Weierstraß-Institut
für Angewandte Analysis und Stochastik
Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

A PDE-constrained optimization approach for topology optimization of strained photonic devices

Lukáš Adam[1], Michael Hintermüller[1,2], Thomas M. Surowiec[3]

submitted: February 24, 2017

[1] Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
Germany
E-Mail: adam@utia.cas.cz
hint@math.hu-berlin.de

[2] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: michael.hintermueller@wias-berlin.de

[3] Philipps-Universität Marburg
FB12 Mathematik und Informatik
Hans-Meerwein Straße 6, Lahnberge
35032 Marburg
Germany
E-Mail: surowiec@mathematik.uni-marburg.de

No. 2377
Berlin 2017

2010 *Mathematics Subject Classification.* 49J20, 35Q93, 35Q74, 90C46.

Key words and phrases. Semiconductor lasers, Germanium, Topology optimization, Optimization with PDE constraints, Elasticity, Phase-field.

This work was carried out in the framework of the DFG under grant no. HI 1466/7-1 “Free Boundary Problems and Level Set Methods” as well as the Research Center MATHEON supported by the Einstein Foundation Berlin within projects OT1, SE5 and SE15.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

A PDE-constrained optimization approach for topology optimization of strained photonic devices

Lukáš Adam, Michael Hintermüller, Thomas M. Surowiec

Abstract

Recent studies have demonstrated the potential of using tensile-strained, doped Germanium as a means of developing an integrated light source for (amongst other things) future microprocessors. In this work, a multi-material phase-field approach to determine the optimal material configuration within a so-called Germanium-on-Silicon microbridge is considered. Here, an "optimal" configuration is one in which the strain in a predetermined minimal optical cavity within the Germanium is maximized according to an appropriately chosen objective functional. Due to manufacturing requirements, the emphasis here is on the cross-section of the device; i.e. a so-called aperture design. Here, the optimization is modeled as a non-linear optimization problem with partial differential equation (PDE) and manufacturing constraints. The resulting problem is analyzed and solved numerically. The theory portion includes a proof of existence of an optimal topology, differential sensitivity analysis of the displacement with respect to the topology, and the derivation of first and second-order optimality conditions. For the numerical experiments, an array of first and second-order solution algorithms in function-space are adapted to the current setting, tested, and compared. The numerical examples yield designs for which a significant increase in strain (as compared to an intuitive empirical design) is observed.

1 Introduction

Over the last several decades, the reduction in size of microprocessors has led to a significant increase in computational performance. Until recent times, this increase has essentially followed Moore's law, which states that the number of components per integrated circuit doubles every other year. However, further rises in performance will require new technologies. In particular, in order to benefit from higher switching speeds within microprocessors, an increase in the bandwidth of on-chip data transfer (currently limited by electrical wiring) is needed; cf. the discussion in [18].

One promising approach is to employ lasers to communicate between the individual parts of the microprocessor, see [44, 13, 43, 51]. Unfortunately, the base material used for integrated circuits, Silicon (Si), is an indirect-bandgap semiconductor, i.e., when an electron recombines with a hole, a photon is never released. Hence, it cannot be used to make a laser. In contrast, it has been observed that Germanium (Ge), a material with very similar properties to Si, can be used. Although Ge is also by nature an indirect-bandgap semiconductor, its band structure can be altered through the application of high tensile stress and doping, see e.g., [43]. Some recent studies concentrating on modeling these effects on the electronical and optical properties are [17]. In particular, we note that a failure of significant gain within the device is more dependent on tensile strain than the doping profile.

Several suggestions for the shape and topology as well as the composition of materials exist for the construction of a Ge-on-Si laser: [32, 13, 14, 39]. One common theme is the presence of a so-called "microbridge"created through standard etching and wetting procedures in photolithograpy. In this work,

DOI 10.20347/WIAS.PREPRINT.2377 Berlin 2017

we consider the optimization of the shape, topology, and material configuration of a cross-section of a microbridge. This is known as an aperture design [39]. The question of finding an optimal doping profile will be addressed in future work, see the discussion in Section 6 as well as the recent study [39]. The configuration of materials is essential. Indeed, the microbridge is a static object, thus, the forces (stresses) used to increase strain inside the device can only amount from the position of materials.

The complete mathematical model of a strained photonic device is given by the following system of linear and non-linear partial differential equations (PDEs). This model links mechanical, electronic, and optical properties:

Elasticity :
$$-\operatorname{div}[\mathbb{C}e(u) - F] = f, \tag{1a}$$

Semiconductors :
$$-\operatorname{div}(\varepsilon\nabla\phi) = q(C_{dop} + p - n), \tag{1b}$$
$$\dot{n} - \operatorname{div}(D_n\nabla n - \mu_n n\nabla\phi) = -R_{net}(n,p,e(u)), \tag{1c}$$
$$\dot{p} - \operatorname{div}(D_p\nabla p - \mu_p p\nabla\phi) = -R_{net}(n,p,e(u)), \tag{1d}$$

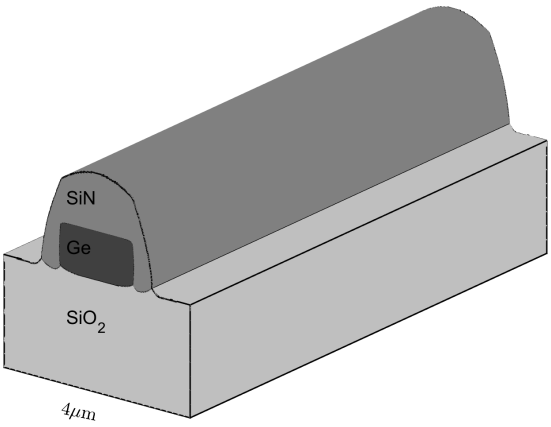
Optics :
$$\left[\nabla^2 + \frac{\omega^2}{c^2}\left(n_r + \imath\frac{c}{2\omega}(g-\alpha)\right)^2\right]\Xi_i = \beta_i^2\Xi_i, \tag{1e}$$
$$\dot{S}_i - v_{g,i}(2\operatorname{Im}\beta_i - \alpha_c)S_i - \dot{S}_{\mathsf{sp},i} = 0. \tag{1f}$$

Although our focus in this paper is on (1a), we briefly detail the full model below.

Equation (1a) is the standard model of linear elasticity, where $\mathbb{C}$ is the elasticity tensor, $e(\mathbf{u})$ the symmetric gradient of the displacement, F inner forces such as eigenstrain and f reflects body forces. In our model, both $\mathbb{C}$ and F depend on several different materials.

Equations (1b)-(1d) form the van Roosbroeck system for semiconductors, which was introduced in [47] and under some assumptions derived in this form in [35]. This drift-diffusion system also exhibts strong similarities to the classical Poisson-Nernst-Planck (PNP) system. However, the recombination rates/reaction terms are specific to this setting. Here, ϕ is the electrostatic potential, ε the permittivity tensor, n the concentration of (negatively charged) conduction electors, q is the charge, C_{dop} the doping profile and $J_n := qD_n\nabla n - q\mu_n n\nabla\phi$ is the conduction current density caused by electrons, the first summand is the diffusion part and the second one the drift part. The remaining term $R_{net}(n,p,e(u))$ can be understood as the difference of rates at which the electrons and holes recombine and at which they are generated. For this reason, R_{net} is called the recombination-generation rate. The same quantities as for electrons are present for (positively charged) holes and correspond to quantities with index p.

The last two equations (1e)-(1f) correspond to the optics. The waveguide equation (1e) is an eigenvalue problem for the optical mode Ξ_i and in the photon rate equation (1f) $\dot{S}_i$ denotes the emission and depends on the eigenstate i of the wave equation. For the optics, see also the theory of the Helmholtz equation in, e.g., [16]. In future works, we will strive to include the electronics and optics into the topology optimization process.



Some work for model (1) has been already performed in [14, 15, 30, 39, 38, 40], where two possible designs were studied. It has been found that both of them have a lower lasing threshold than the ones reported in the literature before. This is important to prevent thermal damage. The idea was to propose the materials in such a way that the insulating materials drive the current directly to the middle of the

optical cavity.

The idea to only consider (1a) in this paper stems from the observations in [43] that show a direct relation between the radiative (stimulated) recombination mechanism, denoted by R_{rad} above, and the biaxial strain within the optical cavity. Note that the optical cavity more or less corresponds to the part occupied by germanium. The secondary reason for only considering (1a) is of practical nature and ultimately stems from the fact that the rates derived in [43, 48] are not derived in closed form and therefore unsuitable for our numerical/algorithmic study. It is, however, assumed that it is monotone in $\mathrm{tr}(e(u))$ for small strains.

Summarizing, the goal of this paper is to determine an optimal composition of several materials in a given domain Ω such that the strain generated in the optical cavity $D \subset \Omega$ is maximized. This is a problem of structural optimization which is a general field, where one tries to distribute several materials into a device such that a given objective is minimized. There are many methods for structural optimization, for example we mention the boundary variation method, the free material optimization or the level set method, see [4] and references therein. The ultimate goal of the structural optimization is to find the boundaries where individual materials come into contact. However, every method handles the boundaries in a different way. For example, the boundary is described as a function in the boundary variation method while it is described as a level set of a function in the level set method. Due to its modelling flexibility, we have decided to follow a multi-material phase-field approach, suggested as in, e.g., [7, 12, 45].

The rest of the paper is organized as follows. Section 2 is devoted to mathematical modelling. This includes a discussion of the appropriate forward problem, additional constraints, the objective functional and possible extensions. In Section 3, sensitivity results, existence of an optimal topology and first-and second-order optimality conditions are derived. The differential sensitivity results play a direct role in the development of function-space-based algorithms. Note that the restriction to aperture designs allows us to work in $\mathbb{R}^2$. As such we can make use of pre-existing regularity results for linear elliptic PDEs. This allows us to work in a Hilbert space setting. In Section 4, we present several algorithmic approaches. In particular, we discuss a popular gradient flow approach, projected gradients, and interior-point methods. The performance and viability of these methods in practice is given in Section 5. Ultimately, the methods are able to provide new designs that suggest a 15% increase in strain within the optical cavity, when compared to the (empirically determined) benchmarks from [39]. As an additional service, we present a brief numerical parametric study of the topologies with respect to the regularization parameter.

2 Model

In this section, we motivate the forward model and the overall optimization problem. Before introducing the rigorous mathematical framework, we discuss the desired properties of design variables. A similar model was considered in [7], where the authors allowed for $\Omega \subset \mathbb{R}^3$. However, we restrict ourselves to 2D as the aperture design considered here appears to be the most relevant to the application. From a mathematical perspective, this restriction allows us to obtain somewhat stronger results than those in the above-mentioned paper. In particular, we obtain the differentiability of the control-to-state operator as a mapping from $H^1(\Omega, \mathbb{R}^N)$ and not only more restrictive $H^1(\Omega, \mathbb{R}^N) \cap L^\infty(\Omega, \mathbb{R}^N)$. This fact is essential for the development of function-space-based numerical methods, as we may then remain in a Hilbert space setting. Nevertheless, the proofs of existence and first-order optimality conditions closely follow those in [7].

2.1 Forward problem

As stated above, we seek to choose N materials inside a given domain Ω so that the strain in a fixed region $D \subset \Omega$ is maximized. In the context of Ge-on-Si microbridges, D is often referred to as the "optical cavity". To each material $i \in \{1, \dots, N\}$ we assign a phase-field function φ_i, whose support $\operatorname{supp} \varphi_i$ denotes the regions where material i should appear. We use the notation $\boldsymbol{\varphi} : \Omega \to \mathbb{R}^N$ to denote the vector of concentrations/phases. The components φ_i arise as parameters in a linear elliptic PDE, which describes a model of small strain elasticity. The solution of this PDE is a displacement mapping $\mathbf{u} : \Omega \to \mathbb{R}^2$. The strong form of this elasticity model is given by

$$\begin{aligned} -\operatorname{div}[\mathbb{C}(\boldsymbol{\varphi})e(\boldsymbol{u}) - F(\boldsymbol{\varphi})] = 0 \quad &\text{in} \quad \Omega, \\ \boldsymbol{u} = 0 \quad &\text{on} \quad \partial\Omega, \end{aligned} \tag{2}$$

where $e(\mathbf{u}) := \frac{1}{2}(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^\top)$ is the symmetric strain of the displacement vector $\mathbf{u}$, $\mathbb{C}(\boldsymbol{\varphi})$ is a fourth-order tensor and

$$F(\boldsymbol{\varphi}) := \epsilon_0 \mathbb{C}(\boldsymbol{\varphi}) \begin{pmatrix} \varphi_{Ge} & 0 \\ 0 & \varphi_{Ge} \end{pmatrix} - \sigma_0 \begin{pmatrix} \varphi_{SiN} & 0 \\ 0 & \varphi_{SiN} \end{pmatrix}, \tag{3}$$

incorporates the effect of the eigenstrain generated by Ge and the thermal (pre-)stress generated by SiN. For simplicity, we consider only the Dirichlet boundary condition. However, it would present no major difficulties to include Neumann or mixed boundary conditions; this would require an additional investigation of the optimal regularity of the associated solutions.

2.2 Phase-field constraints

In what follows, we list several properties that should be fulfilled by the phase-field function φ_i. First, the phases should be non-negative and normalized

$$\varphi_i \geq 0 \text{ a.e. on } \Omega, \ i = 1, \dots, N, \quad \text{and} \quad \sum_{i=1}^{N} \varphi_i = 1 \text{ a.e. on } \Omega. \tag{4}$$

In addition, we would ideally have only pure phases, i.e.,

$$\varphi_i \varphi_j = 0 \text{ a.e. on } \Omega \text{ for } i, j = 1, \dots, N \text{ with } i \neq j, \tag{5}$$

and we assume that there are some manufacturing restrictions, i.e. certain phases are fixed at the domains $\Pi_i \subset \Omega$, $i = 1, \dots, N$

$$\varphi_i = 1 \text{ a.e. on } \Pi_i, \ i = 1, \dots, N. \tag{6}$$

Finally, the coincidence sets $\{\varphi_i = 0\}$ should have finite perimeter.

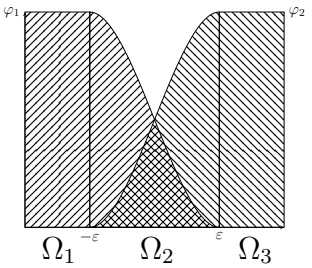


Figure 1: Phase-field model. On Ω_1 and Ω_3 the phases are pure, on Ω_2 they mix.

2.3 Optimization problem

To formulate the problem mathematically, we denote the negative strain functional by $J^0(\boldsymbol{u})$. Then we may summarize the above verbal formulation into the optimization problem

$$\begin{aligned} &\min J^0(\boldsymbol{u}) \\ &\text{s. t. } (\boldsymbol{\varphi}, \boldsymbol{u}) \text{ satisfies (2)}, \\ &\qquad \boldsymbol{\varphi} \text{ satisfies (4), (5) and (6)}. \end{aligned}$$

Since there is no term controlling the perimeter of phases, the "optimal solution" may exhibit fractal behavior, i.e., the solution may not exists in a Sobolev space setting. As a remedy, we add a perimeter term to the objective, and, with a fixed parameter $\alpha > 0$, we thus arrive at

$$\begin{aligned} &\min J^0(\boldsymbol{u}) + \alpha \sum_{i=1}^{N} \mathcal{P}(\{\varphi_i = 1\}; \Omega) \\ &\text{s. t. } (\boldsymbol{\varphi}, \boldsymbol{u}) \text{ satisfies (2)}, \\ &\qquad \boldsymbol{\varphi} \text{ satisfies (4), (5) and (6)}, \end{aligned} \tag{7}$$

where

$$\mathcal{P}(E; \Omega) := \sup \left\{ \int_E \operatorname{div} T(x) dx \,\middle|\, T \in \mathcal{C}_c^\infty(\Omega; \mathbb{R}^n),\ \|T\|_{L^\infty(\Omega)} \leq 1 \right\} \tag{8}$$

is the perimeter of $E \subset \Omega$ with respect to Ω. Note that the constraints (4) and (5) force φ_i to take only binary values, and thus $\mathcal{P}(\{\varphi_i = 1\}; \Omega)$ equals the total variation of φ_i due to the co-area formula [19, Theorem I].

