.
In Figure 16 the development of the residual $r := \sqrt{\varepsilon\gamma}\|\nabla v_k\|_{L^2}$, where v_k is the search direction, is depicted and Figure 17 shows the number of line search iterations in the corresponding Armijo backtracking. Recall that $r = 0$ if and only if a stationary point is found. For $\lambda_k = 1$ it can be seen that the method needs less iterations if linear interpolation is used. On the other hand, no line search is needed for quadratic interpolation, i.e. the full step is always accepted except for the first few iterations, whereas up to three backtracking steps are needed when using linear interpolation. Also the PDAS method for the solution of the projection type subproblem needs more iterations when using linear interpolation (not depicted in the figure). This leads to the fact that a single iteration for linear interpolation is more expensive than an iteration for quadratic interpolation resulting in a total computation time of about 4.5 hours for linear interpolation and 2.5 hours for quadratic interpolation (see Table 3). Thus for quadratic interpolation the method is faster although more iterations are needed.
A more severe difference can be observed by choosing $\lambda = (\gamma\varepsilon)^{-1}$. The method needs significantly less iterations for quadratic interpolation. Moreover, the method for linear interpolation breaks down already at a residual of $r = 10^{-4}$. Again, for quadratic interpolation no line search has to be performed except for the first few steps, whereas for linear interpolation up to 10 backtracking steps have to be done. Thus the method is far more



(a) $\lambda = 1$ (b) $\lambda = (\varepsilon\gamma)^{-1}$ (c) λ updated by (280)

Figure 16: Development of the residual r for different choices of λ.

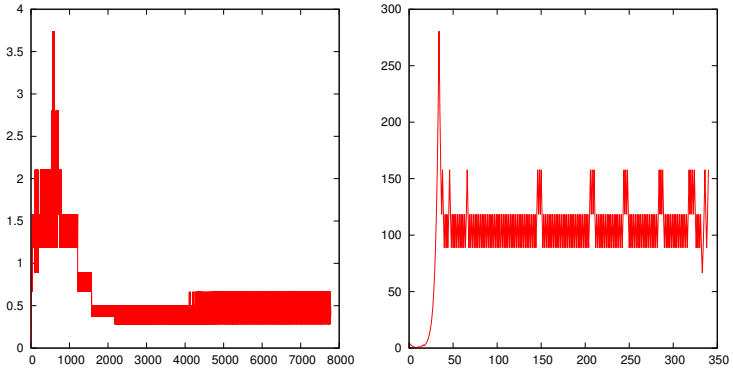


(a) $\lambda = 1$ (b) $\lambda = (\varepsilon\gamma)^{-1}$ (c) λ updated by (280)

Figure 17: Number of line search iterations.

efficient if a quadratic interpolation is used.
In the last case where λ_k is updated by (280) the method for quadratic interpolation is even faster (about 4 minutes of computation time, see Table 3). Also the method for linear interpolation works better compared to $\lambda = (\gamma\varepsilon)^{-1}$ and the number of line search steps is much lower. Now also for quadratic stiffness a line search has to be performed, which is by construction of the update (280). The step lengths λ_k which are calculated by the update (280) are shown in Figure 18. It can be observed that for linear interpolation a value of $\lambda_k \approx 0.5$ is obtained from iteration 2000 on. This explains why the method performs similar to the first case where $\lambda_k = 1$ is chosen. For quadratic interpolation λ_k eventually oscillates between 89 and 118. Note that for this experiment we have $(\gamma\varepsilon)^{-1} = 50$. The reason why the update (280) produces larger step lengths is that somehow the largest possible λ_k is chosen within the discrete possible values. In Figure 19 the values of j across the H^1-projection arc $\lambda \mapsto \mathcal{P}_{a,\lambda}(\varphi_{200})$ in the 200th iteration are depicted, where $\mathcal{P}_{a,\lambda}$ denotes the solution operator of the projection type subproblem (18). The value for λ which generates the largest descent in the energy is $\lambda^* \approx 64$. We see that $(\gamma\varepsilon)^{-1}$ is slightly below λ^*, whereas the update (280) produces a larger λ. In particular this shows that $\lambda = (\gamma\varepsilon)^{-1}$ is a very good choice.

From these experiments we conclude that the H^1-gradient method gives rise to a much better search direction if quadratic stiffness interpolation is used instead of linear inter-



(a) Linear stiffness interpolation. (b) Quadratic stiffness interpolation.

Figure 18: Development of λ_k using (280).

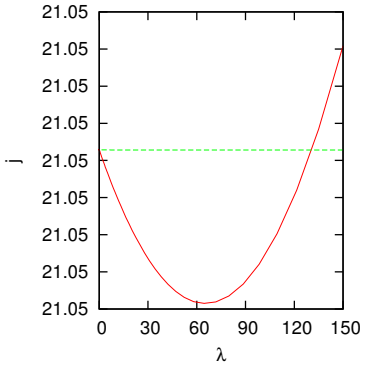


Figure 19: Values for j along the projection arc for quadratic stiffness interpolation.

	$\lambda = 1$	$\lambda = (\varepsilon\gamma)^{-1}$	λ_k updated by (280)
linear interpolation	4h 29m	19h 13m	2h 1m
quadratic interpolation	2h 41m	5.1m	4.3m

Table 3: CPU time for various choices of λ and interpolations $\boldsymbol{C}(\varphi)$.

polation. Thus the performance of the numerical method can be improved by choosing an adequate interpolation. Moreover, the performance of the numerical method can be further enhanced if a suitable value for λ is chosen, such as $\lambda = (\gamma\varepsilon)^{-1}$ or the values generated by the update scheme (280). This is especially important for small ε, since the minimum of j along the H^1-projection arc scales with ε^{-1}, see also Figure 10. We note that also for the H^1-BFGS method the quadratic interpolation behaves better than the linear interpolation (80 vs. 434 iterations).

We try to give an explanation why using quadratic stiffness interpolation works better than linear. Assume that the profile of φ is a sine function, which is typically the case. Then the interpolation $\boldsymbol{C}(\varphi(x))$ is also a sin-function in the linear case and like a $\sin^2$-function in the quadratic case as depicted in Figure 20. Thus the stiffness of the interface is in the latter case lower than in the former case. This leads to a 'penalization' of the interface for quadratic interpolation in the sense that the interface consumes mass from the constraint $\fint_\Omega \varphi = \mathfrak{m}$ but on the other hand it doesn't increase the stiffness of the structure much. This is the same idea as in the SIMP model where the interpolation $\boldsymbol{C}_1 + \rho^p(\boldsymbol{C}_2 - \boldsymbol{C}_1)$ is used to get a clear 0-1 pattern. Here, $\rho(x) \in [0,1]$ is the density of the material and p the penalization parameter, cf. Section 5. Because of this 'penalization' of the interface, the mean compliance part F of the cost functional favors structures without interface. On the other hand the Ginzburg Landau part E of the cost functional favors an interface thickness of $\varepsilon\pi$ (cf. [BE91b]). Using linear interpolation, a structure with large interfacial area can be stiffer than a structure with thin interface. Thus, for quadratic interpolation both terms E and F favor a thin interface (resp. no interface), whereas for linear interpolation, E favors a small interface and F maybe a large interface, which are contrary goals. Therefore the problem behaves better for quadratic interpolation. We note that for quadratic interpolation the demixing of the phases is already done by the compliance. Thus the potential term in the Ginzburg-Landau energy is not necessarily needed for demixing (though it is needed to obtain a reasonable sharp interface limit). As an example we refer to the experiment in Figure 36, where γ is taken very small and the phases already demix for $\varepsilon = 5000$. We also refer to [YINT10, TP13], where only the gradient term of the Ginzburg-Landau energy is incorporated into the cost functional, but not the potential term. In [TP13] it is even stated that the usage of an interpolation of the stiffness tensors as in the SIMP method is a good substitution for the potential term. However, to obtain a reasonable sharp interface limit for $\varepsilon \to 0$ the potential term is mandatory. Also note that the consideration about the interface 'penalization' is only valid for the mean compliance problem. For other problems, such as the compliant mechanism problem, the potential term plays an essential role in demixing the phases.

We support the preceding considerations about interface penalization with the following numerical experiments. We repeat the cantilever beam simulation for varying values of γ (and fixed $\varepsilon = 0.04$). Thereby we can study the influence of the Ginzburg Landau energy on the final design. Figures 21 and 22 show the result. For linear stiffness interpolation it can be observed that only the value $\gamma = 0.5$ gives rise to a physically meaningful structure. For smaller values of γ the final structure consists of a large interfacial area. With decreasing γ this interfacial area grows and also the area occupied by non-void expands

(non-red regions in the picture). This is in agreement with the considerations above. Note that even in the solution for $\gamma = 0.5$ the hole in the middle of the structure consists of interface, i.e. it holds $|\varphi| < 1$. To obtain physically meaningful solutions one has to decrease ε considerably, cf. Figure 39.
A very different behavior can be observed when using quadratic stiffness interpolation. For any γ the solution consists of two pure phases which are separated by a thin interfacial layer as it is desired. We perceive two different trends with decreasing γ. The length of the interface grows and the interface thickness decreases. The latter shows that the mean compliance term F using quadratic stiffness interpolation favors structures without interface as claimed above. Figure 23 shows how the interface thickness depends on γ. The three points on the left hand side approximately lie on a curve $c\gamma^{0.4}$. It can be observed that the interface thickness is always below the thickness $\varepsilon\pi \approx 0.126$, which is favored by the Ginzburg-Landau energy. The fact that the length of the interface grows is also plausible since the Ginzburg Landau energy approximates the perimeter of the structure. We note that for both kinds of interpolation the value for F in the minimum decreases and the value for E increases when decreasing γ. Moreover, the final designs using linear interpolation are always stiffer than the designs obtained by using a quadratic interpolation, although they are physically not meaningful. In particular the design in Figure 21d exhibits a lower mean compliance than the design in Figure 22d.
In Lemma 6.81 and 6.82 we proved that the mean compliance problem using linear stiffness interpolation and $\gamma = 0$ is strictly convex and well posed. In Figure 21 it seems that the solutions for $\gamma \to 0$ converge to the unique global solution for the problem with $\gamma = 0$. In the literature one often uses such 'unpenalized' solutions (as in Figure 21d) as initial guess for methods which penalize intermediate densities. By this continuation technique one hopes to find global optima, since the initial guess is also a global optimum of some relaxed problem. We refer to [PS98] for the SIMP method and to [All02] for the homogenization method. We also note that the solution in Figure 21d can be interpreted as a design of a sheet with variable thickness, where the thickness is given by the phase field φ (up to rescaling), see e.g. [BK88].

We conclude that using a quadratic stiffness interpolation gives rise to physically meaningful final designs already for large ε, which is not the case for linear stiffness interpolation. This is independent of the numerical method and thus can be seen as an improvement of the model. For linear interpolation distinct phases can be obtained by decreasing ε. We refer to Figure 39, where it can be seen that ε has to be chosen very small in order to obtain a full phase separation. However, it is desirable to choose ε as large as possible, since ε determines the weights of the convex and concave parts of the Ginzburg-Landau energy. The larger ε is, the larger is the weight of the convex part, which behaves good in the numerics, see also the experiments in Section 6.13.5.

6.13.3 Mesh independency and h-nested iteration

In this section we investigate how the VMPT method applied to the discrete problem depends on the mesh parameter h. First of all we show the mesh independency of the iteration numbers, as well as of the residual $\sqrt{\gamma\varepsilon}\|\nabla \boldsymbol{v}_k\|_{L^2}$, the step length α_k and the error $|j(\boldsymbol{\varphi}_k) - j(\boldsymbol{\varphi}^*)|$ during the iteration. Moreover, also the update of λ_k (see (280)) is mesh independent. For the solution of the projection type subproblem we use the PDAS method, which is not mesh independent. However, we show that the number of inner PDAS iterations increases only mildly as the mesh is refined, leading to an efficient method for

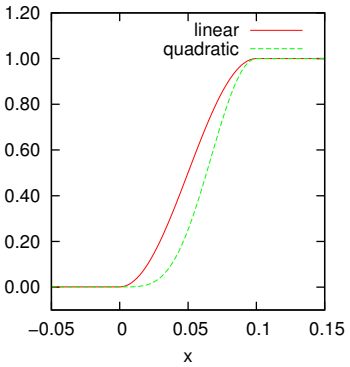
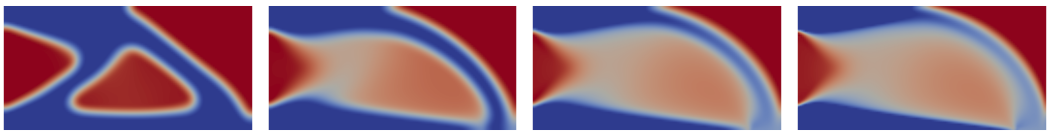


Figure 20: Stiffness $\boldsymbol{C}(\varphi(x))$ of the material across the interface.



(a) $\gamma = 0.5$ (b) $\gamma = 0.1$ (c) $\gamma = 0.02$ (d) $\gamma = 0.004$

Figure 21: Solutions using linear stiffness interpolation. Material in blue.

(a) $\gamma = 0.5$ (b) $\gamma = 0.1$ (c) $\gamma = 0.02$ (d) $\gamma = 0.004$

Figure 22: Solutions using quadratic stiffness interpolation. Material in blue.

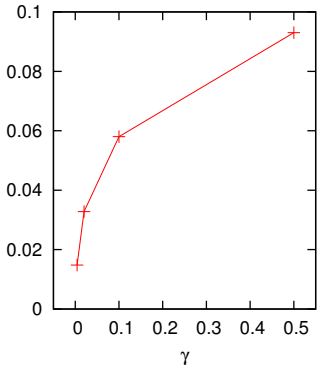


Figure 23: Interface thickness for varying γ using quadratic stiffness interpolation and fixed ε.

moderate mesh sizes. In the second part of the section we show how the computation time of the VMPT method can be reduced by a nested iteration in h. Therefor we start the optimization with a coarse mesh and refine the mesh during the iteration. By this procedure also the number of PDAS iterations can be reduced drastically. The basis of mesh independency is that the method is well defined in the Banach space setting. Thus the iterates of the discrete method can approach the iterates of the continuous method as $h \to 0$. Note that mesh independency is not trivial, which can be seen in Section 6.13.11, where it is shown that the projected L^2-gradient method is not mesh independent, since the corresponding continuous method is not well defined.

Of course mesh independency only holds if h is chosen small enough. The finer the features of the final design are, and the smaller ε is, respectively, the smaller h has to be chosen in order to observe mesh independency. Therefore we choose γ large to get coarser features in the final design, which can already be resolved by coarse meshes. We also choose ε quite large to obtain a broad interfacial region which can be resolved by a coarse mesh.
We show mesh independency of the VMPT method using the H^1-metric, the H^1-BFGS metric and the metric including second order information. In most experiments we use the cantilever beam setup from Example 6.83, quadratic stiffness interpolation and 2 phases. However, to exclude that mesh independency holds only for these special choices, we also include various experiments using the bridge setup (Example 6.84), linear stiffness interpolation and 3 phases.

In all experiments Q-linear convergence of $j(\boldsymbol{\varphi}_k)$ to $j(\boldsymbol{\varphi}^*)$ is observed, where $\boldsymbol{\varphi}^*$ denotes the local minimizer computed by the VMPT method. This can be shown under certain regularity assumptions for the projected gradient method in Hilbert space [Dun81] and also for variable metric projected gradient methods in Hilbert space if finitely many constraints are present [GD88]. We use this result to compute an approximation of $j(\boldsymbol{\varphi}^*)$ by extrapolation. Since this is done for the discrete method, the assumption of finitely many constraints for a problem in a Hilbert space is fulfilled. Let $e_k := |j(\boldsymbol{\varphi}_k) - j(\boldsymbol{\varphi}^*)|$ for all $k \in \mathbb{N}$. From the linear convergence rate we get

$$\frac{e_k}{e_{k-1}} \approx C$$

for some $0 < C < 1$ and for $k \geq k_0$ with k_0 large enough, thus

$$e_k \approx e_{k_0} C^{k-k_0}.$$

Taking the logarithm on both sides leads to

$$\log e_k \approx \log e_{k_0} - k_0 \log C + k \log C. \tag{289}$$

Thus $\log e_k$ is affine linear in k with slope $\log C$. Let iteration numbers $n < m < o$ be given. We use $j(\boldsymbol{\varphi}_n)$ and $j(\boldsymbol{\varphi}_m)$ to calculate the slope and use $j(\boldsymbol{\varphi}_o)$ as an approximation for $j(\boldsymbol{\varphi}^*)$, i.e. in (289) we set $k := o$, $k_0 := m$ and use $j(\boldsymbol{\varphi}^*) \approx j(\boldsymbol{\varphi}_o)$ on the right hand side.

This gives the extrapolation

$$\begin{aligned} j(\varphi^*) &\approx j(\varphi_o) - 10^{(o-m)\frac{\log(j(\varphi_m)-j(\varphi_o))-\log(j(\varphi_n)-j(\varphi_o))}{m-n}+\log(j(\varphi_m)-j(\varphi_o))} \\ &= j(\varphi_o) - (j(\varphi_m)-j(\varphi_o))\left(\frac{j(\varphi_m)-j(\varphi_o)}{j(\varphi_n)-j(\varphi_o)}\right)^{\frac{o-m}{m-n}}. \end{aligned}$$

Here we also used that $j(\varphi_k)$ is decreasing monotonically and thus $j(\varphi_k)-j(\varphi^*)>0$. For the index o the last iterate of the VMPT method is used, and n and m have to be chosen near the end of the iteration. The described extrapolation proves to be very reliable in practice.

In the following experiments we set $\varepsilon = 0.04$ and $\gamma = 0.5$. Moreover, $\mathfrak{m} = 0$, i.e. 50% material and 50% void is prescribed. As initial guess we use the homogeneous mixture $\varphi_0 \equiv \mathfrak{m}$ and use $tol = 10^{-5}$ for the stopping criterion.

We start by showing mesh independency for the H^1-gradient method with λ_k updated as in (280) with $\lambda_0 = 2$, $\lambda_{min} = 10^{-10}$ and $\lambda_{max} = 10^{10}$. Therefor we perform the cantilever beam experiment for various mesh sizes, ranging from a coarse mesh with $h = 2^{-4}$ to a fine mesh with $h = 2^{-8}$. In Figure 24 certain values and parameters of the method are shown in dependence of the iteration number k. In Figure 24a the value of the cost functional $j(\varphi_k)$ is plotted. The different lines can hardly be distinguished except for the coarsest mesh where the discretization error is rather high. For a better insight we also plot the error $|j(\varphi_k)-j(\varphi^*)|$ on a logarithmic scale in Figure 24b, where the extrapolation described above is used to estimate $j(\varphi^*)$. It can be observed that the cost functional values converge fastest for the first (coarsest) mesh. For the second mesh the convergence is slower, but gets better when the mesh is further refined. The curves for the last two meshes are almost identical, which shows that $j_h(\varphi_k^h)$ converges as $h \to 0$. In Figure 24c we examine the residual $\sqrt{\gamma\varepsilon}\|\nabla v_k\|_{L^2}$. The same qualitative behavior as for the cost functional can be observed. The convergence is best for the coarsest mesh, for the finer mesh the method takes longer and the curves for the two finest meshes are almost identical. Thus, also $\sqrt{\gamma\varepsilon}\|\nabla v_k^h\|_{L^2}$ converges as $h \to 0$. The step length α_k which is chosen by the Armijo rule is depicted in Figure 24d. Here also the curves for the different meshes are not distinguishable. For the first few iterations up to three backtracking steps are necessary, whereas from iteration 50 on only up to one backtracking step is necessary. We conclude that the Armijo rule is mesh independent. Figure 24e shows the scaling parameter λ_k chosen by the update scheme (280). We see that also the proposed update scheme is independent of the mesh parameter h. At the beginning of the iteration λ_k is decreased, then again increased until the maximal value of $\lambda_k \approx 250$ is reached. Then it is decreased again until iteration 50, from where on $\lambda_k \approx 80$ is chosen. Table 4 shows the number of VMPT iterations needed to reach the given tolerance. It can be observed that the number of iterations converges to about 265 as $h \to 0$.
Thus the whole outer loop of the VMPT method is mesh independent. The inner loop, i.e. the iterative solution of the projection type subproblem, is implemented here using the PDAS method as described in Section 6.10. As already mentioned, this method is not well defined on the continuous level and thus mesh dependent behavior can be expected. This is confirmed in Figure 24f, where the number of PDAS iterations for the solution of the kth projection type subproblem is shown. One observes that the finer the mesh is the more PDAS iterations are needed. Another observation is that on a fixed mesh, more PDAS iterations are needed in the beginning of the iteration, since the control changes

drastically within the iterations. In advanced iterations the change in the control gets less which entails that the active set of the previous iterate is a good initial guess for the active set of the current iterate and thus less PDAS iterations are needed. This behavior is qualitatively the same for all meshes. We emphasize that the mesh dependence is only moderate. Except for the first few iterations the number of PDAS iterations stays below 10 for all mesh sizes. From iteration 120 on only 1–2 PDAS iterations are needed for all mesh sizes. Thus the PDAS method is still a good solver for the subproblem in spite of the slight mesh dependency. In the simulations later we will see that the high number of PDAS iterations in the beginning can be overcome by using a nested approach in the mesh parameter h.

h	iterations
2^{-4}	111
2^{-5}	407
2^{-6}	320
2^{-7}	275
2^{-8}	269

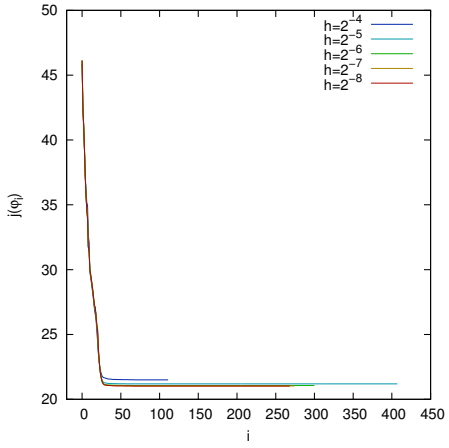
Table 4: Iteration numbers for H^1-gradient method (cantilever beam experiment, cf. Figure 24).

We perform the same experiment using the bridge setup (Example 6.84). Figure 25 shows the result. The same mesh independent behavior can be observed as in the previous experiment. For the bridge setup the mesh independency is even more evident. Already for the coarse mesh with $h = 2^{-6}$ the VMPT method behaves the same as on the finest mesh with $h = 2^{-8}$. It is astonishing that $h = 2^{-6}$, which corresponds to about 3–4 mesh points across the interface, is already sufficient to resolve the movement of the interface. For evolution equations involving phase fields one usually need much more points, e.g. 6 in [BNS04], 7 in [BE93], 8 in [BBG11] or even 15 in [ES03]. Since we don't solve an evolution equation here, but are only interested in a stationary state, less points across the interface are sufficient. We note that also the parameter α_k and λ_k (not depicted) are mesh independent as in the previous experiment. The number of VMPT iterations needed to reach the given tolerance is shown in Table 5. One observes that the number of iterations converges to about 75 as $h \to 0$. The current experiment shows that the method performs well also on very coarse meshes, which can be used to save degrees of freedom and thus computation time.

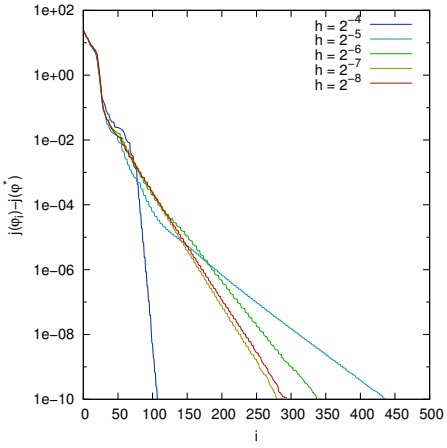
h	iterations
2^{-5}	44
2^{-6}	80
2^{-7}	78
2^{-8}	76

Table 5: Iteration numbers for H^1-gradient method (bridge experiment, cf. Figure 25).

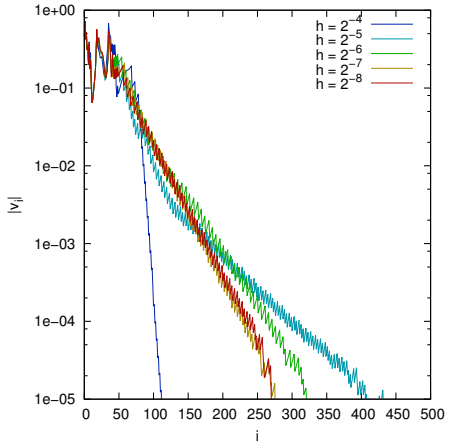
For the next cantilever beam experiment we consider a different metric for the VMPT method. We use the BFGS update of the scaled H^1-inner product and we set $\lambda_{max} = 1$, which corresponds to the full BFGS step. Moreover, we start with $\lambda_0 = 0.001$. Because of the excellent performance of the BFGS method we are able to consider an additional



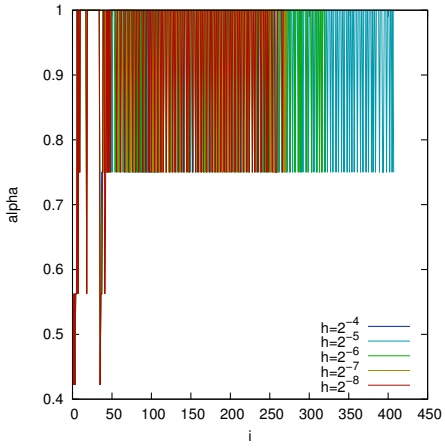
(a) Cost functional.



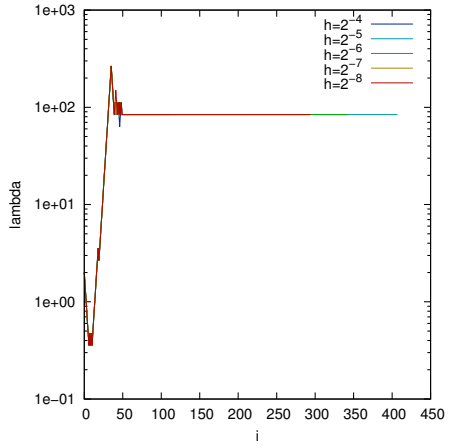
(b) Error in cost functional.



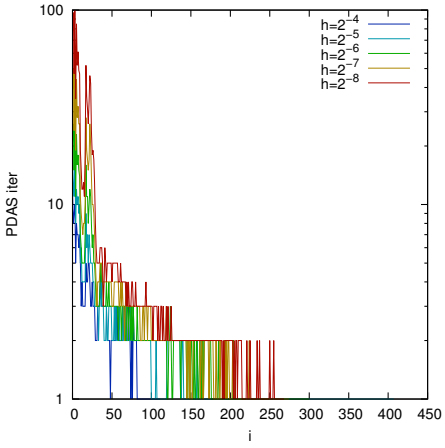
(c) Residual.



(d) Step length α_k.

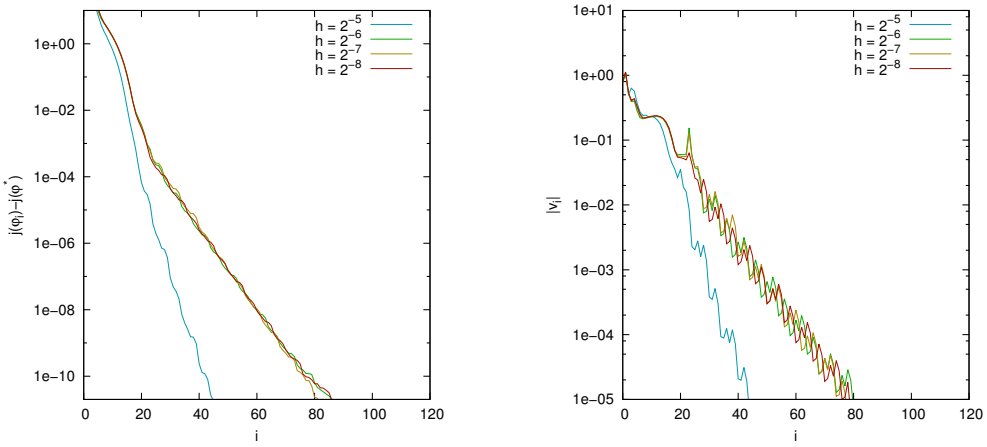


(e) Scaling parameter λ_k.



(f) PDAS iterations in the inner loop.

Figure 24: Cantilever beam experiment using H^1-gradient method for different equidistant meshes (quadratic stiffness, λ_k updated).



(a) Error in cost functional.

(b) Residual.

Figure 25: Bridge experiment using H^1-gradient method for different equidistant meshes. (quadratic stiffness, λ_k updated).

refinement level with $h = 2^{-9}$ and $5 \cdot 10^5$ degrees of freedom. The results are depicted in Figure 26. For all mesh sizes the H^1-BFGS method behaves the same, even on the coarsest mesh with $h = 2^{-5}$, which corresponds to 2–3 mesh points across the interface. Except for the first VMPT step, $\alpha_k = 1$ is accepted for all meshes (not depicted). Thus λ_k is increased until λ_{max} is reached. The figures in Table 6 indicate that 85 H^1-BFGS iterations are needed as $h \to 0$. Again the inner PDAS iteration is mesh dependent.

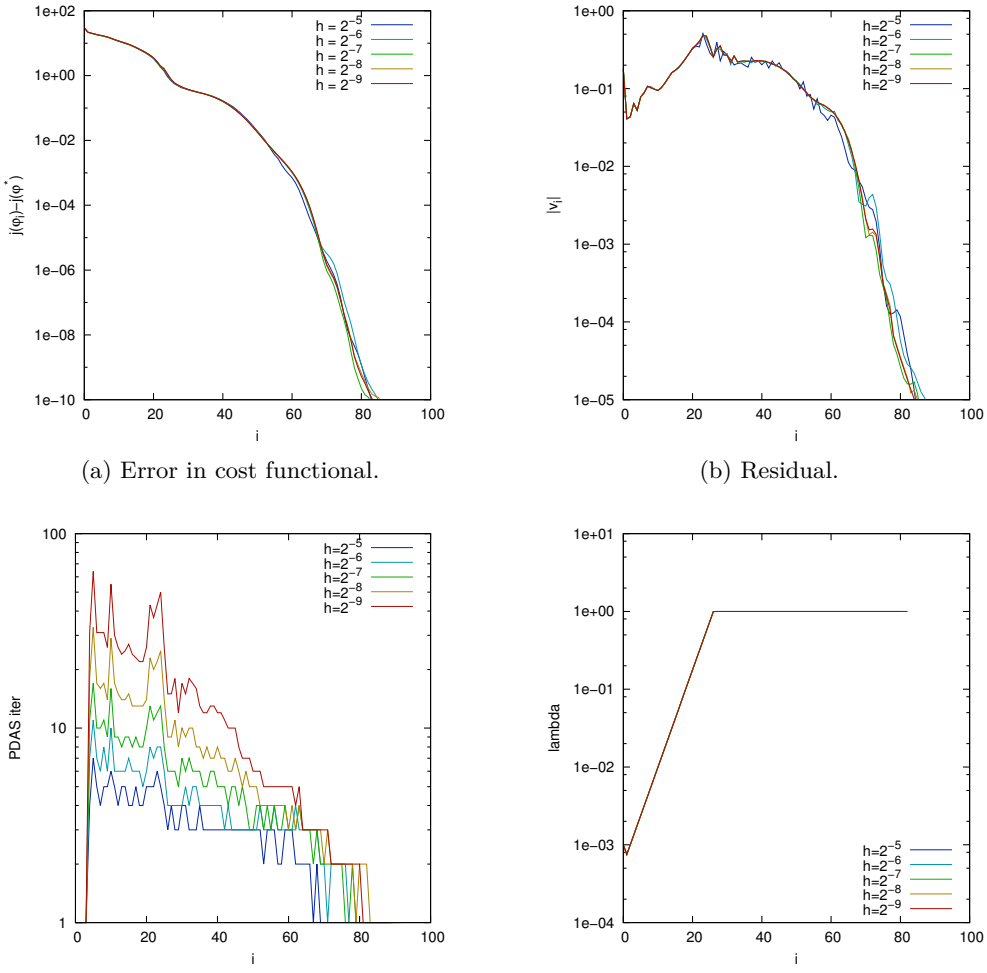
h	iterations
2^{-5}	85
2^{-6}	88
2^{-7}	86
2^{-8}	85
2^{-9}	85

Table 6: Iteration numbers for H^1-BFGS method (cf. Figure 26).

As a third choice for the metric a_k we consider the second order VMPT method. Again we use the update scheme for λ_k starting with $\lambda_0 = 0.3$ and setting $\lambda_{max} = 1$. The result is shown in Figure 27. It can be seen that the error in the cost functional as well as the residual is almost the same for the three finest meshes. We note that the step length $\alpha_k = 1$ is always accepted (not depicted) and thus λ_k is increased until λ_{max} is reached. The inner PDAS iteration is mesh dependent. The VMPT method on the finest mesh already breaks down at a residual of 10^{-4} because of approximation errors.

We also include an experiment using linear interpolation of the stiffness tensors. Since the H^1-gradient method takes more iterations compared to quadratic interpolation (cf. Section 6.13.2), we increase ε to $\varepsilon = 0.05$. Moreover, we set $\lambda_{min} = \lambda_{max} = 1$. The result in Figure 28 shows mesh independent behavior.

Finally we present a cantilever beam experiment with 3 phases, namely a stiff material, a more elastic material and void. We choose the Lamé constants $\lambda = \mu = 5000$ for the



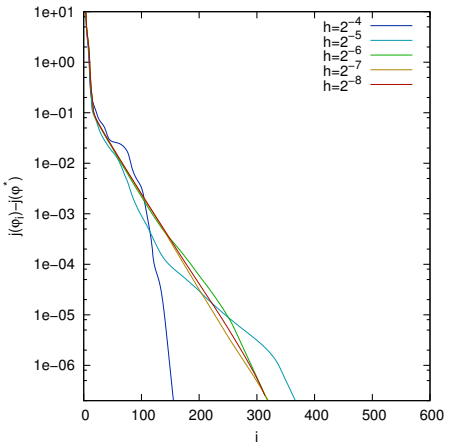
(a) Error in cost functional.

(b) Residual.

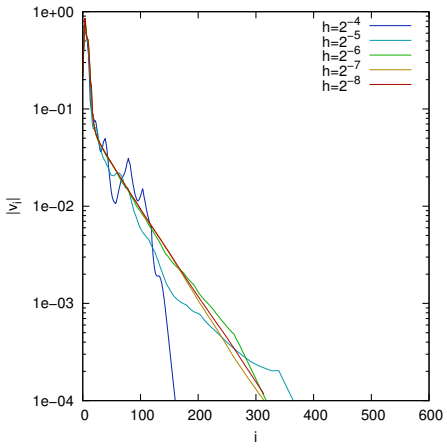
(c) PDAS iterations in the inner loop.

(d) Scaling parameter λ_k.

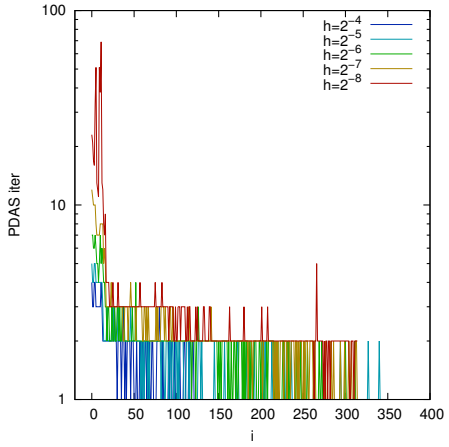
Figure 26: Cantilever beam experiment using H^1-BFGS method for different equidistant meshes. (quadratic stiffness, λ_k updated with $\lambda_{max} = 1$).



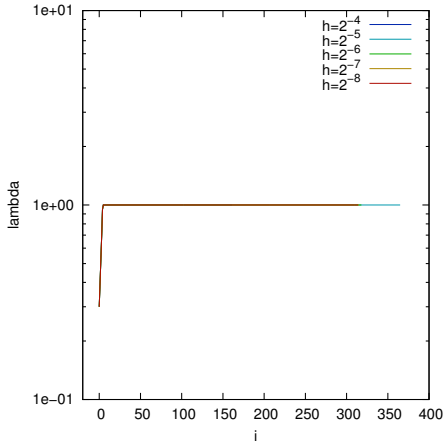
(a) Error in cost functional.



(b) Residual.

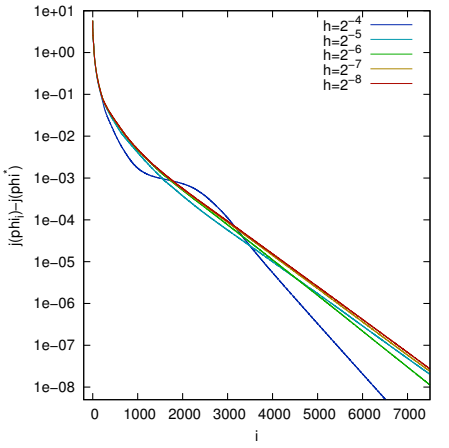


(c) PDAS iterations in the inner loop.

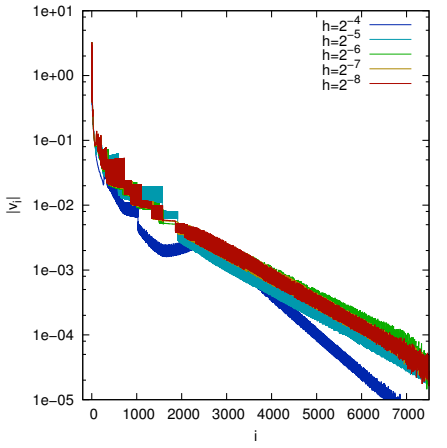


(d) Scaling parameter λ_k.

Figure 27: Cantilever beam experiment using the second order VMPT method for different equidistant meshes. (quadratic stiffness, λ_k updated).



(a) Error in cost functional.



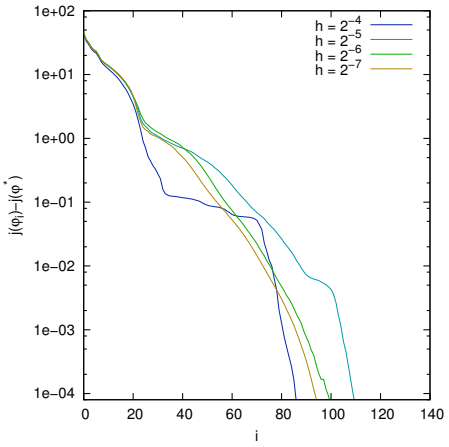
(b) Residual.

Figure 28: Cantilever beam experiment using H^1-gradient method for different equidistant meshes (linear stiffness, $\lambda = 1$, $\varepsilon = 0.05$).

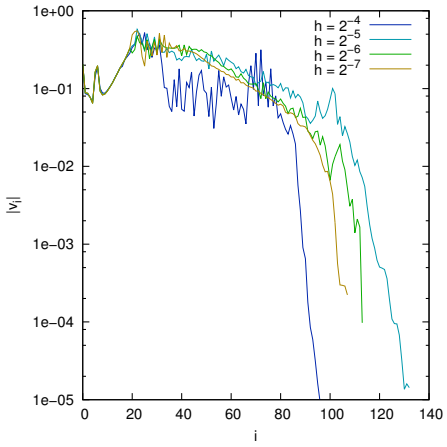
stiff material, $\lambda = \mu = 2500$ for the elastic material and $\lambda = \mu = 10$ for the void phase. The respective mass fraction is given by $\mathfrak{m} = (0.3, 0.2, 0.5)^T$. Since we now are in the vector-valued setting, we have to double the value of γ, i.e. we set $\gamma = 1$, see Section 6.1.3. The results using the H^1-BFGS method are shown in Figure 29. It can be observed that the development of the error in the cost functional is almost the same for the finest two meshes. Also the residual behaves similar. Note that only on the coarsest mesh the given tolerance $tol = 10^{-5}$ is reached. For the finer meshes the method breaks down earlier due to approximation errors. The finer the mesh, the earlier this breakdown occurs. On the finest mesh the final residual is $2 \cdot 10^{-4}$. Also note that a phase field describing 3 phases has 3 times the degrees of freedom of a scalar valued phase field. Also the number of PDAS iterations is much higher than for the 2-phase cantilever beam. In fact a lot of fine tuning was necessary in order for the PDAS method to converge at all on the finest mesh. We had to increase the accuracy for the linear solver during the PDAS iteration and we also had to disable the damping described in Section 6.10.2 at the end of the PDAS iteration. Without disabling the damping the Newton residual would be 10^{-8} after few iterations and then the damping starts to produce very small step sizes such that the PDAS method does not converge within 200 iterations. However, without damping the PDAS method converges within few steps. Also note that it is not possible to stop the PDAS iteration at a Newton residual of 10^{-8}, since the iterates of the PDAS method are unfeasible and the VMPT method could thus break down, see Section 4.9. As in the previous experiments the initial scaling $\lambda_0 = 0.001$ is increased until the full step $\lambda_{max} = 1$ is reached.