Constraint system (5) together with the first part of (4) forms the so-called complementarity system. Consequently,(7) belongs to the class of mathematical problems with complementarity constraints (MPCCs), which are usually difficult to handle even in finite dimensions. This is due to the fact that standard constraint qualifications such as the linear independence constraint qualification (LICQ) or the Mangasarian-Fromovitz constraint Qualification (MFCQ) are violated at all feasible points. For the analysis of such problems in infinite dimension see, e.g., [36, 2] or more recent work by [26, 24] and the references therein.

Even though problem (7) admits an optimal solution as can be seen from Lemma A.1 in the appendix, the numerical handling of the perimeter term may be difficult. We thus replace it by the Ginzburg-Landau energy

$$f_{\mathrm{GL}}(\boldsymbol{\varphi}) := \int_\Omega \frac{\varepsilon}{2} \nabla \boldsymbol{\varphi} : \nabla \boldsymbol{\varphi} + \frac{1}{2\varepsilon} \boldsymbol{\varphi} \cdot (1 - \boldsymbol{\varphi}) dx. \tag{9}$$

It is well-known that for $\varepsilon \to 0$, the Ginzburg-Landau energy Γ-converges to the perimeter functional associated with the sets $\{\varphi_i = 1\}$; see [37]. Moreover, minimization of the second term in (9) aims to force the phases to be pure, in particular as $\varepsilon \to 0$. For this reason, we are also able to omit the complementarity constraints (5) in (7). This leads us to the following model:

$$\begin{aligned} &\min J^0(\boldsymbol{u}) + \alpha f_{\mathrm{GL}}(\boldsymbol{\varphi}) \qquad \text{over } \boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N),\ \boldsymbol{u} \in H_0^1(\Omega, \mathbb{R}^2) \\ &\text{s. t. } (\boldsymbol{\varphi}, \boldsymbol{u}) \text{ satisfies (2)}, \\ &\qquad \boldsymbol{\varphi} \in \mathcal{G}_{ad}, \end{aligned} \tag{10}$$

where we collect the constraints (4) and (6) in the Gibbs simplex $\mathcal{G}$ and the set of feasible solutions $\mathcal{G}_{ad}$:

$$\begin{aligned} \mathcal{G} &:= \left\{ \boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N) \,\middle|\, \boldsymbol{\varphi} \geq 0,\ \sum_{i=1}^{N} \boldsymbol{\varphi} = 1 \text{ a.e. on } \Omega \right\}, \\ \mathcal{G}_{ad} &:= \{ \boldsymbol{\varphi} \in \mathcal{G} \,|\, \varphi_i = 1 \text{ a.e. on } \Pi_i,\ i = 1, \ldots, N \}. \end{aligned} \tag{11}$$

Note that the Ginzburg-Landau energy requires $\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$.

Since $H^1(\Omega)$ functions do not allow jumps over 1D-manifold (recall $\Omega \subset \mathbb{R}^2$), problem (10) will, in general, not produce pure phases. On the other hand, the Ginzburg-Landau energy Γ-converges to the perimeter functional and thus, problem (10) is an approximation of problem (7). Moreover, by resorting

to this approximation, we gain several advantages: (i) Problem (10) is an optimal control problem with qualified constraints rather than a degenerate MPCC; (ii) and we are able to work in a Hilbert space setting rather than in a nonreflexive space of functions of bounded variation, which has advantages also from a numerical point of view.

2.4 Remarks on the objective J^0

Concerning the desired objective $J^0(\boldsymbol{u})$, we propose several alternatives, namely

$$J_1^0(\boldsymbol{u}) := \int_D |\operatorname{tr}(e(\boldsymbol{u}))-e_d|^2 dx, \quad J_2^0(\boldsymbol{u}) := -\int_D |\operatorname{tr}(e(\boldsymbol{u}))|^2 dx, \quad J_3^0(\boldsymbol{u}) := -\int_D \operatorname{tr}(e(\boldsymbol{u}))dx. \tag{12}$$

Each J_i^0 has advantages and disadvantages. Functional J_1^0 is the simplest one since it is convex and bounded below. However, an "optimal" or "desired" strain e_d is not always available. Minimizing J_2^0 corresponds to globally maximizing the strain in D. However, this functional is not bounded from below and is concave. Whereas the boundedness can be obtained by restricting $\boldsymbol{\varphi}$ to the feasible set, a lack of compactness properties inhibit us from providing an existence proof for (10). Finally, J_3^0 is both convex and continuous with respect to the weak-topology on $H_0^1(\Omega, \mathbb{R}^2)$, and thus avoids the problems of J_2^0.

Introducing the $\boldsymbol{\varphi}$-dependent bilinear form $a : H^1(\Omega, \mathbb{R}^N) \times H_0^1(\Omega, \mathbb{R}^2) \times H_0^1(\Omega, \mathbb{R}^2) \to \mathbb{R}$ defined by

$$a(\boldsymbol{\varphi}, \boldsymbol{u}, \boldsymbol{v}) := \int_\Omega \mathbb{C}(\boldsymbol{\varphi})e(\boldsymbol{u}) : e(\boldsymbol{v})dx \tag{13a}$$

along with the $\boldsymbol{\varphi}$-dependent linear form $\ell : H^1(\Omega, \mathbb{R}^N) \times H_0^1(\Omega, \mathbb{R}^2) \to \mathbb{R}$ given by

$$\ell(\boldsymbol{\varphi}, \boldsymbol{v}) := \int_\Omega F(\boldsymbol{\varphi}) : e(\boldsymbol{v})dx, \tag{13b}$$

we arrive at the expected weak/distributional form of the forward problem: find $\boldsymbol{u} \in H_0^1(\Omega, \mathbb{R}^2)$ such that

$$E(\boldsymbol{\varphi}, \boldsymbol{u})(\boldsymbol{v}) := a(\boldsymbol{\varphi}, \boldsymbol{u}, \boldsymbol{v}) - \ell(\boldsymbol{\varphi}, \boldsymbol{v}) = 0, \text{ for all } v \in H_0^1(\Omega, \mathbb{R}^2). \tag{13c}$$

Note that this makes use of the homogeneous Dirichlet boundary condition and Korn's inequality.

3 Existence and optimality conditions

In this section, we prove the existence of an optimal solution to (10) and we derive first- and second-order optimality conditions. We note that the sensitivity results for the control-to-state operator are based strongly on the results in [7] while ours represent several improvements for the 2D case.

3.1 Existence of an optimal topology

In the sequel, we make the following standing assumptions:

(A1) $\Omega \subset \mathbb{R}^2$ and $\Pi_i \subset \Omega$ are open bounded sets with Lipschitz boundary and Π_i are strictly separable, meaning that $\operatorname{cl}\Pi_i \cap \operatorname{cl}\Pi_j = \emptyset$ for all $i \neq j$.

(A2) Objective $J^0 : H_0^1(\Omega, \mathbb{R}^2) \to \mathbb{R}$ is finite on the feasible set, weakly lower semicontinuous and bounded below on bounded sets.

(A3) Tensor $\mathbb{C}$ is a Nemytskii/superposition operator, i.e. there is some tensor-valued mapping $\widehat{\mathbb{C}} : \mathbb{R}^N \to \mathbb{R}^{2\times2\times2\times2}$ such that $\mathbb{C}(\boldsymbol{\varphi})(x) = \widehat{\mathbb{C}}(\boldsymbol{\varphi}(x))$ almost everywhere on Ω. Moreover, it satisfies:

- There exist constants $c_2 > c_1 > 0$ such that for every $\phi \in \mathbb{R}^N$ and $E_1, E_2 \in \mathbb{R}^{2\times2} \setminus \{0\}$ we have
$$c_1\|E_1\|^2_{\mathbb{R}^{2\times2}} \leq \widehat{\mathbb{C}}(\phi)E_1 : E_1,$$
$$\widehat{\mathbb{C}}(\phi)E_1 : E_2 \leq c_2\|E_1\|_{\mathbb{R}^{2\times2}}\|E_2\|_{\mathbb{R}^{2\times2}},$$
where the matrix product is understood as $A : B = \sum_i \sum_j a_{ij}b_{ij}$.
- $\widehat{\mathbb{C}} \in \mathcal{C}^{1,1}(\mathbb{R}^N, \mathbb{R}^{2\times2\times2\times2})$, i.e., $\widehat{\mathbb{C}}$ is continuously differentiable with global Lipschitz derivative. Moreover, $\widehat{\mathbb{C}}$ is globally Lipschitz as well.

We will briefly comment on assumption (A3). Ideally, $\mathbb{C}(\boldsymbol{\varphi})$ would have the form $\mathbb{C}(\boldsymbol{\varphi}) = \sum_{i=1}^N \varphi_i\mathbb{C}_i$, where $\mathbb{C}_i$ are elasticity tensors corresponding to individual materials. Unfortunately, this would not satisfy the uniform ellipticity assumption. To deal with this difficulty, we need to add a cutoff function $\mathrm{cut} : \mathbb{R} \to \mathbb{R}$, which is uniformly positive, and set $\mathbb{C}(\boldsymbol{\varphi}) = \sum_{i=1}^N \mathrm{cut}(\varphi_i)\mathbb{C}_i$. We will mention a specific example of the cutoff function in the numerical section. In the following text, c will denote a general bounding constant. We omit the proof of the following lemma, see for example [21, Theorem 7].

Lemma 3.1. *Under assumptions (A1) and (A3) mapping $\mathbb{C} : H^1(\Omega, \mathbb{R}^N) \to L^p(\Omega, \mathbb{R}^{2\times2\times2\times2})$ and $F : H^1(\Omega, \mathbb{R}^N) \to L^p(\Omega, \mathbb{R}^N)$ are locally Lipschitz continuous and continuously Fréchet differentiable for all $p \in [1, \infty)$.*

The result used throughout the paper is the higher regularity of displacement.

Lemma 3.2. *Under assumptions (A1) and (A3) there exists $p > 2$ such that for every $\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$ the unique solution of (2) lies in $W_0^{1,p}(\Omega, \mathbb{R}^2)$. Moreover, there exists some $M > 0$ such that $\|\boldsymbol{u}\|_{W_0^{1,p}(\Omega,\mathbb{R}^2)} \leq M$ whenever $(\boldsymbol{\varphi}, \boldsymbol{u})$ solves (2) and $\boldsymbol{\varphi} \in \mathcal{G}$. Finally, the solution mapping $S : H^1(\Omega, \mathbb{R}^N) \to W_0^{1,p}(\Omega, \mathbb{R}^2)$, which assigns the state variable $\boldsymbol{u}$ to the control variable $\boldsymbol{\varphi}$, is locally Lipschitz continuous.*

Proof. From Lemma 3.1 we know that $F(\boldsymbol{\varphi}) \in L^4(\Omega, \mathbb{R}^N)$ for all $\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$. Then [5, Theorem 2.1] implies that there exists some $q > 2$ and a constant $c > 0$ such that for every $\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$ the elasticity equation (2) has a unique solution $u \in W_0^{1,q}(\Omega, \mathbb{R}^2)$ and that estimate
$$\|\boldsymbol{u}\|_{W_0^{1,q}(\Omega,\mathbb{R}^2)} \leq c\|F(\boldsymbol{\varphi})\|_{L^4(\Omega,\mathbb{R}^N)} \tag{14}$$
holds true. Note that even though this result was presented for the scalar-valued case, it may be generalized into the vector-valued case, which is needed here. Moreover, even though $\boldsymbol{\varphi}$ enters the differential operator, constant c from the previous estimate is independent of $\boldsymbol{\varphi}$ because we have uniform ellipticity from (A3). The first statement follows from the simple form of F.

Consider now $\boldsymbol{\varphi}^1, \boldsymbol{\varphi}^2 \in H^1(\Omega, \mathbb{R}^N)$ and the corresponding $\boldsymbol{u}^1, \boldsymbol{u}^2 \in W_0^{1,q}(\Omega, \mathbb{R}^2)$. Then we have
$$\int_\Omega \mathbb{C}(\boldsymbol{\varphi}^1)e(\boldsymbol{u}^1-\boldsymbol{u}^2) : e(\boldsymbol{v})dx = \int_\Omega (F(\boldsymbol{\varphi}^1)-F(\boldsymbol{\varphi}^2)) : e(\boldsymbol{v})dx - \int_\Omega (\mathbb{C}(\boldsymbol{\varphi}^1)-\mathbb{C}(\boldsymbol{\varphi}^2))e(\boldsymbol{u}^2) : e(\boldsymbol{v})dx.$$

Using again [5, Theorem 2.1] we obtain existence of some $p \in (2, \frac{q}{2})$ and another $c > 0$ (again independent of the choice of $\boldsymbol{\varphi}^1$ and $\boldsymbol{\varphi}^2$) such that

$$\begin{aligned}\|\boldsymbol{u}^1 - \boldsymbol{u}^2\|_{W_0^{1,p}(\Omega,\mathbb{R}^2)} &\leq c\|F(\boldsymbol{\varphi}^1) - F(\boldsymbol{\varphi}^2)\|_{L^{\frac{q}{2}}(\Omega,\mathbb{R}^N)} + c\|(\mathbb{C}(\boldsymbol{\varphi}^1) - \mathbb{C}(\boldsymbol{\varphi}^2))e(\boldsymbol{u}^2)\|_{L^{\frac{q}{2}}(\Omega,\mathbb{R}^{2\times 2})} \\ &\leq c\|F(\boldsymbol{\varphi}^1) - F(\boldsymbol{\varphi}^2)\|_{L^q(\Omega,\mathbb{R}^N)} + c\|(\mathbb{C}(\boldsymbol{\varphi}^1) - \mathbb{C}(\boldsymbol{\varphi}^2))\|_{L^q(\Omega,\mathbb{R}^{2\times 2})}\|e(\boldsymbol{u}^2)\|_{L^q(\Omega,\mathbb{R}^{2\times 2})} \\ &\leq c(1 + \|e(\boldsymbol{u}^2)\|_{L^q(\Omega,\mathbb{R}^{2\times 2})})\|\boldsymbol{\varphi}^1 - \boldsymbol{\varphi}^2\|_{H^1(\Omega,\mathbb{R}^N)} \\ &\leq c(1 + \|F(\boldsymbol{\varphi}^2)\|_{L^4(\Omega,\mathbb{R}^N)})\|\boldsymbol{\varphi}^1 - \boldsymbol{\varphi}^2\|_{H^1(\Omega,\mathbb{R}^N)} \\ &\leq c(1 + \|\boldsymbol{\varphi}^2\|_{H^1(\Omega,\mathbb{R}^N)})\|\boldsymbol{\varphi}^1 - \boldsymbol{\varphi}^2\|_{H^1(\Omega,\mathbb{R}^N)}.\end{aligned}$$

where we have used Lemma 3.1 and (14). This finishes the proof. □

We are now ready to show that the optimal control problem (10) admits an optimal solution. For notational simplicity we denote its objective function by

$$J(\boldsymbol{\varphi}, \boldsymbol{u}) := J^0(\boldsymbol{u}) + \frac{\alpha}{2}\int_\Omega \left(\varepsilon|\nabla\boldsymbol{\varphi}|^2 + \frac{1}{\varepsilon}\boldsymbol{\varphi}\cdot(1-\boldsymbol{\varphi})\right) dx.$$

Lemma 3.3. *Under assumptions (A1)-(A3) problem* (10) *admits an optimal solution.*

Proof. Since sets Π_i can be separated due to assumption (A1), there exists some $\boldsymbol{\varphi} \in \mathcal{G}_{ad}$. The existence of $\boldsymbol{u} \in W_0^{1,p}(\Omega, \mathbb{R}^2)$ such that pair $(\boldsymbol{\varphi}, \boldsymbol{u})$ satisfies the elasticity equation (2) follows from Lemma 3.2. Thus, problem (10) admits a feasible solution. Let $\{(\boldsymbol{\varphi}^k, \boldsymbol{u}^k)\}$ be an infimizing sequence. By Lemma 3.2 we have that $\boldsymbol{u}^k$ is uniformly bounded in $W_0^{1,p}(\Omega, \mathbb{R}^2)$. Since J^0 is bounded below on bounded sets, Ginzburg-Landau energy ensures that $\{\boldsymbol{\varphi}^k\}$ is bounded in $H^1(\Omega, \mathbb{R}^N)$. Thus, there exist (along a subsequence) $\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$ and $\boldsymbol{u} \in W_0^{1,q}(\Omega, \mathbb{R}^2)$ such that $\boldsymbol{\varphi}^k \rightharpoonup \boldsymbol{\varphi}$ in $H^1(\Omega, \mathbb{R}^N)$ and $\boldsymbol{u}^k \rightharpoonup \boldsymbol{u}$ in $W_0^{1,q}(\Omega, \mathbb{R}^2)$. Moreover, we can the subsequences such that $\boldsymbol{\varphi}^k \to \boldsymbol{\varphi}$ in $L^{\frac{2p}{p-2}}(\Omega, R^N)$.

Due to the pointwise convergence of $\boldsymbol{\varphi}^k$, we have $\boldsymbol{\varphi} \in \mathcal{G}$. Concerning the elasticity (2), we know that

$$\int_\Omega \mathbb{C}(\boldsymbol{\varphi}^k)e(\boldsymbol{v}) : e(\boldsymbol{u}^k)dx = \ell(\boldsymbol{\varphi}^k, \boldsymbol{v}) \tag{15}$$

for all $\boldsymbol{v} \in H_0^1(\Omega, \mathbb{R}^2)$. Moreover, we have

$$|\ell(\boldsymbol{\varphi}, \boldsymbol{v}) - \ell(\boldsymbol{\varphi}^k, \boldsymbol{v})| \leq \int_\Omega |(F(\boldsymbol{\varphi}) - F(\boldsymbol{\varphi}^k)) : e(\boldsymbol{v})|dx \leq \|F(\boldsymbol{\varphi}) - F(\boldsymbol{\varphi}^k)\|_{L^2(\Omega,\mathbb{R}^{2\times 2})}\|\boldsymbol{v}\|_{H_0^1(\Omega,\mathbb{R}^2)} \to 0,$$

where the convergence follows from the form of F and the strong convergence $\boldsymbol{\varphi}^k \to \boldsymbol{\varphi}$ in $L^{\frac{2p}{p-2}}(\Omega, R^N)$. Further we have

$$\int_\Omega \mathbb{C}(\boldsymbol{\varphi}^k)e(\boldsymbol{u}^k) : e(\boldsymbol{v})dx \to \int_\Omega \mathbb{C}(\boldsymbol{\varphi})e(\boldsymbol{u}) : e(\boldsymbol{v})dx$$

for any $\boldsymbol{v} \in H_0^1(\Omega, \mathbb{R}^2)$ due to $e(\boldsymbol{u}^k) \rightharpoonup e(\boldsymbol{u})$ in $L^q(\Omega, \mathbb{R}^2)$ and $\mathbb{C}(\boldsymbol{\varphi}^k) \to \mathbb{C}(\boldsymbol{\varphi})$ in $L^{\frac{2p}{p-2}}(\Omega, R^{2\times 2\times 2\times 2})$. But coupling these equations with (15) implies that $(\boldsymbol{\varphi}, \boldsymbol{u})$ is a feasible point of problem (10).

Since the Ginzburg-Landau energy is weakly lower continuous on $H^1(\Omega, \mathbb{R}^N)$ due to the Rellich-Kondrachov theorem and since J^0 possesses the same property due to assumption (A2), the whole objective J is weakly lower semicontinuous. Since $(\boldsymbol{\varphi}^k, \boldsymbol{u}^k)$ is a minimizing sequence of problem (10), there exists a sequence $\varepsilon^k \downarrow 0$ such that

$$J(\boldsymbol{\varphi}, \boldsymbol{u}) \leq \liminf_{k\to\infty} J(\boldsymbol{\varphi}^k, \boldsymbol{u}^k) \leq \liminf_{k\to\infty} \inf_{(\tilde{\boldsymbol{\varphi}}, \tilde{\boldsymbol{u}}) \text{ feasible}} J(\tilde{\boldsymbol{\varphi}}, \tilde{\boldsymbol{u}}) + \varepsilon^k = \inf_{(\tilde{\boldsymbol{\varphi}}, \tilde{\boldsymbol{u}}) \text{ feasible}} J(\tilde{\boldsymbol{\varphi}}, \tilde{\boldsymbol{u}}),$$

and thus $(\boldsymbol{\varphi}, \boldsymbol{u})$ is indeed a minimum of problem (10). □

3.2 First-order optimality conditions

The first-order optimality conditions here serve as the basis for the gradient flow and projected gradient methods as well as the interior point method developed in Section 4. To derive them, we need to show the differentiability of the control-to-state mapping, which is presented in the next lemma. As noted, we use the higher regularity of $\boldsymbol{u}$ to obtain a stronger differentiability result than in [7].

Lemma 3.4. *Assume that (A1)-(A3) hold true. Then the control-to-state mapping $S : H^1(\Omega, \mathbb{R}^N) \to H_0^1(\Omega, \mathbb{R}^2)$ is continuously Fréchet differentiable. Its directional derivative equals to $S'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi} = \boldsymbol{q}$, where $\boldsymbol{q} \in H_0^1(\Omega, \mathbb{R}^2)$ solves*

$$\int_\Omega \mathbb{C}(\boldsymbol{\varphi})e(\boldsymbol{q}) : e(\boldsymbol{v})dx = -\int_\Omega [\mathbb{C}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}]e(\boldsymbol{u}) : e(\boldsymbol{v})dx + \int_\Omega F'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi} : e(\boldsymbol{v})dx \tag{16}$$

for all $\boldsymbol{v} \in H_0^1(\Omega, \mathbb{R}^2)$.

Proof. Consider any $\boldsymbol{\varphi}$, $\delta\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$. To show the Fréchet differentiability of S, we need to show that

$$\lim_{\|\delta\boldsymbol{\varphi}\|_{H^1(\Omega,\mathbb{R}^N)}\downarrow 0} \frac{\|S(\boldsymbol{\varphi}+\delta\boldsymbol{\varphi}) - S(\boldsymbol{\varphi}) - S'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}\|_{H_0^1(\Omega,\mathbb{R}^2)}}{\|\delta\boldsymbol{\varphi}\|_{H^1(\Omega,\mathbb{R}^N)}} = 0 \tag{17}$$

Defining $\boldsymbol{u}^2 := S(\boldsymbol{\varphi}+\delta\boldsymbol{\varphi})$, $\boldsymbol{u}^1 := S(\boldsymbol{\varphi})$ and using $\boldsymbol{q}$ from (16), we have for all $\boldsymbol{v} \in H_0^1(\Omega, \mathbb{R}^2)$

$$\begin{aligned}
\int_\Omega \mathbb{C}(\boldsymbol{\varphi}+\delta\boldsymbol{\varphi})e(\boldsymbol{u}^2) : e(\boldsymbol{v})dx &= \int_\Omega F(\boldsymbol{\varphi}+\delta\boldsymbol{\varphi}) : e(\boldsymbol{v})dx,\\
-\int_\Omega \mathbb{C}(\boldsymbol{\varphi})e(\boldsymbol{u}^1) : e(\boldsymbol{v})dx &= -\int_\Omega F(\boldsymbol{\varphi}) : e(\boldsymbol{v})dx,\\
-\int_\Omega \mathbb{C}(\boldsymbol{\varphi})e(\boldsymbol{q}) : e(\boldsymbol{v})dx &= \int_\Omega [\mathbb{C}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}]e(\boldsymbol{u}^1) : e(\boldsymbol{v})dx - \int_\Omega F'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi} : e(\boldsymbol{v})dx.
\end{aligned}$$

Summing these three inequalities and rearranging the terms results in

$$\begin{aligned}
\int_\Omega \mathbb{C}(\boldsymbol{\varphi})\left[e(\boldsymbol{u}^2) - e(\boldsymbol{u}^1) - e(\boldsymbol{q})\right] : e(\boldsymbol{v})dx = &-\int_\Omega \left[\mathbb{C}(\boldsymbol{\varphi}+\delta\boldsymbol{\varphi}) - \mathbb{C}(\boldsymbol{\varphi}) - \mathbb{C}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}\right]e(\boldsymbol{u}^2) : e(\boldsymbol{v})dx\\
&+\int_\Omega \left[F(\boldsymbol{\varphi}+\delta\boldsymbol{\varphi}) - F(\boldsymbol{\varphi}) - F'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}\right] : e(\boldsymbol{v})dx\\
&+\int_\Omega \mathbb{C}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}\left[e(\boldsymbol{u}^1) - e(\boldsymbol{u}^2)\right] : e(\boldsymbol{v})dx.
\end{aligned}$$

Now we set $\boldsymbol{v} = \boldsymbol{u}^2 - \boldsymbol{u}^1 - \boldsymbol{q}$, apply Korn's lemma [53, Corollary 62.13] coupled with the ellipticity assumption from (A3) on the left-hand side to obtain

$$\begin{aligned}
\|\boldsymbol{u}^2 - \boldsymbol{u}^1 - \boldsymbol{q}\|^2_{H_0^1(\Omega,\mathbb{R}^2)} \le\ & c\|\mathbb{C}(\boldsymbol{\varphi}+\delta\boldsymbol{\varphi}) - \mathbb{C}(\boldsymbol{\varphi}) - \mathbb{C}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}\|_X \|\boldsymbol{u}^2\|_{W_0^{1,p}(\Omega,\mathbb{R}^2)} \|\boldsymbol{u}^2 - \boldsymbol{u}^1 - \boldsymbol{q}\|_{H_0^1(\Omega,\mathbb{R}^2)}\\
&+ c\|F(\boldsymbol{\varphi}+\delta\boldsymbol{\varphi}) - F(\boldsymbol{\varphi}) - F'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}\|_{L^2(\Omega,\mathbb{R}^{2\times 2})} \|\boldsymbol{u}^2 - \boldsymbol{u}^1 - \boldsymbol{q}\|_{H_0^1(\Omega,\mathbb{R}^2)}\\
&+ c\|\mathbb{C}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}\|_X \|\boldsymbol{u}^1 - \boldsymbol{u}^2\|_{W_0^{1,p}(\Omega,\mathbb{R}^2)} \|\boldsymbol{u}^2 - \boldsymbol{u}^1 - \boldsymbol{q}\|_{H_0^1(\Omega,\mathbb{R}^2)},
\end{aligned}$$

where $p > 2$ is the exponent mentioned from Lemma 3.3 and $X := L^q(\Omega, \mathbb{R}^{2\times 2\times 2\times 2})$ with $q := \frac{2p}{p-2}$. Dividing both sides by $\|\boldsymbol{u}^2 - \boldsymbol{u}^1 - \boldsymbol{q}\|_{H_0^1(\Omega,\mathbb{R}^2)}\|\delta\boldsymbol{\varphi}\|_{H^1(\Omega,\mathbb{R}^N)}$, we realize that the left-hand side coincides with the difference quotient in (17) and the right-hand side converges to zero due to Lemmas 3.1 and 3.2. Thus, we have shown that S is Fréchet differentiable. The continuity of the derivative may be shown similarly as in Lemma 3.3. □

Recall that the objective of problem (10) is denoted by J and define further the reduced functional

$$\mathcal{J}(\boldsymbol{\varphi}) := J(\boldsymbol{\varphi}, S(\boldsymbol{\varphi})).$$

Since S is differentiable, it is not surprising that $\mathcal{J}$ possesses the same property.

Lemma 3.5. *Assume that (A1)-(A3) hold true and consider $\boldsymbol{\varphi} \in \mathcal{G}$. If $J^0 : H_0^1(\Omega, \mathbb{R}^2) \to \mathbb{R}$ is (continuously) Fréchet differentiable at $\boldsymbol{\varphi}$, then $\mathcal{J} : H^1(\Omega, \mathbb{R}^N) \to \mathbb{R}$ is (continuously) Fréchet differentiable at $\boldsymbol{\varphi}$ as well. For any $\delta\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$ its directional derivative equals to*

$$\begin{aligned}\mathcal{J}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi} = \alpha \int_\Omega \left(\varepsilon \nabla\boldsymbol{\varphi} : \nabla\delta\boldsymbol{\varphi} + \frac{1}{2\varepsilon}(1 - 2\boldsymbol{\varphi})\cdot\delta\boldsymbol{\varphi} \right) dx \\ + \int_\Omega [\mathbb{C}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}]e(\boldsymbol{u}) : e(\boldsymbol{p})dx - \int_\Omega F'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi} : e(\boldsymbol{p})dx\end{aligned} \tag{18}$$

where $\boldsymbol{p} \in H_0^1(\Omega, \mathbb{R}^2)$ is the solution to the adjoint equation

$$\begin{aligned} -\operatorname{div} \mathbb{C}(\boldsymbol{\varphi})e(\boldsymbol{p}) &= -(J_u^0)'(\boldsymbol{\varphi}, \boldsymbol{u}) \qquad \text{in} \quad \Omega, \\ \boldsymbol{p} &= 0 \qquad\qquad\qquad\quad \text{on} \quad \partial\Omega. \end{aligned} \tag{19}$$

Proof. The (continuous) differentiability of $\mathcal{J}$ at $\boldsymbol{\varphi}$ follows from the chain rule. By the standard technique, see [29, Section 1.6.2], we obtain that

$$\mathcal{J}'(\boldsymbol{\varphi}) = J'_\varphi(\boldsymbol{\varphi}, \boldsymbol{u}) + E'_\varphi(\boldsymbol{\varphi}, \boldsymbol{u})^*\boldsymbol{p}, \tag{20}$$

where $\boldsymbol{p} \in H_0^1(\Omega, \mathbb{R}^2)$ is the solution of the adjoint equation $E'_u(\boldsymbol{\varphi}, \boldsymbol{u})^*\boldsymbol{p} = -(J_u^0)'(\boldsymbol{\varphi}, \boldsymbol{u})$. Due to the linearity of $E(\boldsymbol{\varphi}, \cdot)$ and the symmetry of $a(\boldsymbol{\varphi}, \cdot, \cdot)$, the adjoint equation simplifies into (19). Hence,

$$\mathcal{J}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi} = J'_\varphi(\boldsymbol{\varphi}, \boldsymbol{u})\delta\boldsymbol{\varphi} + \langle E'_\varphi(\boldsymbol{\varphi}, \boldsymbol{u})^*\boldsymbol{p}, \delta\boldsymbol{\varphi}\rangle = J'_\varphi(\boldsymbol{\varphi}, \boldsymbol{u})\delta\boldsymbol{\varphi} + \langle E'_\varphi(\boldsymbol{\varphi}, \boldsymbol{u})\delta\boldsymbol{\varphi}, \boldsymbol{p}\rangle,$$

from which (18) follows by substitution. □

With the previous lemma at hand, it is not difficult to derive the necessary optimality conditions.