Performing a nested iteration in the mesh parameter h has two advantages. First of all a local minimum on a coarse mesh can be obtained quickly since less degrees of freedom have to be determined. The solution on the coarse mesh can then be used as an initial guess for the optimization process on the finer mesh. Since we start with the homogeneous mixture $\varphi_0 \equiv \mathfrak{m}$, it takes some iterations for the interface to form. During this process the control typically consists only of low frequencies in the Fourier space, which can already be resolved on a coarse mesh. In the previous experiments this can be observed in the residual plots. In the beginning of the iteration the curves are identical and they begin to fork as soon as the interface is formed. Thus it is uneconomic to use a fine mesh in the beginning of the iteration. We also note that the cost of an iteration at the beginning is higher since more PDAS iterations are needed as can be seen in the plots. Moreover, we solve the PDAS system only for degrees of freedom on the interface. Since whole Ω consists of interface in the beginning, the PDAS system is larger, which makes the initial iterations even more expensive. Thus the expensive iterations on a fine mesh in the beginning are replaced by cheap iterations on a coarse mesh, which saves computation time considerably. The second advantage of a nesting in h is that the mesh dependency of the inner PDAS iteration can be controlled thereby as can be seen in the following experiments.

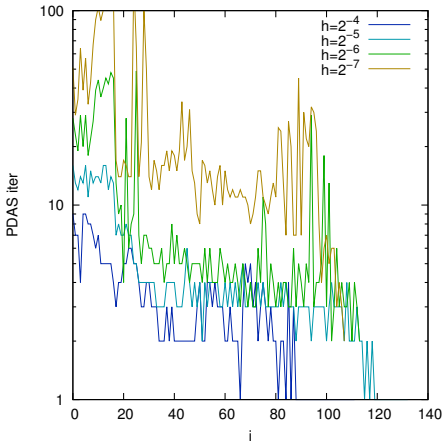
First, we consider the H^1-gradient method for the 2-phase cantilever beam with parameters as above. We start with a coarse mesh with $h = 2^{-4}$ and compute a solution up to a tolerance of $tol = 10^{-2}$. Then we refine the mesh and decrease the tolerance as shown in Table 7. The tolerance on the finest mesh is $tol = 3 \cdot 10^{-5}$, since for lower tolerances the method breaks down. In the last line of Table 7 the computation time and iteration count for the mesh $h = 2^{-8}$ without nesting is given for comparison. We observe that the total computation time is only 16% of the unnested iteration. However, the total number of iterations stays the same. We also see that the main computational effort is on the finest mesh. The residual and the number of inner PDAS iterations is shown in Figure



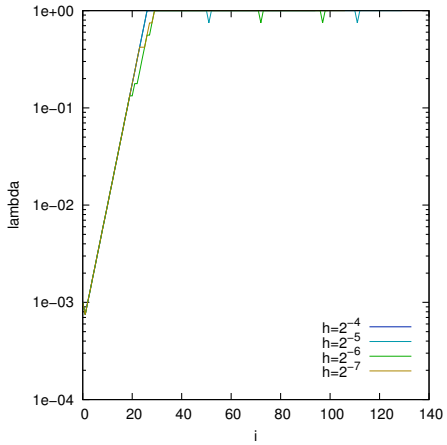
(a) Error in cost functional.



(b) Residual.

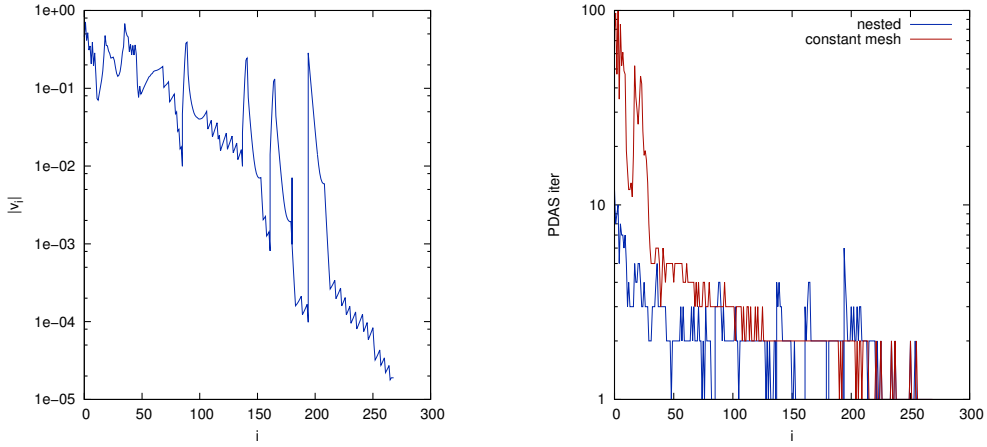


(c) PDAS iterations in the inner loop.



(d) Scaling parameter λ_k.

Figure 29: 3-phase cantilever beam experiment using the H^1-BFGS method for different equidistant meshes. (quadratic stiffness, λ_k updated).



(a) Residual.

(b) PDAS iterations in the inner loop.

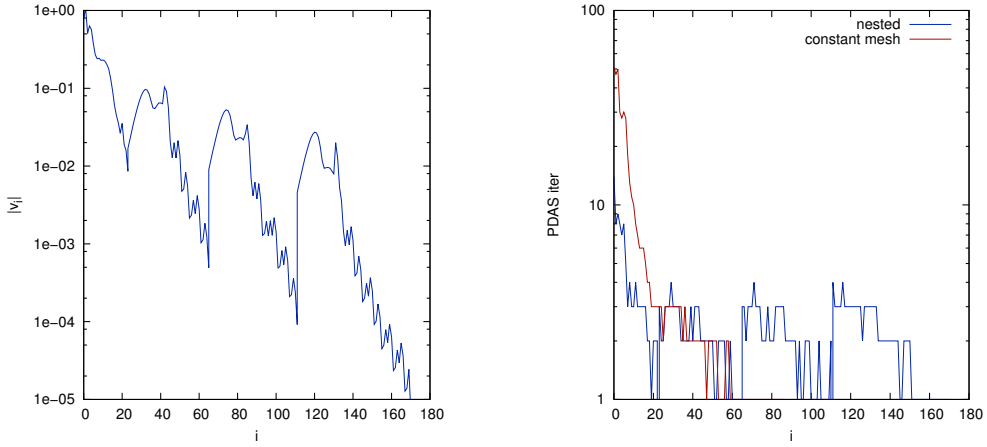
Figure 30: h-nested iteration for cantilever beam.

30. It can be observed that the residual decreases until the tolerance of the current level is reached. Then it jumps up as soon as the mesh is refined and again decreases until the next tolerance is reached. In Figure 30b also the number of PDAS iterations for the unnested case is shown for comparison. Clearly, the number of PDAS iterations is reduced drastically in the beginning. For the unnested iteration up to 109 PDAS iterations are needed, whereas in the nested iteration the PDAS iterations always stays below 10. After the first few iterations the number of PDAS iterations even stays below 5. Thus, the increased number of PDAS iterations, which comes from the mesh dependency, can be reduced by performing a nested iteration in h.

Level	h	DOFs	tol	CPU	iterations
0	2^{-4}	561	10^{-2}	4s	85
1	2^{-5}	2145	10^{-2}	7s	52
2	2^{-6}	8385	10^{-3}	14s	24
3	2^{-7}	33153	10^{-4}	111s	33
4	2^{-8}	131841	$3 \cdot 10^{-5}$	25m	63
			total:	28m	257
			unnested:	2h 57m	259

Table 7: Nested iteration in h for H^1-gradient method (cantilever beam). See also Figure 30.

We perform the same nested iteration using the bridge setup. The results are shown in Table 8 and Figure 31, respectively. In contrast to the cantilever beam experiment the total number of iterations for the nested approach is much higher than in the unnested case. We suppose that this is due to higher discretization errors, i.e. the solution for $h = 2^{-7}$ is still quite far away from the solution for $h = 2^{-8}$. However, the computation time can nevertheless be reduced to 58%. We note that the tolerances on the respective meshes have to be chosen adequately in order to obtain an efficient method. Since most of the computation time is spent on the finest mesh it is preferable to take a lower tolerance on the coarser meshes. Then the total number of iterations probably increases, but the computation time may still decrease if less work has to be done on the finest mesh. As in



(a) Residual.

(b) PDAS iterations in the inner loop.

Figure 31: h-nested iteration for bridge experiment.

the previous experiment the number of PDAS iterations can be reduced drastically in the beginning. Except for the first few iterations, the number of PDAS iterations stays below 4.

Level	h	DOFs	tol	CPU	iterations
0	2^{-5}	2145	10^{-2}	4s	23
1	2^{-6}	8385	10^{-3}	34s	42
2	2^{-7}	33153	10^{-4}	3m 4s	46
3	2^{-8}	131841	10^{-5}	24m 17s	59
			total:	28m	170
			unnested:	48m	76

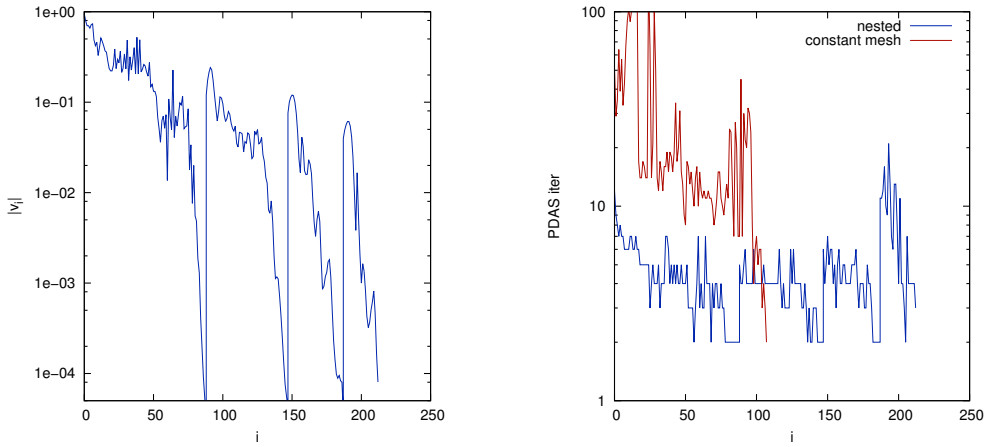
Table 8: Nested iteration in h for H^1-gradient method (bridge). See also Figure 31

We finally show how the h-nesting can be used to increase the performance of the VMPT method for the 3 phase cantilever beam. The parameters are the same as for the unnested experiment above. Table 9 shows the result. Note that this time we also choose a low tolerance on the coarse meshes to reduce the iterations needed on the fine meshes. Thereby the iterations needed on each mesh decrease per level, see the last column of Table 9. On the finest mesh only 25 iterations are needed. By choosing low tolerances the total number of iterations increases from 107 without nesting to 212 with nesting. However, the overall computation time decreases drastically. Only 1.2% of the original computation time is needed when a nested approach is used. On the finest mesh the method breaks down at a residual of $8 \cdot 10^{-5}$. Recall that for the unnested simulation this happened at a residual of $2 \cdot 10^{-4}$. In Figure 32 the development of the residual and the number of PDAS iterations is shown. As in the previous experiments the number of PDAS iterations is reduced drastically. However, when switching to the finest mesh up to 20 PDAS iterations are needed in the first few VMPT steps.

Also for other experiments the h-nesting performs very well. Especially when the number of needed optimization steps is high, e.g. for low values of γ and ε, the efficiency of the method is enhanced considerably.

Level	h	DOFs	tol	CPU	iterations
0	2^{-4}	1683	$5 \cdot 10^{-5}$	1m 26s	88
1	2^{-5}	6435	$5 \cdot 10^{-5}$	2m 44s	59
2	2^{-6}	25155	$5 \cdot 10^{-5}$	12m 29s	40
3	2^{-7}	99459	$8 \cdot 10^{-5}$	1h 49m	25
			total:	2h 5m	212
			unnested:	168h 31m	107

Table 9: Nested iteration in h for the H^1-BFGS method (3 phases). See also Figure 32.



(a) Residual.

(b) PDAS iterations in the inner loop.

Figure 32: h-nested iteration for 3 phase cantilever beam using the H^1-BFGS method.

6.13.4 Comparison of inner products

In this section we compare the different inner products we use for the VMPT method, i.e. we compare the projected H^1-gradient method with different choices of λ_k, the H^1-BFGS method and the second order VMPT method. We want to compare iteration numbers, CPU time and the quality of the obtained minimizer of the respective methods.

We first show how we can improve the VMPT method by a good choice of the stiffness interpolation and the choice of the scaling parameter λ_k. As already seen in Section 6.13.2, the VMPT method works much better when interpolating $\boldsymbol{C}(\boldsymbol{\varphi})$ quadratically rather than linearly. Especially in combination with the developed update scheme for λ_k the VMPT is very efficient. This can also be seen by the following experiment. We use the cantilever beam setup with 2 phases, $\varepsilon = 0.04$ and $\gamma = 0.5$ and perform a H^1-gradient iteration using linear interpolation of $\boldsymbol{C}(\varphi)$ and $\lambda_k = 1$. We repeat the experiment using quadratic interpolation of $\boldsymbol{C}(\varphi)$ and updating λ_k according to (280). The results for various equidistant meshes can be seen in the first 6 columns of Table 10. It can be observed that the number of iterations is reduced drastically from 19000 iterations for the first experiment to 270 for the second experiment. This can also be seen in terms of computation time. On the finest mesh the first experiment took 23 days whereas the second experiment is finished after 3 hours. By performing a nested iteration in h as described in the previous section, the CPU time can be further reduced to 28 minutes (see last row of Table 10), which is 0.08% of the original computation time. In the last two columns of Table 10 the results of the H^1-BFGS method are listed. The iteration number

can be further reduced to 85 (unnested), whereas the CPU time increases. The reason for the higher CPU time is that we use the MINRES solver for the linear systems arising in the PDAS method. When using the H^1-gradient method, the matrix in the linear system can be explicitly assembled and thus we are able to use a direct solver, which is much faster than the MINRES solver, especially at the beginning of the optimization when the interface is not yet formed. Moreover, in the BFGS method a recursion has to be performed to evaluate the BFGS operator. However, this is negligible compared to the higher computational cost when using the MINRES solver instead of a direct solver. By a nested iteration in h the computation time of the H^1-BFGS method can be reduced to 9 minutes, which is much less than the unnested CPU time since the expensive iterations in the beginning of the optimization are performed on a coarse mesh. We thus see that an adequate choice of the stiffness tensor interpolation, the scaling λ_k and the inner product a_k has a large impact on the VMPT method.
Table 11 shows the preceding comparison for the bridge setup instead of the cantilever beam setup. The same qualitative behavior can be observed. However, the difference is less drastic. The computation time can be reduced from 12 hours to 28 minutes.

		$\boldsymbol{C}(\varphi)$ lin., $\lambda = 1$, H^1-gradient		$\boldsymbol{C}(\varphi)$ quadr., λ upd., H^1-gradient		$\boldsymbol{C}(\varphi)$ quadr., λ upd., H^1-BFGS	
h	DOFs	CPU	iterations	CPU	iterations	CPU	iterations
2^{-4}	561	12m	9956	5s	112	—	—
2^{-5}	2145	2h 25m	14590	1m	408	25s	85
2^{-6}	8385	20h 40m	16936	4m	321	2m 34s	88
2^{-7}	33153	3d 20h 28m	19416	21m	276	21m 35s	86
2^{-8}	131841	23d 15h 0m	18891	3h	270	3h 48m	85
			nested:	28m	257	9m	123

Table 10: Comparison of the methods for the cantilever beam.

		$\boldsymbol{C}(\varphi)$ lin., $\lambda = 1$, H^1-gradient		$\boldsymbol{C}(\varphi)$ quadr., λ upd., H^1-gradient	
h	DOFs	CPU	iterations	CPU	iterations
2^{-5}	2145	2m 24s	662	8s	44
2^{-6}	8385	15m 25s	796	1m 11s	80
2^{-7}	33153	1h 32m	832	6m 22s	78
2^{-8}	131841	11h 53m	851	47m 48s	76
			nested:	28m	170

Table 11: Comparison of the methods for the bridge.

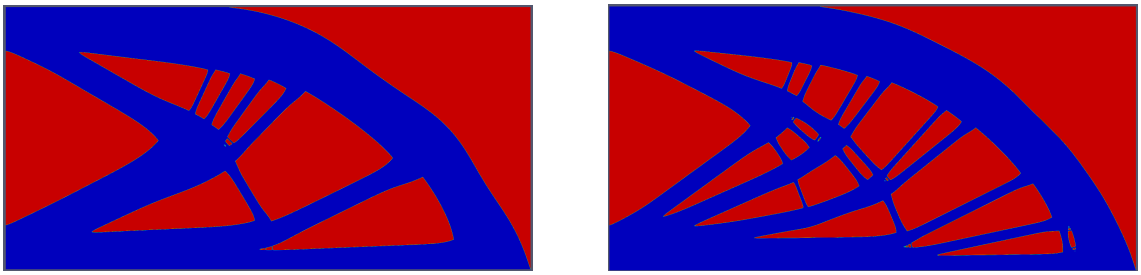
Next we compare the projected H^1-gradient method together with the update scheme for λ_k to the H^1-BFGS method. In both cases we use a quadratic interpolation of the stiffness tensors. Therefor we perform three experiments. We take the 2-phase cantilever beam setup from above, once with $\gamma = 0.5$ and once with $\gamma = 0.002$. Afterwards we switch to a 3-phase cantilever beam setup with $\gamma = 0.5$, where the masses and stiffness tensors are as in the experiment corresponding to Figure 29. The results for a h-nested iteration are shown in Table 12. We use a final mesh of $h = 2^{-8}$ for the first experiment and $h = 2^{-7}$ for the other two experiments. As already seen above, for 2 phases and $\gamma = 0.5$ the H^1-

BFGS method takes 48% of the iteration number and 32% of the CPU time compared to the H^1-gradient method. The difference between the methods becomes larger when γ is reduced to $\gamma = 0.002$. The H^1- BFGS method then takes 21% of the iteration numbers and 23% of the CPU time. This is plausible since for smaller γ the weight of the H^1-regularization term in the cost functional is reduced, thus both methods take longer. However, the secant information used by the BFGS method becomes more important the smaller γ is and thus the loss in efficiency is less for the BFGS method. As seen in Section 6.13.3, the computation time of an h-nested iteration rather depends on the number of iterations on the finest mesh than on the total number of iterations. Here we note that the number of iterations on the finest mesh is much higher for $\gamma = 0.002$ than for $\gamma = 0.5$, where the solution on the coarse mesh is already a good approximation. On the other hand, a single iteration for $\gamma = 0.002$ on the finest mesh is quite cheap, since we use only $h = 2^{-7}$ for the finest mesh, and moreover the interface is thinner for $\gamma = 0.002$ (cf. Figure 22), leading to even less mesh points on the interface and therefor less degrees of freedom in the Newton system, which is only solved on the interface (cf. Section 6.10). We also note that the BFGS method converged to another local minimizer than the gradient method for $\gamma = 0.002$.
The result for 3 phases is surprising at first glance. The BFGS method takes only 27% of the iteration numbers, but takes 338% of the CPU time. The reason is again that we have to use the MINRES solver for the BFGS method, whereas a direct solver can be used for the gradient method. To investigate this more closely we perform the same H^1-gradient iteration using MINRES instead of the direct solver. A single VMPT iteration is then in average about 26 times as expensive on the finest mesh, leading to an estimated CPU time of 11h instead of 37m, which is shown in parenthesis in Table 12. Moreover, we experience that the gradient method using the MINRES solver breaks down earlier than the method using a direct solver. Thus the direct solver is not only faster, but gives even higher accuracy solutions of the linear systems than the MINRES solver. Here, it would be reasonable to introduce a preconditioner to enhance the performance of the MINRES solver.
We conclude that the number of iterations is lower for the H^1-BFGS method, especially for low γ. The computation time in most cases is also lower, but can be higher depending on the used linear solver. We remark that for the compliant mechanism functional the BFGS method is even 100 times faster than the gradient method, see Section 6.14.

	2 phases				3 phases	
	$\gamma = 0.5$		$\gamma = 0.002$		$\gamma = 0.5$	
	CPU	iterations	CPU	iterations	CPU	iterations
H^1-BFGS	9m	123	1h 25m	1610	2h 5m	212
H^1-gradient	28m	257	6h 3m	7603	37m (11h)	776

Table 12: h-nested iteration for the cantilever beam.

To compare the second order VMPT method with the other methods we first look at our standard 2-phase cantilever beam setup with $\gamma = 0.5$, $\varepsilon = 0.04$ and $h = 2^{-8}$. In Figure 27 we see that the second order VMPT method needs a little bit more iterations than the H^1-gradient method. However, the needed CPU time is much higher. The second order method takes 26.5 hours whereas the H^1-gradient method is finished after 3 hours. This is plausible since in the second order method the linearized state equation has to be solved in each inner MINRES iteration. Because a single iteration of the second order method is



(a) projected H^1-gradient method (b) second order VMPT method

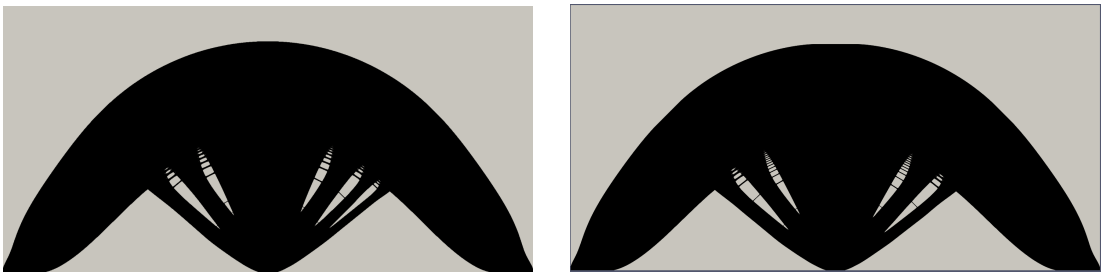
Figure 33: Local minima obtained by different methods for the 2-phase cantilever beam.

much more expensive than an iteration of the H^1-gradient method, the former is efficient only if much less iterations are needed.
Efficiency is not the only criterion one should use to evaluate a numerical method. Since the considered optimization problem is not convex many different local minima can appear and the quality of the obtained minimizer should also be a criterion. In fact we experience that the cost functional value of a minimizer obtained by the second order method is in the majority of cases lower than the cost of the minimizer obtained by the other methods. As an example we use the 2-phase cantilever beam experiment with $\gamma = 0.002$ and $\varepsilon = 0.001$. Note that for small γ the optimal structure becomes finer and thus more local minima are possible. We nest the iteration in h, where we use an adaptive mesh on the finest level with $h_{max} = 2^{-6}$ in the bulk and $h_{min} = 2^{-10}$ on the interface. Moreover we nest the iteration in ε. The final designs using the H^1-gradient method and the second order VMPT method are depicted in Figure 33. We observe that the perimeter of the latter solution is higher than the perimeter of the former solution. In Table 13 the corresponding figures are given. In this experiment the H^1-gradient method took much more iterations and even more CPU time. The final cost functional value $j(\varphi^*)$ is about 0.53% lower for the second order VMPT method. Also the contributions of the Ginzburg-Landau energy $E(\varphi^*)$ and the compliance $F(\varphi^*)$ to the final cost are given in Table 13. It can be observed that the Ginzburg-Landau energy of the second order solution is higher, whereas the compliance is lower. Due to the small weight γ of the Ginzburg-Landau energy the total cost is lower. This is plausible, since the H^1-gradient method uses the H^1-inner product, which contains second order information of the Ginzburg-Landau energy only, whereas the second order VMPT method uses an inner product which additionally contains second order information of the compliance. Thus the H^1- gradient method tends to converge to local minimizers with low Ginzburg-Landau energy, i.e. less perimeter in the limit $\varepsilon \to 0$.
To compare the second order VMPT method also to the H^1-BFGS method we compute

	iterations	CPU time	$j(\varphi^*)$	$F(\varphi^*)$	$E(\varphi^*)$
H^1-gradient	11189	42h 12min	15.07	15.03	20.79
second order VMPT	851	19h	14.99	14.93	30.12

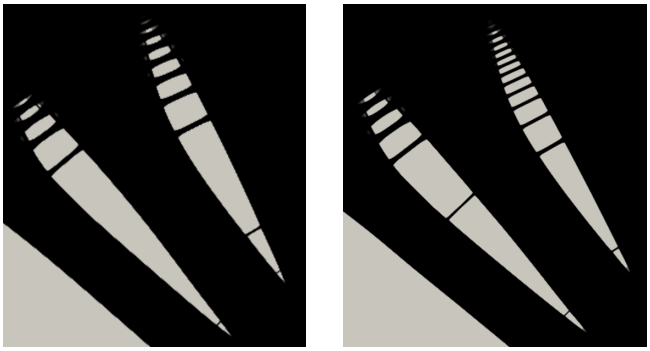
Table 13: Data for the local minima in Figure 33.

local minima of the 2-phase bridge setup with $\gamma = 0.002$ and $\varepsilon = 0.001$ using both methods. We again nest in h and ε and take $h_{max} = 2^{-5}$ and $h_{min} = 2^{-10}$ on the finest level. The results are given in Figure 34 and Table 14. We first of all note that both solutions are not symmetric, although the second order solution looks quite symmetric. It can be observed that the BFGS solution has an additional hole on the right hand side of the middle part.



(a) H^1-BFGS (b) Second order VMPT

Figure 34: Local minima obtained by different methods for the 2-phase bridge.



(a) H^1-BFGS (b) Second order VMPT

Figure 35: Details of the two holes on the left hand side of Figure 34.

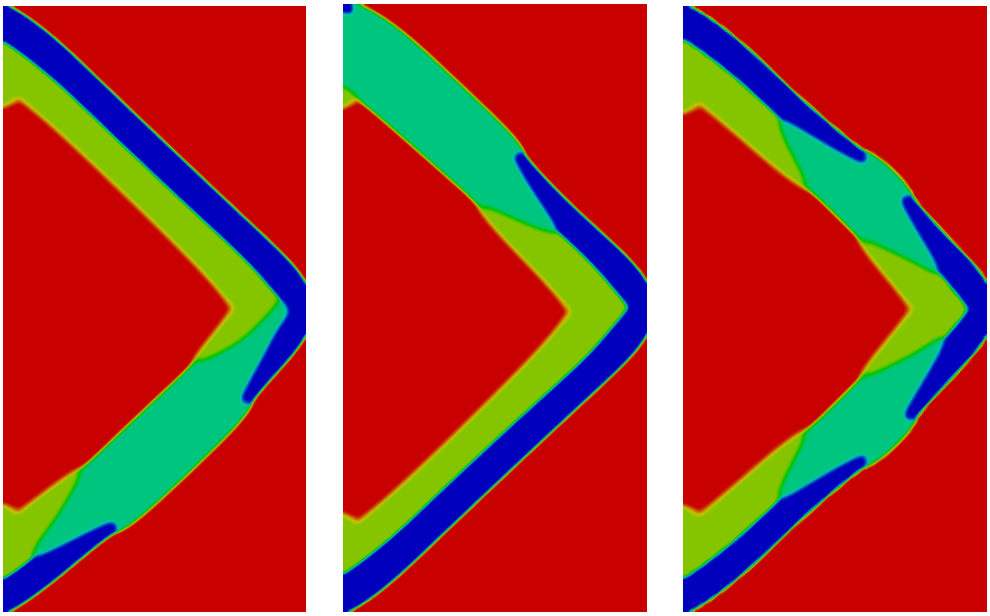
Nevertheless, the perimeter of the second order solution is higher, since finer structures in the holes are present, see Figure 35. In Table 14 it can be seen that the second order method needs less iterations than the BFGS method. However, the computation time is much higher. As in the experiment above we get out that the total cost functional value for the second order solution is lower (about 0.07%), the compliance is lower and the Ginzburg-Landau energy is higher.

	iterations	CPU time	$j(\varphi^*)$	$F(\varphi^*)$	$E(\varphi^*)$
H^1-BFGS	7799	27h	45.81	45.77	20.37
second order VMPT	2512	4d 7h	45.78	45.74	21.33

Table 14: Data for the local minima in Figure 34.

Finally we consider an example with multiple phases. We take the 4-phase short cantilever beam setup described in [WW04]. We use $\Omega = (0,1) \times (-1,1)$, $\Gamma_D = \{x_1 = 0\}$, $\Gamma_g = \{x_1 = 1\} \cap \{|x_2| \leq 1/20\}$, $\boldsymbol{g} \equiv (0,-160)^T$ and $\boldsymbol{f} \equiv \mathbf{0}$. For the distinct materials we take the Lamé constants $\mu = \lambda = 5000$, $\mu = \lambda = 2500$, $\mu = \lambda = 1250$ and $\mu = \lambda = 5$. The last material approximates a void phase. The prescribed volume fraction is $\mathfrak{m} = (0.1, 0.1, 0.1, 0.7)$. Here we take a very low value for the weight of the Ginzburg-Landau energy $\gamma = 2.5 \cdot 10^{-9}$. Since the width of the interface decreases with γ when using quadratic stiffness interpolation (cf. Figure 22 and the corresponding discussion), we have to compensate this by taking a larger value for $\varepsilon = 5000$ to get a reasonable interface thickness. We take an adaptive mesh with $h_{max} = 1/80$ and $h_{min} = 1/320$. Note that our coarse mesh corresponds to 80×160 mesh points, whereas in [WW04] only 27×62 mesh points are used. In [ADDM14]

they use 80×160 mesh points, which is also lower, since in addition we refine the mesh on the interface. We perform an h-nested iteration using the H^1-BFGS method, the H^1-gradient method and the second order VMPT method. However, since the latter is very expensive, we use the second order metric only on the coarsest mesh (410 iterations on a 40×80 mesh) to determine the topology, and then switch to the cheaper BFGS method. The final designs for $tol = 4 \cdot 10^{-5}$ are depicted in Figure 36a and 36b, respectively. We compare the results also to the following fourth variant. Because of symmetry in the used short cantilever beam setup, the iterates of the VMPT method stay symmetric if the initial guess $\boldsymbol{\varphi}_0$ and the inner product a_k is chosen symmetric. Here, the symmetry axis is the $\{x_2 = 0\}$-axis and it holds $j(\boldsymbol{\varphi}(x_1, x_2)) = j(\boldsymbol{\varphi}(x_1, -x_2))$ as well as $\boldsymbol{u}(x_1, x_2) = \tilde{\boldsymbol{u}}(x_1, -x_2)$, where $\boldsymbol{u} = S(\boldsymbol{\varphi}(x_1, x_2))$ and $\tilde{\boldsymbol{u}} = S(\boldsymbol{\varphi}(x_1, -x_2))$. For the discrete VMPT method the iterates are symmetric only if the chosen mesh also exhibits this symmetry. Usually we use an unsymmetric mesh where the hypotenuse of the triangles is oriented from the bottom left to the top right, cf. Figure 7. Here we use now a symmetric mesh by considering 'crossed' diagonals. Thus the H^1-BFGS method generates symmetric iterates and hence also the final design is symmetric, which is depicted in Figure 36c. First of all we observe that all obtained local minima have approximately the same void phase distribution. Only the material distribution within the beam differs. Note that because of symmetry, $\boldsymbol{\varphi}(x_1, -x_2)$ is a solution whenever $\boldsymbol{\varphi}$ is a solution, thus solutions 36a and 36b differ practically only in one half of the beam. In [WW04] it is stated that the solution of the problem is analytically known (in case of a truss structure and only a single material), consisting of two beams at an angle of 45° to the wall, which can be observed for all three minima. It is interesting that no fine structures are present in the final designs although the parameter γ is chosen very small. Thus we suppose that the limit problem for $\gamma \to 0$ exists and is well defined, and the sharp interface problem may have a solution also without perimeter penalization. This is not always the case, cf. Figure 22. Moreover, it can be seen that the material-void boundary is not straight, but has some curvature, in particular in points where two material phases meet. A possible explanation could be the angle condition (286) for triple junctions in the sharp interface limit. However, in this experiment the weight of the potential term ψ_0 in the cost functional is very low ($\gamma/\varepsilon = 5 \cdot 10^{-13}$) and thus it is more likely that the curved boundary contributes to the stiffness of the structure. We note that the symmetric solution is not stable. When applying a small unsymmetric perturbation, the H^1-BFGS method converges to the minimum in Figure 36a. Thus we can assume that the symmetric solution is only a local minimum if the problem is restricted to symmetric designs, but it is a saddle point if also unsymmetric admissible designs are considered. This is no contradiction to the global convergence result (Theorem 4.14), since it is only stated that accumulation points of the VMPT method are stationary points and thus also saddle points are possible limits. The figures for a quantitative comparison of the considered methods can be found in Table 15. As before, the computation time for the second order VMPT method is higher than for the H^1-BFGS method. Here, also the number of iterations is higher. The highest number of iterations is needed by the H^1-gradient method, which takes even more CPU time than the second order VMPT method. The symmetric method is much faster in terms of computation time, since the used special symmetric mesh has only half of the mesh points compared to the mesh used by the other three methods. For all four methods the iteration numbers are not very high compared to the number of iterations needed by pseudo time stepping methods, which usually need far more than 10000 iterations even for large values of γ. Thus the VMPT method works well also for a very small regularization parameter γ. When comparing the total cost of the minima, we see that the second order VMPT method provides the



(a) H^1-BFGS / H^1-gradient (b) Second order VMPT (c) symmetric solution

Figure 36: Different stationary points for the 4-phase cantilever beam. The design on the right hand side is a saddle point. From elastic to stiff: red, yellow, green, blue.

best solution, being 2.4% better than the BFGS solution. As already mentioned, the symmetric solution is only a saddle point with higher cost than the BFGS solution. For the second order solution the compliance is better than for the BFGS solution as before, but here even the Ginzburg-Landau energy is smaller. One notices that the Ginzburg-Landau energy is very high for all three solutions. Essentially this comes from the gradient term. The contribution of the potential term $\int 1/\varepsilon\psi_0$ in the Ginzburg-Landau energy is only of magnitude 10^{-6}. The reason therefor is that the interface thickness of the final designs is much lower than the optimal thickness favored by the Ginzburg-Landau energy. For $\varepsilon = 5000$ the optimal thickness would be of magnitude 10^4, cf. Figure 11 (which shows the optimal transition for $\varepsilon = 1$). However, in the final designs the phase field $\boldsymbol{\varphi}$ is much steeper and thus the gradient term is very high. The thin interface is due to the quadratic stiffness interpolation and the low value for γ, cf. Section 6.13.2.
This experiment also shows that it is not always advantageous to restrict the optimization to symmetric designs, since unsymmetric designs can have lower compliances. However, there are also applications where symmetric solutions are preferred.

	iterations	CPU time	$j(\boldsymbol{\varphi}^*)$	$F(\varphi^*)$	$E(\boldsymbol{\varphi}^*)$
H^1-BFGS	839	9h 17m	0.23101	0.22505	$2.3860 \cdot 10^6$
H^1-gradient	3318	21h 55m	”	”	”
second order VMPT	1110	21h 18m	0.23049	0.22479	$2.2815 \cdot 10^6$
symmetric	550	1h 18m	0.23344	0.22726	$2.4708 \cdot 10^6$

Table 15: Data for the stationary points in Figure 36.

Also in other experiments not shown here the second order VMPT method converged to better solutions than the H^1-BFGS method or the projected H^1-gradient method. Only in a single experiment the solution of the BFGS method was better.

Finally, we briefly comment on the correlation between second order VMPT method and SQP method. We concentrate on the case that $\boldsymbol{C}$ is linear in $\boldsymbol{\varphi}$ and $\boldsymbol{f}$ as well as $\boldsymbol{g}$ are independent of $\boldsymbol{\varphi}$. In Section 6.7 we have already seen that in this case the second order metric a_k coincides with the Hessian $j''(\varphi_k)$ up to the potential term $\frac{\gamma}{\varepsilon}\psi_0(\boldsymbol{\varphi})$. Thus, we can expect that the methods behave similarly for small γ and large ε. In Lemma 6.81 we proved that the topology optimization problem is strictly convex in case of $\gamma = 0$. However, we cannot use $\gamma = 0$ in the second order VMPT method, since the H^1-coercivity of the second order metric a_k is needed for global convergence. Note that in the case $\gamma = 0$ the second order VMPT method would indeed coincide with the SQP method, since the potential term would vanish. For the numerical experiment we use the small value $\gamma = 0.0002$. We take a variant of the cantilever beam experiment and refer to the description of Figure 46 for the geometric setup. We take $\varepsilon = 0.2$, $\mathfrak{m} = 0.5$ (25% material) and an equidistant mesh with $h = 2^{-5}$. The second order VMPT method converges within 11 iterations up to a residual of 10^{-7}. We note that the final design consists mainly of interface, similar to Figure 21d, because of linear interpolation of the stiffness tensors and a rather high value of ε. The residuals $r_k := \sqrt{\gamma\varepsilon}\|\nabla \boldsymbol{v}_k\|_{L^2}$ are listed in Table 16. Superlinear convergence of the (discrete) second order VMPT method can be observed. Note that in other experiments one observes only linear convergence (cf. Figure 27), since the second order metric is only a good approximation of the Hessian with the data used here.

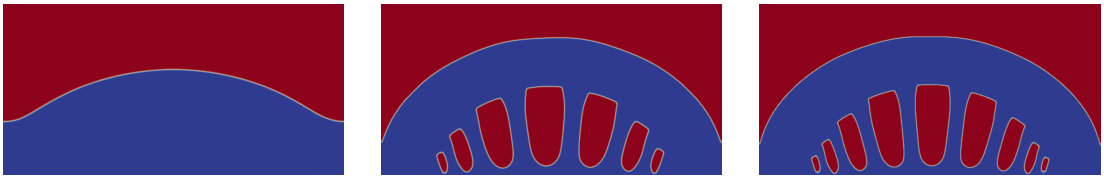
k	r_k	r_k/r_{k-1}
0	$4.51 \cdot 10^{-2}$	
1	$5.90 \cdot 10^{-2}$	$1.31 \cdot 10^{0}$
2	$6.21 \cdot 10^{-2}$	$1.05 \cdot 10^{0}$
3	$2.27 \cdot 10^{-2}$	$3.65 \cdot 10^{-1}$
4	$5.72 \cdot 10^{-3}$	$2.52 \cdot 10^{-1}$
5	$4.43 \cdot 10^{-3}$	$7.74 \cdot 10^{-1}$
6	$3.84 \cdot 10^{-3}$	$8.67 \cdot 10^{-1}$
7	$2.49 \cdot 10^{-3}$	$6.48 \cdot 10^{-1}$
8	$1.39 \cdot 10^{-3}$	$5.56 \cdot 10^{-1}$
9	$3.57 \cdot 10^{-4}$	$2.58 \cdot 10^{-1}$
10	$2.03 \cdot 10^{-5}$	$5.67 \cdot 10^{-2}$
11	$8.14 \cdot 10^{-8}$	$4.02 \cdot 10^{-3}$

Table 16: Superlinear convergence of the second order VMPT method.

6.13.5 Dependency on γ and ε

In this section we discuss how the numerical method as well as the shape of local minimizers of the topology optimization problem depend on the model parameters γ and ε.