Theorem 3.6. *Assume (A1)-(A3) hold and let $(\boldsymbol{\varphi}, \boldsymbol{u})$ be an optimal solution to* (10)*. Then the following first-order optimality condition holds*

$$\begin{aligned}\alpha \int_\Omega \left(\varepsilon \nabla\boldsymbol{\varphi} : (\nabla\hat{\boldsymbol{\varphi}} - \nabla\boldsymbol{\varphi}) + \frac{1}{2\varepsilon}(1 - 2\boldsymbol{\varphi})\cdot(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) \right) dx \\ + \int_\Omega [\mathbb{C}'(\boldsymbol{\varphi})(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})]e(\boldsymbol{u}) : e(\boldsymbol{p})dx - \int_\Omega F'(\boldsymbol{\varphi})(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) : e(\boldsymbol{p})dx \geq 0 \text{ for all } \hat{\boldsymbol{\varphi}} \in \mathcal{G}_{ad},\end{aligned} \tag{21}$$

where $\boldsymbol{p}$ solves the adjoint equation (19)*.*

Proof. The variational inequality (21) arises directly from the standard first-order necessary optimality condition $\mathcal{J}'(\boldsymbol{\varphi})(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) \geq 0$ for all $\hat{\boldsymbol{\varphi}} \in \mathcal{G}_{ad}$. The rest follows from Lemma 3.5. □

Formally, the Karush-Kuhn-Tucker (KKT) conditions for (10) would take the following form: there exist $\boldsymbol{\lambda} \in H^1(\Omega, \mathbb{R}^N)^*$ and $\mu \in H^1(\Omega)^*$ such that $\boldsymbol{\lambda} \geq 0$, $\langle \boldsymbol{\lambda}, \boldsymbol{\varphi} \rangle = 0$ and

$$\begin{aligned} J'_{\varphi}(\boldsymbol{\varphi}, \boldsymbol{u}) + E'_{\varphi}(\boldsymbol{\varphi}, \boldsymbol{u})^* \boldsymbol{p} - \lambda + \mathbf{1}\mu &= 0, \\ \int_\Omega \mathbb{C}(\boldsymbol{\varphi}) e(\boldsymbol{p}) : e(\boldsymbol{v}) dx + \langle (J_u^0)'(\boldsymbol{\varphi}, \boldsymbol{u}), \boldsymbol{v} \rangle &= 0 \qquad \text{for all } \boldsymbol{v} \in H_0^1(\Omega, \mathbb{R}^2), \\ \int_\Omega \mathbb{C}(\boldsymbol{\varphi}) e(\boldsymbol{u}) : e(\boldsymbol{v}) dx - \int_\Omega F(\boldsymbol{\varphi}) : e(\boldsymbol{v}) dx &= 0 \qquad \text{for all } \boldsymbol{v} \in H_0^1(\Omega, \mathbb{R}^2), \\ \sum_{i=1}^N \varphi_i - 1 &= 0, \\ \langle \boldsymbol{\lambda}, \boldsymbol{\varphi} \rangle &= 0 \end{aligned} \tag{22}$$

However, the usual method of deriving the existence of such multipliers via the constraint qualification (cf. [10, 41, 54]):

$$0 \in \operatorname{int}\{\sum_{i=1}^N \varphi_i - 1 |\ \boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N),\ \boldsymbol{\varphi} \geq 0 \text{ a.e. on } \Omega\}$$

fails due to the discrepancy between the H^1 and L^∞ norms. Nevertheless, we can still use the discrete form of (22) to develop a numerical method, see Section 4.

3.3 Second-order optimality conditions

In order to understand the stability of local minima and derive error estimates for finite element discretizations, we typically require second-order optimality conditions. To this aim, we first show that the control-to-state mapping is twice continuously differentiable. The proof of this result basically copies the one of Lemma 3.4. In this section, we need to strengthen assumption (A3) and assume that $\widehat{\mathbb{C}}$ is twice continuously differentiable with second derivative being globally Lipschitz.

Moreover, we note that our method is inspired by the general approach presented in [10, Chapters 3.2, 3.3]. However, there are some differences, which we detail as they come. Though there may be more general conditions, the results in Theorems 3.8 and 3.9 capture the basic forms.

Lemma 3.7. *Assume that (A1)-(A3) hold true. Then the control-to-state mapping $S : H^1(\Omega, \mathbb{R}^N) \to H_0^1(\Omega, \mathbb{R}^2)$ is twice Fréchet differentiable. Its directional derivative equals to $[S''(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^1]\delta\boldsymbol{\varphi}^2 = \boldsymbol{r}$, where $\boldsymbol{r} \in H_0^1(\Omega, \mathbb{R}^2)$ solves*

$$\begin{aligned} \int_\Omega \mathbb{C}(\boldsymbol{\varphi}) e(\boldsymbol{r}) : e(\boldsymbol{v}) dx = &- \int_\Omega [\mathbb{C}''(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^1]\delta\boldsymbol{\varphi}^2 e(\boldsymbol{u}) : e(\boldsymbol{v}) dx - \int_\Omega \mathbb{C}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^1 e(\boldsymbol{q}^2) : e(\boldsymbol{v}) dx \\ &- \int_\Omega \mathbb{C}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^2 e(\boldsymbol{q}^1) : e(\boldsymbol{v}) dx + \int_\Omega [F''(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^1]\delta\boldsymbol{\varphi}^2 : e(\boldsymbol{v}) dx \end{aligned}$$

for all $\boldsymbol{v} \in H_0^1(\Omega, \mathbb{R}^2)$. Here we have denoted $\boldsymbol{q}^1 = S'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^1$ and $\boldsymbol{q}^2 = S'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^2$.

Proof. Since $S' \in \mathcal{L}(H^1(\Omega, \mathbb{R}^N), \mathcal{L}(H^1(\Omega, \mathbb{R}^N), H_0^1(\Omega, \mathbb{R}^2)))$, we need to show that

$$\lim_{\|\delta\boldsymbol{\varphi}^1\|_{H^1} \downarrow 0} \sup_{\|\delta\boldsymbol{\varphi}^2\|_{H^1}=1} \frac{\|S'(\boldsymbol{\varphi} + \delta\boldsymbol{\varphi}^1)\delta\boldsymbol{\varphi}^2 - S'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^2 - [S''(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^1]\delta\boldsymbol{\varphi}^2\|_{H_0^1(\Omega,\mathbb{R}^2)}}{\|\delta\boldsymbol{\varphi}^1\|_{H^1(\Omega,\mathbb{R}^N)}} = 0.$$

Defining $\hat{\boldsymbol{q}} := S'(\boldsymbol{\varphi}+\delta\boldsymbol{\varphi}^1)\delta\boldsymbol{\varphi}^2$ and $\hat{\boldsymbol{u}} := S(\boldsymbol{\varphi}+\delta\boldsymbol{\varphi}^1)$, from Lemma 3.4 we know that for all $\boldsymbol{v} \in H_0^1(\Omega,\mathbb{R}^2)$ we have

$$\begin{aligned}
\int_\Omega \mathbb{C}(\boldsymbol{\varphi}+\delta\boldsymbol{\varphi}^1)e(\hat{\boldsymbol{q}}) : e(\boldsymbol{v})dx &= -\int_\Omega \mathbb{C}'(\boldsymbol{\varphi}+\delta\boldsymbol{\varphi}^1)\delta\boldsymbol{\varphi}^2 e(\hat{\boldsymbol{u}}) : e(\boldsymbol{v})dx + \int_\Omega F'(\boldsymbol{\varphi}+\delta\boldsymbol{\varphi}^1)\delta\boldsymbol{\varphi}^2 : e(\boldsymbol{v})dx \\
-\int_\Omega \mathbb{C}(\boldsymbol{\varphi})e(\boldsymbol{q}^2) : e(\boldsymbol{v})dx &= \int_\Omega \mathbb{C}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^2 e(\boldsymbol{u}) : e(\boldsymbol{v})dx - \int_\Omega F'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^2 : e(\boldsymbol{v})dx \\
-\int_\Omega \mathbb{C}(\boldsymbol{\varphi})e(\boldsymbol{r}) : e(\boldsymbol{v})dx &= \int_\Omega [\mathbb{C}''(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^1]\delta\boldsymbol{\varphi}^2 e(\boldsymbol{u}) : e(\boldsymbol{v})dx + \int_\Omega \mathbb{C}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^1 e(\boldsymbol{q}^2) : e(\boldsymbol{v})dx \\
&\quad + \int_\Omega \mathbb{C}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^2 e(\boldsymbol{q}^1) : e(\boldsymbol{v})dx - \int_\Omega [F''(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^1]\delta\boldsymbol{\varphi}^2 : e(\boldsymbol{v})dx.
\end{aligned}$$

Summing these three equalities and rearranging the terms results in

$$\begin{aligned}
&\int_\Omega \mathbb{C}(\boldsymbol{\varphi})(e(\hat{\boldsymbol{q}}) - e(\boldsymbol{q}^2) - e(\boldsymbol{r})) : e(\boldsymbol{v})dx \\
&\quad = \int_\Omega \left(F'(\boldsymbol{\varphi}+\delta\boldsymbol{\varphi}^1)\delta\boldsymbol{\varphi}^2 - F'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^2 - [F''(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^1]\delta\boldsymbol{\varphi}^2\right) : e(\boldsymbol{v})dx \\
&\quad - \int_\Omega (\mathbb{C}(\boldsymbol{\varphi}+\delta\boldsymbol{\varphi}_1) - \mathbb{C}(\boldsymbol{\varphi}) - \mathbb{C}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}_1)e(\hat{\boldsymbol{q}}) : e(\boldsymbol{v})dx \\
&\quad - \int_\Omega \mathbb{C}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}_1(e(\hat{\boldsymbol{q}}) - e(\boldsymbol{q}^2)) : e(\boldsymbol{v})dx \\
&\quad - \int_\Omega (\mathbb{C}'(\boldsymbol{\varphi}+\delta\boldsymbol{\varphi}_1)\delta\boldsymbol{\varphi}_2 - \mathbb{C}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}_2 - [\mathbb{C}''(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}_1]\delta\boldsymbol{\varphi}_2)\, e(\hat{\boldsymbol{u}}) : e(\boldsymbol{v})dx \\
&\quad - \int_\Omega [\mathbb{C}''(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}_1]\delta\boldsymbol{\varphi}_2(e(\hat{\boldsymbol{u}}) - e(\boldsymbol{u})) : e(\boldsymbol{v})dx \\
&\quad - \int_\Omega \mathbb{C}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^2(e(\hat{\boldsymbol{u}}) - e(\boldsymbol{u}) - e(\boldsymbol{q}^1)) : e(\boldsymbol{v})dx
\end{aligned}$$

Now set $\boldsymbol{v} = \hat{\boldsymbol{q}} - \boldsymbol{q}^2 - \boldsymbol{r}$, apply Korn's lemma [53, Corollary 62.13] coupled with the ellipticity assumption from (A3) on the left-hand side and estimates on the right-hand side to obtain

$$\lim_{\|\delta\boldsymbol{\varphi}^1\|_{H^1}\downarrow 0} \sup_{\|\delta\boldsymbol{\varphi}^2\|_{H^1}=1} \frac{\|\hat{\boldsymbol{q}} - \boldsymbol{q}^2 - \boldsymbol{r}\|_{H_0^1(\Omega,\mathbb{R}^2)}}{\|\delta\boldsymbol{\varphi}_1\|_{H^1(\Omega,\mathbb{R}^N)}} = 0,$$

which is precisely what is needed to show that S is twice differentiable. □

For simplicity we will work only with the linear objective J_3^0 and define linear functional

$$\tau_D(\boldsymbol{v}) := \int_D \operatorname{tr}(e(\boldsymbol{v}))dx$$

Before stating the second-order conditions, we recall some results of convex analysis. We cannot work with the explicit multipliers for the non-negativity and normalization constraints as in (22). But realizing that (21) is nothing else than $\mathcal{J}'(\boldsymbol{\varphi})(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) \geq 0$ for all $\hat{\boldsymbol{\varphi}} \in \mathcal{G}$, due to to the convexity of $\mathcal{G}$, it is possible to write (21) equivalently as

$$0 \in \mathcal{J}'(\boldsymbol{\varphi}) + N_{\mathcal{G}}(\boldsymbol{\varphi}),$$

where N denotes the normal cone to a convex set. Hence, for the multiplier $\boldsymbol{\lambda}$ associated with the whole Gibbs simplex (and not with the individual constraints), we will have $\boldsymbol{\lambda} = -\mathcal{J}'(\boldsymbol{\varphi})$. Further, we recall the definition of the radial cone

$$R_{\mathcal{G}}(\boldsymbol{\varphi}) = \{\boldsymbol{d} \in H^1(\Omega, \mathbb{R}^N) | \text{ there exists } t > 0 : \boldsymbol{\varphi} + t\boldsymbol{d} \in \mathcal{G}\}$$

and the annihilator $\{\cdot\}^\perp$ using the dual pairing on $H^1(\Omega, \mathbb{R}^N)$ and $H^1(\Omega, \mathbb{R}^N)^*$. Finally we define linear operator $A : H^1(\Omega, \mathbb{R}^N) \to H^1(\Omega, \mathbb{R}^N)^*$ by

$$\langle A\boldsymbol{\varphi}, v\rangle = \int_\Omega \nabla\boldsymbol{\varphi} : \nabla\boldsymbol{v} dx, \quad \boldsymbol{v} \in H^1(\Omega, \mathbb{R}^N).$$

Theorem 3.8. *Let (A1)-(A3) be satisfied and let $\boldsymbol{\varphi} \in \mathcal{G}$ be a local minimum of* (10)*. Then*

$$0 \leq \alpha\varepsilon \int_\Omega |\nabla\boldsymbol{d}|^2 dx - \frac{\alpha}{\varepsilon}\int_\Omega \sum_{i=1}^N d_i^2 dx - \tau_D([S''(\boldsymbol{\varphi})\boldsymbol{d}]\boldsymbol{d}) \tag{23}$$

for all $\boldsymbol{d} \in \overline{R_{\mathcal{G}}(\boldsymbol{\varphi}) \cap \{-\mathcal{J}'(\boldsymbol{\varphi})\}^\perp}$.

Proof. Set $\boldsymbol{\lambda} := -\mathcal{J}'(\boldsymbol{\varphi})$ and fix any $\boldsymbol{d} \in R_{\mathcal{G}}(\boldsymbol{\varphi}) \cap \{\boldsymbol{\lambda}\}^\perp$. Then we may write

$$\boldsymbol{v} := -\alpha\varepsilon A\boldsymbol{\varphi} + \frac{\alpha}{2\varepsilon}\mathbf{1} - \frac{\alpha}{\varepsilon}\boldsymbol{\varphi} - \boldsymbol{r} + \boldsymbol{\lambda} = 0, \tag{24}$$

where $\boldsymbol{r} := (\boldsymbol{S}'(\boldsymbol{\varphi}))^*\tau_D$.