The parameter γ is the weight of the Ginzburg-Landau energy and — in the limit $\varepsilon \to 0$ — the weight of the perimeter of the structure. Thus, we expect that for lower values of γ the length of the interface is larger, if the stiffness of the structure can thereby be enhanced. This can be observed in the experiment corresponding to Figure 22, where the number of holes in the cantilever beam increases as γ decreases. We confirm this behavior by the following modified bridge experiment: We take the parameters of Example 6.84, except that we change the traction to $\boldsymbol{g} \equiv (0, -50)^T$ on the whole bottom boundary $\Gamma_g = \{x_2 = 0\} \setminus \Gamma_D$. We take $m = 0$, i.e. 50% material. The optimal designs computed

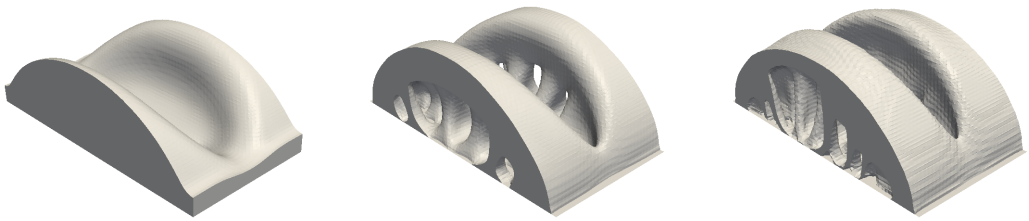


(a) $\gamma = 0.5$ (b) $\gamma = 0.005$ (c) $\gamma = 0.0025$

Figure 37: Local minima for a bridge setup for different γ.

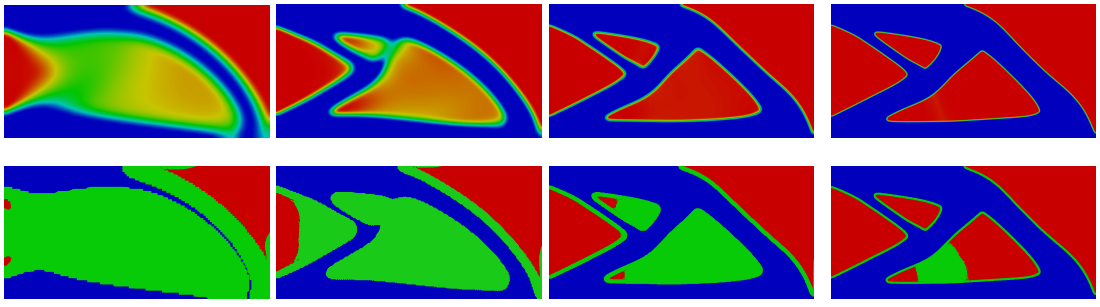
by the H^1-BFGS method for different γ and fixed $\varepsilon = 0.005$ are depicted in Figure 37. For $\gamma = 0.5$ no holes are present in the structure, whereas for $\gamma = 0.005$ there are 7 holes, and 9 holes for $\gamma = 0.0025$. This is the same behavior as for the cantilever beam. We also perform a 3D version of the bridge experiment: We take $\Omega = (-1,1) \times (0,1) \times (0,1)$ and extend all data constantly in x_3-direction, i.e. Γ_D, Γ_g and $\boldsymbol{g}$. The computed local minima for 25% material are depicted in Figure 38. In this experiment it can again be observed that the structure gets finer as γ decreases. This can hardly be seen for the last two structures, but for $\gamma = 0.1$ there are up to two rows of supports on each side, whereas for $\gamma = 0.01$ there are up to three rows. The same behavior of increasing perimeter is also observed in the literature. For a hard perimeter inequality constraint see e.g. [Jog02], for a penalization of an interface energy e.g. [YINT10, TP13] and for a phase field model e.g. [WR12].
However, the perimeter grows only if the compliance of the structure can be lowered thereby. As an example we refer to the experiment corresponding to Figure 36, where the very small value $\gamma = 2.5 \cdot 10^{-9}$ is used and nonetheless the final structures are quite coarse. Another thing to keep in mind is that the previous consideration is only true if ε is chosen small enough. When using a quadratic interpolation of $\boldsymbol{C}(\varphi)$, the growth of the perimeter can be already observed for quite large values of ε. However, if the stiffness tensors are interpolated linearly and γ is decreased for fixed ε, then one does not observe that the perimeter gets larger, but rather that the interfacial region grows as can be seen in Figure 21. In this case one would have to decrease γ and ε simultaneously to see the increase of the perimeter while keeping the phase field structure of φ. As an example consider the cantilever beam setup using the linear interpolation of the stiffness tensors with $\gamma = 0.1$ and varying ε. The computed minima are shown in Figure 39. One observes that for $\varepsilon = 0.04$ almost the whole structure consists of interface. For decreasing ε the mixed phase begins to separate, but only slowly. Even for the very small value of $\varepsilon = 0.005$ there is still a part of Ω where the phases are not separated. Thus the parameter ε has to be chosen very small in order to get the desired phase field structure consisting of pure phases, which are separated by an interface of thickness proportional to ε. Using a quadratic stiffness interpolation, the phases already demix for large values of ε. For instance $\varepsilon = 0.04$ is usually sufficiently small, cf. Figure 22. Note that the sharp interface problem does not depend on the type of interpolation used, thus the obtained shapes should in both cases be similar for ε small enough. However, for positive ε the local minimizers using the quadratic stiffness interpolation are a better approximation of the sharp interface solution.

The parameter ε is part of the phase field model. Its purpose is to control the width of the interface. From Γ-convergence theory of the Ginzburg-Landau energy with obstacle potential one gets that the interface width decreases linearly in ε in the limit $\varepsilon \to 0$ [BE91b]. This can be well observed e.g. in Figure 39, where the width of the interface on the right hand side of the beam is roughly halved when ε is halved.



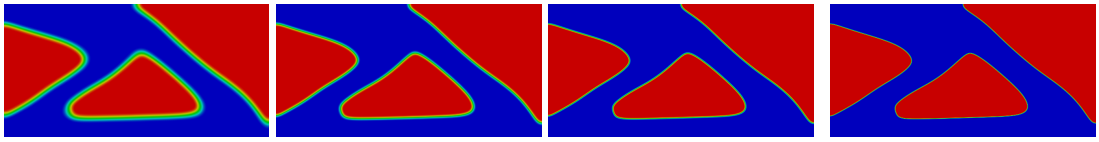
(a) $\gamma = 1$ (b) $\gamma = 0.1$ (c) $\gamma = 0.01$

Figure 38: Local minima for a 3D bridge setup for different γ. Shown are the level sets $\{\varphi \leq 0\}$.



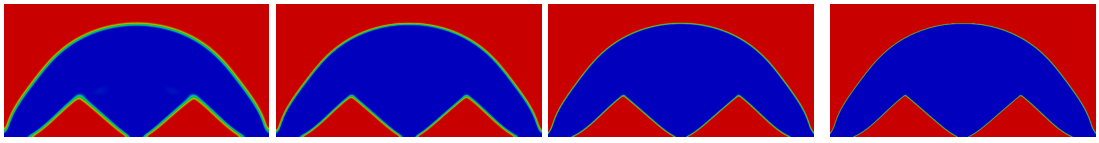
(a) $\varepsilon = 0.04$ (b) $\varepsilon = 0.02$ (c) $\varepsilon = 0.01$ (d) $\varepsilon = 0.005$

Figure 39: Cantilever beam using linear stiffness interpolation for different ε and fixed $\gamma = 0.1$. The top row shows a continuous coloring, the bottom row shows the sets $\{\varphi = -1\}$ in blue, $\{\varphi = 1\}$ in red and $\{-1 < \varphi < 1\}$ in green.



(a) $\varepsilon = 0.03$ (b) $\varepsilon = 0.015$ (c) $\varepsilon = 0.0075$ (d) $\varepsilon = 0.00375$

Figure 40: Cantilever beam using quadratic stiffness interpolation for different ε and fixed $\gamma = 0.5$.



(a) $\varepsilon = 0.04$ (b) $\varepsilon = 0.02$ (c) $\varepsilon = 0.01$ (d) $\varepsilon = 0.005$

Figure 41: Bridge using quadratic stiffness interpolation for different ε and fixed $\gamma = 0.5$.

The transition $\varepsilon \to 0$ for quadratic stiffness interpolation using the cantilever beam and bridge setup is shown in Figure 40 and Figure 41. Also here it can be observed that the interface thickness is roughly halved when ε is halved. In Figure 42 we plot the interface thickness as a function of ε for the cantilever beam experiment. It can be clearly seen that for $\varepsilon \to 0$ the thickness approaches $\pi\varepsilon$, which is the expected interface thickness for the standard obstacle potential [BE91b]. For both experiments the zero level set $\{\varphi = 0\}$ is even for large ε a good approximation of the sharp interface for $\varepsilon \to 0$. The same cantilever beam experiment can be found in [BGHR15], where it is numerically shown that the error $\|\varphi_\varepsilon - \varphi_0\|_{L^1}$, with φ_0 being the L^1-limit of $(\varphi_\varepsilon)_\varepsilon$, is dominated by the diffusion of the sharp interface rather than by its position. As $\varepsilon \to 0$, also the cost functional values $j_\varepsilon(\varphi_\varepsilon)$ as well as the Lagrange multipliers λ_ε for the mass constraint converge, see [BGHR15] for details.
We again emphasize that the thickness of the interface does not only depend on ε, but also on γ. When using quadratic stiffness interpolation the interface thickness decreases monotonically in γ, see Figure 23.

The zero level set of φ_ε is not always a good approximation of the sharp interface. To illustrate this we perform the following MBB (Messerschmitt-Bölkow-Blohm) beam

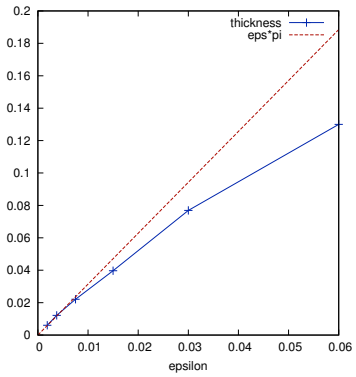
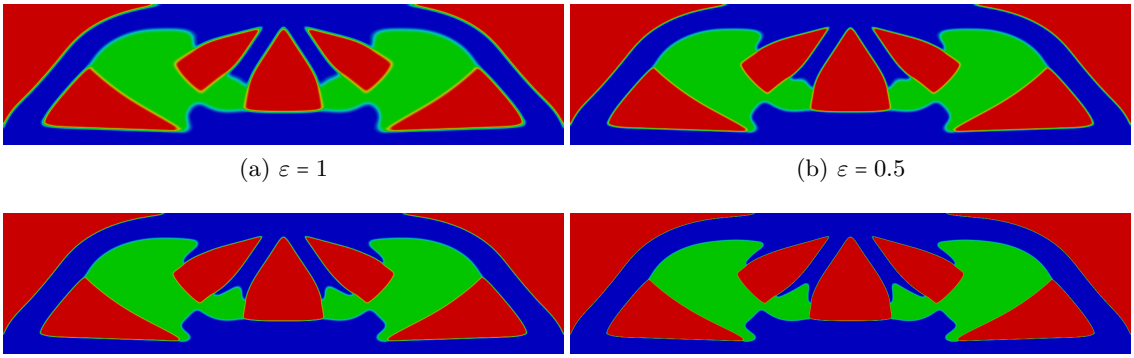


Figure 42: Interface thickness depending on ε for the cantilever beam with fixed $\gamma = 0.5$ and the reference line $\pi\varepsilon$.



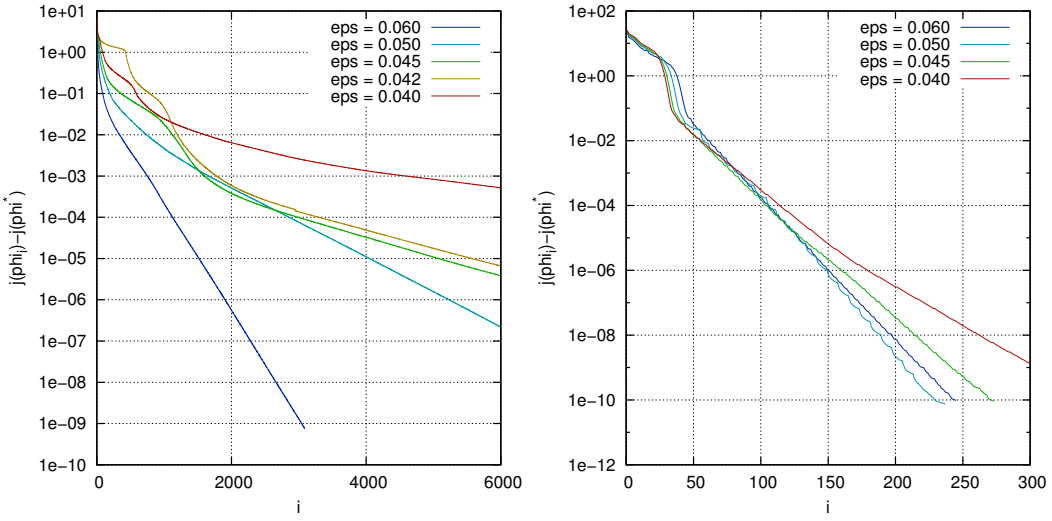
(a) $\varepsilon = 1$ (b) $\varepsilon = 0.5$

(c) $\varepsilon = 0.25$ (d) $\varepsilon = 0.125$

Figure 43: MBB beam for different ε. Strong material in blue, weak material in green, void in red.

experiment. We take $\Omega = (-96, 96) \times (0, 48)$ with $\Gamma_D = \{-96 \leq x_1 \leq -94\} \cap \{x_2 = 0\}$, $\Gamma_g = \{-2 \leq x_1 \leq 2\} \cap \{x_2 = 48\}$, $\boldsymbol{g} \equiv (0, -0.1)^T$ and $\boldsymbol{f} \equiv \boldsymbol{0}$. Moreover, we impose the boundary condition $u_2 = 0$ for the second component of $\boldsymbol{u}$ in the set $\Gamma_{D2} = \{94 \leq x_1 \leq 96\} \cap \{x_2 = 0\}$. Since we want to consider only symmetric solutions, we restrict our computations to the right half of Ω, which is common in the literature. Therefor the boundary condition on Γ_D is dropped and another boundary condition $u_1 = 0$ on $\Gamma_{D1} = \{x_1 = 0\} \cap \{0 \leq x_2 \leq 48\}$ is introduced. This is not really equivalent to the original problem, but common practice, cf. [WZ04a]. We note that in this setting Korn's inequality still holds, although no Dirichlet boundary condition for $\boldsymbol{u}$ is present, see Theorem 7.1. We want to distribute three materials within Ω with Lamé constants $\lambda = 1.154$, $\mu = 0.7692$ for the strong material, $\lambda = 0.5769$, $\mu = 0.3846$ for the weak material and $\lambda = 5.77 \cdot 10^{-10}$, $\mu = 3.84 \cdot 20^{-10}$ for the material approximating void. We take the masses $\mathfrak{m} = (0.4, 0.2, 0.4)$ for the distinct materials and choose $\gamma = 1.25 \cdot 10^{-4}$. The local minimizers computed by the H^1-BFGS method for different ε are shown in Figure 43. Here one observes that the shape of the level sets change considerably as $\varepsilon \to 0$. For example consider the interface between the strong and the weak material within the two beams in the middle of the structure. For $\varepsilon = 1$ the interface seems to be quite straight, whereas it is heavily curved for $\varepsilon = 0.125$. A similar observation can be made for the interface between strong and weak material at the bottom of the weak material region. For $\varepsilon = 1$ it has the form of a semicircle, whereas for $\varepsilon = 0.125$ it rather looks like a rounded rectangle. In general we observe that the curvature of the level sets increases in certain areas as $\varepsilon \to 0$. This is plausible, since ε weights the smoothing gradient term in the Ginzburg-Landau energy and thus higher curvatures are possible for lower ε.
We also refer to the compliant mechanism experiment in Figure 75, where the level sets change drastically as $\varepsilon \to 0$.

In the following we investigate how the numerical method depends on the parameter ε. Since the smoothing, convex gradient term in the Ginzburg-Landau energy is weighted by ε, whereas the potential term, which is concave here, is weighted by ε^{-1}, we expect that the convergence of the VMPT method is worse the smaller ε is. For a numerical example we consider the cantilever beam setup with two phases. We vary ε from 0.06 to 0.04 and look at the development of the error $|j(\varphi_k) - j(\varphi^*)|$ in the cost functional using linear stiffness interpolation and $\lambda_k = 1$ (Figure 44a) and quadratic stiffness interpolation and λ_k updated by (280) (Figure 44b). In both cases a_k was chosen as unscaled H_0^1 inner



(a) linear stiffness, $\lambda_k = 1$

(b) quadratic stiffness, λ_k updated

Figure 44: Development of the error in the cost functional for different ε.

product. Our speculation is confirmed by the numerical results. The smaller ε is chosen, the flatter are the error curves. However, there is a huge difference between the two choices for λ_k and the stiffness interpolation. For the choice $\lambda_k = 1$, the convergence of the method depends heavily on ε. For instance the method for $\varepsilon = 0.05$ takes almost three times of the iteration number of the method for $\varepsilon = 0.06$. For the smallest chosen value $\varepsilon = 0.04$ the curve is already very flat. When using the update scheme (280) for λ_k, which scales λ_k as $\mathcal{O}(\varepsilon^{-1})$ (see Figure 10), the dependence of the method on ε is much lower. There is not even a factor of 2 between the iteration numbers for the highest and lowest value of ε. This also confirms that the scaling of λ_k as ε^{-1} is numerically advantageous together with a quadratic interpolation of the stiffness tensors.

During the numerical simulations we experienced that the convergence of the VMPT method is slower for decreasing values of γ. This can again be justified, since in addition to ε also γ weights the smoothing gradient term in the cost functional. However, the convergence speed rather depends on the optimal shape than on the value of γ. In general we noticed the tendency that the convergence is worse the finer the structures in the final design are. This also depends indirectly on γ, since for smaller values of γ finer structures are possible. In Figure 36 however, the final structure is quite coarse, although the used parameter $\gamma = 2.5 \cdot 10^{-9}$ is very small. The convergence of the VMPT method is in this case rather good.

6.13.6 Computation of local minima and comparison to the literature

In this section we present some local minima computed by the VMPT method and compare them to other optimal designs obtained in the literature. Moreover we compare the VMPT method to selected similar methods used in the literature.

First of all we want to mention that it is hard to compare results from different authors, since the exact parameters of the setup are often not given in the literature. Moreover, often point loads or Dirichlet boundary conditions in a single point are used, especially

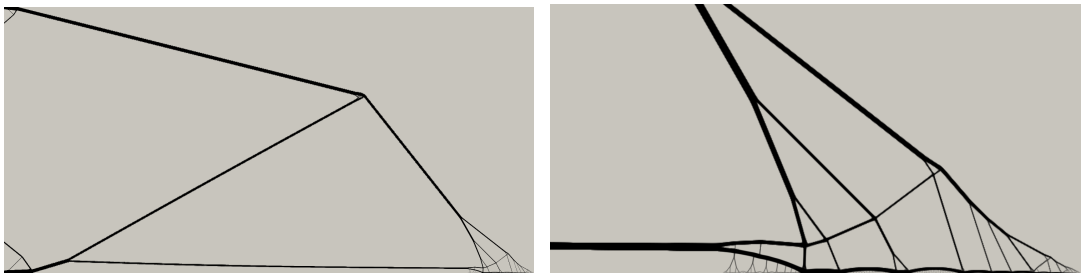


Figure 45: Solution with low volume fraction and zoom of right hand side.

in the engineering community, which is not well defined in a variational sense. Also comparing the numerical method is hard, since often no stopping criteria are used or not mentioned. Comparing computation time is often not meaningful, since it depends on the implementation used, e.g. if the code is parallelized or if the code is written in MATLAB or in C, etc., and of course it also depends on the underlying hardware. The comparison of iteration numbers is only reasonable if the cost of a single step of the two considered methods is the same, see e.g. Section 6.13.8, where the SQP method needs much less steps than the VMPT method, but a single step is much more expensive.

As a proof of concept, we give an example of a topology optimization problem where only few mass is available to distribute in the design container and therefore very fine structures have to appear to gain the desired stiffness of the design. For that purpose we use the cantilever beam setup from Example 6.83 with $\lambda = \mu = 5000$ for the hard material, $\lambda = \mu = 0.1$ for the weak material (approximating void), $\varepsilon = 0.3$ and $\gamma = 5 \cdot 10^{-4}$. We choose $m = 0.95$, corresponding to 2.5% material. The final design computed by the H^1-BFGS method is shown in Figure 45. Also an enlargement of the lower right corner is given, where the structure is very fine. The highly branched structure supports the area Γ_g where the boundary traction acts. In this experiment the final design also depends heavily on the stiffness of the weak material phase as opposed to the experiment using 50% material. Thus we have to choose a very small value for this stiffness here. Note that we didn't perform the optimization until the stopping criterion is fulfilled, but we stopped the iteration earlier, since the movement of the interface becomes very slow. In addition the phases within the branches at the bottom boundary of Ω are not yet well developed and one would have to further reduce ε to obtain a clear phase separation. The final residual when we stop the iteration is $4 \cdot 10^{-2}$. The computation time amounts to 111 hours. This example shows that the H^1-BFGS method is robust and can also handle little material masses. A reasonable structure is produced, which can withstand the given boundary traction. In literature we didn't find experiments resulting in such fine structures, in particular not in the context of phase field models. However, in [PRW12, Figure 13] also branched structures are obtained on a larger length scale by using a phase field model in nonlinear elasticity.

For the cantilever beam example using 50% material, almost the same topology as in our H^1-gradient method solution in Figure 33a is obtained in [WR12, Figure 2c] using a phase field model and a pseudo time stepping method. However, they do not get the solution 33b of our second order VMPT method, which is a lower minimum for the parameters used here. This lower minimum can to our knowledge not be found in the literature. The solution in Figure 40 for higher regularization parameter γ is also obtained in [BGS+12] using the same phase field model and a pseudo time stepping method. By the same pseudo

time stepping method the bridge design in Figure 41 is obtained in [BFGS14]. The same bridge design is also obtained in [XS93, Fig. 8b] using the ESO method.

The short cantilever beam experiment including 4 phases (Figure 36) is also performed in [WW04, Figure 11f.] and [ADDM14, Figure 16] using a level-set method. In both cases the material-void distribution is the same as for our solution, i.e. two bars at an angle of 45° to the wall, but the material distribution within the bars differs considerably. In [WW04] the first solution is similar to our symmetric solution Figure 36c in the sense that the hard material trends to be on the outside boundary and the soft material at the inside of the contact to the wall and at the tip. However, the topology and also the size of the different regions is still different. The second minimum given in [WW04] resembles more our second order VMPT solution (Figure 36b), but it is still quite different. The solution given in [ADDM14] coincides with our symmetric solution (Figure 36c) in the region around the tip. The rest is again quite different. The same outline of the shape is obtained by the ESO method [XS93, Fig. 5e] and the homogenization method [SK91, Fig. 4] using material and void.

The MBB experiment with 3 phases we performed in Figure 43 can also be found in [TM14, Figure 7] and [WZ07, Figure 10], where in both cases a phase field model is used and the first reference corresponds to '$\gamma = 0$', i.e. there is no Ginzburg-Landau energy in the cost functional and therefore they have to do a filtering of the sensitivities in order to get a mesh independent solution. The topology obtained in [TM14] is basically the same as for our solution except for 2 additional holes in the weak material. Since the resolution of the final design in [TM14] is quite low compared to our solution, the delicate curvature of the inner material interface, which we get for low values of ε, cannot be seen. The material-void topology in [WZ07] is exactly the same as for our solution, but again the material distribution within the structure differs much. The numerical method in [TM14] takes 200 iterations to compute the final design, in [WZ07] 120000 iterations (10 hours CPU time) of the time stepping method are needed, and in our computation 485 iterations (1 hour CPU time) of the H^1-BFGS method are performed for $\varepsilon = 1$. It should be mentioned that in [WZ07] the elasticity equation is only solved once every 10 time steps to save computation time. For these methods the following stopping criteria are used: In [TM14] the method is just stopped after 200 iterations, in [WZ07] the method seems to be stopped if the solution visibly doesn't change anymore, and we use our rigorous stopping criterion with $tol = 10^{-6}$. As already mentioned, iteration numbers and CPU times are hard to compare.

As next experiment we consider another variant of the cantilever beam which is widely used in the literature. We change the boundary traction to $\Gamma_g = \{x_1 = 1\} \cap \{|x_2 - 0.5| \leq 1/32\}$, $\boldsymbol{g} \equiv (0, -250)^T$, so the force now doesn't act on the bottom boundary, but in the middle of the right boundary. The other parameters stay unchanged. The solution for 25% material, $\gamma = 0.5$ and $\varepsilon = 0.02$ is shown in Figure 46. The same solution is obtained in [YINT10, Figure 6] for high regularization parameter using a level-set method. In other literature finer structures are obtained, which is due to the absence of the Ginzburg-Landau energy or perimeter penalization in the cost functional. Finer structures (i.e. smaller γ) are compared in the following experiment.

For the so-called long cantilever beam setup the aspect ratio of the design domain Ω is changed to $\Omega = (-2, 2) \times (0, 1)$. The remaining parameters are as in Example 6.83,

Figure 46: Cantilever beam with boundary traction on the right boundary.

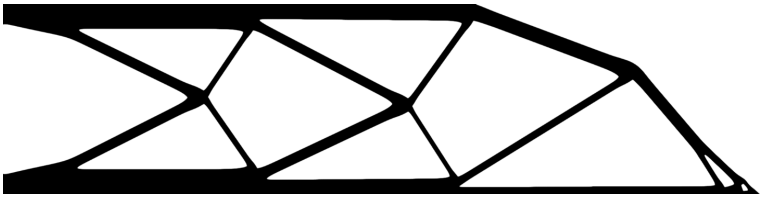


Figure 47: Long cantilever beam.

where $\Gamma_D = \{x_1 = -2\} \cap \{0 \le x_2 \le 1\}$ and $\Gamma_g = \{1.75 \le x_1 \le 2\} \cap \{x_2 = 0\}$ are amended appropriately. The solution for 25% material, $\gamma = 0.17$ and $\varepsilon = 0.002$ is depicted in Figure 47. The final design features the typical truss-like structures, which are obtained by various other models and methods, see e.g. [YINT10, ADDM14, AJT04, AJ05, Jog02, BS03].

We also compare the results for a 3D version of the cantilever beam. We take $\Omega = (-1,1) \times (0,1) \times (0,1)$, $\Gamma_D = \{x_1 = -1\} \cap \{0 \le x_2 \le 1\} \cap \{0 \le x_3 \le 1\}$, $\Gamma_g = \{0.75 \le x_1 \le 1\} \cap \{x_2 = 0\} \cap \{0 \le x_3 \le 1\}$ with $\boldsymbol{g} \equiv (0, -250, 0)^T$ and material constants as in Example 6.83. The solution for 50% mass, $\gamma = 0.5$ and $\varepsilon = 0.038$ is shown in Figure 48. This solution is different to the one obtained in [BFGS14, Figure 8] although the same phase field model is used, but a pseudo time stepping method as numerical method. On the other hand, our solution almost coincides with the SIMP solution presented in [WZ07, Figure 19b] and is also quite similar to the solution in [DJD00, Figure 5]. However, the Cahn-Hilliard solution in [WZ07, Figure 19a] is again different. It is interesting that the Allen-Cahn solution in [BFGS14, Figure 8] also coincides with the Cahn-Hilliard solution in [WZ07, Figure 19a]. Perhaps these are simply two different local minima.

We consider again the 2D cantilever beam (Example 6.83) with 3 phases. We take

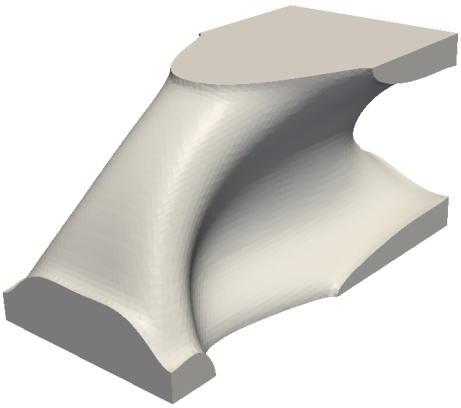
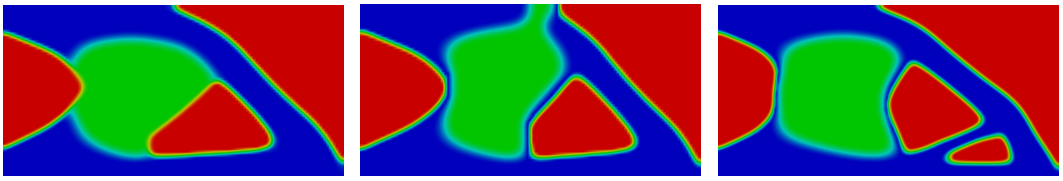


Figure 48: 3D cantilever beam. The level set $\{\varphi \le 0\}$ is shown.



(a) without penalization

(b) penalization of green-red interface

(c) penalization of green-red and green-boundary interface

Figure 49: Solutions including penalization terms. Hard material in blue, soft material in green, void in red.

$\mu = \lambda = 5000$ for the strong material, $\mu = \lambda = 2500$ for the weak material and $\mu = \lambda = 8$ for the void phase and volume fractions $\mathfrak{m} = (0.4, 0.2, 0.4)$. The obtained local minimum for $\gamma = 0.5$ and $\varepsilon = 0.02$ is shown in Figure 49a. Except for the bottom right corner it is the same solution as obtained in [BFGS14, Figure 6]. In [WZ07, Figure 6] a different solution is given, where 200000 pseudo time steps and 17 hours computation time are needed. For our solution the H^1-BFGS method took 187 steps and 17 minutes computation time, where we also use a finer mesh. Again note that in [WZ07] the elasticity equation is only solved every 10 time steps. A completely other solution is given in [TM14, Figure 2]. However, they use a quite different model, where no perimeter penalization or Ginzburg-Landau energy is present. The topology of the solution given in [Tav14, Figure 4c] is the same as for our solution, but the shape is very different.

We now want to demonstrate that for the phase field model used in this work it is very simple to control which materials are allowed to touch in the final design. This is very useful, if e.g. a certain material is prone to rust and may therefore only be put in the interior of the structure and is not allowed to touch the void phase. For instance assume that one wants to prevent a common boundary of the phases 2 and 3 (soft material and void). Then the term

$$\beta \int_\Omega \varphi_2 \varphi_3$$

can be appended to the cost functional with $\beta > 0$ being a penalization parameter (cf. (117)). We note that for fixed parameters γ and ε, the introduction of such a penalization is equivalent to changing the potential ψ_0. To obtain a reasonable sharp interface limit as $\varepsilon \to 0$ one also has to consider the right ε-scaling of the penalization term. The obtained solution for $\beta = 2475$ is depicted in Figure 49b. It can be observed that the weak material now doesn't touch the void phase anymore and also that the topology is different from the solution without penalization. However, the result is still not satisfactory, since there is only a thin strip of blue material between the weak phase and the void. In fact this strip consists of interface and we suppose that we changed only the profile of the 2-3 phase transition. Another difficulty in the application may be that the soft material phase now touches the boundary of Ω, which can be unwanted if there is also void outside of Ω. To overcome this one can introduce an additional penalization term

$$\beta_2 \int_{\partial\Omega} \varphi_2,$$

which penalizes the presence of phase 2 at the boundary of Ω (cf. (118)). The corresponding solution for $\beta_2 = 1$ is shown in Figure 49c. Again the topology is different and the

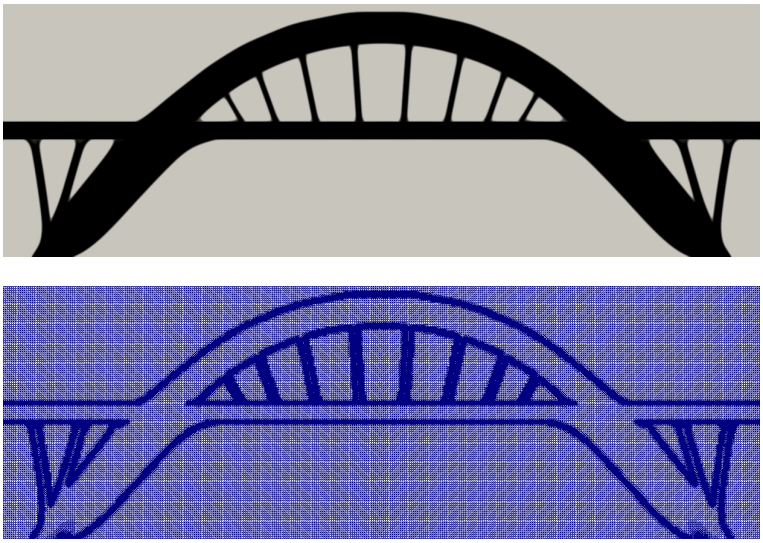


Figure 50: Fixed deck bridge.

weak material now doesn't touch $\partial\Omega$.

As another application we consider the construction of a bridge across a valley with fixed road. As design domain we choose $\Omega = (-90, 90) \times (0, 60)$ with $\Gamma_D = \{76 \leq |x_1| \leq 90\} \cap \{x_2 = 0\}$, $\Gamma_g = \varnothing$ and $\boldsymbol{g} \equiv \mathbf{0}$. We fix material at the location $S_1 = \{-90 \leq x_1 \leq 90\} \cap \{28 \leq x_2 \leq 32\}$ of the road as an equality constraint for φ. The force exerted by the traffic on the road is modeled by the body force $\boldsymbol{f} = (0, -10)^T \chi_{S_1}$. The material constants are $\lambda = \mu = 5000$ for the material and $\mu = \lambda = 70$ for the void. The bridge is allowed to occupy $1/3$ of Ω. Since we only consider symmetric bridges we restrict the computation to the right half of Ω. Here we use a locally refined mesh, which is fine on the interface and we additionally refine the mesh based on the DWR error estimator for the state equation (cf. Section 6.11). Due to the error estimator the mesh is fine in the region where the bridge touches the ground. The result for $\gamma = 0.05$ and $\varepsilon = 2$ together with the resulting adaptive mesh is shown in Figure 50. This design is called a tied arch bridge. For instance the Hoan Bridge across the Milwaukee River in Wisconsin, USA looks very similar to our solution. A similar solution can be found in [ASON13, Figure 3c], although the thickness of the structure below the road is higher. Also the solution obtained in [LS02, Figure 10d] is very similar to ours, although they only have 6 bars below the arch and only one bar on each side of the arch.

Now that we compared the obtained optimal shapes to the ones obtained in the literature, we also want the compare the VMPT method to other methods used in literature. We restrict ourselves to the methods used in [TP13] and [Tav14], since they seem to be quite similar to the VMPT method. We want to reveal similarities and differences to these methods.

We start by the method used in [TP13]. The objective they consider is

$$\begin{aligned} \min\ & F(\boldsymbol{u}) + \lambda \int_\Omega \varphi + \frac{\beta}{2} \int_\Omega |\nabla \varphi|^2, \\ & \varphi \in [0, 1], \end{aligned}$$

where $F(\boldsymbol{u})$ is the compliance. They use mass penalization instead of a hard mass equality constraint. Moreover, the H^1 semi norm of the design variable is appended as a regularization term to get well-posedness. Note that this term is also part of the Ginzburg-Landau energy. However, they don't have a potential term in the cost functional and thus the limit problem for $\beta \to 0$ probably doesn't exist. The applied optimization method can in short form formally be written as

$$\varphi_{k+1} = \mathcal{P}_{A_k}((I - \beta\tau\Delta)^{-1}(\varphi_k - \tau j_N'(\varphi_k)), \tag{290}$$

where $j_N = F(S(\varphi)) + \lambda \int_\Omega \varphi$ is the reduced cost functional without regularization term and $\mathcal{P}_{A_k}$ is the orthogonal projection on the set $A_k = \{\max\{\varphi_k - m, 0\} \le \varphi \le \min\{\varphi_k + m, 1\}\}$ with respect to the inner product $(.,.)_{L^2} + \alpha(.,.)_{H_0^1}$, where $m > 0$ is a move limit, $\alpha > 0$ a weight and $\tau > 0$ is the (fixed) time step size. The move limit is used to stabilize the algorithm. As stopping criterion the relative change in the reduced cost functional is used, i.e.

$$\frac{|j(\varphi_k) - j(\varphi_{k-1})|}{|j(\varphi_{k-1})|} \le tol, \tag{291}$$

where j is the full reduced cost functional (including regularization term) and $tol = 10^{-6}$ is used in the numerics.
We now consider only the case $m = 1$, i.e. the move limit is ignored, and the case $\alpha = \beta\tau$. We show that the algorithm using these special parameters belongs to the family of VMPT methods (up to the line search). We define the inner product

$$a_k(u, v) = \frac{1}{\tau}(u, v)_{L^2} + \beta(\nabla u, \nabla v)_{L^2},$$

which is independent of k. By a calculation similar to those performed in Section 6.8 we obtain that the iterate φ_{k+1} is given as the solution of the projection type subproblem

$$\min_{y \in A_k} \frac{1}{2}\|y - \varphi_k\|_{a_k}^2 + \left\langle j'(\varphi_k), y - \varphi_k \right\rangle,$$

which is exactly in the form used in the VMPT method. Also note that in the case $m = 1$ we have $A_k = \{0 \le \varphi \le 1\}$. Thus the method used in [TP13] with parameters $m = 1$ and $\alpha = \beta\tau$ coincides with the VMPT method, without Armijo backtracking, for the above metric a_k and scaling $\lambda_k = 1$. In [TP13] no convergence result for the method is given. However, by the deduced equivalence to the VMPT method we can by the techniques used in this thesis show global convergence of the method in $H^1 \cap L^\infty$ if in addition to the method in [TP13] an Armijo line search is included. We suppose that the authors in [TP13] introduced the move limit m as a substitute for a line search to get convergence. However, a move limit always leads to the restriction $\|\varphi_k - \varphi_{k-1}\|_{L^\infty} \le m$, i.e. in a single optimization step the values of φ_k cannot change from 0 to 1. This is a serious restriction and may lead to slow progress of the method, especially in the beginning where the design variable usually changes much. Using an Armijo backtracking instead does not lead to this kind of restriction and may therefore be advantageous. However, using Armijo is more expensive.
We note that for $\alpha \neq \beta\tau$, the gradient in (290) is taken with respect to a different inner product than the projection. Thus this is not a VMPT method and convergence of the method is unclear. In fact there are certain counterexamples where convergence for this type of optimization methods is not given, see [Ber99, sec. 2.4]. However, the numerical

results in [TP13] indicate mesh independency also for $\alpha \neq \beta\tau$.
We also note that the method used in [TP13] is equivalent to the pseudo time stepping method of Allen-Cahn type performed e.g. in [BFGS14]. Recall that the inner product corresponding to the semi-implicitly discretized L^2 gradient flow is given by (cf. Section 6.8)

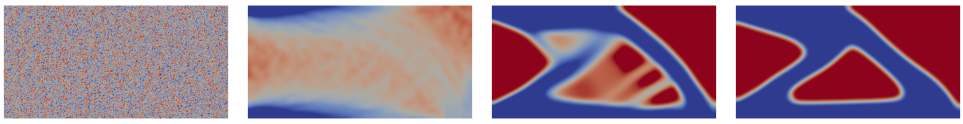
$$a_k^{AC}(u,v) = \frac{\varepsilon}{\tau^{AC}}(u,v)_{L^2} + \gamma\varepsilon(\nabla u, \nabla v)_{L^2}.$$

Thus, for $\beta = \gamma\varepsilon$, $\alpha = \beta\tau$, $m = 1$ and $\tau^{AC} = \varepsilon\tau$, the methods coincide (of course only the numerical methods coincide, but the used cost functionals are different).
We also comment on the used stopping criterion. For the cantilever beam experiment we performed in Figure 24, the stopping criterion was $\sqrt{\gamma\varepsilon}\|\nabla v_k\|_{L^2} \leq tol = 10^{-5}$, which is reached at iteration 269 (for the finest mesh). The relative change in the cost functional (291), which is used as stopping criterion in [TP13], is for our iterate 269 given as $2 \cdot 10^{-12}$. If we use the same stopping criterion as in [TP13] with the same tolerance $tol = 10^{-6}$, then our optimization in Figure 24 would already stop at iteration 76 instead of iteration 269. On the other hand it holds for iteration $k = 76$ that $\sqrt{\gamma\varepsilon}\|\nabla v_k\|_{L^2} = 6.6 \cdot 10^{-2}$. Here one sees that the choice of the stopping criterion is essential and influences how the performance of the method is presented. Using the stopping criterion in [TP13], our method 'is more than three times faster'. As a second example consider the cantilever beam with few available mass in Figure 45. In this experiment the relative difference of 10^{-6} in the cost functional is reached very early in the optimization process, where the current iterate is still very far away from the depicted design. Thus the stopping criterion is in this case too liberal and thus inappropriate. One would have to decrease the tolerance for a reasonable stopping criterion.

Finally we consider the method used in [Tav14]. As basis for the objective function the same multi phase field model as in the present thesis is used. However, certain changes are performed in order to simplify calculations. The used topology optimization problem reads

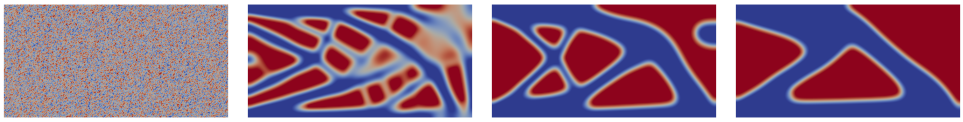
$$\begin{aligned} \min\ & F(\boldsymbol{u}) + \gamma \int_\Omega \left\{ \frac{\varepsilon}{2} \sum_{i=1}^{N-1} |\nabla\varphi_i|^2 + \frac{1}{\varepsilon} \sum_{i=1}^{N-1} \psi_0(\varphi_i) \right\}, \\ & 0 \leq \boldsymbol{\varphi} \leq 1 \\ & 0 \leq 1 - \sum_{i=1}^{N-1} \varphi_i \leq 1 \\ & \fint \boldsymbol{\varphi} = \mathfrak{m}, \end{aligned}$$

where $F(\boldsymbol{u})$ denotes the compliance and $\boldsymbol{\varphi}$ is a vector valued phase field with $(N-1)$ components. The idea behind the simplification is that by the sum constraint the last component φ_N of the phase field can be eliminated by setting $\varphi_N = 1 - \sum_{i=1}^{N-1} \varphi_i$. Substituting this into the Ginzburg-Landau energy leads to certain coupling terms, which complicate the calculation. Therefore these coupling terms are neglected, which is the reason why the sums in the Ginzburg-Landau energy only involve the first $(N-1)$ components of the phase field variable. No justification of these simplifications is given in



(a) t=0 (b) t=0.000625 (c) t=0.00625 (d) t=0.2983

Figure 51: Pseudo time stepping of Allen-Cahn type with fixed time step size.



(a) t=0 (b) t=0.000625 (c) t=0.00625 (d) t=0.1125

Figure 52: Pseudo time stepping of Cahn-Hilliard type with fixed time step size.

[Tav14]. The applied numerical method can be formally written as

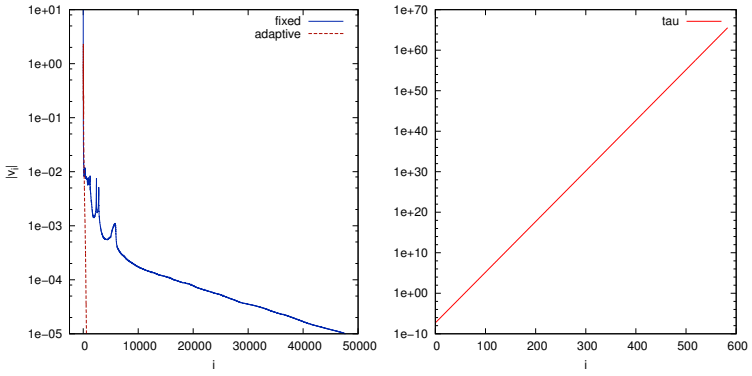
$$\begin{aligned} \boldsymbol{v}_k &= P(\boldsymbol{\varphi}_k - j_N'(\boldsymbol{\varphi}_k)) - \boldsymbol{\varphi}_k \\ \boldsymbol{\varphi}_{k+1} &= (I - \alpha_k \gamma \varepsilon \Delta)^{-1}(\boldsymbol{\varphi}_k + \alpha_k \boldsymbol{v}_k), \end{aligned}$$

where $j_N(\boldsymbol{\varphi}) = F(\boldsymbol{u}) + \frac{\gamma}{\varepsilon}\int_\Omega \sum_{i=1}^{N-1} \psi_0(\varphi_i)$ is the reduced cost functional without the gradient term, $\alpha_k = 0.5$ is the used step size and P denotes the projection onto the feasible set with respect to the Euclidean inner product in the discrete setting. The projection is taken with respect to the Euclidean inner product since the author uses a very efficient method therefor, which involves successive pointwise projections (alternating projection method). The last multiplication by $(I - \alpha_k \gamma \varepsilon \Delta)^{-1}$ is necessary since otherwise checkerboard patterns occur. The method is stopped after a fixed number of 1000 iterations. No convergence analysis is given.
Due to the simplifications in the cost functional and the use of an Euclidean projection, the method is very fast. For a computation involving 8 phases only 15 minutes of computation time is needed. However, the final designs don't look satisfactory, e.g. in [Tav14, Figure 3k] the MBB beam is disconnected. This can be either due to bad modelling because of neglecting the coupling terms in the Ginzburg-Landau energy, or due to the non-convergence of the numerical method.
To compare this to our approach, we already mentioned that no justification for the simplification done in the cost functional is given in [TP13]. Thus it is unknown if the limit $\varepsilon \to 0$ exists. The numerical method looks very similar to the one in [TP13] discussed in the previous paragraph (if α_k is set to 1). The major difference is that the projection and the application of $(I - \alpha_k \gamma \varepsilon \Delta)^{-1}$ is interchanged and that an Euclidean projection is used instead of an H^1-type projection as in [TP13]. Due to the Euclidean projection the method in [Tav14] cannot be equivalent to some VMPT method and convergence cannot be shown in this way.

We also compare the VMPT method to the pseudo time stepping methods used in [BGS+12, BFGS14]. We do this in more detail and thus devote it an extra section.

6.13.7 Comparison to pseudo time stepping methods

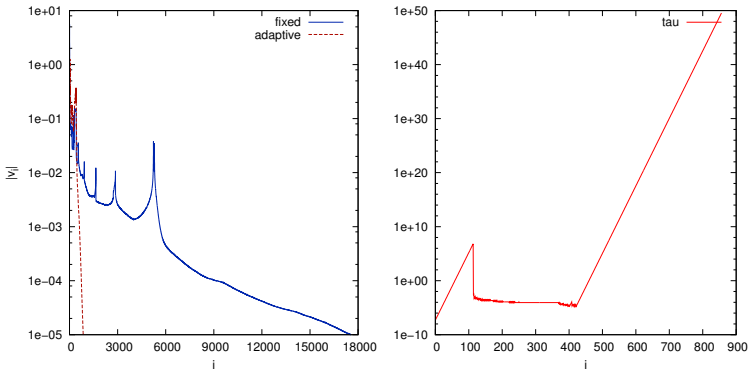
In this section we compare the pseudo time stepping methods used in [BGS+12, BFGS14] with constant time step size to the adaptive time step size strategy developed in Section



(a) Residual for constant and adaptive time step size.

(b) Adaptive time step sizes.

Figure 53: Pseudo time stepping of Allen-Cahn type.



(a) Residual for constant and adaptive time step size.

(b) Adaptive time steps.

Figure 54: Pseudo time stepping of Cahn-Hilliard type.

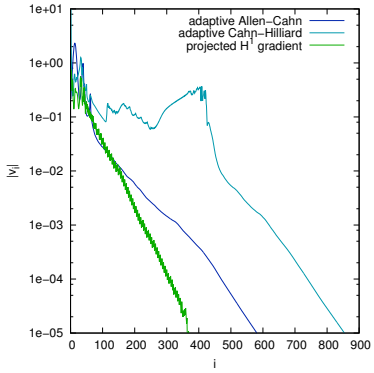


Figure 55: Comparison of the residual for Allen-Cahn and Cahn-Hilliard with adaptive time steps and the projected H^1-gradient method.

6.8. Moreover we study how the adaptive methods perform in comparison to the projected H^1-gradient method.

As experiment we choose the cantilever beam example with $\varepsilon = 0.03$, $\gamma = 0.5$ and a fixed equidistant mesh with mesh parameter $h = 2^{-7}$. As initial guess we use random values between -1 and 1 with a mass of 50%. We perform the Allen-Cahn pseudo time stepping (175) with constant time step size $\tau_k = 6.25 \cdot 10^{-6}$ as in [BGS+12, BFGS14] and the Cahn-Hilliard pseudo time stepping (178) with the same τ_k. The iterates at certain pseudo times can be seen in Figures 51 and 52. As typical for Allen-Cahn type evolutions the length scale of the phases stays constant during the evolution, whereas in the Cahn-Hilliard process the decomposition at the beginning leads to finer structures, which then merge slowly to the final structure. Note that the phase separation in the evolutions takes rather a long time since ε is chosen quite large. Afterwards we perform the same calculations but with τ_k chosen adaptively according to Algorithm 6.1, starting with $\tau_{-1} = 5 \cdot 10^{-8}$ and using $\beta = 0.75$ for increasing/decreasing the time step size. The comparison of the residual $\sqrt{\gamma\varepsilon}\|\nabla \boldsymbol{v}_k\|_{L^2} = \sqrt{\gamma\varepsilon}\|\nabla(\boldsymbol{\varphi}_{k+1} - \boldsymbol{\varphi}_k)\|_{L^2}$ together with the adaptive τ_k is depicted in Figures 53 and 54. At the peaks of the residual for fixed τ_k, topological changes occur. It can be clearly seen that using adaptive time steps leads to a huge boost in performance. For the Allen-Cahn method, the adaptive choice of τ_k is about 80 times faster, whereas for Cahn-Hilliard, adaptivity is still 20 times faster. The time step size τ_k for Allen-Cahn can be increased monotonically until at the end it holds $\tau_k > 10^{60}$. For the Cahn-Hilliard experiment this is not the case. During the first 100 iterations τ_k can be increased, but then it has to be chosen smaller again ($\tau_k \approx 10^{-4}$). At this time the 4 holes depicted in Figure 52c are melting together. As soon as the topological change is completed, τ_k can be increased again until $\tau_k = 10^{50}$ at the end. For small time step size the Allen-Cahn and Cahn-Hilliard pseudo time stepping methods are an approximation of the continuous L^2 and H^{-1} gradient flow of the cost functional. However, when choosing an adaptive time step size which is very large as in this experiment, the pseudo time stepping methods may not approximate the respective gradient flows anymore. Thus the pseudo time stepping can converge to another local minimizer than the continuous flow.
We note that in [BC03] a similar update scheme for the time step size is used. However, the maximum time step size they get is $5 \cdot 10^{-2}$, which is much lower than our time step size. In [DBH12] also an adaptive time step size for the Cahn-Hilliard equation is used. There, the final time step size is 10^{13} times the initial time step size, which resembles more our result.
The pseudo time stepping methods using an equidistant temporal mesh need many iterations here to converge. For the Allen-Cahn method almost 50000 iterations are needed, for the Cahn-Hilliard simulation almost 18000 iterations are needed. Note that also in other literature pseudo time stepping methods usually need many iterations to converge. For instance in [WZ07] a Cahn-Hilliard flow is considered, where up to 370000 iterations are needed. Also for the simulation with fastest convergence 3000 iterations are needed.
In the present experiment it also turns out that it is important to have a rigorous stopping criterion for the method. In [BGS+12, BFGS14] it is mentioned that the iteration is stopped if the phase field visually does not change anymore. In the Allen-Cahn experiment with fixed time step size (Figure 53a), this would be the case after 9000 iterations, since the phase field does not change visibly within 200 iterations. However, the interface still moves very slowly, yielding that the position of the hole in the final structure at iteration 47729 is quite different from the position at iteration 9000, although the topology does not change anymore. Thus stopping is not reasonable at iteration 9000.

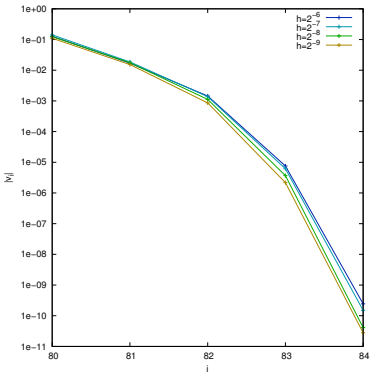
Finally we compare the adaptive Allen-Cahn and Cahn-Hilliard evolutions to the projected H^1-gradient method. The respective residuals are depicted in Figure 55. One can see that although the adaptive Allen-Cahn and Cahn-Hilliard methods work very well, the projected H^1-gradient method is still faster, even more than twice as fast as the Cahn-Hilliard method. As shown in Lemma 6.62, the Allen-Cahn and Cahn-Hilliard methods both approach the projected H^1-gradient method with $\lambda_k = 1$ as $\tau \to \infty$. This can be also seen in Figure 55, where at the end of the iteration the residual lines of both methods become parallel. The curve of the residual of the projected H^1-gradient method is slightly steeper since λ_k is chosen by the update (280), which leads to better performance. Also a huge difference between the methods is that the H^1-gradient method produces the optimal topology as soon as the phases are separated, whereas the Cahn-Hilliard method usually produces a very fine topology after the spinodal decomposition, which then slowly merges to the final coarse topology (see e.g. [WZ07]).

6.13.8 Comparison to the SQP method

We investigate the SQP method from Section 6.9 numerically. As already discussed, $j''(\boldsymbol{\varphi})$ is not positive definite and thus only local convergence of the SQP method can be expected without further globalization techniques. In this section we use the VMPT method as a possible globalization, i.e. we perform a VMPT iteration until the iterate is close enough to the minimizer and then we switch the inner product a_k in the VMPT subproblem to the second order derivative $j''(\boldsymbol{\varphi}_k)$, which defines the SQP method. We will see that the SQP method is mesh independent and converges at least with Q-superlinear rate. Moreover, it exhibits the drawbacks that the convergence radius is very small, especially for small ε and that the SQP subproblem is very hard to solve. As already mentioned we use the PDAS method described in Section 6.10 as a solver for the SQP subproblem.

We use the cantilever beam experiment (Example 6.83) with $\varepsilon = 0.06$. First we perform 80 iterations of the H^1-gradient method and then we switch to the SQP method. Since we do not know how close the iterate has to be to the minimum a priori, we have to obtain the necessary 80 VMPT iterations by trial and error. The experiment is repeated for different mesh sizes, from $h = 2^{-6}$ to $h = 2^{-9}$, where an adaptive mesh is used on the finest level. The results are shown in Figure 56. The Figure on the left hand side shows the distances $\sqrt{\varepsilon\gamma}\|\varphi_{k+1} - \varphi_k\|_{H^1}$, which clearly are mesh independent. The SQP method converges within 4 steps to a tolerance of 10^{-9} for all meshes. The iteration on the finest mesh is also shown in Table 17. One clearly observes that $\|v_k\|_{H^1}/\|v_{k-1}\|_{H^1}$ converges to zero, thus the rate is at least Q-superlinear. Since $\|v_k\|_{H^1}/\|v_{k-1}\|_{H^1}^2$ doesn't become too large, we can also assume Q-quadratic convergence. Note that for Q-superlinearly convergent sequences it holds $\|v_k\| = \|\varphi_{k+1} - \varphi_k\| \approx \|\varphi_k - \varphi^*\|$ for large k, hence we can also assume that $\varphi_k \to \varphi^*$ Q-quadratically in H^1.
In this experiment the initial guess for the SQP method has to be very close to the minimum as can be seen in Figure 56b and 56c. For an initial guess further away from the solution, the SQP method won't converge. Thus even for this large ε, the convergence radius of the SQP method is very small.

If one wants to obtain an optimal shape for small ε, one usually starts with a larger ε and decreases it slowly, since the numerics works better for larger ε. The idea is now to compute a minimizer for large ε using some VMPT method and use this minimizer as



(a) $\sqrt{\varepsilon\gamma}\|\varphi_{k+1} - \varphi_k\|_{H^1}$.

(b) Initial guess.

(c) Solution.

Figure 56: SQP method for $\varepsilon = 0.06$.

k	$\sqrt{\varepsilon\gamma}\|v_k\|_{H^1}$	$\sqrt{\varepsilon\gamma}\|v_k\|_{H^1}/(\sqrt{\varepsilon\gamma}\|v_{k-1}\|_{H^1})$	$\sqrt{\varepsilon\gamma}\|v_k\|_{H^1}/(\sqrt{\varepsilon\gamma}\|v_{k-1}\|_{H^1})^2$
80	1.10e-1	-	-
81	1.53e-2	1.39e-1	1.27
82	8.60e-4	5.63e-2	3.68
83	2.17e-6	2.53e-3	2.94
84	2.79e-11	1.28e-5	5.89

Table 17: Quadratic convergence of $v_k := \varphi_{k+1} - \varphi_k$ on the finest mesh.

initial guess for the SQP method for smaller ε. In this manner the fast local convergence of the SQP method can be exploited. However, we will see that even this is not possible. We start with $\varepsilon_0 = 0.06$ and compute a minimizer of the respective cantilever beam experiment using the H^1-BFGS method. When we decrease ε to $\varepsilon_1 = 0.75\varepsilon_0 = 0.045$, the SQP method converges within 7 steps. Another decrease to $\varepsilon_2 = 0.75^2\varepsilon_0 = 0.03375$ lets the SQP method converge within 6 steps. However, after another decrease to $\varepsilon_3 = 0.75^3\varepsilon_0 \approx 0.025$ the SQP method doesn't converge anymore. The only difference between the initial guess (which is the solution for ε_2) and the solution for ε_3 is that the interface is slightly thinner, see Figure 57a and 57b. Nonetheless, the SQP method does not converge with this seemingly good initial guess. After one SQP iteration, the control is very far from the solution (Figure 57c), and does not lead to a decrease in j, nor does it define a descent direction. After the second SQP iteration, the control looks even worse. We note that the PDAS method described in Section 6.10 is not able to solve the SQP subproblem in this case. Thus we have to perform 200 iterations of the projected H^1-gradient method applied to the SQP subproblem to get a good initial guess of the active set for the PDAS method, which is very expensive.

We want to explain why the SQP method doesn't converge. From the analysis of the Josephy-Newton method we get that under the assumption that $\overline{\varphi}$ is a strongly regular solution of the variational inequality

$$\varphi \in \Phi_{ad}, \quad \left\langle j'(\varphi), \eta - \varphi \right\rangle \geq 0 \quad \forall \eta \in \Phi_{ad}$$

in the sense of Robinson, the iterates φ_k of the Josephy-Newton method, which is equivalent to the SQP method, converge superlinearly to $\overline{\varphi}$ if $\|\varphi_0 - \overline{\varphi}\|_{H^1 \cap L^\infty}$ is sufficiently small. The convergence radius has to be measured in the $(H^1 \cap L^\infty)$-norm, since j' is differentiable with respect to this norm. On the other hand one can expect that the minimizers

(a) Initial guess. (b) Solution. (c) After one SQP step.

Figure 57: SQP method for $\varepsilon \approx 0.025$.

φ_ε of j_ε converge in L^1 as $\varepsilon \to 0$ due to the Γ-convergence result (cf. Section 6.4), but neither in H^1 nor in L^∞. In fact, it typically holds $\|\nabla\varphi_\varepsilon\|_{L^2} \sim \frac{1}{\sqrt{\varepsilon}}$, cf. (279). Assume that we decrease ε by a factor $0 < C < 1$ as above, then we get the sequence $\varepsilon_i := C^i\varepsilon_0$ and it holds for the H^1-distance of two consecutive minimizers

$$\begin{aligned}\|\nabla(\varphi_{\varepsilon_{i+1}} - \varphi_{\varepsilon_i})\|_{L^2} \geq \|\nabla\varphi_{\varepsilon_{i+1}}\|_{L^2} - \|\nabla\varphi_{\varepsilon_i}\|_{L^2} \sim \frac{1}{\sqrt{\varepsilon_{i+1}}} - \frac{1}{\sqrt{\varepsilon_i}} = \frac{1}{\sqrt{C^{i+1}\varepsilon_0}} - \frac{1}{\sqrt{C^i\varepsilon_0}} \\ = \frac{1-\sqrt{C}}{\sqrt{C^{i+1}\varepsilon_0}} \to \infty \quad \text{as } i \to \infty.\end{aligned}$$

If we assume that the convergence radius of the SQP method does not grow as $\varepsilon \to 0$, then φ_{ε_i} is eventually outside the area of local convergence for ε_{i+1}.
On the other hand one could choose the sequence ε_i such that $\|\nabla(\varphi_{\varepsilon_{i+1}} - \varphi_{\varepsilon_i})\|_{L^2}$ is small enough. However, even when we take the solution for $\varepsilon = 0.027$ as an initial guess for $\varepsilon = 0.0265$ the SQP method does not converge, since we again end up with a SQP subproblem minimizer similar to Figure 57c. Note that as ε decreases also the solution of the SQP subproblem gets more expensive. Here we had to perform 1500 projected H^1-gradient iterations followed by 1 PDAS iteration for a single SQP step.
A similar problem arises if the initial guess for the SQP method is a phase field with slightly shifted interface rather than the solution of the problem with higher ε. Assume that the interfaces of the solution $\overline{\varphi}$ and the initial guess φ_0 do not touch, then a similar calculation yields

$$\|\nabla(\varphi_0 - \overline{\varphi})\|_{L^2}^2 = \|\nabla\varphi_0\|_{L^2}^2 + \|\nabla\overline{\varphi}\|_{L^2}^2 \sim \frac{1}{\varepsilon} \to \infty \quad \text{as } \varepsilon \to 0.$$

Thus for small ε the initial guess φ_0 lies outside the convergence area of the SQP method, although $\|\varphi_0 - \overline{\varphi}\|_{L^1}$ might be very small.

To further investigate this issue we consider the SQP subproblem at the minimizer, i.e. at $\varphi_k = \overline{\varphi}$ and we take again $\varepsilon = 0.06$. We thus solve the subproblem

$$\begin{aligned}\min\ & \frac{1}{2}j''(\overline{\varphi})[y-\overline{\varphi}, y-\overline{\varphi}] + \left\langle j'(\overline{\varphi}), y-\overline{\varphi}\right\rangle \\ & y \in \Phi_{ad}.\end{aligned}$$

As already discussed, the solution of the SQP subproblem may not be unique, since $j''(\overline{\varphi})$ is not positive definite. Although $y = \overline{\varphi}$ is always a solution (at least a stationary point), there may be additional local minimizers. To obtain local minimizers other than $y = \overline{\varphi}$ we calculate solutions of the SQP subproblem by means of the projected H^1-gradient method

Figure 58: Multiple solutions of the SQP subproblem in the minimum $\overline{\varphi}$.

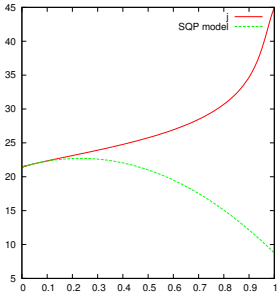


Figure 59: Plot of j and the SQP model between two local minima of the SQP model.

for different initial guesses. Thereby we are able to compute 5 different local minima which are depicted in Figure 58, where the first picture is the trivial minimum $y = \overline{\varphi}$. It can be computed that the minimum φ_1 in the second picture is lower than $\overline{\varphi}$, as can be seen in Figure 59, where the values of j and its second order Taylor polynomial in $\overline{\varphi}$ are shown along the line segment $\alpha \mapsto \overline{\varphi} + \alpha(\varphi_1 - \overline{\varphi})$. Note that values of α outside $[0,1]$ define infeasible points. It can be clearly seen that $j''(\overline{\varphi})$ is negative in this direction. Since φ_1 is a lower minimum of the SQP subproblem than $\overline{\varphi}$, we can also conclude that the regularity condition of Dunn [Dun80], i.e.

$$\exists m > 0 : \quad \frac{1}{2} j''(\overline{\varphi})[y - \overline{\varphi}, y - \overline{\varphi}] + \left\langle j'(\overline{\varphi}), y - \overline{\varphi} \right\rangle \geq \frac{m}{2} \|y - \overline{\varphi}\|_{\mathbb{X}}^2 \quad \forall y \in \Phi_{ad},$$

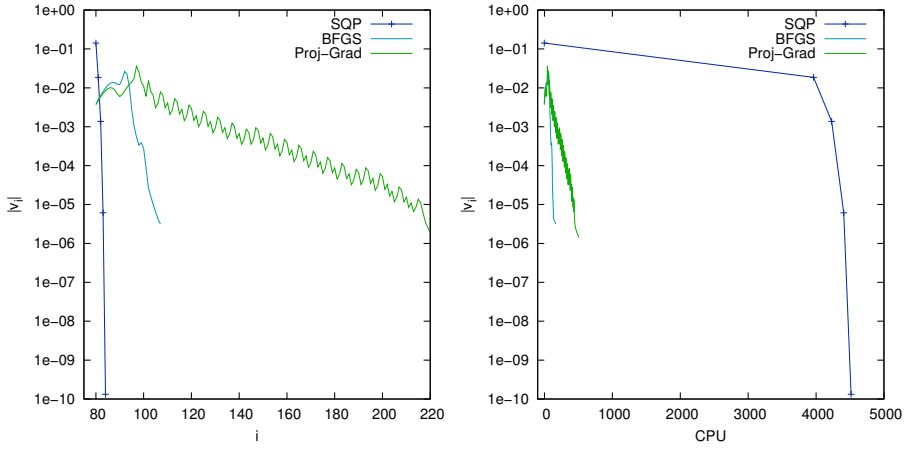
cannot be fulfilled, since $\overline{\varphi}$ is not the global minimizer of the SQP subproblem. On the other hand, Robinson's strong regularity condition may still be fulfilled, since only locally unique solvability of the SQP subproblem is required.
In the SQP algorithm one has to take the SQP subproblem minimizer which is closest to φ_k in order to gain convergence, which is not an easy task. Perhaps the minimizer shown in Figure 57c is not the right one and thus the SQP method does not converge. However, we were not able to calculate another model minimizer which might be closer to φ_k. Either the numerical method is incapable of calculating it, or no closer minimizer exists at all.

We have already seen that the computation of a solution of the SQP subproblem becomes more expensive the smaller ε is. We experience that for very small ε the numerical methods considered here are not able to compute a solution of the SQP subproblem at all. For example we use $\varepsilon = 0.005$ and apply the H^1-BFGS method to the mean compliance problem until it breaks down with a residual of $\sqrt{\varepsilon\gamma}\|v_k\|_{H^1} \approx 5 \cdot 10^{-6}$. Using this solution as an initial guess for the SQP method, we expect convergence within one or two steps. Moreover, a good initial guess for the SQP subproblem is available since it will not be far away from the current iterate. However, it is not possible to compute a minimizer of the SQP subproblem. The projected H^1-gradient method applied to the subproblem breaks down in the first step, since no step length can be found due to approximation errors. Also the PDAS method applied to subproblem does not converge. Without damping

the method oscillates between certain active sets. With damping the damping parameter converges to zero. We also tried to solve the state equation and linearized state equation more accurately to decrease approximation errors and we also tried to refine the mesh around the interface. However, no improvement could be seen. We note that also the PDAS method applied directly to the mean compliance problem has the same difficulties, see Section 6.13.9. Thus for small ε the SQP method is not practicable at all. On the other hand, we have no problems to solve the subproblem in the projected H^1-gradient method or BFGS method. The PDAS method always converges within one or two steps near the minimum, even for very tiny ε.

We return to large values of ε, where the SQP method works well. In terms of iteration numbers the SQP method is more efficient compared to the projected H^1-gradient or H^1-BFGS method because of its superlinear rate of convergence. However, a single SQP iteration is very expensive since PDEs have to be solved each time the Hessian $j''(\varphi)$ is evaluated in some direction as described in Section 6.10.4. Thus it is reasonable to compare the methods also in terms of CPU time. We repeat the experiment from above using $\varepsilon = 0.06$ and starting with the same initial guess near the solution and use a mesh with $h = 2^{-7}$. The resulting iterations of the SQP, H^1-BFGS and H^1-gradient method are depicted in Figure 60, where the latter methods are stopped when they break down because of approximation errors near the minimum. On the left hand side the methods are compared in terms of iteration numbers whereas on the right hand side the computation time is shown. The SQP method finds the minimum within 4 steps consuming 75 minutes of computation time. For the H^1-BFGS method we have 25 iterations in 2.3 minutes and for the H^1-gradient method we have 138 iterations in 7.6 minutes (see also Table 18). Thus the H^1-BFGS method only needs 3% of the CPU time of the SQP method and is therefore more efficient. At first glance the SQP method provides a more accurate solution than the H^1-BFGS method, since the final SQP residual is 10^{-10}, whereas the H^1-BFGS residual is larger than 10^{-6}. However, this is not true as can be seen by comparing the final (discrete) cost functional values. For the SQP method we get $j(\varphi^*) = 21.4238044157962$ and for the H^1-BFGS method $j(\varphi^*) = 21.4238044157949$, which is even a bit better. On the other hand, the last digits are not very reliable, see Figure 6.
Note that comparing CPU times is always implementation depending. Thus another solver for the subproblems may lead to other CPU times. However, solving the SQP subproblem will always involve the expensive solution of PDEs, which is not the case for the H^1-BFGS method.

We can conclude the numerical study of the SQP method as follows. For large values of ε the SQP method locally works fine and at least superlinear convergence and mesh independency can be observed. As ε gets smaller the computation of a minimizer of the SQP subproblem gets more expensive. For very small ε it is not possible to calculate a minimizer with the developed numerical methods. The convergence radius of the SQP method is measured in the $H^1 \cap L^\infty$-norm, which leads to difficulties since $\|\varphi_\varepsilon\|^2_{H^1} \sim \frac{1}{\varepsilon}$, and thus initial guesses with slightly shifted interfaces or the solution for larger ε lie outside the area of convergence for small ε. Hence the position of the interface has to be almost known to get convergence of the SQP method. On the other hand if the position of the interface is known then one can stop the iteration and the SQP iteration is unnecessary. Moreover, since j'' is not positive definite near the minimum, the SQP subproblem has many local minimizers and it is difficult to obtain the 'right' one numerically. In terms of computation time the projected H^1-gradient method and the corresponding BFGS update



(a) No. of iterations. (b) CPU time.

Figure 60: Residual $\sqrt{\varepsilon\gamma}\|v_k\|_{H^1}$ for SQP, H^1-BFGS and H^1-gradient method.

are more efficient than the SQP method. Because of the preceding difficulties, the SQP method is not appropriate for the mean compliance problem.

In the special case the $\boldsymbol{C}$ interpolates linearly and γ is small whereas ε is large, the SQP method behaves similarly to the second order VMPT method, which converges globally. We refer to the end of Section 6.13.4 for a numerical experiment.

6.13.9 Comparison to the semismooth Newton method

	SSN	SQP	H^1-gradient	H^1-BFGS
iterations	10	4	138	25
CPU time	18m	75m	7.6m	2.3m
# of solved Newton systems	10	52	223	57

Table 18: Data for $h = 2^{-7}$ using a good initial guess.

In the SQP method as well as in the various considered VMPT methods we use the semismooth Newton (SSN) method, or equivalently the PDAS method, to solve the corresponding subproblem. A natural question arising is: Why don't we apply the SSN method directly to the overall optimization problem? We will see that the SSN method suffers from the same drawbacks as the SQP method and that it is in addition mesh dependent.

We already observed the moderate mesh dependency of the SSN method when applied to the projection type subproblem, for instance see Figure 26c. To show that the moderate mesh dependency also applies when the SSN method is used for the solution of the overall topology optimization problem we perform a numerical experiment using the cantilever beam setup from Example 6.83 with $\gamma = 0.5$ and $\varepsilon = 0.06$. This is the same experiment used to show the mesh independency of the SQP method (Figure 56). Since only local convergence of the SSN method can be expected we have to choose a good initial guess. As for the SQP method we take the 80th iterate of the H^1-gradient method, see Figure 56b. As opposed to the SQP method we need additionally initial guesses for the Lagrange

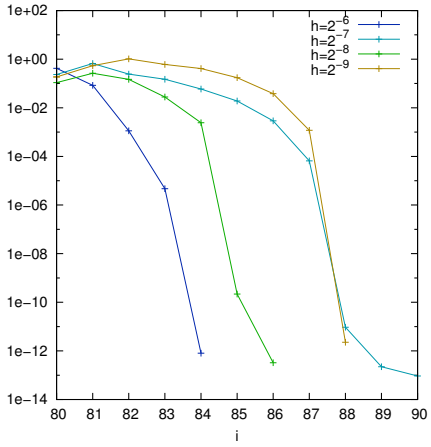


Figure 61: Newton residual for different mesh sizes reveals mesh dependency.

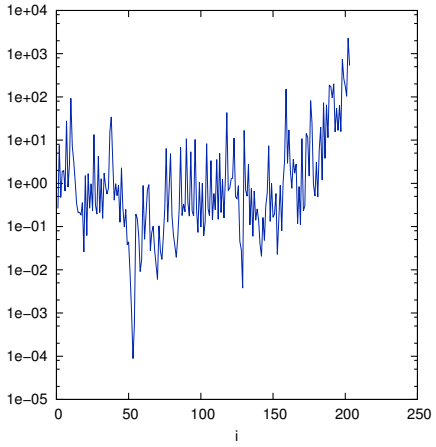


Figure 62: Newton residual for $\varepsilon \approx 0.025$ starting with the solution for $\varepsilon \approx 0.034$. SSN does not converge.

multipliers λ and μ for the mass constraint and the box constraints. We compute an initial guess for λ by the formula (145), which we used to show uniqueness of the Lagrange multiplier. The initial μ is chosen such that the gradient equation in the KKT system is fulfilled. The result for different mesh sizes is depicted in Figure 61. To be precise we plot the Euclidean norm of the discrete KKT System (242)-(244). Of course this norm depends on the mesh size h, thus it would be better to consider $\|\varphi_k - \varphi_{k-1}\|_{H_0^1}$ instead. In Figure 61 the moderate mesh dependency can be seen. On the finest mesh the SSN method takes double the number of iterations of the coarsest mesh. The number of iterations does not increase monotonically. For $h = 2^{-8}$ the method is faster than for $h = 2^{-7}$. However, there is a clear mesh dependency compared to the behavior of the SQP method (Figure 56). As already discussed in Section 6.10 the mesh dependency stems from the fact that the SSN method is not well defined in the continuous setting due to the lack of semismoothness.

For a comparison of the SSN method to the other considered methods we refer to Table 18. Since the SSN method depends on the mesh parameter h, we choose $h = 2^{-7}$ for the comparison. The first row shows the number of iterations needed by the SSN, SQP, H^1-gradient and H^1-BFGS method. The SSN method takes more iterations than the SQP method, but less than the H^1-BFGS method. Since the comparison of iteration numbers is not meaningful here we also consider the corresponding computation time, which is given in the second row of Table 18. One observes that even though the SSN method needs more iterations than the SQP method, the CPU time is less. The reason clearly is that a single SQP step is more expensive, since a quadratic optimization problem has to be solved, whereas a single step of the SSN method only involves the solution of a linear system. However, the SSN method still needs more computation time than the H^1-BFGS method. This is again plausible since we solve the linearized state equation in each MINRES step within a SSN iteration. For the H^1-BFGS method no linearized state equation has to be solved. In the third row of Table 18 we also give the cumulative number of solved Newton systems. For the SSN method this is just the number of SSN steps. For the other methods we sum up the number of SSN steps used to solve the corresponding subproblem. One observes that the solution of a Newton system in the SSN method is in average more expensive (1.8m/solve) than in the SQP method (1.44m/solve).

We note that the iteration numbers shown in Table 18 are implementation independent in contrast to the CPU time and the number of linear solves. The SQP method can still be very good if a better solver for the SQP subproblem is found. Also the SSN method can be improved e.g. by introducing a preconditioner for the Newton system.

We give an example to show that the convergence radius of the SSN method is very small similar to the SQP method. Therefor we perform the same experiment as for the SQP method in Figure 57, i.e. we use the solution of the cantilever beam problem for $\varepsilon = 0.03375$ as initial guess for the same problem with $\varepsilon \approx 0.025$. The result is shown in Figure 62. The SSN method seems to converge around iteration 50, but then the residual increases again until it explodes at iteration 200. Thus the unglobalized SSN method doesn't converge for the seemingly good initial guess.

Due to the small convergence radius the SSN method is unsuitable for the solution of the overall optimization problem. However, it performs very well as a solver for the projection type subproblems in the VMPT method. The reason is that we have a good initial guess from the last VMPT step and that the projection type subproblem is much easier to solve since it is a convex, quadratic optimization problem and has a unique minimizer. Moreover, the SSN method can be used to compute the Lagrange multipliers.

We finally note that the SSN method is successfully applied to other shape optimization problems, which don't involve a phase field model. For instance in [KU14] the considered shape optimization problem is approximated by a sequence of regularized problems. For the subproblems semismoothness in the function space setting can be shown together with the superlinear convergence of the SSN method.

6.13.10 Computation and discussion of Lagrange multipliers

In this section we give an example where the Lagrange multipliers are functionals rather than functions as expected from the theoretical results in Theorem 6.33. Moreover, we show that the Lagrange multipliers are in another example functions which can have singularities. We also present an experiment where the Lagrange multiplier for the constraint $\varphi \leq 1$ is a scaled characteristic function.

As first example we consider the binary cantilever beam with volume force $\boldsymbol{f} \equiv 0$ and boundary traction $\boldsymbol{g}(x,\varphi) = \frac{1+\varphi}{2}\boldsymbol{g}_1(x) + \frac{1-\varphi}{2}\boldsymbol{g}_2(x)$, where $\boldsymbol{g}_1(x) = (0,-350)^T$ is the force acting on the void and $\boldsymbol{g}_2(x) = (0,-250)^T$ is the force acting on the material. We use the H^1-BFGS method to compute the discrete solution φ_h and then apply the SSN method to compute the associated discrete Lagrange multipliers $\boldsymbol{\lambda}$ and μ_h. By using equations (145) and (135) to compute an initial guess for the Lagrange multipliers, the SSN method is able to compute the final solution $(\varphi_h, \boldsymbol{\lambda}, \mu_h)$ of the KKT system within 2 steps. Here, we denote by μ the combined Lagrange multiplier $\mu_2 - \mu_1$, where $\mu_1 \geq 0$ is the Lagrange multiplier for $\varphi \geq -1$ and $\mu_2 \geq 0$ is the Lagrange multiplier for $\varphi \leq 1$ as in the KKT system (150). Figure 63 shows the resulting μ_h for $h = \frac{1}{64}$, $h = \frac{1}{128}$ and $h = \frac{1}{256}$. It can be seen that it holds $\|\mu_h\|_\infty = \mathcal{O}(\frac{1}{h})$ and that the large values of $|\mu_h|$ are located around Γ_g in the lower right part of Ω. The plot of μ_h in y-direction for $x = 0.9$ is displayed in Figure 64. There it can be seen that the region where $|\mu_h|$ is large is of size h in y-direction. This indicates that the large part of μ_h converges to a measure concentrated on Γ_g. This justifies the

Figure 63: Cantilever beam: Lagrange multiplier μ_h for $h = \frac{1}{64}$, $h = \frac{1}{128}$ and $h = \frac{1}{256}$. Its modulus grows in the bottom right corner.

ansatz

$$\langle \mu, \eta \rangle = \int_\Omega \mu^d \eta + \int_{\Gamma_g} \mu^b \eta$$

for the Lagrange multiplier of the continuous problem. To compute an approximation of μ^b we assume that μ^b is constant and neglect μ^d. Recall that μ is discretized by $(\overline{\mu}_h)_i = \frac{\langle \mu, \chi_i \rangle}{m_i}$ and $\mu_h = \sum_i (\overline{\mu}_h)_i \chi_i$, where χ_i is the standard nodal basis function of the piecewise linear finite element space and $m_i = \int_\Omega \chi_i$ is its mass. Let now χ_i be the basis function corresponding to a mesh point x_i, which lies in the interior of Γ_g. We compute

$$\mu_h(x_i) = (\overline{\mu}_h)_i = \frac{\langle \mu, \chi_i \rangle}{m_i} = \frac{\int_{\Gamma_g} \mu^b \chi_i}{m_i} = \frac{\mu^b}{m_i} \int_{\Gamma_g} \chi_i = \frac{2\mu^b}{h},$$

where we used that $m_i = \frac{1}{2}h^2$ and $\int_{\Gamma_g} \chi_i = h$, which follows from simple calculations. We take the value $\mu_h(x_i)$ from the plot in Figure 63 for $x_i = \binom{0.9}{0}$. For all three values of h one thus computes $\mu^b \approx -13.6$. Recall the equation (141) in the KKT system:

$$\gamma \varepsilon \partial_\nu \overline{\varphi} + \chi_{\Gamma_g} \nabla_\varphi \boldsymbol{g}(\overline{\varphi}) \boldsymbol{u} + \mu^b = 0 \quad \text{ on } \partial\Omega$$

Note that the sign of μ^b is changed since we consider $\mu = \mu_2 - \mu_1$ here instead of $\mu_1 - \mu_2$. Using $\partial_\nu \overline{\varphi} = 0$, which is the case in this numerical example, we get

$$\nabla_\varphi \boldsymbol{g}(\overline{\varphi}) \boldsymbol{u} = -\mu^b \quad \text{ on } \Gamma_g.$$

From $\boldsymbol{u}(x_i) \approx (-0.0849, -0.2721)^T$ we can compute $\nabla_\varphi \boldsymbol{g}(\overline{\varphi}(x_i)) \boldsymbol{u}(x_i) \approx 13.6$, which coincides with $-\mu^b$ computed above as expected. We see that in this numerical example the boundary contribution μ^b has to be present in order to compensate the value of $\nabla_\varphi \boldsymbol{g}(\overline{\varphi}) \boldsymbol{u}$ on Γ_g.