Since $\boldsymbol{\varphi}$ is a local minimum of (10), due to (24) we have for any $t > 0$ small enough

$$\begin{aligned} 0 \leq \frac{\mathcal{J}(\boldsymbol{\varphi} + t\boldsymbol{d}) - \mathcal{J}(\boldsymbol{\varphi})}{\frac{1}{2}t^2} &= \frac{\mathcal{J}(\boldsymbol{\varphi} + t\boldsymbol{d}) - \mathcal{J}(\boldsymbol{\varphi}) - t\langle \boldsymbol{v}, \boldsymbol{d}\rangle}{\frac{1}{2}t^2} \\ &= \frac{\mathcal{J}(\boldsymbol{\varphi} + t\boldsymbol{d}) - \mathcal{J}(\boldsymbol{\varphi}) - t\langle -\alpha\varepsilon A\boldsymbol{\varphi} + \frac{\alpha}{2\varepsilon}\mathbf{1} - \frac{\alpha}{\varepsilon}\boldsymbol{\varphi} - \boldsymbol{r} + \boldsymbol{\lambda}, \boldsymbol{d}\rangle}{\frac{1}{2}t^2}. \end{aligned} \tag{25}$$

We now group like-terms in order to simplify (25). Consider first the terms in the Ginzburg-Landau energy:

$$\frac{\alpha\varepsilon}{2}\int_\Omega |\nabla\boldsymbol{\varphi} + t\nabla\boldsymbol{d}|^2 dx - \frac{\alpha\varepsilon}{2}\int_\Omega |\nabla\boldsymbol{\varphi}|^2 dx - t\langle -\alpha\varepsilon A\boldsymbol{\varphi}, \boldsymbol{d}\rangle = \frac{\alpha\varepsilon t^2}{2}\int_\Omega |\nabla\boldsymbol{d}|^2 dx, \tag{26a}$$

$$\frac{\alpha}{2\varepsilon}\int_\Omega \sum_{i=1}^N (\varphi_i + td_i) dx - \frac{\alpha}{2\varepsilon}\int_\Omega \sum_{i=1}^N \varphi_i dx - \frac{\alpha t}{2\varepsilon}\int_\Omega \sum_{i=1}^N d_i dx = 0, \tag{26b}$$

$$-\frac{\alpha}{2\varepsilon}\int_\Omega \sum_{i=1}^N (\varphi_i + td_i)^2 dx + \frac{\alpha}{2\varepsilon}\int_\Omega \sum_{i=1}^N (\varphi_i)^2 dx + \frac{\alpha t}{\varepsilon}(\boldsymbol{\varphi}, \boldsymbol{d})_{L^2(\Omega;\mathbb{R}^N)} = -\frac{\alpha t^2}{2\varepsilon}\int_\Omega \sum_{i=1}^N d_i^2 dx. \tag{26c}$$

Now, substituting (26a), (26b), and (26c) into (25), we have:

$$0 \leq \alpha\varepsilon \int_\Omega |\nabla\boldsymbol{d}|^2 dx - \frac{\alpha}{\varepsilon}\int_\Omega \sum_{i=1}^N d_i^2 dx - \frac{\tau_D(S(\boldsymbol{\varphi} + t\boldsymbol{d})) - \tau_D(S(\boldsymbol{\varphi})) - t\tau_D(S'(\boldsymbol{\varphi})\boldsymbol{d})}{\frac{1}{2}t^2} - \frac{t\langle\boldsymbol{\lambda}, \boldsymbol{d}\rangle}{\frac{1}{2}t^2}. \tag{27}$$

Due to Lemma 3.7, the solution mapping $S(\boldsymbol{\varphi})$ has a second-order expansion of the type

$$S(\boldsymbol{\varphi}+t\boldsymbol{d}) = S(\boldsymbol{\varphi}) + tS'(\boldsymbol{\varphi})\boldsymbol{d} + \frac{t^2}{2!}[S''(\boldsymbol{\varphi})\boldsymbol{d}]\boldsymbol{d} + o(t^2), \tag{28}$$

Since we assumed $\langle \boldsymbol{\lambda}, \boldsymbol{d}\rangle = 0$, this allows us to reduce (27) into

$$0 \leq \alpha\varepsilon \int_\Omega |\nabla \boldsymbol{d}|^2 dx - \frac{\alpha}{\varepsilon}\int_\Omega \sum_{i=1}^N \boldsymbol{d}_i^2 dx - \tau_D([S''(\boldsymbol{\varphi})\boldsymbol{d}]\boldsymbol{d}) + \frac{o(t^2)}{t^2}, \tag{29}$$

Passing to the limit in t, we obtain (23). Since $S(\boldsymbol{\varphi})$ is twice differentiable, we may pass to the closure of $R_{\mathcal{G}}(\boldsymbol{\varphi}) \cap \{\boldsymbol{\lambda}\}^\perp$. □

It is known that no "curvature" should appear in either second-order necessary or sufficient optimality conditions if the underlying constraint set is polyhedric in the sense of Haraux, i.e., if $T_M(\varphi)$ is the tangent cone and $\lambda \in N_M(\varphi)$, then M is polyhedric provided

$$\overline{R_M(\varphi) \cap \{\lambda\}^\perp} = T_M(\varphi) \cap \{\lambda\}^\perp.$$

If we have two phases, then we obtain the polyhedricity of the Gibbs simplex by similar arguments as in [36]. The general case of N phases goes beyond the scope and purpose of this text. Note that the results of [50] cannot be used here and that if M is not polyhedric, then $\overline{R_{\mathcal{G}}(\boldsymbol{\varphi}) \cap \{-\mathcal{J}'(\boldsymbol{\varphi})\}^\perp}$ might be too small, i.e., $\{0\}$.

For the sufficient second-order conditions, we define cone

$$\mathcal{K}_\eta(\boldsymbol{\varphi}, \boldsymbol{\lambda}) := \left\{ \boldsymbol{d} \in H^1(\Omega, \mathbb{R}^N) \mid \boldsymbol{d} \in R_{\mathcal{G}}(\boldsymbol{\varphi}) \text{ and } -\eta\|d\|_{H^1} \leq \langle \boldsymbol{\lambda}, \boldsymbol{d}\rangle \leq 0 \right\}.$$

This is strongly reminiscent of the *approximate critical cone* used in [10]. However, we there are several key differences:

1 Here, $\boldsymbol{\varphi}$ is assumed to be a stationary point. For [10], the approximate critical cone is defined for any feasible point.

2 We make direct use of dual information, i.e., $\boldsymbol{\lambda} \in N_{\mathcal{G}}(\boldsymbol{\varphi})$, in the definition of $\mathcal{K}_\eta(\boldsymbol{\varphi}, \boldsymbol{\lambda})$.

3 Here, $\boldsymbol{d} \in \mathcal{K}_\eta(\boldsymbol{\varphi}, \boldsymbol{\lambda})$ is required to be in the radial cone $R_{\mathcal{G}}(\boldsymbol{\varphi})$, whereas in [10] $\boldsymbol{d}$ is taken to be "close" to the linearization cone.

In particular, 3. means that $\mathcal{K}_\eta(\boldsymbol{\varphi}, \boldsymbol{\lambda})$ is potentially smaller than the approximate critical cone used in [10].

Theorem 3.9. *Assume that (A1)-(A3) holds and let $\boldsymbol{\varphi}$ be a stationary point and assume that the following growth condition holds: there exist $\eta > 0$ and some $\beta > 0$ such that*

$$\alpha\varepsilon \int_\Omega |\nabla \boldsymbol{d}|^2 dx - \frac{\alpha}{\varepsilon}\int_\Omega \sum_{i=1}^N d_i^2 dx - \tau_D([u''(\boldsymbol{\varphi})\boldsymbol{d}]\boldsymbol{d}) \geq \beta\|\boldsymbol{d}\|_{H^1}^2, \ \forall \boldsymbol{d} \in \mathcal{K}_\eta(\boldsymbol{\varphi}, -\mathcal{J}'(\boldsymbol{\varphi})). \tag{30}$$

Then $\boldsymbol{\varphi}$ is a strong local minimum of (10) *meaning that there exists $\delta > 0$ and a neighborhood $\mathcal{U}$ of $\boldsymbol{\varphi}$ such that for all $\hat{\boldsymbol{\varphi}} \in \mathcal{U} \cap \mathcal{G}$ we have*

$$\mathcal{J}(\hat{\boldsymbol{\varphi}}) - \mathcal{J}(\boldsymbol{\varphi}) \geq \delta\|\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|^2. \tag{31}$$

Proof. Assume that (31) is not true. Then there exists some $\boldsymbol{\varphi}_n \in \mathcal{G}$ such that

$$\mathcal{J}(\boldsymbol{\varphi}_n) - \mathcal{J}(\boldsymbol{\varphi}) < \frac{\beta}{4}\|\boldsymbol{\varphi}_n - \boldsymbol{\varphi}\|^2. \tag{32}$$

Defining $\boldsymbol{d}_n = \frac{\boldsymbol{\varphi}_n - \boldsymbol{\varphi}}{\|\boldsymbol{\varphi}_n - \boldsymbol{\varphi}\|_{H^1(\Omega,\mathbb{R}^N)}}$ and $t_n := \|\boldsymbol{\varphi}_n - \boldsymbol{\varphi}\|_{H^1(\Omega,\mathbb{R}^N)}$, we obtain $\boldsymbol{\varphi}_n = \boldsymbol{\varphi} + t_n\boldsymbol{d}_n$, $\|\boldsymbol{d}_n\|_{H^1(\Omega,\mathbb{R}^N)} = 1$ and $t_n > 0$ with $t_n \downarrow 0$. Define further $\boldsymbol{\lambda} := -\mathcal{J}'(\boldsymbol{\varphi})$. Since $\boldsymbol{\varphi}$ is a stationary point, we have $\boldsymbol{\lambda} \in N_{\mathcal{G}}(\boldsymbol{\varphi})$ and

$$\langle \boldsymbol{\lambda}, \boldsymbol{d}_n\rangle = -\mathcal{J}'(\boldsymbol{\varphi})\boldsymbol{d}_n = -\frac{\mathcal{J}(\boldsymbol{\varphi}_n) - \mathcal{J}(\boldsymbol{\varphi})}{t_n} + \frac{o(t_n)}{t_n} > -\frac{\beta}{4}t_n + \frac{o(t_n)}{t_n}, \tag{33}$$

where we used the differentiability of $\mathcal{J}$ and (32). Moreover, $\boldsymbol{\lambda} \in N_{\mathcal{G}}(\boldsymbol{\varphi})$ and $\boldsymbol{\varphi}_n \in \mathcal{G}$ imply that $0 \geq \langle \boldsymbol{\lambda}, \boldsymbol{\varphi}_n - \boldsymbol{\varphi}\rangle = t_n\langle \boldsymbol{\lambda}, \boldsymbol{d}_n\rangle$. This yields the inequality $\langle \boldsymbol{\lambda}, \boldsymbol{d}_n\rangle \leq 0$, which together with (33) implies that for large enough n we have $\boldsymbol{d}_n \in \mathcal{K}_\eta(\boldsymbol{\varphi}, \boldsymbol{\lambda})$. But then by similar arguments as in the proof of Theorem 3.8 we obtain

$$\begin{aligned}
\frac{\beta}{2} &\geq \frac{\mathcal{J}(\boldsymbol{\varphi}_n) - \mathcal{J}(\boldsymbol{\varphi})}{\frac{1}{2}t_n^2} = \frac{\mathcal{J}(\boldsymbol{\varphi}_n) - \mathcal{J}(\boldsymbol{\varphi}) - t_n\langle 0, \boldsymbol{d}_n\rangle}{\frac{1}{2}t_n^2} \\
&= \alpha\varepsilon\int_\Omega |\nabla \boldsymbol{d}_n|^2 dx - \frac{\alpha}{\varepsilon}\int_\Omega \sum_{i=1}^N (d_n)_i^2 dx - \tau_D([u''(\boldsymbol{\varphi})\boldsymbol{d}_n]\boldsymbol{d}_n) + \frac{o(t_n^2)}{t_n^2} - \frac{t_n\langle \boldsymbol{\lambda}, \boldsymbol{d}_n\rangle}{\frac{1}{2}t^2} \\
&\geq \beta + \frac{o(t_n^2)}{t_n^2} - \frac{t_n\langle \boldsymbol{\lambda}, \boldsymbol{d}_n\rangle}{\frac{1}{2}t^2}
\end{aligned}$$

Since $\boldsymbol{\lambda} \in N_{\mathcal{G}}(\boldsymbol{\varphi})$, we have $-t_n\langle \boldsymbol{\lambda}, \boldsymbol{d}_n\rangle = -\langle \boldsymbol{\lambda}, \boldsymbol{\varphi}_n - \boldsymbol{\varphi}\rangle \geq 0$. But then by passing to the limit, we obtain $\frac{\beta}{2} \geq \beta$, which is a contradiction. □

4 Numerical methods

The explicit form in (21) lends itself nicely to several numerical approaches, e.g., a non-smooth gradient flow, projected gradients, and interior point methods. We will shortly describe these methods in this section. For the simplicity of presentation, we assume that the prescribed domains Π_i are empty. These domains can be incorporated in a simple way by using an affine linear operator from a reduced domain to Ω.

4.1 Gradient flow

Gradient flow is a commonly used technique [3, 9], in which one introduces an artificial dependence of $\boldsymbol{\varphi}$ on time. Adding $\alpha\varepsilon\frac{\partial\boldsymbol{\varphi}}{\partial t}$ to (21), a semi-implicit discretization is considered in which the material tensors and their sensitivities remain fixed at each time step. This drastically reduces the difficulty of the original variational inequality as the material tensor $\mathbb{C}(\boldsymbol{\varphi}^k)$ is used instead of $\mathbb{C}(\boldsymbol{\varphi}^{k+1})$. One then arrives at the following variational inequality:

$$\begin{aligned}
\int_\Omega \alpha\varepsilon\frac{\boldsymbol{\varphi}^{t+1} - \boldsymbol{\varphi}^t}{\delta t}\cdot(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}^{t+1})dx + \alpha\int_\Omega \left(\varepsilon\nabla\boldsymbol{\varphi}^{t+1} : (\nabla\hat{\boldsymbol{\varphi}} - \nabla\boldsymbol{\varphi}^{t+1}) + \frac{1}{2\varepsilon}(1 - 2\boldsymbol{\varphi}^t)\cdot(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}^{t+1})\right)dx \\
+ \int_\Omega [\mathbb{C}'(\boldsymbol{\varphi}^t)(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}^{t+1})]e(\boldsymbol{u}^t) : e(\boldsymbol{p}^t)dx - \int_\Omega F'(\boldsymbol{\varphi}^t)(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}^{t+1}) : e(\boldsymbol{p}^t)dx \geq 0 \text{ for all } \hat{\boldsymbol{\varphi}} \in \mathcal{G},
\end{aligned} \tag{34}$$

where $\boldsymbol{u}^t$ and $\boldsymbol{p}^t$ are solutions of the elasticity and adjoint equation, respectively, with $\boldsymbol{\varphi} = \boldsymbol{\varphi}^t$. At each time step, we are required to solve a variational inequality, which in the current setting is equivalent to the H^1-projection onto the Gibbs simplex $\mathcal{G}$, for which there exist efficient function-space-based numerical approaches, see [1]. The algorithm should stop when $\boldsymbol{\varphi}^{t+1} \approx \boldsymbol{\varphi}^t$, however it may be very slow, i.e., we may need to solve tens of thousands variational inequalities, and there is no guarantee of convergence.