Next we consider the bridge example, choosing the traction $\boldsymbol{g}$ independent of φ. We compute the discrete Lagrange multipliers as described above. The discrete μ_h is depicted in Figure 65 for $h = \frac{1}{32}$, $h = \frac{1}{256}$ and $h = \frac{1}{2048}$. It can be seen in the first picture, that $|\mu_h|$ exhibits high values at three points on the bottom boundary of Ω. When decreasing h it can be observed that the values at the outer points grow, whereas the value in the middle stays bounded. We therefore use a locally refined mesh for the finest h, which is chosen such that $h = \frac{1}{2048}$ in a neighborhood of the outer singularities and $h = \frac{1}{512}$ elsewhere. As in the cantilever beam experiment above it holds $|\mu_h|_\infty = \mathcal{O}(\frac{1}{h})$. But opposed to above, it does not hold that the region where the values of $|\mu_h|$ are high is of size h. This can be seen in Figure 66, where μ_h is plotted for the finest mesh through the singularity on the right hand side in x- and y-direction. In the picture it holds $h \approx 0.0005$ and $|\mu_h|$ clearly is high in an area of larger length scale. Thus, the large part of μ_h does not converge to a

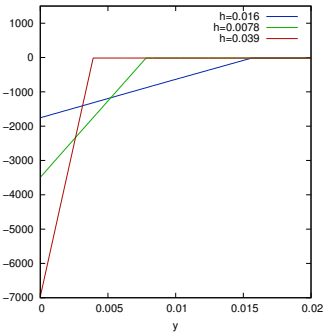


Figure 64: Lagrange multiplier μ_h in y-direction at $x = 0.9$. It can be observed that μ_h steepens to a Dirac measure as $h \to 0$.

Figure 65: Bridge: Lagrange multiplier μ_h for $h = \frac{1}{32}$, $h = \frac{1}{256}$ and $h = \frac{1}{2048}$.

boundary measure, but it holds

$$\langle \mu, \eta \rangle = \int_\Omega \mu^d \eta,$$

where μ^d features two singularities at the bottom boundary of Ω. These singularities are exactly at the boundary of Γ_D and can be interpreted as follows. From the equation (150) in the KKT system we get

$$\begin{aligned} -\gamma\varepsilon\Delta\overline{\varphi} + \frac{\gamma}{\varepsilon}\psi_0'(\overline{\varphi}) - \boldsymbol{C}'(\overline{\varphi})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{u}) - \lambda + \mu^d = 0 \quad &\text{in } \Omega \\ \partial_\nu \overline{\varphi} = 0 \quad &\text{on } \partial\Omega \end{aligned}$$

The equation $\partial_\nu \overline{\varphi} = 0$ can be confirmed in the experiment. Moreover, it holds that the displacement $\boldsymbol{u}$ vanishes in Γ_D by definition and increases rapidly at the inner boundaries of Γ_D. We refer to Figure 7, where the rapid change of $\boldsymbol{u}$ near $\partial\Gamma_D$ can be observed. This leads to singularities of $\mathcal{E}(\boldsymbol{u})$ in these points, which are compensated by the Lagrange multiplier μ^d.

We finally discuss the Lagrange multiplier μ_2 for the constraint $\varphi \leq 1$. We assume that μ_2 is a function, i.e.

$$\langle \mu_2, \eta \rangle = \int_\Omega \mu_2^d \eta.$$

Note that in the cantilever beam experiment above, it was μ_1 which included the boundary measure, but μ_2 still was a function. It turns out that μ_2 has a very simple structure if

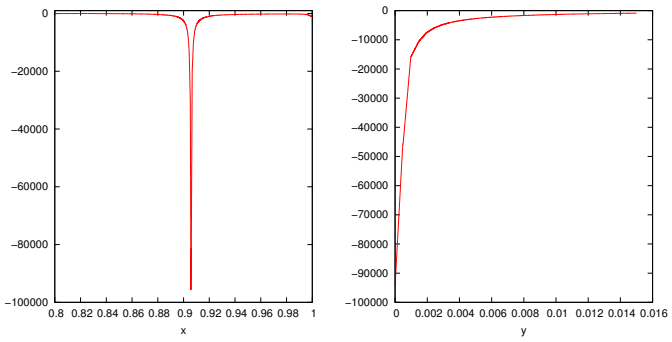


Figure 66: Lagrange multiplier μ_h ($h = \frac{1}{2048}$) in x- and y-direction around the singularity on the right hand side.

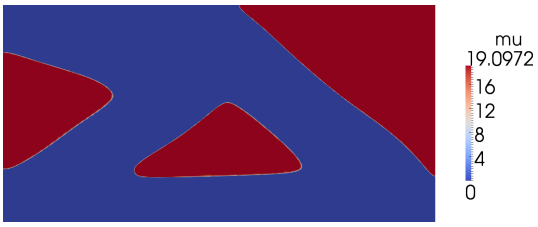


Figure 67: Lagrange multiplier μ_2 for the constraint $\varphi \leq 1$ is a scaled characteristic function.

the stiffness interpolation (110) is used. From the complementarity condition

$$\mu_2^d(1-\overline{\varphi}) = 0,$$

we get that $\mu_2^d = 0$ on $\{\overline{\varphi} < 1\}$. On the remaining subset $\{\overline{\varphi} = 1\}$ of Ω, we get from the equation (150) in the KKT system

$$-\gamma\varepsilon\Delta\overline{\varphi} + \frac{\gamma}{\varepsilon}\psi_0'(\overline{\varphi}) - \boldsymbol{C}'(\overline{\varphi})\mathcal{E}(\boldsymbol{u}) : \mathcal{E}(\boldsymbol{u}) - \lambda + \mu_2^d = 0 \quad \text{in } \Omega$$

It holds $\Delta\overline{\varphi} = 0$ in the interior of $\{\overline{\varphi} = 1\}$, as well as $\boldsymbol{C}'(\overline{\varphi}) = 0$ if $\boldsymbol{C}$ is chosen to interpolate quadratically between -1 and 1 with minimum in $\varphi = 1$ (void). We get

$$\mu_2^d \equiv \lambda - \frac{\gamma}{\varepsilon}\psi_0'(1) \quad \text{in the interior of } \{\overline{\varphi} = 1\}.$$

Thus we showed that μ_2^d is a scaled characteristic function for our special choice of the interpolation $\boldsymbol{C}(\varphi)$. This result can be used in the numerics. For instance to resolve the Lagrange multiplier μ_2^d an adaptive mesh is sufficient which has mesh points only on the transition area between $\{\varphi = 1\}$ and $\{\varphi < 1\}$. The multiplier μ_2^d could even be eliminated from the KKT system to reduce the number of unknowns. To validate the result numerically we look at μ_2 for the cantilever beam experiment with $\boldsymbol{g}$ independent of φ, $\gamma = 0.5$, $\varepsilon = 0.06$ and $\psi_0 = \frac{1}{2}(1-\varphi^2)$. We use a very fine adaptive mesh with $h_{max} = 2^{-8}$ and $h_{min} = 2^{-9}$. For the Lagrange multiplier of the mass constraint we get $\lambda \approx 10.76386$. Thus we calculate $\lambda - \frac{\gamma}{\varepsilon}\psi_0'(1) \approx 19.0972$, which is exactly the value observed in the experiment, see Figure 67.

From the experiments in this section we conclude that the Lagrange multipliers can have contributions from a boundary measure, that they can have singularities and that they can have jumps across a hypersurface. Thus an adequate mesh has to be chosen which can resolve these features. An adaptive mesh which is only fine on the interface is unsuitable in this case.

We also emphasize that the VMPT method is able to handle Lagrange multipliers that are not functions, since the VMPT method only considers the primal variable $\boldsymbol{\varphi}$. Thus we don't have to regularize the problem to get Lagrange multipliers in L^1. However, the problem of low regularity of Lagrange multipliers is transfered to the subproblem.

6.13.11 Counterexamples: Projected L^2-gradient and L^2-BFGS method

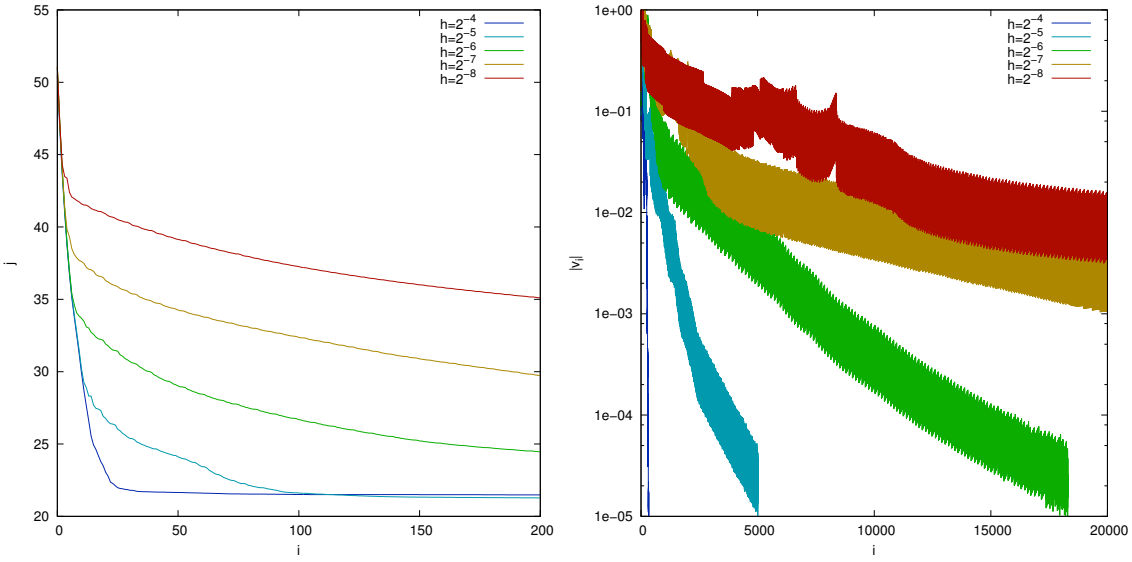
It is very popular to apply the projected gradient method with respect to the L^2 scalar product, since the projection on box constraints can then be computed pointwise, which is very cheap. In this section we apply the projected L^2-gradient method to the mean compliance problem and show that this method is not appropriate for the problem.

As already discussed at the end of Section 6.7, convergence of the L^2 method in function space cannot be shown by the methods developed in this thesis, since j is not differentiable in L^2. Thus the method may not be well defined in the continuous setting. However, for fixed discretization with parameter h the method is well defined, since in finite dimension the L^2 norm is equivalent to the H^1 norm and thus $\mathbb{X} = L^2(\Omega)$ can be used. However, the L^2-Lipschitz constant L_h of j_h' will depend on h and one can expect that $L_h \to \infty$ as $h \to 0$.
It can be shown that the L^2 projection onto an admissible set defined by box constraints, e.g. $\Phi_{ad} = \{\varphi \in L^2(\Omega) \mid \varphi_a \le \varphi \le \varphi_b \text{ a.e. in } \Omega\}$ with $\varphi_a, \varphi_b \in L^\infty(\Omega)$, coincides with the pointwise projection, i.e.

$$P_{L^2,\Phi_{ad}}(u)(x) = P_{[\varphi_a(x),\varphi_b(x)]}(u(x)) \quad \text{a.e. in } \Omega$$

see [Trö09]. In our case Φ_{ad} also contains the nonlocal constraint $\fint \varphi = \mathfrak{m}$, thus the L^2 projection cannot be performed pointwise and a projection problem has to be solved instead, which is given by the subproblem 18 with $a_k(x,y) = (x,y)_{L^2}$. We solve this subproblem by the PDAS method as described in Section 6.10. Moreover, we apply the update (280) for λ_k to improve the performance of the method.
For the numerical example we choose the cantilever beam experiment with parameters as in Example 6.83 with $\varepsilon = 0.03$ and $\gamma = 0.5$.
We perform the experiment for various equidistant meshes with mesh size varying from $h = 2^{-4}$ to $h = 2^{-8}$. The development of the cost $j(\varphi_k)$ for the first 200 iterations and the residual $\sqrt{\varepsilon\gamma}\|\nabla v_k\|_{L^2}$ is depicted in Figure 68. We take the scaled $H_0^1(\Omega)$-norm of v_k instead of the L^2-norm to be able to compare the results to the projected H^1-gradient method. In Table 19 the number of iterations needed to reach $\sqrt{\varepsilon\gamma}\|\nabla v_k\|_{L^2} \le tol = 10^{-5}$ is shown in the second column. The mark (e) indicates that the iteration number is extrapolated by assuming that $\sqrt{\varepsilon\gamma}\|\nabla v_k\|_{L^2}$ converges R-linearly to zero.

One clearly observes a mesh dependent behavior. On the coarsest mesh it takes only 323 iterations, whereas on the finest mesh 172621 iterations are needed. To analyze this behavior we include the plots for the iterate φ_{100} in the 100th step for varying mesh size in Figure 69. One can see that the smaller h gets, the further the iterate φ_{100} is away from the optimum.



(a) Cost functional (b) Residual

Figure 68: Mesh dependencies for the L^2-gradient method

(a) h=2^{-4} (b) h=2^{-5} (c) h=2^{-6} (d) h=2^{-7} (e) h=2^{-8}

Figure 69: L^2-gradient method: Iteration no. 100.

(a) h=2^{-4} (b) h=2^{-5} (c) h=2^{-6} (d) h=2^{-7} (e) h=2^{-8}

Figure 70: L^2-gradient method: Iterates for pseudo time $t_0 = 3.77$.

h	iter (for $tol = 10^{-5}$)	$\bar{\lambda}$	$\frac{\bar{\lambda}_{2h}}{\bar{\lambda}_h}$	$\frac{\bar{\lambda}}{h^2}$	$\frac{T}{6h^2}$	$iter_k/iter_{k-1}$
2^{-4}	323	$3.8 \cdot 10^{-2}$	-	9.65	1138	
2^{-5}	5015	$5.7 \cdot 10^{-3}$	6.65	5.80	4550	15.5
2^{-6}	18200	$1.3 \cdot 10^{-3}$	4.29	5.41	18200	3.6
2^{-7}	(e) 57630	$3.2 \cdot 10^{-4}$	4.12	5.25	72800	3.2
2^{-8}	(e) 172621	$8.0 \cdot 10^{-5}$	4.01	5.24	291198	3.0

Table 19: L^2-gradient method.

In the following we want to explain this mesh dependent behavior by considering the optimization method as a gradient flow. In Section 4.11.1 we showed that the projected L^2-gradient method is equivalent to a pseudo time stepping approach, resulting from an explicit time discretization of the L^2-gradient flow. The step length λ_k then corresponds to the time step size τ_k. The L^2-gradient flow is given by the variational inequality

$$(\partial_t \varphi, \eta - \varphi)_{L^2(\Omega)} + \left\langle j'(\varphi), \eta - \varphi \right\rangle \geq 0 \quad \forall t \geq 0,\ \eta \in \Phi_{ad}.$$

Inserting the derivative of j formally gives

$$(\partial_t \varphi - \varepsilon\gamma\Delta\varphi + \frac{\gamma}{\varepsilon}\psi_0'(\varphi) - C'(\varphi)\mathcal{E}(u) : \mathcal{E}(u), \eta - \varphi)_{L^2(\Omega)} \geq 0 \quad \forall t \geq 0,\ \eta \in \Phi_{ad},$$

which is a parabolic variational inequality. It is well known that for explicit time discretization of parabolic equations, the time step size τ has to fulfill the stability condition $\tau = \mathcal{O}(h^2)$. For example to solve the parabolic equation

$$\partial_t \varphi - \varepsilon\gamma\Delta\varphi = f$$

on an equidistant rectangular mesh in 2D one can show that $\tau \leq \frac{1}{4\varepsilon\gamma}h^2$ is needed for the explicit scheme the to be stable [RM67, ch. 8.7]. For the values of ε and γ used in the experiment this gives

$$\tau \leq 12.5h^2. \tag{292}$$

We check if this behavior can be observed in the projected L^2-gradient method. Therefor we consider the parameter λ_k, which corresponds to the pseudo time step size τ_k, and which is depicted in Figure 71 for the first 100 iterations and different mesh sizes. Recall that λ_k is determined by the update (280), which is based on the Armijo rule for α. It can be observed that the mean value of λ_k decreases as h decreases. The mean value $\bar{\lambda}$ of the values λ_k is shown in the third column of Table 19 as well as the ratio $\bar{\lambda}_{2h}/\bar{\lambda}_h$ in the fourth column. One clearly sees that $\bar{\lambda}$ approximately quarters as the mesh size h halves, thus $\bar{\lambda} = \mathcal{O}(h^2)$ as expected from the stability condition for the time step size τ. The constant can be read off the fifth column of Table 19. For small h it holds $\bar{\lambda} \approx 5.24h^2$, which is roughly half of the time step size in (292). A similar time step restriction is observed in [BC03]. Note that this scaling of λ is done automatically in the algorithm by the Armijo rule for α and the update (280) for λ. We also note that the values of λ_k are mesh dependent for the L^2-gradient method (see Figure 71), whereas the values of λ_k generated by the H^1-gradient method are mesh independent (cf. Section 6.13.3).
To illustrate that the iterates of the method really correspond to an L^2-gradient flow, we introduce the pseudo time t and denote the solution of the gradient flow by $\varphi(t)$. By the

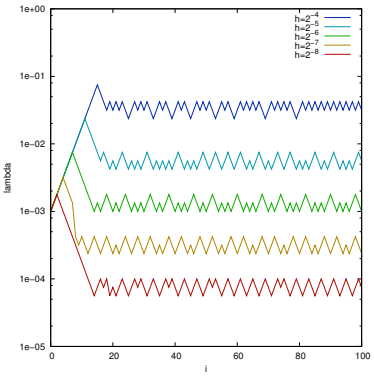


Figure 71: L^2-gradient method: λ_k.

approximation $t \approx k\bar{\lambda}$, we can consider the iterate φ_k as an approximation of $\varphi(k\bar{\lambda})$. We now fix the time $t_0 = 3.77$. The iterates approximating $\varphi(t_0)$ for different h are depicted in Figure 70, i.e. we plot φ_k for $k = \frac{t_0}{\bar{\lambda}}$. In contrast to Figure 69, the iterates look almost the same, which confirms our assumption.

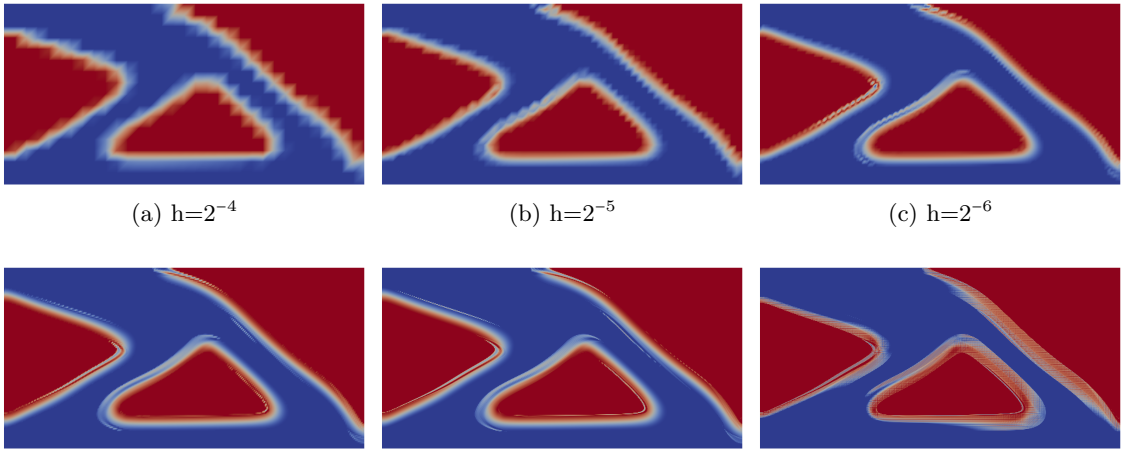
It is interesting to investigate which step in the global convergence proof of the VMPT method goes wrong when the L^2 inner product is used. It turns out that the projection type subproblem doesn't have a solution in Φ_{ad}, which is due to the lack of H^1-coercivity of the L^2-inner product. We show this by the following numerical experiment. We compute $y_h := \mathcal{P}_h(I_h\overline{\varphi})$ for varying mesh sizes h, where $\mathcal{P}_h$ denotes the solution operator of the discrete projection type subproblem with fixed $\lambda = 0.05$, I_h is the interpolation operator on the respective discrete space and $\overline{\varphi}$ is an approximation of the solution of the mean compliance problem. For smallest h we use an adaptive mesh, which is only fine on the interface to save memory. The other meshes are taken equidistantly. Note that the discrete projection type subproblem always has a unique solution for any $h > 0$. The functions y_h are depicted in Figure 72. One clearly observes that y_h oscillates on a length scale of the mesh size h. The vector y_h gives rise to the discrete search direction $v_h := y_h - I_h\overline{\varphi}$. The H_0^1 norm and the L^2 norm of v_h are listed in Table 20 together with the optimal value $g_h(y_h) = \min_{y \in S_h \cap \Phi_{ad}} g_h(y)$ of the cost functional g_h of the discrete projection type subproblem. These values are also plotted in Figure 73. We observe that $\|\nabla v_h\|_{L^2}$ and $\|v_h\|_{L^2}$ increase and $g_h(y_h)$ decreases as $h \to 0$. This indicates that $\|\nabla v_h\|_{L^2} \to \infty$ and $g_h(y_h) \to -\infty$ as $h \to 0$. Note that it has to hold $\|v_h\|_{L^2} \leq 2\sqrt{|\Omega|} = 2\sqrt{2}$ because of the constraint $-1 \leq y_h \leq 1$. Thus, the cost functional g of the continuous projection type subproblem is unbounded from below. Since the L^2 inner product is not H^1-coercive, we only get the estimate

$$g(y) \geq c_1\|y\|_{L^2}^2 - c_2\|y\|_{H^1} - c_3$$

instead of (24) for the projected H^1-gradient method, i.e.

$$g(y) \geq c_1\|y\|_{H^1}^2 - c_2\|y\|_{H^1} - c_3 \geq -c.$$

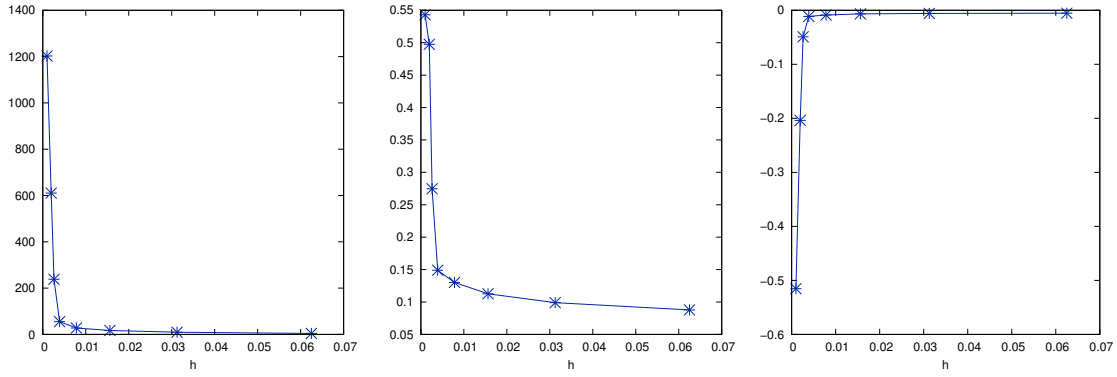
In the first estimate we now get $c_1\|y_h\|_{L^2}^2 - c_2\|y_h\|_{H^1} - c_3 \to -\infty$ as $h \to 0$. The growth of the H^1 norm is due to the oscillations of y_h. In the projected H^1-gradient method these oscillations are prevented by the smoothing term $\int_\Omega |\nabla(y - \varphi_k)|^2$ within g.



(a) h=2^{-4} (b) h=2^{-5} (c) h=2^{-6}

(d) h=2^{-7} (e) h=2^{-8} (f) h=2^{-9}

Figure 72: L^2-gradient method: Mesh dependent solution of the discrete projection type subproblem for varying mesh sizes.



(a) $\|\nabla v_h\|_{L^2}$ (b) $\|v_h\|_{L^2}$ (c) $g_h(y_h)$

Figure 73: L^2-gradient method: Mesh dependency of the data in Table 20.

h	$\|\nabla v_h\|_{L^2}$	$\|v_h\|_{L^2}$	$g_h(y_h)$
2^{-4}	4.34	0.088	-0.0053
2^{-5}	9.47	0.099	-0.0058
2^{-6}	16.99	0.113	-0.0067
2^{-7}	27.56	0.130	-0.0087
2^{-8}	55.19	0.149	-0.0114
$2^{-8.6}$	238.11	0.275	-0.0491
2^{-9}	610.64	0.497	-0.2038
2^{-10}	1202.50	0.543	-0.5153

Table 20: L^2-gradient method: Mesh dependent solution of the discrete projection type subproblem for varying mesh sizes.

We perform the same numerical experiments using a BFGS update of the L^2 inner product for a_k, as described in (282) except that we take $a_0 = (.,.)_{L^2}$ for the initialization of the BFGS method. The results are shown in Table 21 and Figure 74. The number of iterations is significantly less than for the L^2-gradient method. However, the same mesh dependent behavior can be observed. Also the average value for λ_k decreases as h decreases, although a bit less than for the L^2-gradient method. Thus it is very important how the BFGS method is initialized. Taking the H^1-inner product as initialization leads to a mesh independent method, whereas the L^2-inner product as initialization gives rise to a mesh dependent method.
We remark that in [PRW12], a BFGS method is used to solve a similar optimization problem and they use the identity matrix to initialize the BFGS iteration on the discrete level. They also report a mesh dependent behavior.

h	iter (for $tol = 10^{-5}$)	$\bar{\lambda}$	$\frac{\bar{\lambda}_{2h}}{\bar{\lambda}_h}$
2^{-4}	214	$8.8 \cdot 10^{-2}$	-
2^{-5}	593	$7.2 \cdot 10^{-3}$	12.2
2^{-6}	1085	$1.7 \cdot 10^{-3}$	4.2
2^{-7}	1830	$4.3 \cdot 10^{-4}$	4.0
2^{-8}	2817	$1.2 \cdot 10^{-4}$	3.6

Table 21: L^2-BFGS method for different h.

6.14 Numerical results for the compliant mechanism problem

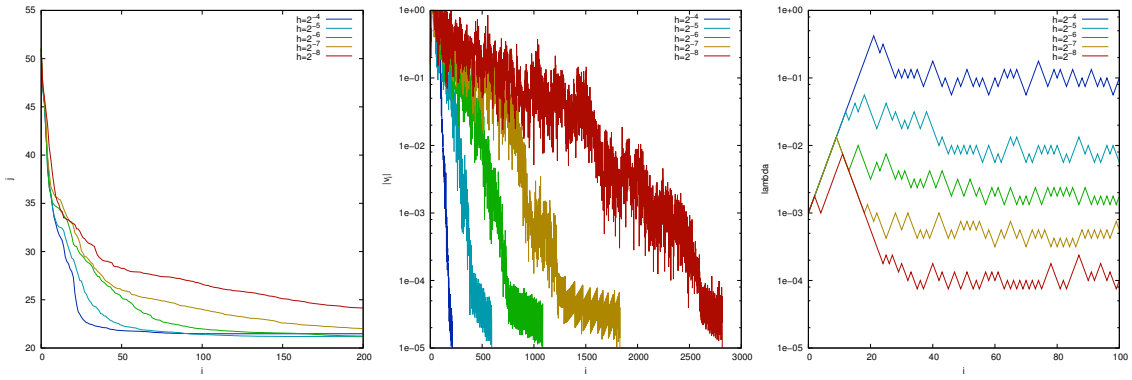
In this section we consider the design of a compliant mechanism. Here we use the cost functionals

$$F(\boldsymbol{\varphi}, \boldsymbol{u}) = \frac{1}{2} \int_\Omega c(x, \boldsymbol{\varphi}) |\boldsymbol{u} - \boldsymbol{u}_\Omega|^2 \tag{293}$$

and

$$F(\boldsymbol{\varphi}, \boldsymbol{u}) = - \int_{\Gamma_{out}} \boldsymbol{g}_{out} \cdot \boldsymbol{u}, \tag{294}$$

see (114) and (116), respectively. The goal is to find an elastic structure, which transfers a given input load $\boldsymbol{g}$ on Γ_g to some output displacement $\boldsymbol{u}_\Omega$ in $\mathrm{supp}(c)$ in case of the first



(a) CostFunctional (b) Residual (c) λ_k

Figure 74: L^2-BFGS method for different h.

cost functional and to some output displacement in direction of $\boldsymbol{g}_{out}$ in Γ_{out} in case of the second cost functional. The tracking type functional (293) is used e.g. in [BFGS14, AD14, AJT04, TNK10] and the linear functional (294) e.g. in [YINT10, Sig01, AKG94, BS03]. In the literature also other cost functionals are used for the compliant mechanism problem. For instance often a workpiece located at the output port is modeled as linear spring which exerts a reaction force on the mechanism. This is done e.g. in [Sig97]. There, the input and output ports are single points, thus $\boldsymbol{g}$ and $\boldsymbol{g}_{out}$ have to be seen as Dirac measures. We use functions as densities $\boldsymbol{g}$ and $\boldsymbol{g}_{out}$ here, since this gives rise to a well defined variational model. The cost functional in [Sig97] is the (negative) mechanical advantage, which is the quotient of output and input forces. If no gap between the mechanism and the elastic workpiece and no volume force is present, the negative mechanical advantage is given as

$$F(\boldsymbol{\varphi}, \boldsymbol{u}, \tilde{\boldsymbol{u}}) = -\frac{\int_{\Gamma_{out}} \boldsymbol{g}_{out} \cdot \boldsymbol{u}}{\int_{\Gamma_{out}} \boldsymbol{g}_{out} \cdot \tilde{\boldsymbol{u}} + 1/K} \tag{295}$$

where $K > 0$ is related to the stiffness of the workpiece and $\tilde{\boldsymbol{u}}$ is the displacement of the mechanism under the unit dummy load $\boldsymbol{g}_{out}$. The same cost functional is used in [WCWM05] and a very similar functional can be found in [YA01]. In [WCWM05] also the geometric advantage is considered, which is the quotient of output and input displacement. We see that for low stiffness K of the workpiece the mechanical advantage (295) divided by K approaches the linear cost functional (294). Thus, (294) models the situation without workpiece. It will turn out in the numerical experiments below that it is not a good idea to neglect the presence of a workpiece when a phase field model is used. The other extreme case of an infinitely stiff workpiece is considered for instance in [NFMK98, Sig97], where the cost functional (295) without the term $1/K$ is used. As alternative cost functional, a weighted sum of the numerator and denominator in (295) is taken in [NFMK98] instead of the quotient.

The linear functional (294) can also be seen as linearization of the tracking type functional $\frac{1}{2}\int_{\Gamma_{out}} |\boldsymbol{u} - \boldsymbol{u}_\Omega|^2$ in direction $-\boldsymbol{g}_{out}$ up to an additive constant.

Note that in the tracking type functional (293) the direction and magnitude of $\boldsymbol{u}$ is optimized, whereas in the linear functional (294) only the component of $\boldsymbol{u}$ along $\boldsymbol{g}_{out}$ is taken into account. This means that $\boldsymbol{u}$ is maximized in the direction of $\boldsymbol{g}_{out}$, but can also have components perpendicular to $\boldsymbol{g}_{out}$. This can lead to unwanted designs if e.g. the final displacement perpendicular to $\boldsymbol{g}_{out}$ dominates, see [BS03]. However, for the experiments

considered here, perpendicular displacements are not an issue.
For simplicity of implementation we use a distributed integral rather than a boundary integral for the functional (294) in the numerics, i.e. we use

$$F(\boldsymbol{\varphi}, \boldsymbol{u}) = -\int_{\Omega_{out}} \boldsymbol{g}_{out} \cdot \boldsymbol{u},$$

with $\Omega_{out} \subset \Omega$.
Important applications of the compliant mechanism problem are micro electro mechanical systems (MEMS) [BS03, Pet82], where the mechanisms are made e.g. of silicon at the length scale of micrometers. It is desirable that the mechanism is made of a single material without hinges. However, hinges appear naturally in the final design, since an elastic deformation of the mechanism implies a loss of energy, which is stored as elastic energy. Thus the energy throughput from input to output port can be maximized when the loss by an elastic deformation is low, which can be achieved if the mechanism behaves like a rigid mechanism by the use of hinges. Many methods to avoid hinges in the final design have been developed. Amongst others there are the MOLE [Pou03] or NoHinge constraints [BS03], the usage of nine node quad elements instead of four node quad elements [GMWG14], and the introduction of an additional soft material, which should replace the hinge [YA01, GMWG14]. In [LLC+08] a nonlocal energy of the interface is introduced, which penalizes bars with a smaller thickness than some constant d_{min}. Also the consideration of uncertainties in the input force can reduce the occurrence of hinges [AD14]. These methods can prevent hinges more or less successfully. Hinges are still a large problem in the design of compliant mechanisms [YKBS04], thus we will ignore the appearance of hinges since this would be out of the scope of this work.

The use of linearized elasticity may not always be adequate for this kind of problem since large displacements can appear and thus nonlinear effects like locking of bars, buckling etc. cannot be neglected [BS03, PRW12]. The obtained results using linearized elasticity have to be interpreted for infinitesimal small forces and displacements. For Sigmund's functional (294) this is no problem since the cost functional is linear in $\boldsymbol{u}$ and thus linear in the forces $\boldsymbol{f}$ and $\boldsymbol{g}$. Thus the obtained optimal structure for some $\boldsymbol{f}$ and $\boldsymbol{g}$ is also optimal for the rescaled forces $\tilde{\boldsymbol{f}} = \alpha \boldsymbol{f}$ and $\tilde{\boldsymbol{g}} = \alpha \boldsymbol{g}$, if the cost functional is rescaled adequately. This is also possible for the tracking type functional (293) if one additionally rescales $\boldsymbol{u}_\Omega$. See also the discussion in [NFMK98].

We will present numerical results for 2D and 3D compliant mechanisms. We want to emphasize here that also 2D structures are of practical interest, since flat mechanisms are manufactured in practice, see e.g. [GMWG14].
Most of the literature deals with compliant mechanisms consisting of a single material (and void). However, also multi-material compliant mechanisms are of current interest [GMWG14, WCWM05, YA01].

For the minimal compliance problem it is important to impose a mass constraint to avoid trivial solutions. This is not the case for the compliant mechanism problem since more mass does not necessarily result in better performance of the mechanism. Therefore we drop the mass constraint in some experiments. In this case we have to choose $\mathbb{X} = H^1(\Omega)^N$ in the analysis of the VMPT method. Note that we don't have a Poincaré inequality in the space $\mathbb{X}$ and thus we have to add some multiple of the L^2-inner product to the variable metrics to obtain H^1-coercivity, which is needed for **(A9)**. For instance

we use $a_k(\boldsymbol{p},\boldsymbol{v}) = \gamma\varepsilon\int_\Omega \nabla\boldsymbol{p}:\nabla\boldsymbol{v} + \gamma\varepsilon\delta\int_\Omega \boldsymbol{p}\cdot\boldsymbol{v}$ instead of $a_k(\boldsymbol{p},\boldsymbol{v}) = \gamma\varepsilon\int_\Omega \nabla\boldsymbol{p}:\nabla\boldsymbol{v}$, where we take $\delta = 0.01$ in the numerics. It would be also consistent to use the full H^1-norm in the stopping criterion of the VMPT method. However, this is not implemented here.

When using the linear functional one often gets optimal designs which are very flexible due to fine structures, and which involve very large displacements. To avoid such structures and to get more stable mechanisms we add the compliance of the structure as penalization term to the cost functional, which is also done in [AD14]. This has a similar effect to imposing a constraint on the displacement at the input port Γ_g, which is done in [Sig97, WCWM05]. The difference to the objective used in [NFMK98] is that we penalize the compliance with respect to the input force $\boldsymbol{g}$ rather than the compliance with respect to the reaction force $\boldsymbol{g}_{out}$. Moreover, in [AD14] also the volume of the structure is penalized in order to avoid disconnected parts. Here, we don't need such a volume penalization because disconnected parts that don't contribute to the performance of the mechanism are automatically removed due to the perimeter penalization and the Ginzburg-Landau energy, respectively.
Note that the results for the choice of the stiffness interpolation scheme discussed in Section 6.13.2 are not valid for the compliant mechanism problem, because a quadratic interpolation does not penalize intermediate values of $\boldsymbol{\varphi}$ here. However, in all experiments performed in this section a quadratic interpolation is used as described in (110).

Summarizing, we solve the following optimization problem,

$$\min F(\boldsymbol{\varphi},\boldsymbol{u}) + \gamma E(\boldsymbol{\varphi}) + \alpha\left(\int_\Omega \boldsymbol{f}(x,\boldsymbol{\varphi})\cdot\boldsymbol{u} + \int_{\Gamma_g}\boldsymbol{g}(x,\boldsymbol{\varphi})\cdot\boldsymbol{u}\right) \tag{296}$$

$$\int_\Omega \boldsymbol{C}(\boldsymbol{\varphi})\mathcal{E}(\boldsymbol{u}):\mathcal{E}(\boldsymbol{\xi}) = \int_\Omega \boldsymbol{f}(x,\boldsymbol{\varphi})\cdot\boldsymbol{\xi} + \int_{\Gamma_g}\boldsymbol{g}(x,\boldsymbol{\varphi})\cdot\boldsymbol{\xi} \quad \forall\boldsymbol{\xi}\in H^1_D \tag{297}$$

$$\boldsymbol{\varphi} \geq 0$$

$$\sum_{i=1}^{N}\varphi_i = 1$$

$$\beta \fint \boldsymbol{\varphi} = \beta\mathfrak{m},$$

where E is the Ginzburg-Landau energy as before, F can be (293) or (294) and $\alpha > 0$ is the weight of the compliance penalization. By setting $\beta = 0$ we can disable the mass constraint. We note that this problem fits in the abstract framework of Section 6.1.1. In particular we get global convergence of the VMPT method in $H^1\cap L^\infty$ and Γ-convergence of the cost functional in $L^1(\Omega)$ if $\boldsymbol{f}$ and $\boldsymbol{g}$ are independent of $\boldsymbol{\varphi}$, see Theorem 6.22.

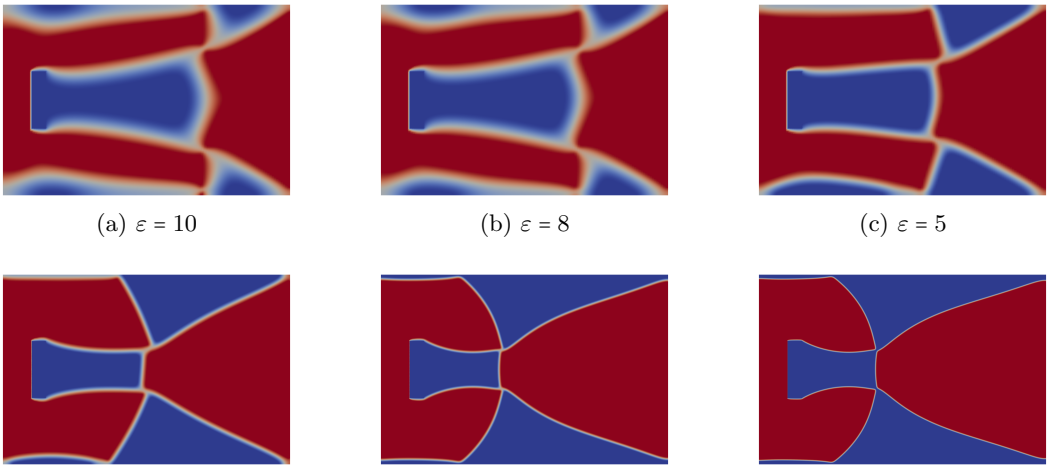
The goals of this section are to check the performance of the VMPT method while comparing various metrics and the two different cost functionals (293) and (294). Additional goals are the evaluation of the behavior of the solutions for $\varepsilon\to 0$ and the comparison of our solutions to the designs obtained in the literature.

If not mentioned otherwise, the VMPT method is stopped if the stopping criterion $\sqrt{\gamma\varepsilon}\|\nabla v_k\|_{L^2} < tol$ is fulfilled. Note that this stopping criterion is rigorous and contrary to many other numerical methods we are thereby able to measure the optimality of a tentative minimizer. In certain cases we also stop the method earlier if the phase field changes only very slowly. Also note that all calculated solutions are in general only local

minimizers and that there is no way to tell if the minimizer is global.
Moreover, we use the potential $\psi_0(\varphi) = \frac{1}{2}(1-\varphi^2)$ for 2 phases and $\psi_0(\boldsymbol{\varphi}) = -\frac{1}{2}\boldsymbol{\varphi}^T A\boldsymbol{\varphi}$ for multiple phases, where we choose $A = \begin{pmatrix} 0 & -0.1 & -1 \\ -0.1 & 0 & -1 \\ -1 & -1 & 0 \end{pmatrix}$, which ensures that in triple junctions the angle for the void phase is larger than the angles for the material phases, see also Section 6.12.

In the numerics, often a trivial solution is found, which mainly minimizes the Ginzburg-Landau energy and not the functional F. For instance void or material is placed everywhere if no mass constraint is used, or the interface is a straight line or part of a circle if a mass constraint is used. In the case $\alpha > 0$ there are also trivial solutions which mainly minimize the compliance of the structure rather than F. To avoid such trivial solutions we usually start with small weights γ and α for the first few iterations and then increase them. In most experiments we also use higher values for ε and h at the beginning of the iteration and decrease them during the optimization process. In the problem description we only state the final values of the parameters. If not mentioned otherwise we start the iteration with the homogeneous mixture $\boldsymbol{\varphi} \equiv \mathfrak{m}$.

The first experiment we consider is a crunching mechanism or push-clamp [Sig97, AKG94, YA01] with 2 phases (material and void). We set $\Omega = (-100, 100) \times (-66, 66)$, $\Gamma_D = \{x_1 = -100\}$, $\Gamma_g = \{x_1 > 90\} \cap \{|x_2| = 66\}$, $\boldsymbol{g}(x) = (0, -\mathrm{sgn}(x_2)9)^T$ and $\boldsymbol{f} \equiv \mathbf{0}$. The Lamé constants of the material are $\mu = 1071$ and $\lambda = 4285$ and the void is modelled with a 1000 times lower stiffness. The tracking type cost functional is used with $\boldsymbol{u}_\Omega \equiv (-10, 0)^T$ and $c = \chi_{\Omega_{obs}}$ where $\Omega_{obs} = (-80, -70) \times (-20, 20)$. We add a constraint, which prescribes material in Ω_{obs} and void in $(-100, -81) \times (-20, 20)$. We set $\beta = 1$ and $\mathfrak{m} = 0.32$ (32% material), $\alpha = 0$, $\gamma = 0.5$. The VMPT method is used to compute a local minimizer with $tol = 10^{-5}$. We perform a nested iteration in ε, i.e. we start with $\varepsilon = 10$ and decrease ε slowly until $\varepsilon = 0.5$ is reached. For each fixed ε we run the VMPT method until the stopping criterion is fulfilled. Together with ε we also refine the mesh on the interface such that there are at least 5 mesh points across the interface. We compare the H^1 metric to the H^1-BFGS metric. Table 22 shows the number of iterations needed by each method, in Figure 75 the corresponding local minima are depicted. Note that both methods converge to the same local minima. For $\varepsilon = 5$ and $\varepsilon = 2$ the mechanisms are not symmetric since we use an unsymmetric mesh. However, when the mesh is refined for smaller ε the solution becomes symmetric. We observe that only for the smallest value of ε the mechanism is connected. For larger ε hinges are present. Thus in this experiment one has to choose ε very small in order to get the final topology.
We also see that the mechanisms for largest and smallest ε differ considerably, not only in the thickness of the interface. This is because the thickness of the interface influences the force transfer within the mechanism. For the mean compliance problem the thickness of the interface has less influence on the compliance of the structure. As a consequence, we suggest to start with a rather small value of ε, since solutions for larger ε may not be a good approximation for the solution with small ε. Furthermore the experiment gives a hint that the optimal designs converge as $\varepsilon \to 0$ to a 0-1-design as it is suggested by the Γ-convergence result. This is not always the case as will be seen in a later experiment.
It can be seen in Table 22 that the H^1-BFGS method is much faster than the H^1-gradient method. Note that a single iteration using the H^1-BFGS metric is on average less then twice as expensive in terms of CPU time as an iteration using the H^1-metric. For $\varepsilon = 2$ there is more than a factor of 100 between the iteration numbers. Hence the H^1-BFGS metric gives rise to an efficient method, which allows us to solve the compliant mechanism



(a) $\varepsilon = 10$ (b) $\varepsilon = 8$ (c) $\varepsilon = 5$

(d) $\varepsilon = 2$ (e) $\varepsilon = 1$ (f) $\varepsilon = 0.5$

Figure 75: Solution for the Cruncher experiment for different ε.

problem in moderate time. Since the projected H^1-gradient method needs many iterations, we didn't compute the minimum for $\varepsilon = 0.5$. Also for the other experiments in this section the projected H^1-gradient method is very slow and thus we don't consider this method in the following. It is well known that the compliant mechanism problem is much harder to solve than the mean compliance problem. Recall that for the mean compliance problem the BFGS method was at most 5 times faster than the gradient method, whereas in the current experiment the difference is much severer.

ε	proj. H^1-grad. iter.	H^1-BFGS iter.
10	16551	1527
8	551	196
5	105948	765
2	174575	1432
1	181076	3344
0.5	-	1454

Table 22: Nested iteration in ε for the Cruncher experiment.

Next we repeat the Cruncher experiment using the linear functional (294) instead of the tracking type functional. We set $\boldsymbol{g}_{out} \equiv (-1, 0)^T$ and $\Gamma_{out} = \Omega_{obs}$. As mentioned before we minimize a distributed integral rather than a boundary integral. Moreover, we perform the computation only in the upper half of Ω, since we only consider symmetric solutions to save computation time. More precisely we set $\Omega = (-100, 100) \times (0, 66)$ and impose in addition the Dirichlet boundary condition $u_2 = 0$ on $\{x_2 = 0\}$, which we put into the space H^1_D. We use the parameters $\gamma = 0.05$, $\varepsilon = 2$, $\alpha = 0.0005$, $\beta = 0$ and $\boldsymbol{g}(x) = (0, -\mathrm{sgn}(x_2)18)^T$ together with an adaptive mesh using $h_{max} = 66/26$ and $h_{min} = h_{max}/4$. The remaining parameters are as above. First we compute a solution using the H^1-BFGS metric. The final design together with the deformed mechanism is shown in Figure 76. We also compute

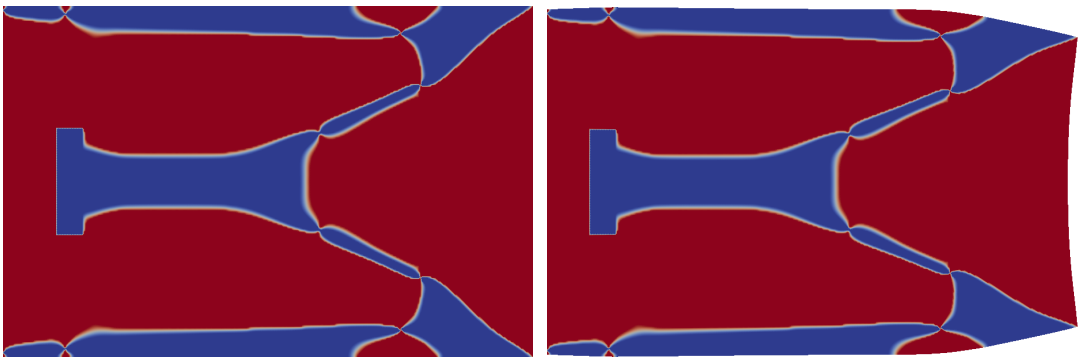


Figure 76: H^1-BFGS solution using the linear functional.

a solution using the metric (167) including second order information, i.e.

$$a_k(\boldsymbol{v}_1, \boldsymbol{v}_2) = \gamma\varepsilon \int_\Omega \nabla \boldsymbol{v}_1 : \nabla \boldsymbol{v}_2 + \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta u}_1) : \mathcal{E}(\boldsymbol{\delta u}_2) + \int_\Omega \boldsymbol{C}(\boldsymbol{\varphi}_k)\mathcal{E}(\boldsymbol{\delta p}_1) : \mathcal{E}(\boldsymbol{\delta p}_2),$$

with added L^2 inner product due to the absence of a mass constraint as discussed above. Recall that $\boldsymbol{\delta u}_i$ and $\boldsymbol{\delta p}_i$ are the solutions of the linearized state and linearized adjoint equations, respectively, in direction $\boldsymbol{v}_i$, $i = 1, 2$. Since the solution of the VMPT subproblem using this metric is very expensive, we switch to the cheaper H^1-BFGS metric as soon as the final topology is formed. The obtained solution is shown in Figure 77. Note that for both solutions only 20% of the deformation is depicted to stay in the regime of linear elasticity. This is justified by the scaling considerations above. In Table 23 the number of iterations, the CPU time and the function values at the minimizer $\boldsymbol{\varphi}^*$ are shown. Note that we don't use the stopping criterion here, but we stop the iteration when the phase field doesn't change visibly. Moreover we perform a nested iteration in ε and h similar to the experiment using the tracking type functional. Thus the iteration numbers may not be completely comparable. However, a rough qualitative comparison should be possible. In this experiment the second order metric needs about twice the iteration numbers of the H^1-BFGS metric. The CPU time is more than tripled, since the second order metric is more expensive than the H^1-BFGS metric. It can be seen that here the H^1-BFGS metric computed a minimizer with lower energy than the second order metric. Also the Ginzburg-Landau energy is lower.
However, the final cost values differ only by 0.5% and no difference in the deformation at Ω_{obs} can be seen. This is the only experiment where we observe this behavior. In all other experiments the minimum using the second order metric is lower. For both minima the displacement in Ω_{obs} perpendicular to $\boldsymbol{g}_{out}$ is very small. This is mainly due to the Dirichlet boundary condition $u_2 = 0$ on $\{x_2 = 0\}$.
We see that the thickness of the interface is not constant throughout Ω. Especially near the hinges the interface is very thin. This is because the weight γ of the Ginzburg-Landau energy is taken quite small. For the solution in Figure 75 the interface thickness is constant, since a much higher value for γ is taken. Due to the same reason the number of hinges is higher for the solutions in Figure 76 and Figure 77. However, the solutions obtained for smaller γ are more similar to the results obtained in [WCWM05] by a level set method and in [Sig97] using a sequential linear programming method on the discrete problem. The solution obtained in [AKG94] by a homogenization method together with an optimality criteria method is still very different.

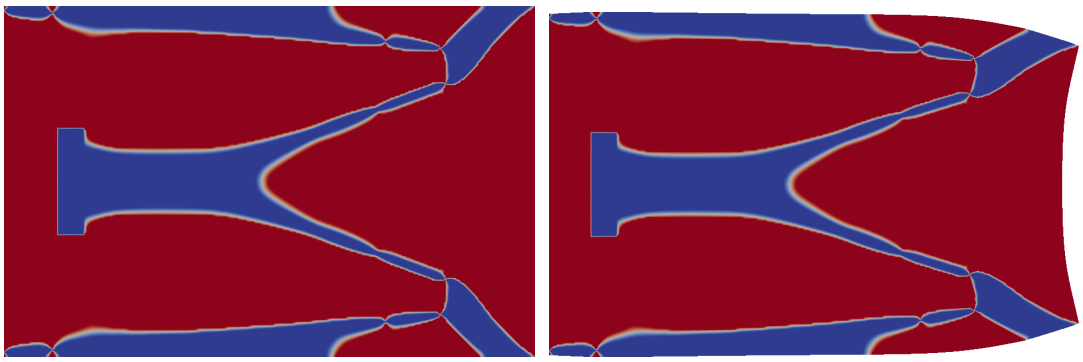


Figure 77: Solution using second order metric and the linear functional.

inner product	iterations	CPU time	$j(\boldsymbol{\varphi}^*)$	$\gamma E(\boldsymbol{\varphi}^*)$	$j(\boldsymbol{\varphi}^*) - \gamma E(\boldsymbol{\varphi}^*)$
H^1-BFGS	599	38min	-3750	92	-3842
second order VMPT	1169	124min	-3732	104	-3836

Table 23: Comparison of two different inner products for the Cruncher experiment.

As next experiment we consider a push-gripper, which is used by many authors as benchmark problem [Sig97, YINT10, WCWM05, YKBS04, AD14]. The objective is to transfer a force coming from the left hand side of the mechanism to the right hand side in order to close two given jaws. The geometry of the problem is taken from [AD14]: We set $\Omega = (0,1) \times (-0.5, 0.5)$, $\Gamma_D = \{x_1 = 0\} \cap \{0.4 \leq |x_2| \leq 0.5\}$, $\Gamma_g = \{x_1 = 0\} \cap \{|x_2| \leq 0.05\}$, $\boldsymbol{g} \equiv (0.1, 0)^T$ and $\boldsymbol{f} \equiv \boldsymbol{0}$. As Lamé constants of the material we take $\mu = 1$ and $\lambda = 1$ and the void is modelled with a 10000 times lower stiffness. We use the linear cost functional with $\Gamma_{out} = \Omega_{obs} = \{0.8 \leq x_1 \leq 1\} \cap \{0.05 \leq |x_2| \leq 0.1\}$ and $\boldsymbol{g}_{out}(x) = \operatorname{sgn}(x_2)(0, -1)^T$. Material is prescribed in Ω_{obs} and in the region $\{0 \leq x_1 \leq 0.05\} \cap \{0.4 \leq |x_2| \leq 0.5\}$ near the Dirichlet boundary, and void in the region $\{0.8 \leq x_1 \leq 1\} \cap \{|x_2| \leq 0.025\}$ between the jaws. We set $\beta = 0$, $\alpha = 0.5$, $\gamma = 0.0002$ and $\varepsilon = 0.005$. We use an adaptive mesh with $h_{max} = 1/80$ on the bulk and $h_{min} = 1/320$ on the interface. Again the computation is restricted to the upper half of the design domain due to symmetry and we stop the VMPT method if the phase field does not change anymore. First we compare the H^1-BFGS metric with the second order metric (167) as in the previous Cruncher experiment. The results are shown in Table 24 and the final designs are depicted in Figure 78 and Figure 79, respectively, where 1% of the deformation is shown. As opposed to the Cruncher experiment the solutions differ strongly. Also the cost functional value $j(\boldsymbol{\varphi}^*)$ is 13% lower for the second order VMPT method, whereas the Ginzburg-Landau energy is almost identical for both solutions.

inner product	iterations	$j(\boldsymbol{\varphi}^*)$	$\gamma E(\boldsymbol{\varphi}^*)$	$j(\boldsymbol{\varphi}^*) - \gamma E(\boldsymbol{\varphi}^*)$
H^1-BFGS	1619	-0.01316	0.001109	-0.01427
second order VMPT	1169	-0.01487	0.001104	-0.01597

Table 24: Comparison of two different inner products for the Gripper experiment.

Next we compute another gripping mechanism with slightly different geometry. We set $\Omega = (-1,1) \times (-1,1)$, $\Gamma_D = \{x_1 = -1\} \cap \{0.8 \leq |x_2| \leq 1\}$, $\Gamma_g = \{x_1 = -1\} \cap \{|x_2| \leq 0.1\}$, $\boldsymbol{g} \equiv (0.005, 0)^T$ and $\boldsymbol{f} \equiv \boldsymbol{0}$. As Lamé constants of the material we take $\mu = 5$ and $\lambda = 5$ and the void is modelled with a 10000 times lower stiffness. We use the linear cost functional

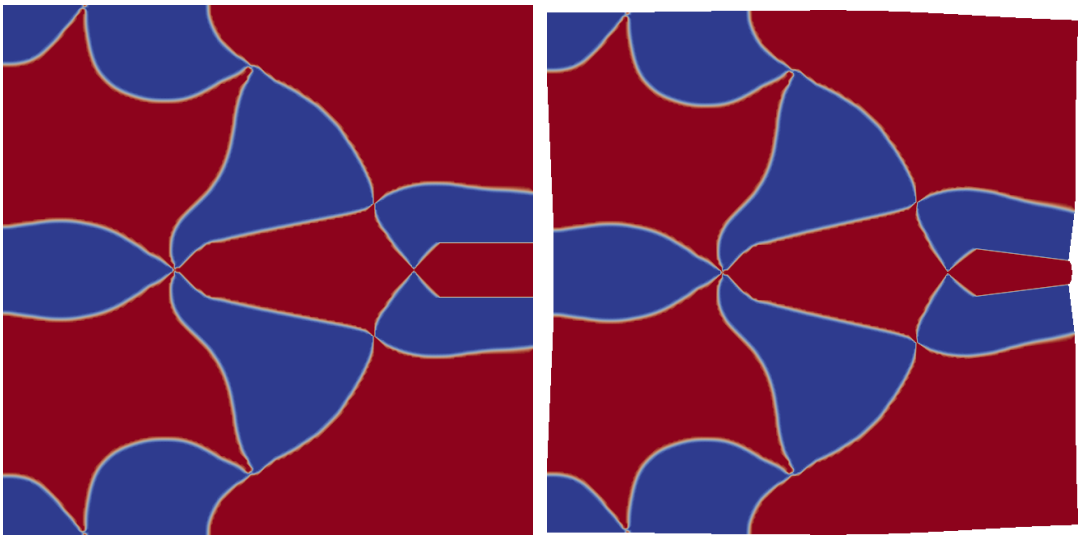


Figure 78: H^1-BFGS solution using the linear functional.

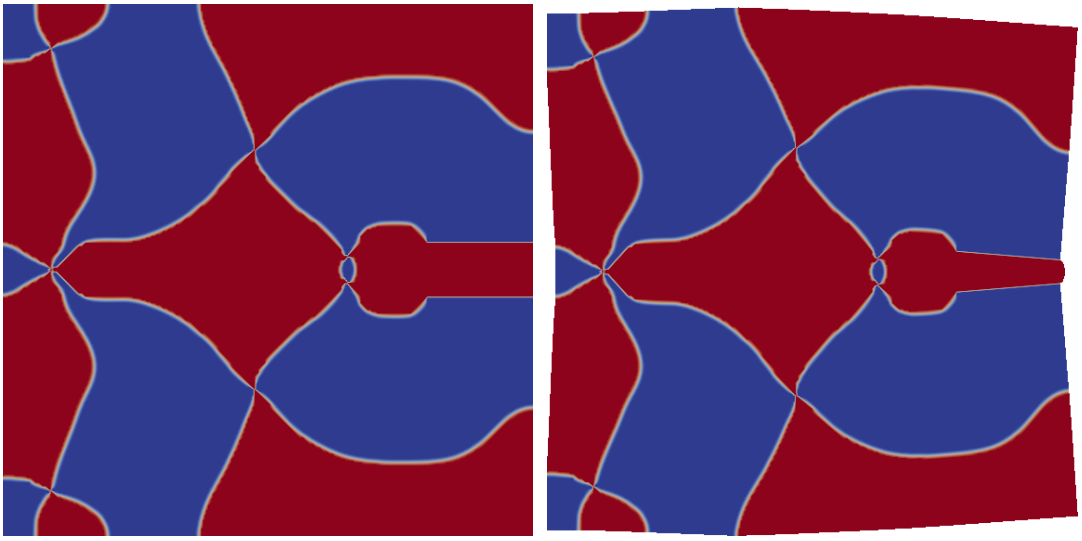


Figure 79: Solution using the second order metric and the linear functional.

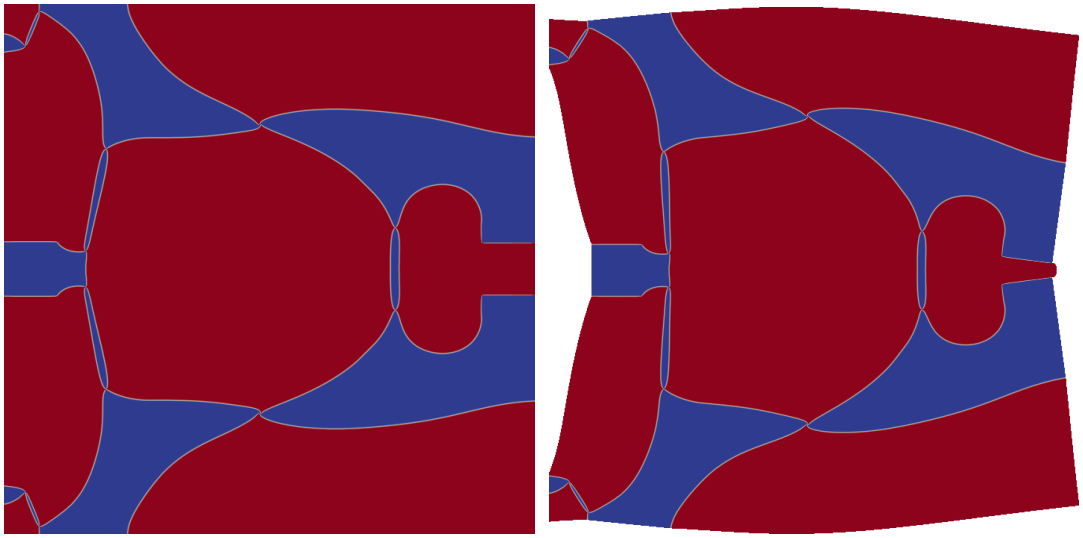


Figure 80: Linear functional, H^1-BFGS with mass constraint.

with $\Gamma_{out} = \Omega_{obs} = \{0.8 \leq x_1 \leq 1\} \cap \{0.1 \leq |x_2| \leq 0.17\}$ and $\boldsymbol{g}_{out}(x) = \operatorname{sgn}(x_2)(0,-1)^T$. Material is prescribed in Ω_{obs} and in the region $\{-1 \leq x_1 \leq -0.8\} \cap \{-0.1 \leq |x_2| \leq 0.1\}$ near the input port, and void in the region $\{0.6 \leq x_1 \leq 1\} \cap \{|x_2| \leq 0.09\}$ between the jaws. We set $\beta = 1$ with $\mathfrak{m} = 0.5$ (i.e. 25% material), $\alpha = 0$, $\gamma = 0.00005$ and $\varepsilon = 0.003$. We use an adaptive mesh with $h_{max} = 1/64$ on the bulk and $h_{min} = 1/1024$ on the interface. Again the computation is restricted to the upper half of the design domain due to symmetry and we stop the VMPT method if the phase field does not change anymore. We note that the final residual was $\sqrt{\gamma\varepsilon}\|\nabla v_k\|_{L^2} \approx 10^{-6}$, whereas the initial residual was of the magnitude of 1. The main difference to the Gripper experiment above is that the length of the jaws is halved, material is prescribed around the input force instead of the Dirichlet domain and that a mass constraint is used. The final design using the H^1-BFGS method is depicted in Figure 80, where 40% of the deformation is shown. We see that the VMPT method is able to compute a reasonable mechanism using different geometries and constraints. The obtained local minima are quite different. They also differ from the designs obtained in the literature.

We present also a result with three phases. We use the geometry of the first gripper experiment with the Lamé coefficients $\mu = 1$, $\lambda = 1$ for the hard material, $\mu = 0.5$, $\lambda = 0.5$ for the soft material and $\mu = 10^{-4}$, $\lambda = 10^{-4}$ for the void. Moreover we take $\alpha = 0.04$, $\beta = 1$ with $\mathfrak{m} = (0.1, 0.1, 0.8)$, $\gamma = 0.002$ and $\varepsilon = 0.002$. The H^1-BFGS method is used with $tol = 10^{-5}$ for the stopping criterion. Figure 81 shows the final design and 0.4% of the displacement, where blue corresponds to hard material, gray to soft material and red to void. This solution is rather different from the multiphase solutions e.g. in [WCWM05, GMWG14], where the soft material is placed around the boundary of the hard material, rather than in large areas. This is probably due to the penalization of the Ginzburg-Landau energy, which prefers areas with short boundary.

A difficulty in the construction of compliant mechanisms using a phase field model is the occurrence of thin bars consisting of interface. We give an example where such bars can be observed. We take the geometry of the first Gripper experiment with Lamé constants $\mu = 5$ and $\lambda = 5$ for the material and $\mu = 0.0005$ and $\lambda = 0.0005$ for the void. This time

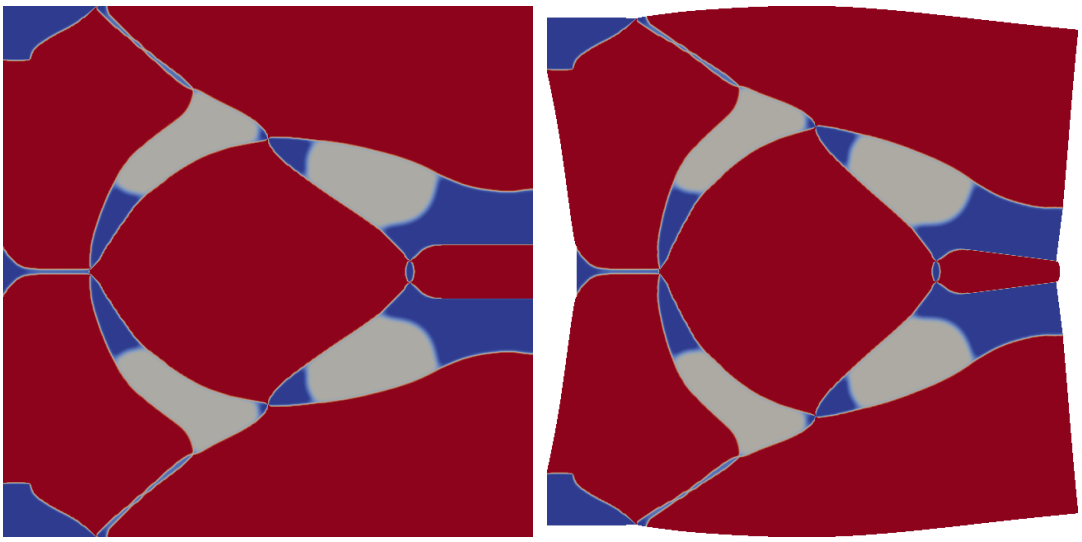
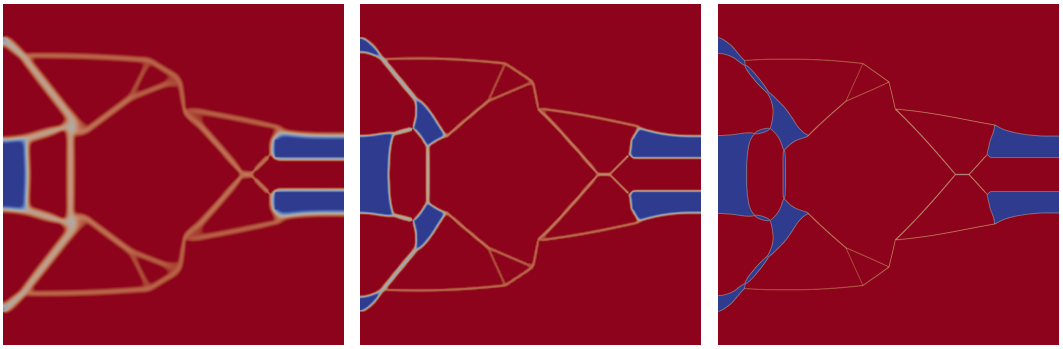


Figure 81: Linear functional, with mass constraint, 3 phases.

we take the tracking type cost functional with $c = \chi_{\Omega_{obs}}$ and $\boldsymbol{u}_\Omega(x) = \operatorname{sgn}(x_2)(0,-0.05)^T$. No material is prescribed around the Dirichlet boundary. Further we take $\alpha = 0.01$, $\beta = 0$ and $\gamma = 0.00001$. Here we also add a mass penalization with weight 10^{-5}. We use the H^1-BFGS method to compute a local minimizer for varying ε with $tol = 10^{-6}$. The result is shown in Figure 82. The thin bars in the middle of the mechanism are not part of the void or material phase, but φ has values in $(-1,1)$. One would expect that for ε small enough φ attains the value -1 (material phase) within the thin bars. However, this is not the case. In the plot on the left hand side of Figure 83 the value of φ across one of the thin bars is shown for $\varepsilon = 0.01$ and $\varepsilon = 0.001$. For the larger ε, the bar is thicker and φ attains values in $[0.5,1]$, where $\varphi = 1$ corresponds to the void-phase. As ε decreases the thickness of the bars also decreases and the values of φ move towards $\varphi = -1$. The latter happens so slowly that the overall L^1-distance of the thin bars to the void phase decreases. On the right hand side of Figure 83 this distance is shown. More precisely we plot $\varepsilon \mapsto \int_{\Omega_{bar}} |1-\varphi_\varepsilon|$ with $\Omega_{bar} = (0.3,0.7)\times(0,0.5)$ and φ_ε are the respective minimizers. We can suppose that $\int_{\Omega_{bar}} |1-\varphi_\varepsilon| \to 0$ as $\varepsilon \to 0$. Thus the L^1- limit φ_0 of φ_ε defines a mechanism for which the input and output ports are not connected. From the Γ-convergence result we get that the limit of a sequence of (global) minimizers is a (global) minimizer of the Γ-limit and that the function values converge, see Theorem 6.20. However, φ_0 is certainly not a minimizer of the sharp interface problem, since the disconnected structure cannot transfer the input force to the output port. The difficulty is probably that φ_ε is only a local minimizer of j_ε and thus nothing can be said about its L^1-limit. We conclude that the solutions obtained here by the phase field relaxation don't approximate a solution of the sharp interface problem and are thus unwanted. We also note that using a lower stiffness for the void phase leads to even thinner bars.

We want to analyze the occurrence of these thin bars from the previous experiment more closely. Therefore we perform the following experiment, where the development of the bars can be observed very well. We take the geometry and parameters of the second gripper setup (with short jaws) and set $\boldsymbol{g} = (0.18,0)^T$, $\alpha = 0$, $\beta = 0$, $\gamma = 0.0005$ and $\varepsilon = 0.02$. We use the tracking type cost functional with $c = 10000\chi_{\Omega_{obs}}$ and $\boldsymbol{u}_\Omega(x) = (0,-\operatorname{sgn}(x_2)0.02)^T$. To solve the optimization problem we perform an H^1-BFGS iteration on an equidistant



(a) $\varepsilon = 0.01$ (b) $\varepsilon = 0.004$ (c) $\varepsilon = 0.001$

Figure 82: Gripper for different ε. Tracking type functional.

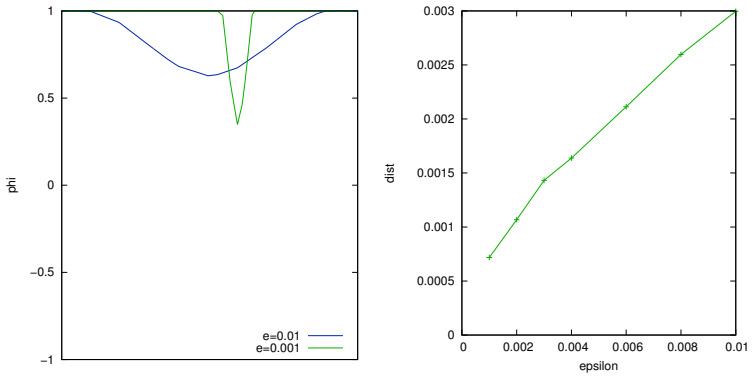
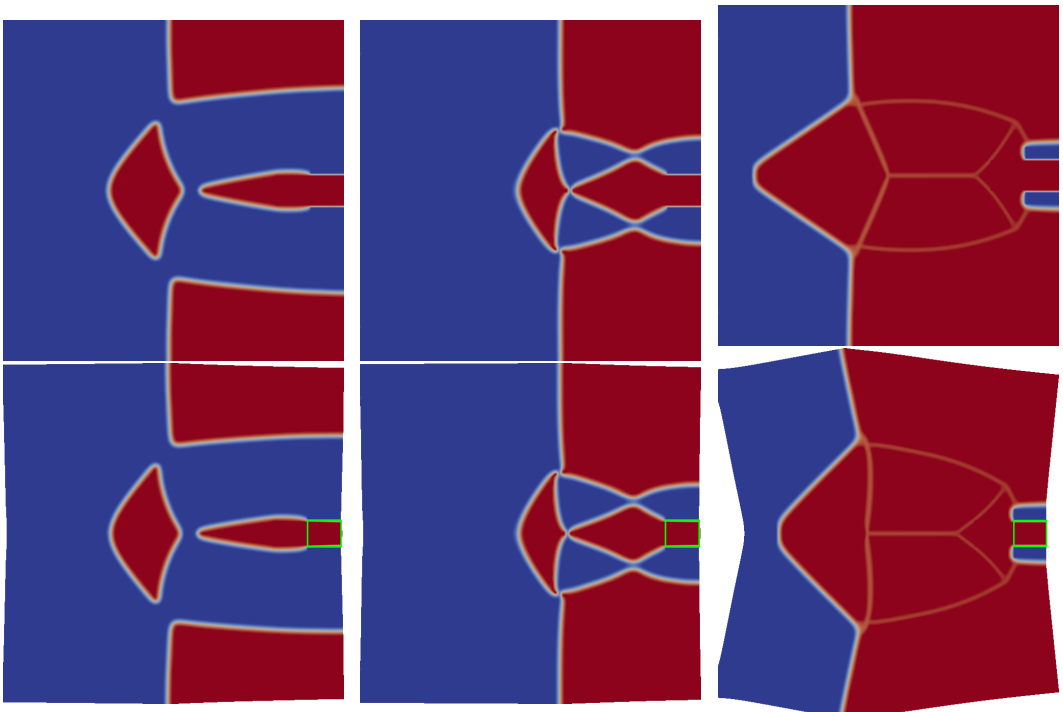


Figure 83: Plot of φ across the interface (left) and L^1 distance of the thin structure in Figure 82 to the void phase (right).

mesh with $h = 1/128$ and use $tol = 2 \cdot 10^{-7}$ for the stopping criterion. The phase field for different iteration numbers during the optimization process is shown in Figure 84, the corresponding development of the scaled Ginzburg-Landau energy γE and the tracking type energy $j - \gamma E$ is plotted in Figure 85. First of all we observe that there is no visible difference in the displacement at Ω_{obs} for all three iteration numbers. The bottom row of Figure 84 shows the corresponding displaced mechanisms, where we included a green colored reference box in the graphics, upon which the jaws should close. Also note that the value 0.0004 of the tracking term at iteration 940 already is very low compared to the initial tracking value of 1.5. Iterate 940 looks like a good solution, since it deforms as demanded, no thin bars and even no hinges are present, and the mechanism is rather stiff, which can be seen from the low displacement at Γ_g. The tracking term is locally minimal at this iterate. Then the Ginzburg-Landau energy decreases and the tracking term increases until hinges appear as can be seen in iterate 3120. As soon as the hinges form the tracking term drops to almost 0. At iteration 3120 a pinching of the hinges near the jaws occurs, which can be seen in the peak in the Ginzburg-Landau energy. After the pinching the thin bars form while the Ginzburg-Landau energy decreases rapidly. During this process the tracking term stays almost 0. The local minimum is found at iteration 29579.
From this experiment we learn that the thin bars occur since the Ginzburg-Landau energy is about 35% lower for the mechanism with bars compared to the mechanism at iteration 940 without bars. Also the tracking term is lower for the mechanism with thin bars. However, the dominant term is the Ginzburg-Landau energy in this experiment, since the value of the weight γ is relatively large. From an engineering point of view, one would rather prefer the mechanism at iterate 940. Anyhow, in the model used here, the energy of the mechanisms at iterate 3120 and 29579 is much lower.
The desired displacement $\boldsymbol{u}_\Omega$ can here be achieved by a mechanism with thin bars since no reaction force is present. There is no workpiece between the jaws which resists the mechanism. Thus only little energy is necessary to close the bars. This consideration also complies with the observation that less stiffness of the void phase leads to even thinner bars, since in this case less energy is needed to compress the void between the jaws.