Algorithm 4.1 Gradient flow

Input: initial point $\boldsymbol{\varphi}^0 \in \mathcal{G}$, $k \leftarrow 0$

1: **repeat**
2: Solve (2) and (19) for $\boldsymbol{u}^k$ and $\boldsymbol{p}^k$, respectively, with $\boldsymbol{\varphi} = \boldsymbol{\varphi}^k$
3: Solve (34) for $\boldsymbol{\varphi}^{k+1}$; set $k \leftarrow k + 1$
4: **until** stopping criterion is satisfied
5: **return** $\boldsymbol{\varphi}^k$

4.2 Projected gradients

The idea of projected gradients goes back to [22, 31]. At each step, we compute the following update

$$\boldsymbol{\varphi}^{k+1} = Proj_{\mathcal{G}}\left(\boldsymbol{\varphi}^k - t^k \mathcal{J}'(\boldsymbol{\varphi}^k)_{Riesz}\right), \tag{35}$$

Note that $\mathcal{J}'(\boldsymbol{\varphi})_{Riesz}$ is the Riesz representation of $\mathcal{J}'(\boldsymbol{\varphi})$ in the primal space. This is computed by solving the following elliptic PDE:

$$\begin{aligned} -\Delta\xi + \xi &= \mathcal{J}'(\boldsymbol{\varphi}) \qquad \text{in} \quad \Omega, \\ \frac{\partial \xi}{\partial n} &= 0 \qquad\qquad \text{on} \quad \partial\Omega. \end{aligned} \tag{36}$$

In an implementation, we make use of the generalized Armijo step size rule [6], where we choose some $t^k > 0$ such that the following inequality

$$\mathcal{J}(\boldsymbol{\varphi}^k) - \mathcal{J}(\boldsymbol{\varphi}^{k+1}) \geq \sigma \frac{\|\boldsymbol{\varphi}^k - \boldsymbol{\varphi}^{k+1}\|^2_{H^1(\Omega,\mathbb{R}^N)}}{t^k} \tag{37}$$

is satisfied. Here $\sigma > 0$ is a given parameter. Note that if $\mathcal{G}$ were the whole space, then (37) would reduce to the classical Armijo rule. For the stopping criterion we select the simplest condition

$$\boldsymbol{\varphi} = Proj_{\mathcal{G}}\left(\boldsymbol{\varphi} - c\mathcal{J}'(\boldsymbol{\varphi})_{Riesz}\right) \tag{38}$$

with $c > 0$ being a fixed constant. Since $\mathcal{G}$ is a convex set, we obtain that once $\boldsymbol{\varphi}^k \in \mathcal{G}$ satisfies the optimality condition (38) for any $c > 0$, then $\boldsymbol{\varphi}^k$ is a fixed point of update (35) for all $t^k > 0$.

We summarize this method in Algorithm 4.2. A convergence proof has been performed for the first time already in [6] for finite dimension. Recently, it was generalized in [8] to an intersection of a Hilbert with a Banach space satisfying certain properties. This paper was motivated by [7], where the authors worked with space $H^1(\Omega, \mathbb{R}^N) \cap L^\infty(\Omega, \mathbb{R}^N)$ and used gradient flow scheme. Note that in our approach, we were able to obtain higher regularity of $\boldsymbol{u}$, which resulted in being able to work with $H^1(\Omega, \mathbb{R}^N)$. Even in this Hilbert space setting, as mentioned in the previous subsection, projecting onto the Gibbs simplex is still a nontrivial task, see [1].

Algorithm 4.2 Projected gradients

Input: initial point $\boldsymbol{\varphi}^0 \in \mathcal{G}$, $k \leftarrow 0$

1: **repeat**
2: Solve (2) and (19) for $\boldsymbol{u}^k$ and $\boldsymbol{p}^k$, respectively, with $\boldsymbol{\varphi} = \boldsymbol{\varphi}^k$
3: Find $t^k > 0$ such that (37) holds
4: Set $\boldsymbol{\varphi}^{k+1} \leftarrow Proj_{\mathcal{G}}(\boldsymbol{\varphi}^k - t^k \nabla \mathcal{J}'(\boldsymbol{\varphi}^k))$ and $k \leftarrow k+1$
5: **until** stopping criterion is satisfied
6: **return** $\boldsymbol{\varphi}^k$

4.3 Interior point method

Although the projected gradient method is largely successful for solving (10), it can in some instances require a high number of steps in order to obtain a reasonable tolerance for the residual (38). In order to remedy this problem, we turn to second-order methods based on a direct solve of (22) or a variant thereof. Aside from the fact that the multipliers $\boldsymbol{\lambda}$ and μ need not exist, our experience with direct solvers for (22) based on Newton's method have exhibited poor performance. Thus, we consider instead interior point methods, which ensure feasibility of $\boldsymbol{\varphi}^k$ throughout. For an excellent review with many references, see [20]. For its applications to optimal control with PDE constraints see [42, 46].

As noted, we cannot assume that $\boldsymbol{\lambda}$ and μ exist. Therefore, we use a Moreau-Yosida regularization of the indicator function for the constraint $\mathbf{1}^\top \boldsymbol{\varphi} - 1 = 0$ with parameter γ. Then remaining system to be solved at each iteration is as follows:

$$\begin{aligned} J'_{\varphi}(\boldsymbol{\varphi}, \boldsymbol{u}) + E'_{\varphi}(\boldsymbol{\varphi}, \boldsymbol{u})^* \boldsymbol{p} - \lambda + \gamma \mathbf{1}(\mathbf{1}^\top \boldsymbol{\varphi} - 1) &= 0, \\ \int_\Omega \mathbb{C}(\boldsymbol{\varphi}) e(\boldsymbol{p}) : e(\boldsymbol{v}) dx + \langle (J_u^0)'(\boldsymbol{\varphi}, \boldsymbol{u}), \boldsymbol{v} \rangle &= 0 \qquad \text{for all } \boldsymbol{v} \in H_0^1(\Omega, \mathbb{R}^2), \\ \int_\Omega \mathbb{C}(\boldsymbol{\varphi}) e(\boldsymbol{u}) : e(\boldsymbol{v}) dx - \int_\Omega F(\boldsymbol{\varphi}) : e(\boldsymbol{v}) dx &= 0 \qquad \text{for all } \boldsymbol{v} \in H_0^1(\Omega, \mathbb{R}^2), \\ \langle \boldsymbol{\lambda}, \boldsymbol{\varphi} \rangle - \beta &= 0, \end{aligned} \tag{39}$$

where $\beta = 0$. The last condition, together with $\lambda \geq 0$ and $\boldsymbol{\varphi} \geq 0$, is nothing more than the complementarity condition for the remaining inequality constraint. At each "inner" loop of the interior point method, we set $\beta > 0$ and drive $\beta \to 0$ until the residual of (39) is sufficiently small. In every iteration we compute a Newton step $\boldsymbol{d}^k$ for (22) or (39). Since we have to keep positivity of variables $\boldsymbol{\varphi}^k$ and $\boldsymbol{\lambda}^k$, we compute the distance to the boundary as follows:

$$\begin{aligned} t_\varphi^k &:= \sup\{t \geq 0 |\ \boldsymbol{\varphi}^k + t \boldsymbol{d}_\varphi^k > 0 \text{ a.e. on } \Omega\}, \\ t_\lambda^k &:= \sup\{t \geq 0 |\ \boldsymbol{\lambda}^k + t \boldsymbol{d}_\lambda^k > 0 \text{ a.e. on } \Omega\}, \end{aligned} \tag{40}$$

where $\boldsymbol{d}_\varphi^k$ and $\boldsymbol{d}_\lambda^k$ are the corresponding components of $\boldsymbol{d}^k$. Then we take the step with stepsize $t^k \leftarrow \min\{c_f \cdot \min\{t_\varphi^l, t_\lambda^l\}, 1\}$. Note that this update may become problematic if $\boldsymbol{d}^k$ is negative and unbounded. Thus, we take a full step whenever we are away from the boundary. But once we are close to the boundary, we take a reduced step, where the reduction is determined by parameter $c_f \in (0, 1)$. After doing so, parameter β is decreased and the process is repeated. We write the interior point method in Algorithm 4.3. By G we denote the left-hand side of (22) or (39). Moreover, denote the combined variable $\boldsymbol{y} := (\boldsymbol{\varphi}, \boldsymbol{u}, \boldsymbol{p}, \boldsymbol{\lambda}, \mu)$ or $\boldsymbol{y} := (\boldsymbol{\varphi}, \boldsymbol{u}, \boldsymbol{p}, \boldsymbol{\lambda})$.

Algorithm 4.3 Interior point method

Input: fraction to the boundary $c_f \in (0,1)$, decrease parameter $c_\beta \in (0,1)$, initial penalization β^0, minimal penalization $\beta_{\min}$, $k \leftarrow 0$

1: **repeat**
2: based on (22) or (39) compute $\boldsymbol{d}^k \leftarrow -G'(\boldsymbol{y}^k)^{-1}G(\boldsymbol{y}^k)$ ▷ direction
3: based on (40) compute $t^k \leftarrow \min\{c_f \cdot \min\{t^l_\varphi, t^l_\lambda\}, 1\}$ ▷ step size
4: $\boldsymbol{y}^{k+1} \leftarrow \boldsymbol{y}^k + t^k \boldsymbol{d}^k$ ▷ new iterate
5: $\beta^{k+1} \leftarrow \max(c_\beta \beta^k, \beta_{\min})$
6: $k \leftarrow k+1$
7: **until** stopping criterion is satisfied
8: **return** $\boldsymbol{\varphi}^k$

5 Numerical results

Since our main application (optimization of strained Ge-on-Si microbridge) is new, we also provide the results of the algorithms for a classical "bridge" problem found throughout the topology optimization literature, see [4, 23]. In both applications, the elasticity tensor is of form

$$\mathbb{C}(\boldsymbol{\varphi}) = \sum_{i=1}^{N} \operatorname{cut}(\varphi_i)\mathbb{C}_i,$$

where $\mathbb{C}_i$ is the standard elasticity tensor associated with material i, thus for $E_1, E_2 \in \mathbb{R}^{2\times 2}$ we have

$$\mathbb{C}_i E_1 : E_2 = \lambda_i \operatorname{tr} E_1 \operatorname{tr} E_2 + 2\mu_i E_1 : E_2,$$

where λ_i and μ_i are Lamé constants of individual materials and $\operatorname{cut} : \mathbb{R} \to \mathbb{R}$ is the cutoff function

$$\operatorname{cut}(x) = \begin{cases} \operatorname{arctg}(x-\delta_2) + \delta_2 & \text{if } x \geq \delta_2, \\ x & \text{if } x \in [\delta_1, \delta_2), \\ x - 2\delta_1(x-\delta_1)^3 - (x-\delta_1)^4 & \text{if } x \in [0, \delta_1), \\ a \operatorname{arctg}(bx) + \delta_1^4 & \text{if } x < 0 \end{cases} \tag{41}$$

for some small $\delta_1 > 0$, large $\delta_2 > 0$ and $a = \frac{\delta_1^4}{\pi}$ and $b = \frac{(1-2\delta_1^3)\pi}{\delta_1^4}$. Note that the cutoff function is twice continuously differentiable, increasing function with $\operatorname{cut}(x) \geq \frac{1}{2}\delta_1^4$ for all $x \in \mathbb{R}$ and thus assumptions (A1)-(A3) are satisfied. We have chosen such cutoff function so that its first and second derivatives approximate those of identity on the interval $[0,1]$ as well as possible.

For the projection onto the Gibbs simplex $\mathcal{G}$, we discretized the problem and used the semismooth Newton's method [25], which is equivalent to a primal-dual active set strategy. Another possibility would be to use the path-following method from [1]. We use the former in all experiments.

5.1 Updating the parameters

The general model contains a number of parameters, whose purpose we list here for convenience:

- N: Number of phases

- α: Penalty parameter which controls the perimeter of phases
- ε: Parameter corresponding to interfacial thickness
- δ_1, δ_2: Cutoff parameters from (41)
- ϵ_0, δ_0: Constants for the eigenstrain generated by Ge and the thermal (pre-)stress generated by SiN, see (3)
- c_f, c_β, $\beta_{\min}$: Parameters for the interior point method (fraction to the boundary, decrease parameter for β and minimal value of β)
- γ: Penalty parameter for $\mathbf{1}^\top \boldsymbol{\varphi} - 1$ constraint in (39)
- tol_{PG}, tol_{IP}: Stopping tolerances for first-order systems (38) and (39), respectively
- $h_{\min}$, $\varepsilon_{\min}$: Width of the smallest triangle in mesh and value of ε on the finest mesh
- δt: Step size for the gradient flow method (34)

We now discuss the refinement process and the parameter values which are summarized in Table 5.1. After solving (10) on a given mesh, we refine every element, where the phases are not pure, thus with $10^{-6} < \varphi_i < 1 - 10^{-6}$ for some i. For refinement we use red refinement, see [11]. Since ε corresponds to the interfacial thickness, the initial ε was chosen to be four times the length of the biggest element and we divide ε by 2 upon every mesh refinement. Meshes were refined three times for the first application and four times for the second one.

Concerning parameters, α was chosen as small as possible, see Section 5.4. Cutoff parameters δ_1 and δ_2 were chosen so that that the cutoff has a negligible effect on interval $(0, 1)$. Fraction to the boundary c_f was chosen close to 1 and c_β close to 0 to promote high convergence. Since the second application is more demanding, we needed to decrease c_f, to increase c_β and to set minimal value $\beta_{\min}$. For γ we chose a relatively high value to obtain small violation of constraints (4). As the convergence for the interior point was fast once the solution was approached and the convergence for the projected gradients was rather slow, we chose the first tolerance small and the other one large, for residual development see Figure 5. Finally, δt was chosen small to ensure small steps for the gradient flow method. Note that even $h_{\min} = \frac{1}{128}$ may seem too large, the mesh is rather fine because Ω is not the unit square.

	α	Ω	N	$h_{\min}$	$\varepsilon_{\min}$	δ_1	δ_2	
Bridge construction	10	$(-1,1)\times(0,1)$	2	$\frac{1}{128}$	$\frac{1}{32}$	10^{-3}	10^{16}	
Microbridge design	$2{\cdot}10^{-4}$	$(-2,2)\times(0,3)$	4	$\frac{1}{128}$	$\frac{1}{32}$	10^{-3}	10^{16}	
	c_f	c_β	$\beta_{\min}$	γ	tol_{IP}	σ	δt	tol_{PG}
Bridge construction	0.9	0.25	$-\infty$	-	10^{-10}	10^{-4}	10^{-3}	10^{-5}
Microbridge design	0.5	0.5	10^{-10}	10^{6}	10^{-10}	10^{-4}	-	10^{-5}

Table 1: List of parameters

For the first application, we compared the performance of gradient flow, projected gradients and interior point when applied on (22). Since the gradient flow performed subpar, we omitted it for the second application. For it we run the projected gradients and interior point applied both on (22) and (39). Since they performed comparably, we show only results for (39).

The method comparison may be skewed for three reasons. First, different residuals are checked. Even though we could theoretically check the residual of (22) for the projected gradients, we do not do so because of the multipliers do not have to exist. Second, one iteration refers to solving the elasticity

and adjoint equations and performing the line search for the gradient flow and the projected gradients, while solving one large system for the interior point method. Third, since the mesh refinement is based on the solution on the coarser mesh, the meshes need not coincide.