A possibility to overcome this issue is perhaps to use the cost functional (295), which takes the presence of a reaction force into account. However, for simplicity and as a first experiment we still use the tracking functional (293), but prescribe a soft material between the jaws instead of void. The soft material models the presence of a workpiece between the jaws and gives rise to a reaction force, which resists the movement of the jaws. Therefor we switch to the multiphase setting with three phases, where we take the Lamé constants $\mu = \lambda = 5$ for the hard material, $\mu = \lambda = 1$ for the soft material (and the workpiece, respectively) and $\mu = \lambda = 5/10000$ for the void phase. We repeat the preceding experiment using $\boldsymbol{g} = (0.4, 0)^T$, $\alpha = 1$, $\beta = 1$ with $\mathfrak{m} = (0.2, 0.1, 0.7)$, $\gamma = 0.0005$ and $\varepsilon = 0.02$. An adaptive mesh is used with $h_{max} = 1/64$ and $h_{min} = 1/128$. We perform the optimization twice, once by prescribing void and once by prescribing soft material between the jaws. The local minima obtained by the H^1-BFGS method are depicted in Figure 86 and Figure 87, respectively, where the hard material is shown in blue, the soft material and the interface in gray and the void phase in red. We note that the thin bars in Figure 86 don't consist of soft material, but are part of the interfacial area between hard material and void. As opposed to the preceding experiment corresponding to Figure 84, a compliance penalization is used here. Thereby we want show that the usage of compliant penalization cannot prevent the occurrence of thin bars in the solution. Compared to the



(a) iter no. 940 (b) iter no. 3120 (c) iter no. 29579

Figure 84: Development of thin bars (tracking type functional)

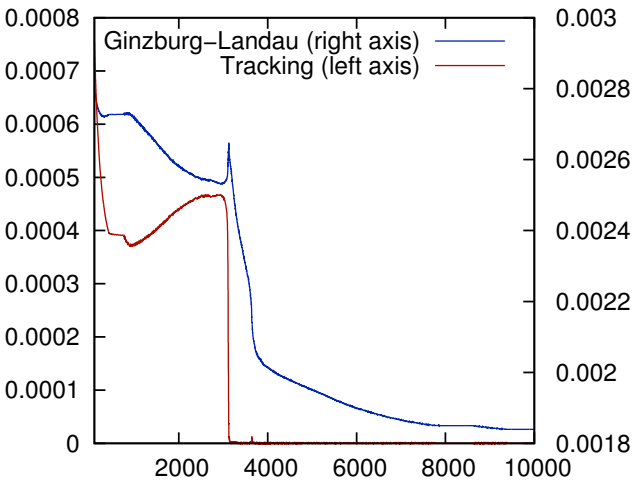


Figure 85: Development of the regularization γE and the tracking term $j - \gamma E$ during the optimization process.

previous experiment, only the region on the left hand side changes to increase the stiffness of the mechanism. However, the thin bars remain. On the other hand, in the solution including the workpiece, no thin bars are present. In fact, the solution looks very similar to designs obtained in the literature as will be discussed below. The thin bars cannot occur since the force that is transfered to the jaws is too low to be able to compress the workpiece. Also note that the Ginzburg-Landau energy is more than doubled compared to the solution without workpiece. Thus we see that the introduction of a reaction force prevents thin bars in the final design. However, the Ginzburg-Landau energy still prefers mechanisms including thin bars. We note that we perform here only a first experiment to be able to explain the occurrence of thin bars. This issue should be pursued in detail in future research. Also note that thin bars also appear often when using the linear cost functional instead of the tracking type functional. For other models and methods, like the level-set method, such bars don't occur, since on the one hand no intermediate phase is allowed, but only material or void and on the other hand because no Ginzburg-Landau energy is used, which favors such bars.
With this experiment also a major difference can be noticed between solutions using the linear functional and the tracking type functional. For the linear functional (Figures 78, 79, 80, 81) the jaws don't close parallelly but more like scissors, whereas the designs using the tracking type functional close parallelly (Figures 84, 86, 87). This scissor-like behavior is also observed in the literature when using the linear functional [YINT10] or when optimizing the displacement only in the outer point of the jaws [BS03, TNK10]. In [AD14] the jaws also close like scissors although a tracking type functional is used. This is probably due to a too large value for $\boldsymbol{u}_\Omega$.
As already mentioned, the solution in Figure 87 is contrary to the other solutions for the gripper experiment very similar to the designs obtained in the literature using only a single material and void. When replacing the soft material in Figure 87 by hard material and when removing the workpiece, the solution is almost identical to the design obtained in [AD14] when no uncertainty is present. Also the gripper in [YINT10] is very similar. However, only for our solution the jaws close parallelly. There are also other grippers in the literature, which are quite different [TNK10, Sig97, WCWM05, LLC+08, YKBS04], although in most cases the left half of the mechanism is very similar.

To confirm that the introduction of a workpiece can improve the final design we consider a second experiment. We compute a force inverter, which is also used as a benchmark problem in the literature [Sig97, GMWG14, TNK10, WC09, WCWM05, YKBS04, AD14]. The goal is to transfer the input force from the left hand side to the right hand side, such that its direction is reversed. The setup is $\Omega = (0,1) \times (-0.5, 0.5)$ with $\Gamma_D = \{x_1 = 0\} \times \{0.4 \leq |x_2| \leq 0.5\}$ and $\Gamma_g = \{x_1 = 0\} \cap \{|x_2| \leq 0.05\}$, $\boldsymbol{g} \equiv (0.2, 0)^T$ and $\boldsymbol{f} \equiv \boldsymbol{0}$. We again use a multiphase setting with three phases, where we set the Lamé constants of the hard and soft material as well as the void phase as above. The tracking type functional is used with $c = 10000\chi_{\Omega_{obs}}$, where $\Omega_{obs} = (0.095, 1) \times (-0.05, 0.05)$, and $\boldsymbol{u}_\Omega \equiv (-0.1, 0)^T$. Hard material is prescribed in Ω_{obs}, in the region $(0, 0.05) \times (0.4, 0.5)$ around the Dirichlet domain as well as in the region $(0, 0.05) \times (-0.05, 0.05)$ around the input force. We set $\alpha = 0.1$, $\beta = 1$ with $\mathfrak{m} = (0.2, 0.11, 0.69)$, $\gamma = 0.001$ and $\varepsilon = 0.02$. The local minimum obtained by the H^1-BFGS method is depicted in Figure 88. Again thin bars are part of the mechanism. We repeat the same experiment, but now we prescribe soft material between the input and output port. By this method we model a reaction force which resists the movement of the output port. Note that we use this model for simplicity to study the influence of a reaction force. There certainly are better ways to model a reaction force, e.g.

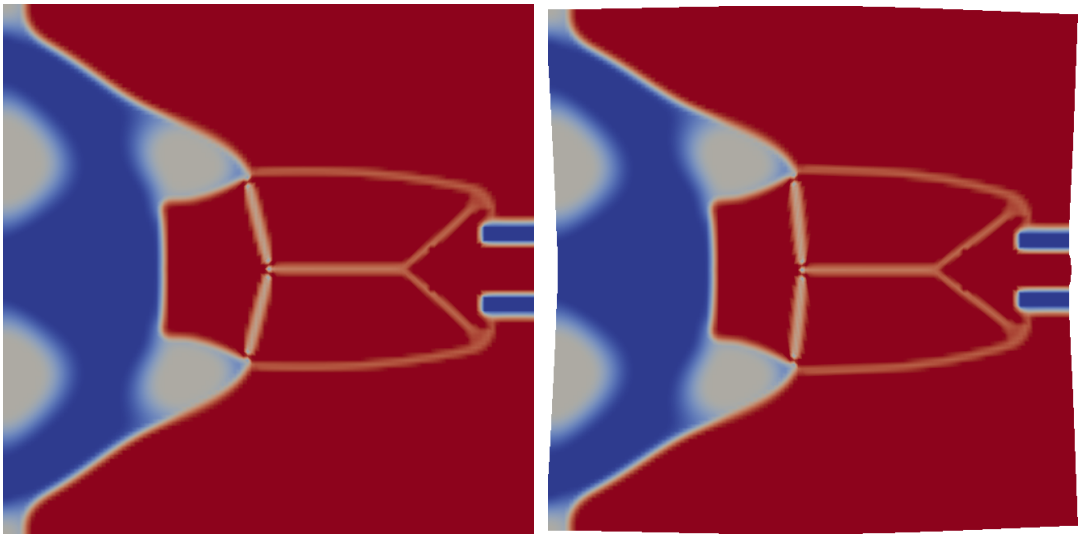


Figure 86: Gripper with 3 phases without workpiece using tracking type functional and a mass constraint.

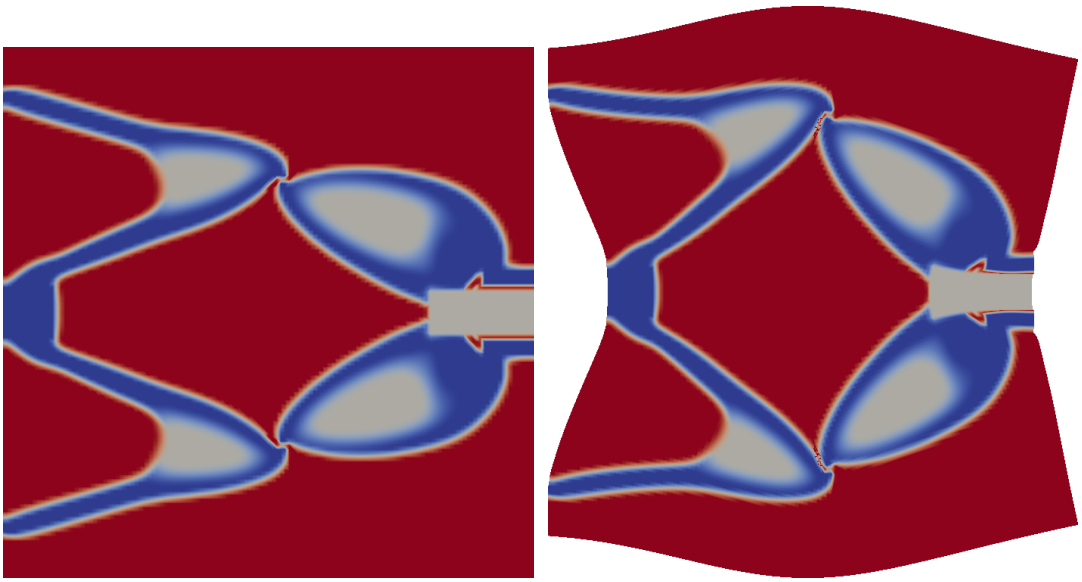


Figure 87: Gripper with 3 phases with workpiece using tracking type functional and a mass constraint.

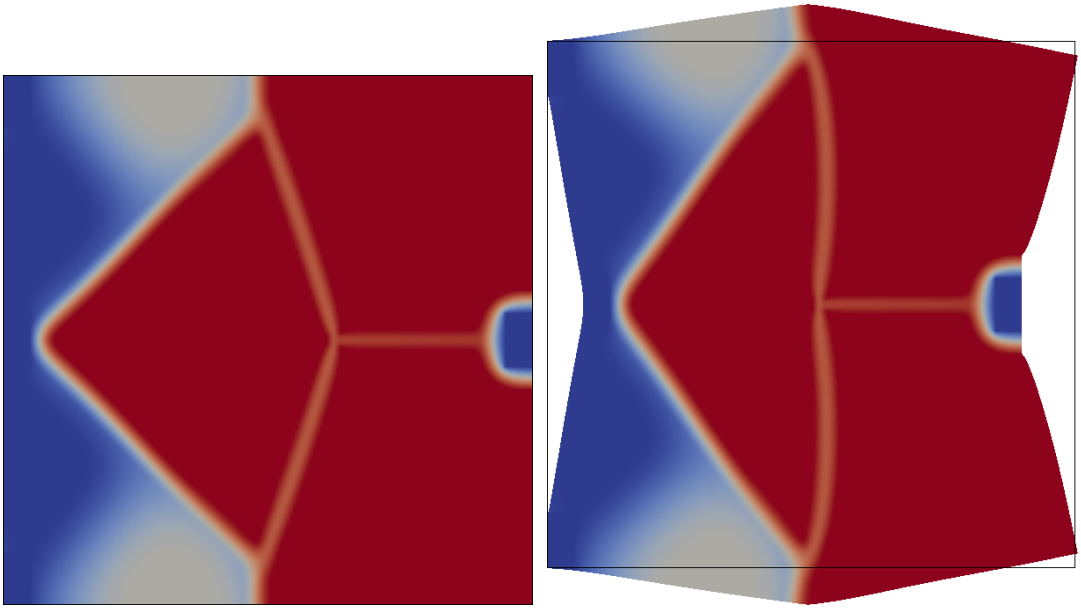


Figure 88: Force inverter with 3 phases without workpiece using tracking type functional and a mass constraint.

by a spring which is located on the right hand side of Ω_{obs}. The final design is depicted in Figure 89. Again, no thin bars are present. Note that there are not even hinges included in the mechanism, which is a good result. However, the performance of the mechanism is far less than for the mechanism without workpiece, where the tracking term is almost 0. Of course this can be improved by increasing the input force and reducing the stiffness of the materials, respectively. The Ginzburg-Landau energy for Figure 89 is also 4 times the energy of Figure 88. As opposed to the gripper and the cruncher experiment, we were not able to compute a solution for the inverter problem without thin bars if no reaction force is present.
Like in the gripper experiment, the design including the workpiece is very similar to designs in the literature. When replacing the soft material by hard material and removing the workpiece and the small bridge attached to the workpiece in the middle, our solution is almost identical to the design obtained in [WCWM05] using a level set method. Similar results can also be found in [Sig97, GMWG14, TNK10, WC09]. Most designs look more or less like a rhombus, where the upper and lower corner is connected to the Dirichlet domain. Partially there are holes in the hard material where there is soft material in our solution to give the mechanism more flexibility. It is interesting that in [YKBS04] they get the same solution except that the soft material is replaced by a checkerboard pattern of the hard material. In another experiment including checkerboard control the checkerboard is replaced by hard material with holes. Thus in our solution the usage of soft material replaces checkerboard patterns and holes in the hard material, respectively, which may lead to a more robust mechanism. A less similar solution can be found in [AD14]. When comparing our multiphase solution to the multiphase solutions obtained in [WCWM05, GMWG14], we observe that in our solution the soft material is concentrated in larger areas, whereas in [WCWM05, GMWG14] the soft material is rather distributed around the boundary of the hard material, which leads to very different designs as it is the case for the gripper experiment.

Finally, we present a result for a compliant mechanism in 3D. Therefor we extend

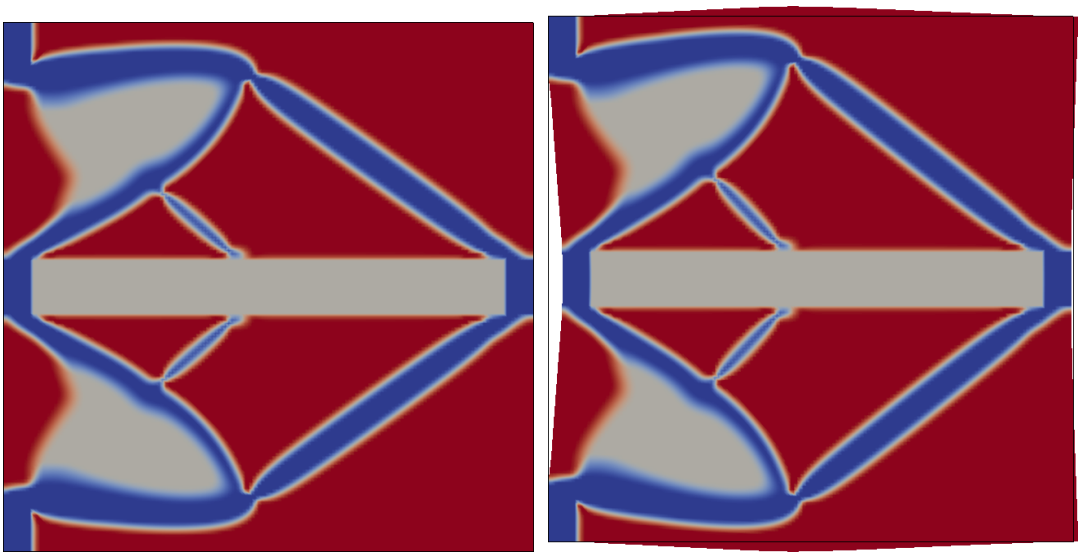


Figure 89: Force inverter with 3 phases with workpiece using tracking type functional and a mass constraint.

the geometry of the second gripper experiment to three space dimensions. We use $\Omega = (-0.5, 0.5) \times (-1, 1) \times (-1, 1)$, $\Gamma_D = \{x_3 = -1\} \cap \{0.8 \leq |x_2| \leq 1\}$, $\Gamma_g = \{x_3 = -1\} \cap \{|x_2| \leq 0.1\}$ with $\boldsymbol{g} \equiv (0, 0, 0.4)^T$ and $\boldsymbol{f} \equiv \mathbf{0}$. We use the Lamé constants $\mu = \lambda = 5$ for the material and $\mu = \lambda = 5/10000$ for the void phase. The cost functional is the tracking type functional with $c = 10000\chi_{\Omega_{obs}}$ where $\Omega_{obs} = \{0.1 \leq |x_2| \leq 0.17\} \cap \{0.8 \leq x_3 \leq 1\}$ and $\boldsymbol{u}_\Omega(x) = (0, -\mathrm{sgn}(x_2)0.02, 0)^T$. We prescribe material in Ω_{obs}. Because of the good experience when including a reaction force, we again prescribe a workpiece between the jaws. However, since we use only two phases here we include the constraint $\varphi = 0.2$ in the region $(-0.5, 0.5) \times (-0.09, 0.09) \times (0.6, 1)$ between the jaws. Taking the interpolation of the stiffness tensor into account this corresponds to a workpiece with a stiffness of $\mu = \lambda = 0.16 \cdot 5$. Moreover we set $\alpha = 1.0$, $\beta = 1$ with $\mathfrak{m} = 0$ (i.e. 50% material), $\gamma = 0.0005$ and $\varepsilon = 0.02$. An adaptive mesh is used with $h_{max} = 1/20$ and $h_{min} = 1/40$. Due to symmetry we restrict the computation to 1/4 of the design domain. The local minimum found by the H^1-BFGS method within 702 iterations and 12 hours computation time is depicted in Figure 90 (the level set $\{\varphi \leq 0\}$ is shown). The final residual is $\sqrt{\gamma\varepsilon}\|\nabla v_k\|_{L^2} = 3 \cdot 10^{-6}$, the final Ginzburg-Landau energy $\gamma E(\varphi^*) = 0.0026$ and the remaining energy is $j(\varphi^*) - \gamma E(\varphi^*) = 0.0059$. It is interesting that the profile of the mechanism on the left and right hand side looks very similar to the 2D solution in Figure 87. Except for the beam in the middle of the mechanism our solution looks similar to the design obtained in [YINT10] using a level set method. The solution obtained in [TNK10] by a phase field method is quite different.

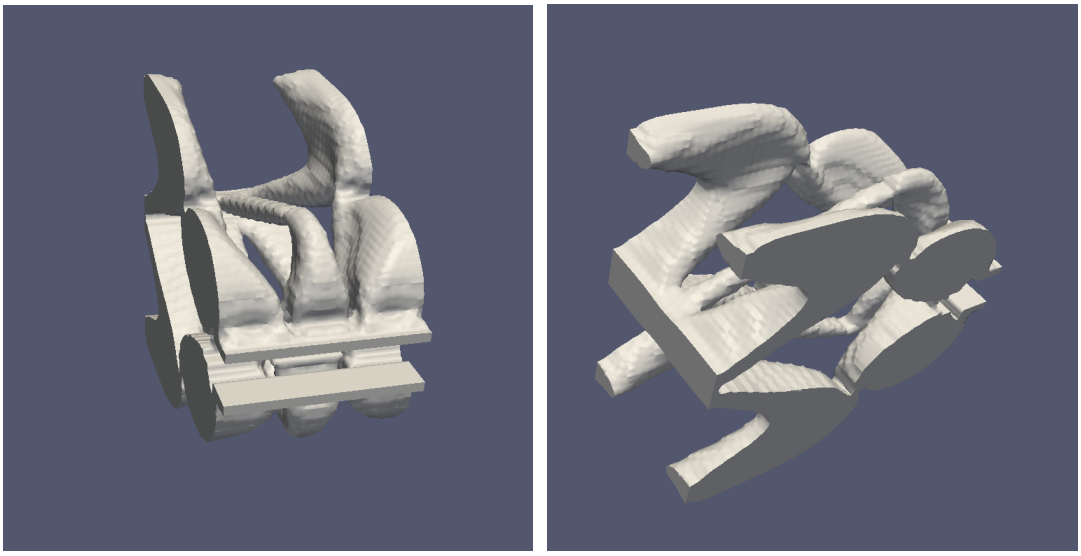


Figure 90: 3D gripping mechanism with mass constraint using the tracking type functional. The workpiece is not included in the graphics.

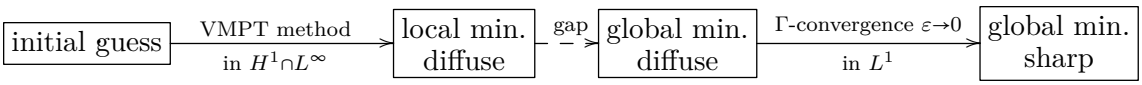
7 Conclusions and perspectives

Due to the generalization of the scaled projected gradient method to a Banach space setting we could enlarge the class of optimization problems which can be solved by the method. Previously the method could be applied only formally to such problems. However, because of our global convergence proof the method is now rigorous even in the Banach space setting. We included two different norms in our analysis which is essential for problems involving two-norm discrepancies. Moreover we achieved thereby to decouple the space in which the optimization problem is differentiable from the space in which the projection is performed. Thus it is not necessary anymore to look for a Hilbert space in which the optimization problem is differentiable, which can be a very small space.
We have shown *global* convergence, which is essential for highly nonlinear problems as we have seen in Section 6.13.8 and Section 6.13.9. The VMPT method is thus an enrichment for globally convergent methods for convexly constrained optimization problems in Banach space. For such general problems only few globally convergent numerical methods are available to date. The VMPT method can also be used to globalize other numerical methods, such as Newton type methods, which may yield a better local convergence rate. Because we weakened the differentiability assumption on the problem, the analysis of the optimization problem can be simplified considerably. It is now sufficient to show the differentiability only with respect to some stronger norm. For instance a projected gradient method can be applied in L^2 even if the optimization problem is only differentiable in L^∞. On the other hand, if the problem is already differentiable in some Hilbert space the results of this thesis allow that the projected gradient method is performed in another space, which can simplify the numerical computations. For instance, if the problem is differentiable in H^1 one can still perform the projected gradient method in L^2. However, in this case it is necessary that the feasible set is bounded in H^1, which we demonstrated by the counterexamples in Section 6.13.11.
Moreover, our analysis implies that the proximal gradient method can be performed in a Banach space setting for nonconvex cost functionals using a variable metric, which is a new result in this combination.

It is still an open question if a curved search alone can be used as globalization. At least this cannot be proved in the considered Banach space setting by standard techniques, since the continuity of the projection arc is given in a norm which is too weak. On the other hand we were not able to find a counterexample. However, the presented hybrid method works in any case since a line search is performed as backup if the curved search fails.

The analysis of the topology optimization problem was performed under structural assumptions rather than for a concrete problem. Thus it is possible to use the VMPT method for any other regular enough cost functional than the considered mean compliance and compliant mechanism functionals. Also the existence and uniqueness of Lagrange multipliers was shown for a general objective. Thus the results apply also to other optimization problems where the unknown is a vector valued (or scalar valued) phase field with constraints coming from a typical obstacle potential and where the cost functional is differentiable in $H^1 \cap L^\infty$. A mass constraint is optionally possible. As examples we refer to the optimization of a Stokes flow in [GH15] or the inverse problem of identifying diffusion coefficients considered in [DES15]. Finally, also the convergence of the presented semismooth Newton method was shown for such general optimization problems. However, this was shown only on the discrete level.
First numerical tests indicate that the VMPT method can also be successfully applied to the topology optimization problem in Stokes flow in [GH15] and to the inverse problem in [DES15]. Especially in the Stokes problem the method seems to be promising. Convergence within 7 steps can be observed for the H^1-BFGS method for a good choice of the parameters, whereas other methods such as the projected L^2-gradient method need more than 3000 iterations for the same accuracy.

In contrast to other numerical methods used in topology optimization the VMPT method is rigorous. We also derived a rigorous stopping criterion, of which other methods lack. Thus it is possible to measure the optimality of the current iterate, hence the optimization is not stopped if the iterates progress slowly. Opposed to other used numerical methods the VMPT method is mesh independent since it is well defined in the infinite dimensional setting. Together with the Γ-convergence result in [BGHR15] the VMPT method is a rigorous tool for solving general topology optimization problems. However, there is still an application gap, since the VMPT method converges to *local* minima, whereas Γ-convergence is concerned with *global* minima:



Numerical experiments showed that this can be an issue when constructing compliant mechanisms. We presented a practical solution by taking reaction forces into account. Then the limit of the obtained solution for $\varepsilon \to 0$ is reasonable. However, this does not close the mentioned gap. Further research in the field of compliant mechanisms is still necessary, starting with a better model which takes reaction forces into account.

We showed that the VMPT method can be used to solve also very complicated problems where other methods fail. Although the method can take a long time to converge in this case, it is still robust and no tuning of the parameters is needed to obtain convergence. We were able to improve the phase field model by a reasonable choice for the potential

and the stiffness tensor interpolation, which enables to take larger values for the phase field parameter ε. This results can also be useful for other numerical methods. We showed that a good choice for the variable metric can enhance the performance of the VMPT method considerably. Moreover, the quality of the obtained minimizer (in the sense of lower energy) can be enhanced by a sophisticated choice of the variable metric. However, this choice is problem specific and has to be derived separately for the considered problems.

The convergence analysis for the VMPT method can also be used to show global convergence for other methods. We demonstrated this by the example of a pseudo time stepping for which no convergence analysis was available before. In this way we also developed a rigorous stopping criterion for the pseudo time stepping. Moreover, we were able to deduce an adaptive choice for the time step sizes based on the Armijo condition, which can be used by pseudo time stepping methods in the future. We showed that the adaptivity allows the time step sizes to grow to infinity, yielding a rapid evolution near the minimum. This result is not bound to topology optimization problems and thus applies also to other optimization problems.

Appendix

In the following theorem we show that $\|\mathcal{E}(.)\|_{L^2}$ can still be a norm equivalent to the H^1-norm in the case that other boundary conditions are used than Dirichlet boundary conditions for both components of $\boldsymbol{u}$. The geometry is taken from the MBB beam example in Section 6.13.5.

Theorem 7.1. *Let $\Omega \subset \mathbb{R}^2$, $\Omega = (0,1)^2$, $\Gamma_1, \Gamma_2 \subset \partial\Omega$ with $\mathcal{H}^1(\Gamma_1) > 0$ and $\mathcal{H}^1(\Gamma_2) > 0$, $\Gamma_1 \subset \{x_1 = 0\}$, $\Gamma_2 \subset \{x_2 = 0\}$, where $\mathcal{H}^1$ denotes the 1D Hausdorff measure. Let*

$$V := \{\boldsymbol{u} \in H^1(\Omega)^2 \mid u_1 = 0 \text{ on } \Gamma_1,\ u_2 = 0 \text{ on } \Gamma_2\}$$

Then there exists $c > 0$ such that

$$\|\boldsymbol{u}\|_{H^1} \leq c\|\mathcal{E}(\boldsymbol{u})\|_{L^2} \quad \forall \boldsymbol{u} \in V.$$

Proof. The key ingredients are the Poincaré inequality in V and that there is only the trivial rigid motion contained in V. The proof is by contradiction analog to [Zei88]. Assume that there exists a sequence $\boldsymbol{u}_n \in V$ (without loss of generality $\|\boldsymbol{u}_n\|_{H^1} = 1$) with

$$\|\boldsymbol{u}_n\|_{H^1} > n\|\mathcal{E}(\boldsymbol{u}_n)\|_{L^2} \quad \forall n > 0.$$

We conclude

$$\|\mathcal{E}(\boldsymbol{u}_n)\|_{L^2} \to 0.$$

Since $\boldsymbol{u}_n$ is bounded in H^1, we can extract a subsequence (denoted again by $\boldsymbol{u}_n$), with $\boldsymbol{u}_n \to \boldsymbol{u}$ weakly in H^1 for some $\boldsymbol{u} \in H^1$. Because the embedding $H^1 \to L^2$ is compact, we get $\boldsymbol{u}_n \to \boldsymbol{u}$ strongly in L^2. For u_1 and u_2 we can apply the Poincaré-Friedrichs inequality [Trö09], leading to

$$\|\boldsymbol{u}\|_{H^1} \leq c\|\nabla \boldsymbol{u}\|_{L^2} \quad \forall \boldsymbol{u} \in V.$$

We use Korn's inequality (Lemma 6.1), i.e.

$$\exists c > 0 : \quad \|\nabla \boldsymbol{u}\|_{L^2}^2 \leq c(\|\mathcal{E}(\boldsymbol{u})\|_{L^2}^2 + \|\boldsymbol{u}\|_{L^2}^2) \quad \forall \boldsymbol{u} \in H^1(\Omega)^2.$$

Since $\mathcal{E}(\boldsymbol{u}_n)$ and $\boldsymbol{u}_n$ are Cauchy sequences in L^2 we also get from preceding inequality and the Poincaré-Friedrichs inequality that $\boldsymbol{u}_n$ is a Cauchy sequence in H^1, thus $\boldsymbol{u}_n \to \boldsymbol{u}$ in H^1. We conclude

$$\mathcal{E}(\boldsymbol{u}_n) \to \mathcal{E}(\boldsymbol{u}) \quad \text{in } L^2,$$

thus $\mathcal{E}(\boldsymbol{u}) = 0$. Hence $\boldsymbol{u}$ is a linearized rigid body motion, i.e. we can write $\boldsymbol{u}(x) = a + b\begin{pmatrix} x_2 \\ -x_1 \end{pmatrix}$ for some $a \in \mathbb{R}^2$ and $b \in \mathbb{R}$ [EGK08]. Since it holds $\boldsymbol{u}_n \in V$ for all n and V is closed in H^1 we get $\boldsymbol{u} \in V$. From the boundary conditions we get $a_1 + bx_2 = 0$ on Γ_1. Subtraction of the equation for different points $(0, x_2)^T \in \Gamma_1$ and $(0, \tilde{x}_2)^T \in \Gamma_1$ gives $b = 0$ and thus $a_1 = 0$. On Γ_2 we have $a_2 - bx_1 = 0$, thus $a_2 = 0$. We proved $\boldsymbol{u} = 0$. On the other hand we get $\|\boldsymbol{u}\|_{H^1} = \|\boldsymbol{u}_n\|_{H^1} = 1$, which is a contradiction. □

Note that the proof above cannot be carried out if e.g. $\Gamma_1 = \{x_2 = 0\}$ and $\Gamma_2 = \{x_1 = 0\}$, since then the linearized rigid body motion $r(x) = \begin{pmatrix} x_2 \\ -x_1 \end{pmatrix}$ (clockwise rotation) is in V and fulfills $\mathcal{E}(r) = 0$.

The next Lemma proves some typical Taylor estimate for $C^{1,\alpha}$-functions.

Lemma 7.2. *Let X be a Banach space, $U \subset X$ convex and $f \in C^{1,\alpha}(U)$ with modulus L for some $0 < \alpha \leq 1$. Let $u \in U$ and $v \in X$ such that $u + v \in U$. Then it holds the estimate*

$$f(u+v) - f(u) \leq \langle f'(u), v \rangle + \frac{L}{1+\alpha} \|v\|^{1+\alpha}.$$

Proof. We use the fundamental theorem of calculus and calculate

$$\begin{aligned} f(u+v) - f(u) &= \int_0^1 \langle f'(u+\eta v), v \rangle - \langle f'(u), v \rangle \, \mathrm{d}\eta + \langle f'(u), v \rangle \\ &\leq \int_0^1 |\langle f'(u+\eta v), v \rangle - \langle f'(u), v \rangle| \mathrm{d}\eta + \langle f'(u), v \rangle \\ &\leq \int_0^1 \|f'(u+\eta v) - f'(u)\| \|v\| \mathrm{d}\eta + \langle f'(u), v \rangle \\ &\leq \int_0^1 L \eta^\alpha \|v\|^{1+\alpha} \mathrm{d}\eta + \langle f'(u), v \rangle \\ &= \langle f'(u), v \rangle + \frac{L}{1+\alpha} \|v\|^{1+\alpha}. \end{aligned}$$

□

We will often use the following argument concerning convergence of sequences.

Lemma 7.3. *Let X be a topological space, $(x_n)_n \subset X$ a sequence and let $x \in X$. Assume that out of any subsequence of $(x_n)_n$ one can extract another subsequence converging to x, then the whole sequence converges to x.*

Proof. Assume that $(x_n)_n$ does not converge to x. Then there is a neighborhood U of x such that infinitely many elements of $(x_n)_n$ lie outside of U. Take a subsequence of $(x_n)_n$, which lies outside of U. By assumption we can extract a subsequence converging to x, which is a contradiction. □

The following two results for functions in $H^1 \cap L^\infty$ are very useful.

Lemma 7.4. *Let Ω be a bounded Lipschitz domain as in Section 6. For all $\boldsymbol{\varphi} \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$ it holds*

$$\|\boldsymbol{\varphi}\|_{L^\infty(\partial\Omega)} \leq \|\boldsymbol{\varphi}\|_{L^\infty(\Omega)}.$$

Proof. See [Trö09] for the scalar case. The vector valued case follows from applying the estimate on each component of $\boldsymbol{\varphi}$. □

In particular, the previous lemma yields that the trace operator $\tau : H^1(\Omega)^N \cap L^\infty(\Omega)^N \to L^\infty(\partial\Omega)^N$ is continuous, even if $H^1(\Omega)^N \cap L^\infty(\Omega)^N$ is equipped with the L^∞ norm.

Theorem 7.5. *Let $\Omega \subset \mathbb{R}^d$ be open. If $f, g \in H^1(\Omega) \cap L^\infty(\Omega)$, then $fg \in H^1(\Omega) \cap L^\infty(\Omega)$ and it holds $\nabla(fg) = \nabla f g + f \nabla g$.*

Proof. See [EG91, p. 129]. □

Note that the previous result does in general not hold for arbitrary H^1-functions.

Theorem 7.6. *Let A be a Hausdorff topological vector space and let X, D be Banach spaces with $X \hookrightarrow A$ and $D \hookrightarrow A$. Suppose that $X \cap D$ is dense in both X and D. Then it holds $(X \cap D)^* = X^* + D^*$.*

Proof. See Theorem 2.7.1 in [BL76]. □

References

[AB93] L. Ambrosio and G. Buttazzo, *An optimal design problem with perimeter penalization*, Calculus of Variations and Partial Differential Equations **1** (1993), no. 1, 55–69 (English).

[AC79] S. M. Allen and J. W. Cahn, *A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening*, Acta Metallurgica **27** (1979), 1085–1095.

[AD14] G. Allaire and C. Dapogny, *A linearized approach to worst-case design in parametric and geometric shape optimization*, Mathematical Models and Methods in Applied Sciences **24** (2014), no. 11, 2199–2257.

[ADDM14] G. Allaire, C. Dapogny, G. Delgado, and G. Michailidis, *Multi-phase structural optimization via a level set method*, ESAIM: COCV **20** (2014), no. 2, 576–611.

[AF03] R.A. Adams and J.J.F. Fournier, *List of spaces and norms*, Sobolev Spaces, Pure and Applied Mathematics, vol. 140, Elsevier, 2003, pp. xii – xiii.

[AJ05] G. Allaire and F. Jouve, *A level-set method for vibration and multiple loads structural optimization*, Computer Methods in Applied Mechanics and Engineering **194** (2005), 3269–3290.

[AJ08] _____, *Minimum stress optimal design with the level set method*, Engineering Analysis With Boundary Elements **32** (2008), 909–918.

[AJT04] G. Allaire, F. Jouve, and A.M. Toader, *Structural optimization using sensitivity analysis and a level-set method*, Journal of Computational Physics **194** (2004), no. 1, 363 – 393.

[AKG94] G.K. Ananthasuresh, S. Kota, and Y. Gianchandani, *A methodical approach to the design of compliant micromechanisms*, Solid-state sensor and actuator workshop **1994** (1994), 189–192.

[Alb96] Y.I. Alber, *Metric and Generalized Projection Operators in Banach Spaces: Properties and Applications*, Theory and Applications of Nonlinear Operators of Accretive and Monotone Type (A. Kartsatos, ed.), Lecture Notes in Pure and Applied Mathematics, Taylor & Francis, 1996, pp. 15–50.

[All80] J.C. Allwright, *A feasible direction algorithm for convex optimization: Global convergence rates*, Journal of Optimization Theory and Applications **30** (1980), no. 1, 1–18 (English).

[All02] G. Allaire, *Shape optimization by the homogenization method*, Applied Mathematical Sciences, Springer-Verlag New York, 2002.

[Alt12] H.W. Alt, *Lineare Funktionalanalysis: Eine anwendungsorientierte Einführung*, Springer-Lehrbuch Masterclass, Springer, 2012.

[ART02] N. Arada, J.P. Raymond, and F. Tröltzsch, *On an Augmented Lagrangian SQP Method for a Class of Optimal Control Problems in Banach Spaces*, Computational Optimization and Applications **22** (2002), no. 3, 369–398 (English).

[ASON13] V. S. Almeida, H. L. Simonetti, and L. Oliveira Neto, *The strut-and-tie models in reinforced concrete structures analysed by a numerical technique*, Revista IBRACON de Estruturas e Materiais **6** (2013), 139 – 157 (en).

[Bal90] S. Baldo, *Minimal interface criterion for phase transitions in mixtures of Cahn-Hilliard fluids*, Annales de l'institut Henri Poincaré (C) Analyse non linéaire **7** (1990), no. 2, 67–90 (eng).

[Bal99] E. Balagurusamy, *Numerical methods*, Tata McGraw-Hill, 1999.

[BB88] J. Barzilai and J.M. Borwein, *Two-Point Step Size Gradient Methods*, IMA Journal of Numerical Analysis **8** (1988), no. 1, 141–148.

[BBG11] L. Blank, M. Butz, and H. Garcke, *Solving the Cahn-Hilliard variational inequality with a semi-smooth Newton method*, ESAIM: Control, Optimisation and Calculus of Variations **17** (2011), 931–954.

[BC00] B. Bourdin and A. Chambolle, *Optimisation topologique de structures soumises à des forces de pression*, Actes du 32ème Congrès National d'Analyse Numérique (SMAI, ed.), 2000.

[BC03] ———, *Design-dependent loads in topology optimization*, ESAIM: Control, Optimisation and Calculus of Variations **9** (2003), 19–48.

[BC06] ———, *The phase-field method in optimal design*, IUTAM Symposium on Topological Design Optimization of Structures, Machines and Materials (MartinPhilip Bendsøe, Niels Olhoff, and Ole Sigmund, eds.), Solid Mechanics and Its Applications, vol. 137, Springer Netherlands, 2006, pp. 207–215 (English).

[BDH12] M. Burger, Y. Dong, and M. Hintermüller, *Exact relaxation for classes of minimization problems with binary constraints*, IFB-Report No. 62 (11/2012), Institute of Mathematics and Scientific Computing, University of Graz, 2012.

[BE91a] J.W. Barrett and C.M. Elliott, *Finite element approximation of a free boundary problem arising in the theory of liquid drops ans plasma physics*, ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique **25** (1991), no. 2, 213–252 (eng).

[BE91b] J.F. Blowey and C.M. Elliott, *The Cahn-Hilliard gradient theory for phase separation with non-smooth free energy Part I: Mathematical analysis*, European Journal of Applied Mathematics **2** (1991), 233–280.

[BE93] ———, *Curvature dependent phase boundary motion and parabolic double obstacle problems*, Degenerate Diffusions (Wei-Ming Ni, L.A. Peletier, and J.L. Vazquez, eds.), The IMA Volumes in Mathematics and its Applications, vol. 47, Springer New York, 1993, pp. 19–60 (English).

[Ben83] M.P. Bendsøe, *On Obtaining a Solution to Optimization Problems for Solid, Elastic Plates by Restriction of the Design Space*, Journal of Structural Mechanics **11** (1983), no. 4, 501–521.

[Ben89] ———, *Optimal shape design as a material distribution problem*, Structural optimization **1** (1989), no. 4, 193–202 (English).

[Ber76] D.P. Bertsekas, *On the Goldstein-Levitin-Polyak gradient projection method*, Automatic Control, IEEE Transactions on **21** (1976), no. 2, 174–184.

[Ber82] ———, *Projected Newton methods for optimization problems with simple constraints*, SIAM Journal on Control and Optimization **20** (1982), no. 2, 221–246.

[Ber99] ———, *Nonlinear programming*, Athena scientific optimization and computation series, Athena Scientific, 1999.

[BFGS14] L. Blank, M.H. Farshbaf-Shaker, H. Garcke, and V. Styles, *Relating phase field and sharp interface approaches to structural topology optimization*, ESAIM: Control, Optimisation and Calculus of Variations **20** (2014), 1025–1058.

[BGHR15] L. Blank, H. Garcke, C. Hecht, and C. Rupprecht, *Sharp interface limit for a phase field model in structural optimization*, ArXiv e-prints (2015).

[BGN08] J.W. Barrett, H. Garcke, and R. Nürnberg, *On sharp interface limits of Allen Cahn/Cahn Hilliard variational inequalities*, Discrete and Continuous Dynamical Systems - Series S **1** (2008), 1–14.

[BGS+12] L. Blank, H. Garcke, L. Sarbu, T. Srisupattarawanit, V. Styles, and A. Voigt, *Phase-field Approaches to Structural Topology Optimization*, Constrained Optimization and Optimal Control for Partial Differential Equations (G. Leugering, S. Engell, A. Griewank, M. Hinze, R. Rannacher, V. Schulz, M. Ulbrich, and S. Ulbrich, eds.), International Series of Numerical Mathematics, vol. 160, Springer Basel, 2012, pp. 245–256 (English).

[BGSS13a] L. Blank, H. Garcke, L. Sarbu, and V. Styles, *Nonlocal Allen-Cahn systems: analysis and a primal-dual active set method*, IMA Journal of Numerical Analysis (2013).

[BGSS13b] ———, *Primal-dual active set methods for Allen-Cahn variational inequalities with nonlocal constraints*, Numerical Methods for Partial Differential Equations **29** (2013), no. 3, 999–1030.

[BI12] D. Butnariu and A.N. Iusem, *Totally Convex Functions for Fixed Points Computation and Infinite Dimensional Optimization*, Applied Optimization, Springer Netherlands, 2012.

[BK88] M.P. Bendsøe and N. Kikuchi, *Generating optimal topologies in structural design using a homogenization method*, Computer Methods in Applied Mechanics and Engineering **71** (1988), no. 2, 197 – 224.

[BL76] J. Bergh and J. Löfström, *Interpolation spaces: an introduction*, Grundlehren der mathematischen Wissenschaften, Springer, 1976.

[BL05] P. Bochev and R. Lehoucq, *On the Finite Element Solution of the Pure Neumann Problem*, SIAM Review **47** (2005), no. 1, 50–66.

[BLNZ95] R.H. Byrd, P. Lu, J. Nocedal, and C. Zhu, *A Limited Memory Algorithm for Bound Constrained Optimization*, SIAM Journal on Scientific Computing **16** (1995), no. 5, 1190–1208.

[BMR00] E.G. Birgin, J.É.M. Martínez, and M. Raydan, *Nonmonotone spectral projected gradient methods on convex sets*, SIAM Journal on Optimization (2000), 1196–1211.

[BMR14] ———, *Spectral Projected Gradient Methods: Review and Perspectives*, Journal of Statistical Software **60** (2014), no. 3, 1–21.

[BN89] R.H. Byrd and J. Nocedal, *A Tool for the Analysis of Quasi-Newton Methods with Application to Unconstrained Minimization*, SIAM Journal on Numerical Analysis **26** (1989), no. 3, 727–739.

[BNS04] J. Barrett, R. Nürnberg, and V. Styles, *Finite Element Approximation of a Phase Field Model for Void Electromigration*, SIAM Journal on Numerical Analysis **42** (2004), no. 2, 738–772.

[Bou01] B. Bourdin, *Filters in topology optimization*, International Journal for Numerical Methods in Engineering **50** (2001), no. 9, 2143–2158.

[BR01] R. Becker and R. Rannacher, *An optimal control approach to a posteriori error estimation in finite element methods*, Acta Numerica **10** (2001), 1–102.

[Bra01] D. Braess, *Finite elements: Theory, fast solvers, and applications in solid mechanics*, Cambridge University Press, 2001.

[Bre09] K. Bredies, *A forward-backward splitting algorithm for the minimization of non-smooth convex functionals in Banach space*, Inverse Problems **25** (2009), no. 1, 015005.

[Bre11] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Universitext, Springer New York, 2011.

[BS99] M.P. Bendsøe and O. Sigmund, *Material interpolation schemes in topology optimization*, Archive of Applied Mechanics **69** (1999), no. 9-10, 635–654 (English).

[BS00] R.S. Burachik and S. Scheimberg, *A Proximal Point Method for the Variational Inequality Problem in Banach Spaces*, SIAM Journal on Control and Optimization **39** (2000), no. 5, 1633–1649.

[BS03] M.P. Bendsøe and O. Sigmund, *Topology optimization: Theory, methods and applications*, Engineering online library, Springer, 2003.

[BS06] M. Burger and R. Stainko, *Phase-Field Relaxation of Topology Optimization with Local Stress Constraints*, SIAM Journal on Control and Optimization **45** (2006), no. 4, 1447–1466.

[BSS12] L. Blank, L. Sarbu, and M. Stoll, *Preconditioning for Allen-Cahn variational inequalities with non-local constraints*, Journal of Computational Physics **231** (2012), no. 16, 5406 – 5420.

[But12] M. Butz, *Computational methods for Cahn-Hilliard variational inequalities*, Ph.D. thesis, Universität Regensburg, März 2012.

[BZ95] D. Bucur and J.P. Zolesio, *N-dimensional shape optimization under capacitary constraint*, Journal of Differential Equations **123** (1995), no. 2, 504 – 522.

[CDN10] M. Costabel, M. Dauge, and S. Nicaise, *Corner Singularities and Analytic Regularity for Linear Elliptic Systems. Part I: Smooth domains.*, 211 pages, hal-00453934v2, February 2010.

[CGT00] A.R. Conn, N.I.M. Gould, and P.L. Toint, *Trust Region Methods*, MPS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics, 2000.

[CH58] J.W. Cahn and J.E. Hilliard, *Free Energy of a Nonuniform System. I. Interfacial Free Energy*, The Journal of Chemical Physics **28** (1958), no. 2, 258–267.

[Che96] X. Chen, *Convergence of the BFGS Method for LC1 Convex Constrained Optimization*, SIAM J. Control Optim. **34** (1996), no. 6, 2051–2063.

[Cho15] P. Cholamjiak, *A generalized forward-backward splitting method for solving quasi inclusion problems in Banach spaces*, Numerical Algorithms (2015), 1–18 (English).

[Cia93] P.G. Ciarlet, *Mathematical elasticity: Three-dimensional elasticity*, Mathematical Elasticity: Three-dimensional Elasticity, no. v. 1, North-Holland, 1993.

[Cia02] _____, *The finite element method for elliptic problems*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, 2002.

[CJ12] C. Clason and B. Jin, *A Semismooth Newton Method for Nonlinear Parameter Identification Problems with Impulsive Noise*, SIAM Journal on Imaging Sciences **5** (2012), no. 2, 505–536.

[CK11] C. Clason and K. Kunisch, *A duality-based approach to elliptic control problems in non-reflexive Banach spaces*, ESAIM: COCV **17** (2011), no. 1, 243–266.

[CK15] _____, *A convex analysis approach to multi-material topology optimization*, Preprint, 2015.

[CO82] K.-T. Cheng and N. Olhoff, *Regularized formulation for optimal design of axisymmetric plates*, International Journal of Solids and Structures **18** (1982), no. 2, 153 – 169.

[CPR14] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, *Variable Metric Forward-Backward Algorithm for Minimizing the Sum of a Differentiable Function and a Convex Function*, Journal of Optimization Theory and Applications **162** (2014), no. 1, 107–132 (English).

[CR97] G.H-G. Chen and R.T. Rockafellar, *Convergence Rates in Forward–Backward Splitting*, SIAM Journal on Optimization **7** (1997), no. 2, 421–444.

[CRT13] E. Casas, C. Ryll, and F. Tröltzsch, *Sparse Optimal Control of the Schlögl and FitzHugh-Nagumo Systems*, Computational Methods in Applied Mathematics **13** (2013), no. 4, 415–442.

[CV14] P.L. Combettes and B.C. Vũ, *Variable metric forward-backward splitting with applications to monotone inclusions in duality*, Optimization **63** (2014), no. 9, 1289–1318.

[DAK13] P.D. Dunning and H. Alicia Kim, *A new hole insertion method for level set based structural topology optimization*, International Journal for Numerical Methods in Engineering **93** (2013), no. 1, 118–134.

[Dav07] T.A. Davis, *UMFPACK Version 5.2.0 User Guide*, University of Florida, November 2007.

[DBH12] L. Dedè, M.J. Borden, and T.J.R. Hughes, *Isogeometric analysis for topology optimization with a phase field model*, Archives of Computational Methods in Engineering **19** (2012), no. 3, 427–465 (English).

[DDE05] K. Deckelnick, G. Dziuk, and C.M. Elliott, *Computation of geometric partial differential equations and mean curvature flow*, Acta Numerica **14** (2005), 139–232.

[DES15] K. Deckelnick, C.M. Elliott, and V. Styles, *Double obstacle phase field approach to an inverse problem for a discontinuous diffusion coefficient*, Preprint, 2015.

[DJD00] G. C. A. DeRose Jr. and A. R. Díaz, *Solving three-dimensional layout optimization problems using fixed scale wavelets*, Computational Mechanics **25** (2000), no. 2-3, 274–285 (English).

[DK10] M. Dambrine and D. Kateb, *On the ersatz material approximation in level-set methods*, ESAIM: Control, Optimisation and Calculus of Variations **16** (2010), 618–634.

[DM74] J.E. Jr. Dennis and J.J. Moré, *A Characterization of Superlinear Convergence and Its Application to Quasi-Newton Methods*, Mathematics of Computation **28** (1974), no. 126, pp. 549–560 (English).

[DM93] G. Dal Maso, *An Introduction to* Γ*-Convergence*, Progress in Nonlinear Differential Equations and Their Applications, Birkhäuser Boston, 1993.

[Don12] A.L. Dontchev, *Generalizations of the Dennis–Moré Theorem*, SIAM Journal on Optimization **22** (2012), no. 3, 821–830.

[DR70] V.F. Demyanov and A.M. Rubinov, *Approximate Methods in Optimization Problems*, Modern Analytic and Computational Methods in Science and Mathematics, Amer. Elsevier, 1970.

[DS95] A. Díaz and O. Sigmund, *Checkerboard patterns in layout optimization*, Structural optimization **10** (1995), no. 1, 40–45 (English).

[Dun80] J.C. Dunn, *Newton's Method and the Goldstein Step-Length Rule for Constrained Minimization Problems*, SIAM Journal on Control and Optimization **18** (1980), no. 6, 659–674.

[Dun81] ———, *Global and Asymptotic Convergence Rate Estimates for a Class of Projected Gradient Processes*, SIAM Journal on Control and Optimization **19** (1981), no. 3, 368–400.

[Dun87] ———, *On the convergence of projected gradient processes to singular critical points*, Journal of Optimization Theory and Applications **55** (1987), no. 2, 203–216 (English).

[Dun88] ———, *A projected Newton method for minimization problems with nonlinear inequality constraints*, Numerische Mathematik **53** (1988), no. 4, 377–409 (English).

[Dun09] ———, *Local attractors for gradient-related descent iterations*, Encyclopedia of Optimization (C.A. Floudas and P.M. Pardalos, eds.), Springer US, 2009, pp. 1911–1919 (English).

[DZ01] M. C. Delfour and J.-P. Zolésio, *Shapes and geometries: Analysis, differential calculus, and optimization*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001.

[EG91] L.C. Evans and R.F. Gariepy, *Measure Theory and Fine Properties of Functions*, Studies in Advanced Mathematics, Taylor & Francis, 1991.

[EGK08] C. Eck, H. Garcke, and P. Knabner, *Mathematische Modellierung*, Springer-Lehrbuch, Springer-Verlag GmbH, 2008.

[EL91] C.M. Elliott and S. Luckhaus, *A generalised diffusion equation for phase separation of a multi-component mixture with interfacial free energy*, Preprint, Sonderforschungsbereich 256, 1991.

[ES03] C. M. Elliott and V. Styles, *Computations of bidirectional grain boundary dynamics in thin metallic films*, J. Comput. Phys. **187** (2003), no. 2, 524–543.

[ET99] I. Ekeland and R. Témam, *Convex analysis and variational problems*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, 1999.

[Fle89] C. Fleury, *CONLIN: An efficient dual optimizer based on convex approximation concepts*, Structural optimization **1** (1989), no. 2, 81–89 (English).

[Gar00] H. Garcke, *On mathematical models for phase separation in elastically stressed solids*, Habilitation thesis, University Bonn, 2000.

[GB82] E.M. Gafni and D.P. Bertsekas, *Convergence of a gradient projection method*, LIDS-P ; 1201, Lab. for Info. and Dec. Systems, M.I.T., 1982.

[GB84] ———, *Two-metric projection methods for constrained optimization*, SIAM Journal on Control and Optimization **22** (1984), no. 6, 936–964.

[GCH09] H. Gomez, V.M. Calo, and T.J.R. Hughes, *Isogeometric Analysis of Phase-Field Models: Application to the Cahn-Hilliard Equation*, ECCOMAS Multidisciplinary Jubilee Symposium (Josef Eberhardsteiner, Christian Hellmich, HerbertA. Mang, and Jacques Périaux, eds.), Computational Methods in Applied Sciences, vol. 14, Springer Netherlands, 2009, pp. 1–16 (English).

[GD88] M. Gawande and J.C. Dunn, *Variable metric gradient projection processes in convex feasible sets defined by nonlinear inequalities*, Applied Mathematics and Optimization **17** (1988), no. 1, 103–119 (English).

[GGM00] S. Garreau, P. Guillaume, and M. Masmoudi, *The Topological Asymptotic for PDE Systems: The Elasticity Case*, SIAM J. Control Optim. **39** (2000), no. 6, 1756–1778.

[GH14] H. Garcke and C. Hecht, *A phase field approach for shape and topology optimization in Stokes flow*, Preprint-Nr.: 09/2014, Universität Regensburg, Mathematik (2014).

[GH15] ———, *Shape and Topology Optimization in Stokes Flow with a Phase Field Approach*, Applied Mathematics & Optimization (2015), 1–48 (English).

[GK02] C. Geiger and C. Kanzow, *Theorie und Numerik restringierter Optimierungsaufgaben*, Springer-Lehrbuch Masterclass, Springer-Verlag GmbH, 2002.

[GKT92] H. Goldberg, W. Kampowsky, and F. Tröltzsch, *On NEMYTSKIJ Operators in L_p-Spaces of Abstract Functions*, Mathematische Nachrichten **155** (1992), no. 1, 127–140.

[GLS05] M. Gugat, G. Leugering, and G. Sklyar, *L^p-Optimal Boundary Control for the Wave Equation*, SIAM Journal on Control and Optimization **44** (2005), no. 1, 49–74.

[GMWG14] A.T. Gaynor, N.A. Meisel, C.B. Williams, and J.K. Guest, *Multiple-material topology optimization of compliant mechanisms created via polyjet three-dimensional printing*, Journal of Manufacturing Science and Engineering **136** (2014), no. 6.

[GNS99] H. Garcke, B. Nestler, and B. Stoth, *A multiphase field concept: Numerical simulations of moving phase boundaries and multiple junctions*, SIAM Journal on Applied Mathematics **60** (1999), no. 1, 295–315.

[Gob62] J. Gobert, *Une inégalité fondamentale de la théorie de l'élasticité*, Bull. Soc. Roy. Sci. Liège **31** (1962), 182–191. MR 0133684 (24 #A3510)

[Gol64] A.A. Goldstein, *Convex programming in Hilbert space*, Bulletin of the American Mathematical Society **70** (1964), no. 5, 709–710.

[Gol65] ———, *On Newton's method*, Numerische Mathematik **7** (1965), no. 5, 391–393 (English).

[Gou06] F. de Gournay, *Velocity Extension for the Level-set Method and Multiple Eigenvalues in Shape Optimization*, SIAM J. Control Optim. **45** (2006), no. 1, 343–367.

[GP12] A.L. Gain and G.H. Paulino, *Phase-field based topology optimization with polygonal elements: a finite volume approach for the evolution equation*, Structural and Multidisciplinary Optimization **46** (2012), no. 3, 327–342 (English).

[GPM76] U.M. Garcia Palomares and O.L. Mangasarian, *Superlinearly convergent quasi-Newton algorithms for nonlinearly constrained optimization problems*, Mathematical Programming **11** (1976), no. 1, 1–13 (English).

[GS81] W.A. Gruver and E. Sachs, *Algorithmic methods in optimal control*, Research notes in mathematics, Pitman Pub., 1981.

[Han76] S.P. Han, *Superlinearly convergent variable metric algorithms for general nonlinear programming problems*, Mathematical Programming **11** (1976), no. 1, 263–282 (English).

[Han77] _____, *A globally convergent method for nonlinear programming*, Journal of Optimization Theory and Applications **22** (1977), no. 3, 297–309 (English).

[Hec14] C. Hecht, *Shape and topology optimization in fluids using a phase field approach and an application in structural optimization*, Ph.D. thesis, Universität Regensburg, 2014.

[HIK02] M. Hintermüller, K. Ito, and K. Kunisch, *The Primal-Dual Active Set Strategy As a Semismooth Newton Method*, SIAM J. on Optimization **13** (2002), no. 3, 865–888.

[HM03] J. Haslinger and R. A. E. Mäkinen, *Introduction to Shape Optimization: Theory, Approximation, and Computation*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2003.

[HPUU08] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE Constraints*, Mathematical modelling, Springer, 2008.

[HT77] E. Huntley and P.R. Turner, *Direct prediction methods in Hilbert space with applications to control problems*, Journal of Optimization Theory and Applications **22** (1977), no. 3, 399–415 (English).

[HUU99] M. Heinkenschloss, M. Ulbrich, and S. Ulbrich, *Global Convergence of Trust-region Interior-point Algorithms for Infinite-dimensional Nonconvex Minimization Subject to Pointwise Bounds*, SIAM Journal on Control and Optimization **37** (1999), no. 3, 731–764.

[HX09] X. Huang and Y.M. Xie, *Bi-directional evolutionary topology optimization of continuum structures with one or multiple materials*, Computational Mechanics **43** (2009), no. 3, 393–401 (English).

[IB97] A.N. Iusem and D. Butnariu, *On a proximal point method for convex optimization in Banach spaces*, Numerical Functional Analysis and Optimization **18** (1997), no. 7-8, 723–744.

[IK96] K. Ito and K. Kunisch, *Augmented Lagrangian-SQP-Methods in Hilbert Spaces and Application to Control in the Coefficients Problems*, SIAM Journal on Optimization **6** (1996), no. 1, 96–125.

[IK08] _____, *Lagrange Multiplier Approach to Variational Problems and Applications*, Advances in Design and Control, Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 2008.

[IKS13] A.F. Izmailov, A.S. Kurennoy, and M.V. Solodov, *The Josephy-Newton Method for Semismooth Generalized Equations and Semismooth SQP for Optimization*, Set-Valued and Variational Analysis **21** (2013), no. 1, 17–45 (English).

[IO01] A.N. Iusem and R.G. Otero, *Inexact versions of proximal point and augmented lagrangian algorithms in banach spaces*, Numerical Functional Analysis and Optimization **22** (2001), no. 5-6, 609–640.

[ISU12] A.F. Izmailov, M.V. Solodov, and E.I. Uskov, *Global Convergence of Augmented Lagrangian Methods Applied to Optimization Problems with Degenerate Constraints, Including Problems with Complementarity Constraints*, SIAM Journal on Optimization **22** (2012), no. 4, 1579–1606.

[Ius03] A.N. Iusem, *On the convergence properties of the projected gradient method for convex optimization*, Computational & Applied Mathematics **22** (2003), 37 – 52 (en).

[Jog02] C. S. Jog, *Topology design of structures using a dual algorithm and a constraint on the perimeter*, International Journal for Numerical Methods in Engineering **54** (2002), no. 7, 1007–1019.

[KA64] L.V. Kantorovich and G.P. Akilov, *Functional analysis in normed spaces*, International series of monographs in pure and applied mathematics, Pergamon Press; [distributed in the Western Hemisphere by Macmillan, New York], 1964.

[Kel99] C.T. Kelley, *Iterative methods for optimization*, Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics, 1999.

[KL94] W. Krabs and G. Leugering, *On boundary controllability of one-dimensional vibrating systems by $W_0^{1,p}$-controls for $p \in [2, \infty]$*, Mathematical Methods in the Applied Sciences **17** (1994), no. 2, 77–93.

[KM92] T. Kilpeläinen and J. Malý, *Supersolutions to Degenerate Elliptic Equations on Quasi open Sets*, Communications in Partial Differential Equations **17** (1992), 371–405.

[Kor94] R. Kornhuber, *Monotone multigrid methods for elliptic variational inequalities I*, Numerische Mathematik **69** (1994), no. 2, 167–184 (English).

[KS87] C.T. Kelley and E.W. Sachs, *Quasi-Newton Methods and Unconstrained Optimal Control Problems*, SIAM Journal on Control and Optimization **25** (1987), no. 6, 1503–1516.

[KS89] ———, *A pointwise quasi-Newton method for unconstrained optimal control problems*, Numerische Mathematik **55** (1989), no. 2, 159–176 (English).

[KS92] ———, *Mesh Independence of the Gradient Projection Method for Optimal Control Problems*, SIAM J. Control Optim. **30** (1992), no. 2, 477–493.

[KS99] ———, *A Trust Region Method for Parabolic Boundary Control Problems*, SIAM Journal on Optimization **9** (1999), no. 4, 1064–1081.

[KU14] M. Keuthen and M. Ulbrich, *Moreau-Yosida regularization in shape optimization with geometric constraints*, Computational Optimization and Applications (2014), 1–36 (English).

[KZ13] W. Krabs and J. Zowe, *Modern Methods of Optimization: Proceedings of the Summer School "Modern Methods of Optimization", held at the Schloß Thurnau of the University of Bayreuth, Bayreuth, FRG, October 1–6, 1990*, Lecture Notes in Economics and Mathematical Systems, Springer Berlin Heidelberg, 2013.

[KZPP76] M.A. Krasnoselskii, P.P. Zabreiko, E.I. Pustylnik, and Sbolevskii P.E., *Integral operators in spaces of summable functions*, Noordhoff International Publishing, 1976.

[LF01] D.H. Li and M. Fukushima, *A modified BFGS method and its global convergence in nonconvex minimization*, Journal of Computational and Applied Mathematics **129** (2001), no. 1-2, 15 – 35, Nonlinear Programming and Variational Inequalities.

[Li05] J. Li, *The generalized projection operator on reflexive Banach spaces and its applications*, Journal of Mathematical Analysis and Applications **306** (2005), no. 1, 55 – 71.

[LK11] Y. Li and J. Kim, *Multiphase image segmentation using a phase-field model*, Computers & Mathematics with Applications **62** (2011), no. 2, 737 – 745.

[LLC+08] J. Luo, Z. Luo, S. Chen, L. Tong, and M.Y. Wang, *A new level set method for systematic design of hinge-free compliant mechanisms*, Computer Methods in Applied Mechanics and Engineering **198** (2008), no. 2, 318 – 331.

[LMMWX12] G. López, V. Martín-Márquez, F. Wang, and H.K. Xu, *Forward-Backward Splitting Methods for Accretive Operators in Banach Spaces*, Abstract and Applied Analysis **2012** (2012), Hindawi Publishing Corporation, 25 pages.

[LMW+12] A. Logg, K.A. Mardal, G.N. Wells, et al., *Automated Solution of Differential Equations by the Finite Element Method*, Springer, 2012.

[LP66] E.S. Levitin and B.T. Polyak, *Constrained minimization methods*, USSR Computational mathematics and mathematical physics **6** (1966), no. 5, 1–50.

[lR15] J.C. De los Reyes, *Numerical PDE-Constrained Optimization*, Springer Briefs in Optimization, Springer International Publishing, 2015.

[LS02] Q.Q. Liang and G.P. Steven, *A performance-based optimization method for topology design of continuum structures with mean compliance constraints*, Computer Methods in Applied Mechanics and Engineering **191** (2002), no. 13-14, 1471–1489.

[LT93] Z.Q. Luo and P. Tseng, *Error bounds and convergence analysis of feasible descent methods: a general approach*, Annals of Operations Research **46-47** (1993), no. 1, 157–178 (English).

[Lus99] M.T. Lusk, *A phase-field paradigm for grain growth and recrystallization*, Proceedings: Mathematical, Physical and Engineering Sciences **455** (1999), no. 1982, pp. 677–700 (English).

[LWH12] A. Logg, G.N. Wells, and J. Hake, *DOLFIN: a C++/Python Finite Element Library*, ch. 10, Springer, 2012.

[MKC95] Z.-D. Ma, N. Kikuchi, and H.-C. Cheng, *Topological design for vibrating structures*, Computer Methods in Applied Mechanics and Engineering **121** (1995), no. 1-4, 259 – 280.

[MM77] L. Modica and S. Mortola, *Un esempio di Γ-convergenza*, Boll. Un. Mat. Ital. B (5) **14** (1977), no. 1, 285–299.

[Mod87] L. Modica, *The gradient theory of phase transitions and the minimal interface criterion*, Archive for Rational Mechanics and Analysis **98** (1987), no. 2, 123–142 (English).

[MQ80] R.V. Mayorga and V.H. Quintana, *A family of variable metric methods in function space, without exact line searches*, Journal of Optimization Theory and Applications **31** (1980), no. 3, 303–329 (English).

[MS76] F. Murat and J. Simon, *Etude de problèmes d'optimal design*, Optimization Techniques Modeling and Optimization in the Service of Man Part 2 (Jean Cea, ed.), Lecture Notes in Computer Science, vol. 41, Springer Berlin Heidelberg, 1976, pp. 54–62 (French).

[MT72] G.P. McCormick and R.A. Tapia, *The Gradient Projection Method under Mild Differentiability Conditions*, SIAM Journal on Control **10** (1972), no. 1, 93–98.

[MXW15] Q. Ma, Z. Xu, and L. Wang, *Recovery of the local volatility function using regularization and a gradient projection method*, Journal of Industrial and Management Optimization **11** (2015), no. 2, 421–437.

[NFMK98] S. Nishiwaki, M.I. Frecker, S. Min, and N. Kikuchi, *Topology optimization of compliant mechanisms using the homogenization method*, International Journal for Numerical Methods in Engineering **42** (1998), no. 3, 535–559.

[NST15] S. Nicaise, S. Stingelin, and F. Tröltzsch, *Optimal control of magnetic fields in flow measurement*, Discrete and Continuous Dynamical Systems - Series S **8** (2015), no. 3, 579–605.

[NW06] J. Nocedal and S. Wright, *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, Springer New York, 2006.

[OI07] R.G. Otero and A.N. Iusem, *Proximal methods in reflexive Banach spaces without monotonicity*, Journal of Mathematical Analysis and Applications **330** (2007), no. 1, 433 – 450.

[OS88] S. Osher and J.A. Sethian, *Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations*, Journal of Computational Physics **79** (1988), no. 1, 12 – 49.

[Pet82] K.E. Petersen, *Silicon as a mechanical material*, Proceedings of the IEEE **70** (1982), no. 5, 420–457.

[Pet99] J. Petersson, *Some convergence results in perimeter-controlled topology optimization*, Computer Methods in Applied Mechanics and Engineering **171** (1999), no. 1-2, 123 – 140.

[PL14] B. Poling and G. Lerman, *A new approach to two-view motion segmentation using global dimension minimization*, International Journal of Computer Vision **108** (2014), no. 3, 165–185 (English).

[Pou03] T.A. Poulsen, *A new scheme for imposing a minimum length scale in topology optimization*, International Journal for Numerical Methods in Engineering **57** (2003), no. 6, 741–760.

[PRW12] P. Penzler, M. Rumpf, and B. Wirth, *A phase-field model for compliance shape optimization in nonlinear elasticity*, ESAIM: Control, Optimisation and Calculus of Variations **18** (2012), no. 1, 229–258 (eng).

[PS75] C. Paige and M. Saunders, *Solution of Sparse Indefinite Systems of Linear Equations*, SIAM Journal on Numerical Analysis **12** (1975), no. 4, 617–629.

[PS98] J. Petersson and O. Sigmund, *Slope constrained topology optimization*, International Journal for Numerical Methods in Engineering **41** (1998), no. 8, 1417–1434.

[QSX98] O. M. Querin, G. P. Steven, and Y. M. Xie, *Evolutionary structural optimisation (ESO) using a bidirectional algorithm*, Engineering Computations **15** (1998), 1031–1048.

[Rie01] A. Rietz, *Sufficiency of a finite exponent in simp (power law) methods*, Structural and Multidisciplinary Optimization **21** (2001), no. 2, 159–163 (English).

[RM67] R.D. Richtmyer and K.W. Morton, *Difference methods for initial-value problems*, Interscience tracts in pure and applied mathematics, Interscience Publishers, 1967.

[Rob80] S.M. Robinson, *Strongly Regular Generalized Equations*, Mathematics of Operations Research **5** (1980), no. 1, pp. 43–62 (English).

[Roc76] R.T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM Journal on Control and Optimization **14** (1976), 877–898.

[Ros60] J.B. Rosen, *The Gradient Projection Method for Nonlinear Programming. Part I. Linear Constraints*, Journal of the Society for Industrial and Applied Mathematics **8** (1960), no. 1, 181–217.

[Ros61] ———, *The Gradient Projection Method for Nonlinear Programming. Part II. Nonlinear Constraints*, Journal of the Society for Industrial and Applied Mathematics **9** (1961), no. 4, 514–532.

[Roz09] G.I.N. Rozvany, *A critical review of established methods of structural topology optimization*, Structural and Multidisciplinary Optimization **37** (2009), no. 3, 217–237 (English).

[Rus84] B. Rustem, *A class of superlinearly convergent projection algorithms with relaxed stepsizes*, Applied Mathematics and Optimization **12** (1984), no. 1, 29–43 (English).

[RZB92] G.I.N. Rozvany, M. Zhou, and T. Birker, *Generalized shape optimization without homogenization*, Structural optimization **4** (1992), no. 3-4, 250–252 (English).

[Sar10] L. Sarbu, *Primal-dual active set methods for Allen-Cahn variational inequalities* , Ph.D. thesis, University of Sussex, November 2010.

[Sch09] A. Schiela, *Barrier Methods for Optimal Control Problems with State Constraints*, SIAM Journal on Optimization **20** (2009), no. 2, 1002–1031.

[Sho97] R.E. Showalter, *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations*, Mathematical Surveys and Monographs, American Mathematical Society, 1997.

[Sig97] O. Sigmund, *On the Design of Compliant Mechanisms Using Topology Optimization*, Mechanics of Structures and Machines **25** (1997), no. 4, 493–524.

[Sig01] ———, *Design of multiphysics actuators using topology optimization — part i: One-material structures*, Computer Methods in Applied Mechanics and Engineering **190** (2001), no. 49-50, 6577 – 6604.

[SK86] G. Strang and R.V. Kohn, *Optimal design in elasticity and plasticity*, International Journal for Numerical Methods in Engineering **22** (1986), no. 1, 183–188.

[SK91] K. Suzuki and N. Kikuchi, *A homogenization method for shape and topology optimization*, Computer Methods in Applied Mechanics and Engineering **93** (1991), no. 3, 291 – 318.

[SKS11] M. Schmidt, D. Kim, and S. Sra, *Projected newton-type methods in machine learning*, pp. 305–330, MIT Press, Cambridge, MA, USA, nov 2011.

[SLS06] F. Schöpfer, A.K. Louis, and T. Schuster, *Nonlinear iterative methods for linear ill-posed problems in banach spaces*, Inverse Problems **22** (2006), no. 1, 311.

[SM12] O. Sigmund and K. Maute, *Sensitivity filtering from a continuum mechanics perspective*, Structural and Multidisciplinary Optimization **46** (2012), no. 4, 471–475 (English).

[SP98] O. Sigmund and J. Petersson, *Numerical instabilities in topology optimization: A survey on precedures dealing with checkerboards, mesh-dependencies and local minima*, Struct Optim **16** (1998), no. 1, 68–75.

[SS01] M. Stolpe and K. Svanberg, *On the trajectories of penalization methods for topology optimization*, Structural and Multidisciplinary Optimization **21** (2001), no. 2, 128–139 (English).

[SS04] A. Shapiro and J. Sun, *Some Properties of the Augmented Lagrangian in Cone Constrained Optimization*, Mathematics of Operations Research **29** (2004), no. 3, 479–491.

[Ste91] P. Sternberg, *Vector-valued local minimizers of nonconvex variational problems*, Rocky Mountain J. Math. **21** (1991), no. 2, 799–807.

[Sva87] K. Svanberg, *The method of moving asymptotes – a new method for structural optimization*, International Journal for Numerical Methods in Engineering **24** (1987), no. 2, 359–373.

[SvdBFM09] M.W. Schmidt, E. van den Berg, M.P. Friedlander, and K.P. Murphy, *Optimizing Costly Functions with Simple Constraints: A Limited-Memory Projected Quasi-Newton Algorithm*, Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009, 2009, pp. 456–463.

[SZ92] J. Sokolowski and J.P. Zolesio, *Introduction to Shape Optimization: Shape Sensitivity Analysis*, Springer Series in Computational Mathematics, Springer Berlin Heidelberg, 1992.

[SZ99] J. Sokolowski and A. Zochowski, *On the Topological Derivative in Shape Optimization*, SIAM J. Control Optim. **37** (1999), no. 4, 1251–1272.

[Tav14] R. Tavakoli, *Multimaterial topology optimization by volume constrained Allen-Cahn system and regularized projected steepest descent method*, Computer Methods in Applied Mechanics and Engineering **276** (2014), 534–565.

[Tav15] ———, *On the coupled continuous knapsack problems: projection onto the volume constrained Gibbs N-simplex*, Optimization Letters (2015), 1–22 (English).

[TDLC15] Q. Tran-Dinh, Y.-H. Li, and V. Cevher, *Composite Convex Minimization Involving Self-concordant-Like Cost Functions*, Modelling, Computation and Optimization in Information Systems and Management Sciences (H.A. Le Thi, T. Pham Dinh, and N.T. Nguyen, eds.), Advances in Intelligent Systems and Computing, vol. 359, Springer International Publishing, 2015, pp. 155–168 (English).

[Tho92] J. Thomsen, *Topology optimization of structures composed of one or two materials*, Structural optimization **5** (1992), no. 1-2, 108–115 (English).

[TM14] R. Tavakoli and S.M. Mohseni, *Alternating active-phase algorithm for multimaterial topology optimization problems: a 115-line matlab implementation*, Structural and Multidisciplinary Optimization **49** (2014), no. 4, 621–642 (English).

[TNK10] A. Takezawa, S. Nishiwaki, and M. Kitamura, *Shape and topology optimization based on the phase field method and sensitivity analysis*, J. Comput. Phys. **229** (2010), no. 7, 2697–2718.

[TP13] C. Talischi and G.H. Paulino, *An operator splitting algorithm for Tikhonov-regularized topology optimization*, Computer Methods in Applied Mechanics and Engineering **253** (2013), no. Complete, 599–608 (eng).

[Trö09] F. Tröltzsch, *Optimale Steuerung partieller Differentialgleichungen: Theorie, Verfahren und Anwendungen*, Vieweg+Teubner Verlag, 2009.

[TYW01] A. Tsai, Jr. Yezzi, A., and A.S. Willsky, *Curve evolution implementation of the mumford-shah functional for image segmentation, denoising, interpolation, and magnification*, Image Processing, IEEE Transactions on **10** (2001), no. 8, 1169–1186.

[Ulb01] M. Ulbrich, *Nonsmooth Newton-like methods for variational inequalities and constrained optimization problems in function spaces.*, Habilitation thesis, Technische Universität München, 2001.

[UU00] M. Ulbrich and S. Ulbrich, *Superlinear Convergence of Affine-Scaling Interior-Point Newton Methods for Infinite-Dimensional Nonlinear Problems with Pointwise Bounds*, SIAM Journal on Control and Optimization **38** (2000), no. 6, 1938–1984.

[UU09] _____, *Primal-dual interior-point methods for PDE-constrained optimization*, Mathematical Programming **117** (2009), no. 1-2, 435–485 (English).

[VC02] L.A. Vese and T.F. Chan, *A multiphase level set framework for image segmentation using the mumford and shah model*, International Journal of Computer Vision **50** (2002), no. 3, 271–293 (English).

[vdW93] J.D. van der Waals, *The thermodynamic theory of capillarity flow under the hypothesis of a continuous variation of density*, Verhandel. Konink. Akad. Weten. Amsterdam **1** (1893), no. 8.

[WC09] M.Y. Wang and S. Chen, *Compliant mechanism optimization: Analysis and design with intrinsic characteristic stiffness*, Mechanics Based Design of Structures and Machines **37** (2009), no. 2, 183–200.

[WCWM05] M.Y. Wang, S. Chen, X. Wang, and Y. Mei, *Design of multimaterial compliant mechanisms using level-set methods*, Journal of Mechanical Design **127** (2005), no. 5, 941–956.

[Wer07] D. Werner, *Funktionalanalysis*, Springer-Lehrbuch, Springer-Verlag Berlin Heidelberg, 2007.

[WIR15] M. Wallin, N. Ivarsson, and M. Ristinmaa, *Large strain phase-field-based multi-material topology optimization*, International Journal for Numerical Methods in Engineering (2015), n/a–n/a.

[WR12] M. Wallin and M. Ristinmaa, *Howard's algorithm in a phase-field topology optimization approach*, International Journal for Numerical Methods in Engineering (2012).

[WW04] M.Y. Wang and X. Wang, *"Color" level sets: a multi-phase method for structural topology optimization with multiple materials*, Comput. Methods Appl. Mech. Engrg. **193** (2004), no. 6-8, 469–496. MR 2033962

[WZ04a] M.Y. Wang and S. Zhou, *Phase field: a variational method for structural topology optimization*, Comput Model Eng Sci **6** (2004), no. 6, 547–566.

[WZ04b] _____, *Synthesis of shape and topology of multi-material structures with a phase-field method*, Journal of Computer-Aided Materials Design **11** (2004), no. 2-3, 117–138 (English).

[WZ07] ———, *Multimaterial structural topology optimization with a generalized cahn–hilliard model of multiphase transition*, Structural and Multidisciplinary Optimization **33** (2007), no. 2, 89–111 (English).

[XS93] Y.M. Xie and G.P. Steven, *A simple evolutionary procedure for structural optimization*, Computers & Structures **49** (1993), no. 5, 885 – 896.

[XS97] ———, *Evolutionary structural optimization*, Springer London, 1997.

[XWK07] N. Xiu, C. Wang, and L. Kong, *A note on the gradient projection method with exact stepsize rule*, Journal of Computational Mathematics **25** (2007), no. 2, 221–230.

[YA01] L. Yin and G.K. Ananthasuresh, *Topology optimization of compliant mechanisms with multiple materials using a peak function material interpolation scheme*, Structural and Multidisciplinary Optimization **23** (2001), no. 1, 49–62 (English).

[YINT10] T. Yamada, K. Izui, S. Nishiwaki, and A. Takezawa, *A topology optimization method based on the level set method incorporating a fictitious interface energy* , Computer Methods in Applied Mechanics and Engineering **199** (2010), no. 45-48, 2876 – 2891.

[YKBS04] G.H. Yoon, Y.Y. Kim, M.P. Bendsøe, and O. Sigmund, *Hinge-free topology optimization with embedded translation-invariant differentiable wavelet shrinkage*, Structural and Multidisciplinary Optimization **27** (2004), no. 3, 139–150 (English).

[Zei85] E. Zeidler, *Nonlinear Functional Analysis and Its Applications I: Fixed point theorems*, Nonlinear Functional Analysis and Its Applications, Springer-Verlag, 1985.

[Zei88] ———, *Nonlinear Functional Analysis and Its Applications IV: Applications to Mathematical Physics*, Nonlinear Functional Analysis and Its Applications, Springer-Verlag, 1988.

[Zil93] C. Zillober, *A globally convergent version of the method of moving asymptotes*, Structural optimization **6** (1993), no. 3, 166–174 (English).

[ZK79] J. Zowe and S. Kurcyusz, *Regularity and stability for the mathematical programming problem in Banach spaces*, Applied Mathematics and Optimization **5** (1979), no. 1, 49–62 (English).

[ZR92] M. Zhou and G.I.N. Rozvany, *DCOC: An optimality criteria method for large systems Part I: theory*, Structural optimization **5** (1992), no. 1-2, 12–25 (English).

[ZR01] ———, *On the validity of ESO type methods in topology optimization*, Structural and Multidisciplinary Optimization **21** (2001), no. 1, 80–83 (English).

[ZXL15] Z. Zhao, F. Xu, and X. Li, *Adaptive projected gradient thresholding methods for constrained l_0 problems*, Science China Mathematics **58** (2015), no. 10, 1–20 (English).