5.2 Bridge construction

In this example, the goal is to find a material distribution that minimizes compliance and occupies fifty percent of the available space. The optimization was performed on domain $\Omega = (-1,1) \times (0,1)$. The material was fixed on $\Gamma_D = (-1,-0.9] \times \{0\} \cup [0.9,1) \times \{0\}$. The force acting in a downward direction on $\Gamma_N = [-0.02, 0.02] \times \{0\}$ was constant on Γ_N and equalled to $\boldsymbol{g} = (0,-5000)$. The problem takes form

$$\min \int_{\Gamma_N} \boldsymbol{g}\cdot\boldsymbol{u}\, dS + \frac{\alpha}{2}\int_\Omega \left(\varepsilon|\nabla\boldsymbol{\varphi}|^2 + \frac{1}{\varepsilon}\boldsymbol{\varphi}\cdot(1-\boldsymbol{\varphi})\right) dx \qquad \text{over } \boldsymbol{\varphi} \in H^1(\Omega,\mathbb{R}^N),\ \boldsymbol{u} \in H^1_0(\Omega,\mathbb{R}^2)$$
$$\text{s.t. } \hat{E}(\boldsymbol{\varphi},\boldsymbol{u}) = 0, \quad \boldsymbol{\varphi} \in \mathcal{G}_{ad}, \quad \int_\Omega \varphi_1 dx = \frac{1}{2}|\Omega|.$$

Here the elasticity $\hat{E}$ was defined in its strong form by

$$\begin{aligned} -\operatorname{div}\mathbb{C}(\boldsymbol{\varphi})e(\boldsymbol{u}) &= 0 && \text{in } \Omega,\\ \boldsymbol{u} &= 0 && \text{on } \Gamma_D,\\ \mathbb{C}(\boldsymbol{\varphi})e(\boldsymbol{u})\boldsymbol{n} &= \boldsymbol{g} && \text{on } \Gamma_N,\\ \mathbb{C}(\boldsymbol{\varphi})e(\boldsymbol{u})\boldsymbol{n} &= \boldsymbol{0} && \text{on } \Gamma\setminus(\Gamma_D\cup\Gamma_N). \end{aligned}$$

For more details we refer to [7, Section 6.1] or to our codes available online.

The obtained bridge shape is presented in Figure 2. While the interior point and projected gradients obtained the same design (left), the gradient flow did not manage to converge to this design (right). However, if the limit on maximal iteration number of 1000 was not imposed, it could converge to the same solution. Further evidence is summarized in Table 2. Every row describes one method. The first column denotes the total objective $\mathcal{J}(\boldsymbol{\varphi})$ and the second column compliance $\int_{\Gamma_N} \boldsymbol{g}\cdot\boldsymbol{u}\, dS$. Even though the total objective is lowest for the interior point, the maximal stiffness was reached by the projected gradients. The next four columns show the number of iterations and the last four columns the number of nodes on all meshes. It is clear that the only method, which shows mesh-independence is the interior point while the number of iterations for the projected gradients approximately triples every mesh refinement. The gradient flow did not manage to converge on any mesh. This is connected with the higher number of nodes.

	Objective		# Iterations per (M)esh				# Nodes per (M)esh			
	$\mathcal{J}(\boldsymbol{\varphi})$	$\int_{\Gamma_N} \boldsymbol{g}\cdot\boldsymbol{u}\, dS$	M1	M2	M3	M4	M1	M2	M3	M4
IP	401.59	366.99	21	20	25	29	629	2233	8038	21317
PG	401.59	366.98	26	76	214	836	629	2233	8038	21325
GF	403.05	367.41	1000	1000	1000	1000	629	2247	8196	24435

Table 2: Numerical evidence for the application described in Subsection 5.2. Rows correspond to interior point method (IP), projected gradients (PG) and gradient flow (GF). First two columns are the values of the objective function and compliance, the next four are the iteration numbers on subsequently refined meshes and the last four columns the number of nodes.

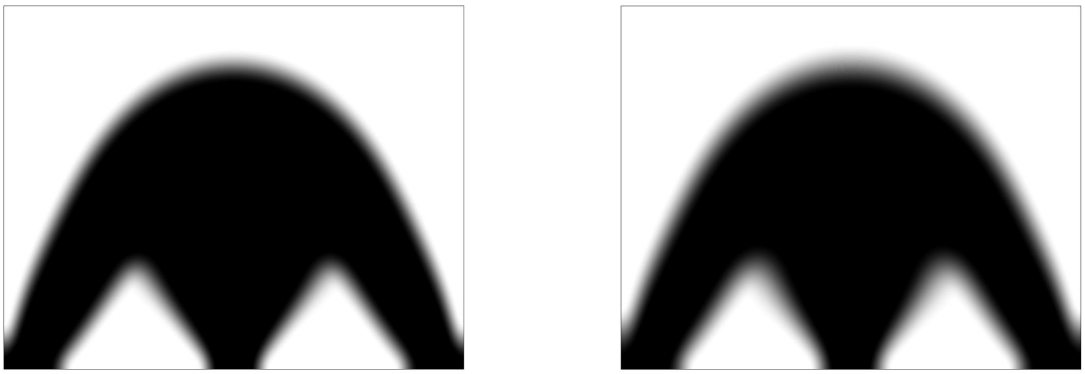


Figure 2: Optimal design for the bridge construction problem for the interior point method and projected gradients (left) and gradient flow (right). There are differences at corners and at the width of the interfacial region.

5.3 Ge-on-Si microbridge design

We now turn our attention back to the model studied thoroughly in this paper, the design of a germanium microbridge. As the domain we chose $\Omega = (-2, 2) \times (0, 3)$, on which we considered three materials (Ge, SiN, SiO_2) and air. The corresponding parameters equalled to $\lambda_{Ge} = 44.279$, $\lambda_{SiN} = 110.369$, $\lambda_{SiO_2} = 16.071$, $\mu_{Ge} = 27.249$, $\mu_{SiN} = 57.813$, $\mu_{SiO_2} = 20.798$, $\epsilon_0 = 2.5 \cdot 10^{-3}$ and $\sigma_0 = -2.5$, see [33, 49, 52]. Concerning the objective function, the weak lower semicontinuity of J^0 from (A2) is satisfied only for J^0_1 and J^0_3 but not for J^0_2, all of them being defined in (12). Since we do not have an estimate for e_d, we have decided to work with J^3_0.

Since we have not shown the existence of multipliers μ and λ, a direct solution of the nonlinear system (22) by Newton's method (i.e. semismooth Newton) may, potentionally, exhibit mesh-dependent behavior. Nevertheless, its performance was almost identical to the function-space conforming interior point method, in which we solve (39) for some large $\gamma > 0$. Moreover, we have compared our solution to the design proposed in [39] which we refer to as the original configuration. We show the design differences in Figure 3. We see that the difference between both designs is significant, mainly the SiN stressor encapsulates the entire section of Ge. The biaxial strain is depicted in Figure 4.

The results are summarized in Table 3 which is very similar to Table 2. This time the best value for the objective was reached by the projected gradients but the strain profile was better for the interior point. In both cases we obtained improvement in the strain of approximately 15% compared to the original configuration. The next five columns denote the number of iterations on individual meshes, which stays approximately constant for the interior point and doubles on the last mesh. For the projected gradients, the number of iterations doubles every mesh refinement. For the residual development on the next-to-last mesh, see Figure 5. Concerning the precise meaning of iteration numbers, please refer to the end of the previous subsection. The last five columns denote the number of nodes. Note that the number of variables is much higher, for example the resulting matrix in system (22) has dimension 929113×929113 on the finest mesh.

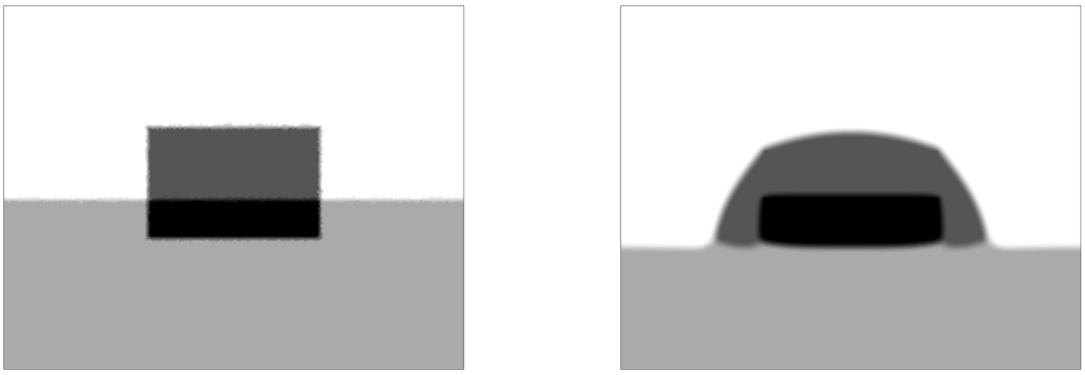


Figure 3: The original configuration (left) and optimal design (right) for the application presented in Subsection 5.3. The colors go as follows: black (Ge), dark gray (SiN) and light gray (SiO_2).

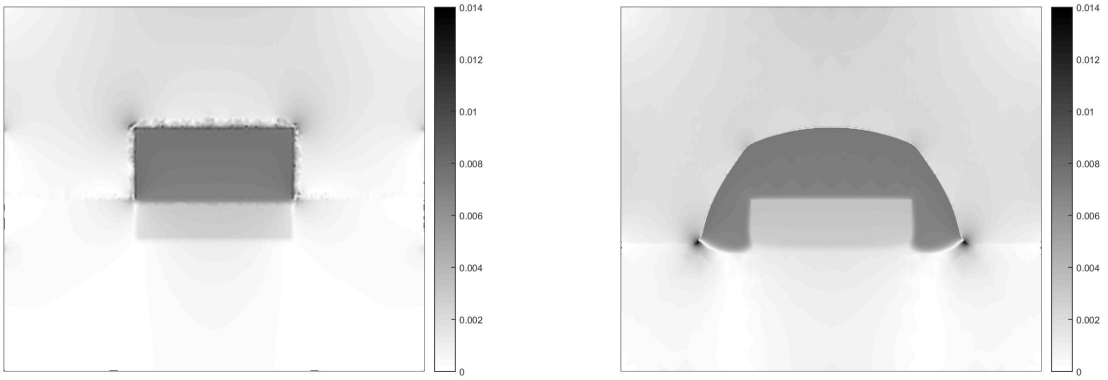


Figure 4: The biaxial strain for the original configuration (left) and optimal design (right) for the application presented in Subsection 5.3.

	Objective			# Iterations per (M)esh					# Nodes per (M)esh				
	$\mathcal{J}(\boldsymbol{\varphi})$	$-J_3^0(\boldsymbol{\varphi})$	Improv	M1	M2	M3	M4	M5	M1	M2	M3	M4	M5
IP	-0.001315	0.002847	14.90%	51	64	58	65	143	1155	2993	9087	28046	87651
PG	-0.001315	0.002847	14.90%	26	23	45	109	288	1155	2947	8764	27045	85861

Table 3: Numerical evidence for the application described in Subsection 5.2. Rows correspond to interior point method (IP) and projected gradients (PG). First two columns are the values of the objective function, then improvement over the initial configuration and the remaining ones are the number of iterations and nodes.

5.4 Parameter sensitivity

In this section, we provide a short, experimental study on the sensitivity of the designs to the parameters in the objective. A full analytical path-following study as in [27, 28] is not possible due to a lack of convexity and uniqueness of the solutions. We first investigate the "strain" objective J_0^3. Let $\boldsymbol{u} = (u_x, u_y)$ and given parameters $a, b \geq 0$ define objective

$$J_4^0(\boldsymbol{u}) := -\int_D \left(a\frac{\partial u_x}{\partial x} + b\frac{\partial u_y}{\partial y} \right) dx.$$

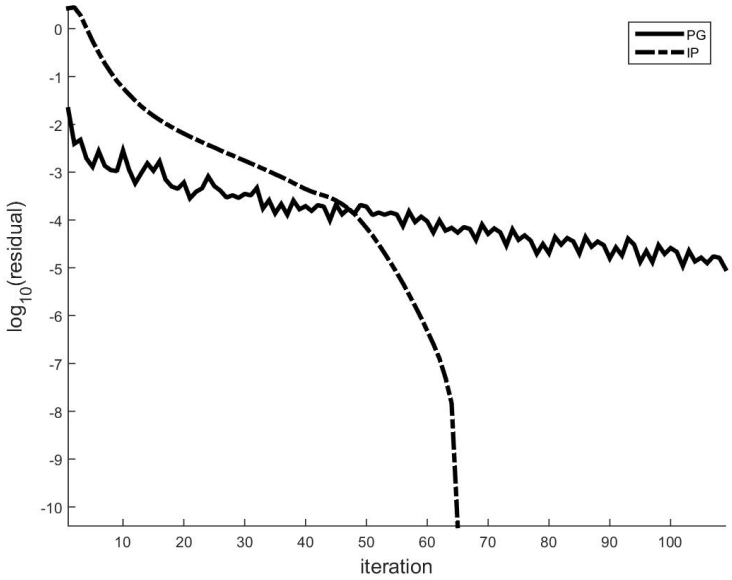


Figure 5: Residual devolepment for the interior point method (IP) and the projected gradients (PG) for the application presented in Subsection 5.3.

Note that we obtain J_0^3 when we choose $a = b = 1$. Letting $a = 10$, $b = 1$, we obtain a functional that puts more emphasis on the strain along the x-axis; this goes analogously along the y-axis when we set $a = 1$ and $b = 10$. The resulting designs appear in Figure 6. F

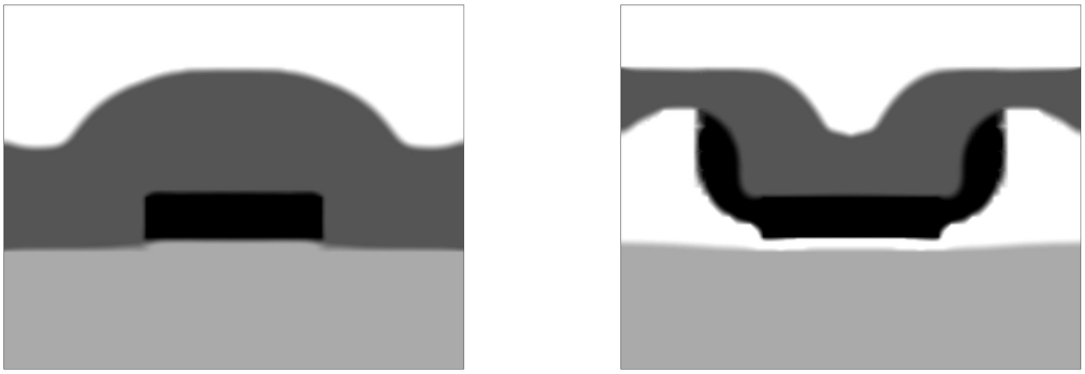


Figure 6: Optimal designs for weighted objective J_0^4 with weights $a = 1$, $b = 10$ (left) and $a = 10$, $b = 1$ (right). The colors go as follows: black (Ge), dark gray (SiN) and light gray (SiO_2).

Keeping $a = b = 1$, we now consider the dependence of the optimal design on the regularization parameter α. The magnitude of J_3^0 is plotted in Figure 7. In addition, we include three vastly different designs. We note that the topological genus of the structure increases as α goes to zero. This is not suprising as the regularization term disappears for $\alpha \to 0$.

Finally, in Figure 8 we perform a similar analysis for the dependence of optimal desing on the eigenstrain parameter ε_0, see (3). In accordance with results in Figure 7 we fix $\alpha = 5 \cdot 10^{-5}$. The top left figure depicts the value of the Ginzburg-Landau energy. It develops in a continuous way but a jump

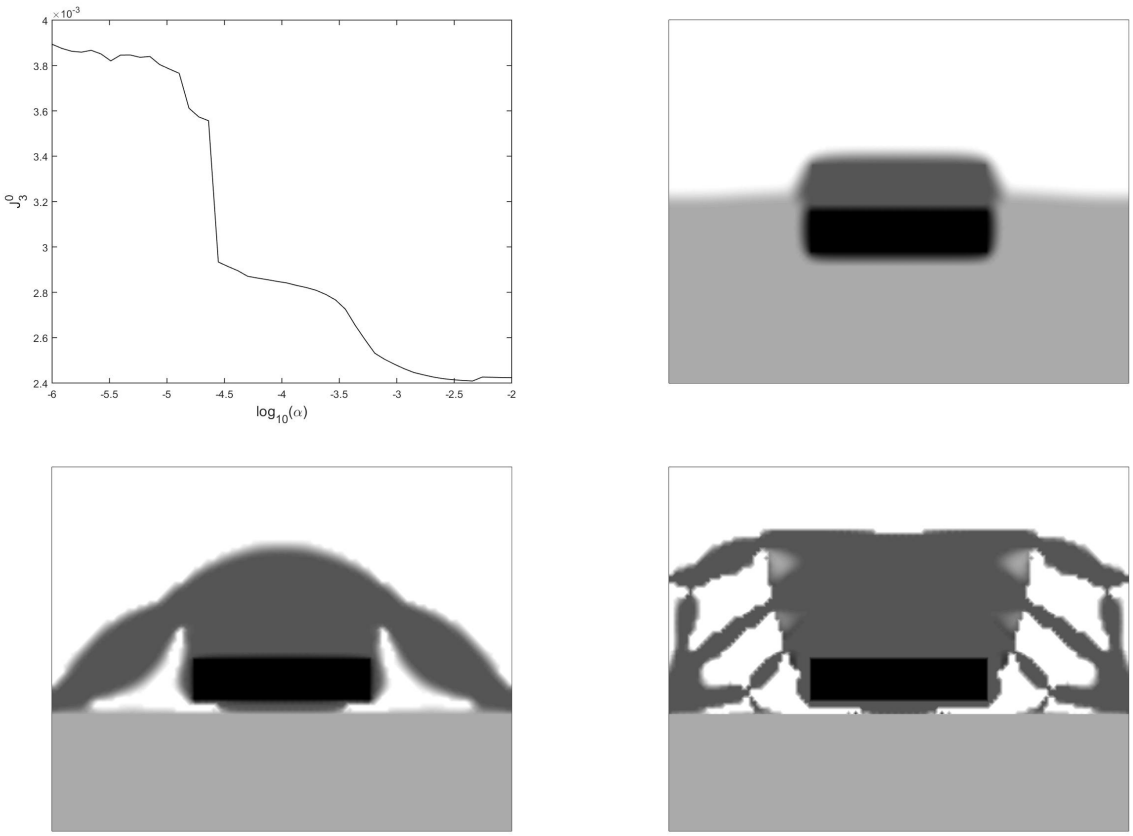


Figure 7: Dependence of the strain on α (top left) and optimal designs for $\alpha = 10^{-2}$ (top right), $\alpha = 10^{-5}$ (bottom left) and $\alpha = 10^{-6}$ (bottom right). The colors go as follows: black (Ge), dark gray (SiN) and light gray (SiO_2).

occurs around $\varepsilon_0 = 0.003185$, which means that the perimeter of the optimal design changed dramatically. The next two figures depict the configurations before the jump and the last one the configuration directly after the jump. Since the last one resembles the bottom left configuration in Figure 7, it may mean that problem (10) contains multiple local (ε-)minima and that the original configuration falls to a different region of attraction after a small change of parameters and thus converge to a different local minimum.

6 Conclusions

In this paper, we investigate the multi-material optimization of the cross-section of a strained photonic device. Though we only include the elasticity equation, this represents an important first step in the design process. Following a recent paper on multi-material topology optimization [7], we formulated the problem using the phase-field approach and derived first- and second-order optimality conditions. On their basis, we compared performance of several (popular) algorithms from nonlinear optimization, namely gradient flow, projected gradients and interior point method. In the end, a device configuration is suggested that adds a significant increase in the amount of strain in the optical cavity.

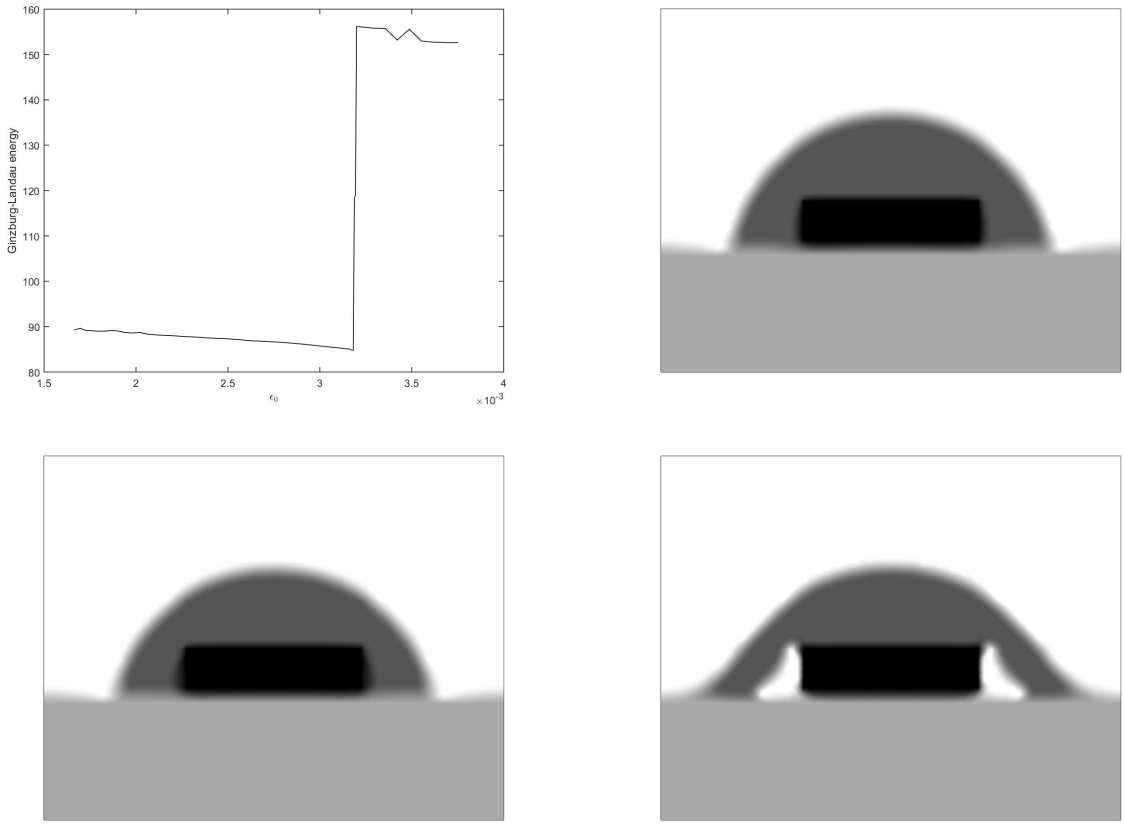


Figure 8: Dependence of the Ginzburg-Landay energy on the eigenstrain parameter ε_0 (top left) and optimal designs for $\varepsilon_0 = 0.001667$ (top right), $\varepsilon_0 = 0.003182$ (bottom left) and $\varepsilon_0 = 0.003188$ (bottom right). The colors go as follows: black (Ge), dark gray (SiN) and light gray (SiO_2).

Acknowledgements

We would especially like to thank Dirk Peschka and Marita Thomas, both from the Weierstrass Institute for Applied Analysis and Stochastics Berlin, for helpful discussions concerning the application to Ge-on-Si microbridges.

References

[1] L. Adam, M. Hintermüller, and T. M. Surowiec. A semismooth Newton method with analytical path-following for the H^1-projection onto the Gibbs simplex. Submitted.

[2] V. Barbu. *Optimal control of variational inequalities*. Research notes in mathematics. Pitman Advanced Pub. Program, 1984.

[3] W. Behrman. *An efficient gradient flow method for unconstrained optimization*. PhD thesis, Stanford University, 1998.

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

[5] A. Bensoussan and J. Frehse. *Regularity Results for Nonlinear Elliptic Systems and Applications*. Springer Berlin Heidelberg, 2002.

[6] D. P. Bertsekas. On the Goldstein-Levitin-Polyak gradient projection method. *IEEE Trans. Automat. Contr.*, 21(2):174–184, 1976.

[7] L. Blank, H. Garcke, M. H. Farshbaf-Shaker, and V. Styles. Relating phase field and sharp interface approaches to structural topology optimization. *ESAIM Control. Optim. Calc. Var.*, 20(02):1025–1058, 2014.

[8] L. Blank and C. Rupprecht. An extension of the projected gradient method to a Banach space setting with application in structural topology optimization. Preprintreihe der Fakultät Mathematik 04/2015, University of Regensburg, 2015.

[9] J. F. Blowey and C. M. Elliott. Curvature dependent phase boundary motion and parabolic double obstacle problems. In W.-M. Ni, L. A. Peletier, and J. L. Vazquez, editors, *Degenerate Diffusions*, pages 19–60. Springer New York, 1993.

[10] J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer, 2000.

[11] S. C. Brenner and C. Carstensen. Finite element methods. In *Encyclopedia of Computational Mechanics*, chapter 4. Wiley Online Library, 2004.

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

[13] R. E. Camacho-Aguilera, Y. Cai, N. Patel, J. T. Bessette, M. Romagnoli, L. C. Kimerling, and J. Michel. An electrically pumped germanium laser. *Opt. Express*, 20(10):11316–11320, 2012.

[14] G. Capellini, C. Reich, S. Guha, Y. Yamamoto, M. Lisker, M. Virgilio, A. Ghrib, M. E. Kurdi, P. Boucaud, B. Tillack, and T. Schroeder. Tensile Ge microstructures for lasing fabricated by means of a silicon complementary metal-oxide-semiconductor process. *Opt. Express*, 22(1):399–410, 2014.

[15] G. Capellini, M. Virgilio, Y. Yamamoto, L. Zimmermann, B. Tillack, D. Peschka, M. Thomas, A. Glitzky, R. Nürnberg, K. Gärtner, T. Koprucki, and T. Schroeder. Modeling of an edge-emitting strained-Ge laser. In *Advanced Solid State Lasers*, 2015.

[16] R. Courant and D. Hilbert. *Methoden der mathematischen Physik*. Verlag von Julius Springer, 1924.

[17] B. Dutt, D. S. Sukhdeo, D. Nam, B. M. Vulovic, Z. Yuan, and K. C. Saraswat. Roadmap to an efficient germanium-on-silicon laser: Strain vs. n-type doping. *IEEE Photonics Journal*, 4(5):2002–2009, 2012.

[18] M. El Kurdi, G. Fishman, S. Sauvage, and P. Boucaud. Band structure and optical gain of tensile-strained germanium based on a 30 band $k \cdot p$ formalism. *Journal of Applied Physics*, 107(1), 2010.

[19] W. H. Fleming and R. Rishel. An integral formula for total gradient variation. *Archiv der Mathematik*, 11(1):218–222, 1960.

[20] A. Forsgren, P. E. Gill, and M. H. Wright. Interior methods for nonlinear optimization. *SIAM Review*, 44(4):525–597, 2002.

[21] H. Goldberg, W. Kampowsky, and F. Tröltzsch. On Nemytskij operators in L^p-spaces of abstract functions. *Math. Nachr.*, 155:127–140, 1992.

[22] A. A. Goldstein. Convex programming in Hilbert space. *Bulletin of the American Mathematical Society*, 70(5):709–710, 1964.

[23] J. Haslinger and P. Neittaanmäki. *Finite Element Approximation for Optimal Shape Design: Theory and Applications*. Wiley, 1988.

[24] R. Herzog, C. Meyer, and G. Wachsmuth. B- and strong stationarity for optimal control of static plasticity with hardening. *SIAM Journal on Optimization*, 23(1):321–352, 2013.

[25] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, 13(3):865–888, 2002.

[26] M. Hintermüller and I. Kopacka. Mathematical programs with complementarity constraints in function space: C- and strong stationarity and a path-following algorithm. *SIAM J. Optim.*, 20(2):868–902, 2009.

[27] M. Hintermüller and K. Kunisch. Feasible and noninterior path-following in constrained minimization with low multiplier regularity. *SIAM Journal on Control and Optimization*, 45(4):1198–1221, 2006.

[28] M. Hintermüller and K. Kunisch. Path-following methods for a class of constrained minimization problems in function space. *SIAM Journal on Optimization*, 17(1):159–187, 2006.

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

[30] T. Koprucki, D. Peschka, and M. Thomas. Towards the optimization of on-chip germanium lasers. In *WIAS Annual Research Report 2015*, 2015.

[31] E. S. Levitin and B. T. Polyak. Constrained minimization methods. *Zh. Vychisl. Mat. Mat. Fiz.*, 6(5):787–823, 1966.

[32] J. Liu, X. Sun, R. Camacho-Aguilera, L. C. Kimerling, and J. Michel. Ge-on-si laser operating at room temperature. *Opt. Lett.*, 35(5):679–681, 2010.

[33] Z. Lu. *Dynamics of wing cracks and nanoscale damage in silica glass*. PhD thesis, University of Southern California, 2007.

[34] F. Maggi. *Sets of Finite Perimeter and Geometric Variational Problems: An Introduction to Geometric Measure Theory*. Cambridge University Press, 2012.

[35] P. A. Markowich. *The stationary semiconductor device equations*. Springer-Verlag Wien New York, 1986.

[36] F. Mignot. Contrôle dans les inéquations variationelles elliptiques. *Journal of Functional Analysis*, 22(2):130 – 185, 1976.

[37] L. Modica. The gradient theory of phase transitions and the minimal interface criterion. *Archive for Rational Mechanics and Analysis*, 98(2):123–142, 1987.

[38] D. Peschka, N. Rotundo, and M. Thomas. Towards doping optimization of semiconductor lasers. *Journal of Computational and Theoretical Transport*, 45(5):410–423, 2016.

[39] D. Peschka, M. Thomas, A. Glitzky, R. Nürnberg, K. Gärtner, M. Virgilio, S. Guha, T. Schroeder, G. Capellini, and T. Koprucki. Modeling of edge-emitting lasers based on tensile strained germanium microstrips. *IEEE Photonics Journal*, 7(3):1–15, 2015.

[40] D. Peschka, M. Thomas, A. Glitzky, R. Nürnberg, M. Virgilio, S. Guha, T. Schroeder, G. Capellini, and T. Koprucki. Robustness analysis of a device concept for edge-emitting lasers based on strained germanium. *Optical and Quantum Electronics*, 48(156), 2016.

[41] S. M. Robinson. First order conditions for general nonlinear optimization. *SIAM Journal on Applied Mathematics*, 30(4):597–607, 1976.

[42] A. Schiela and M. Weiser. Superlinear convergence of the control reduced interior point method for PDE constrained optimization. *Computational Optimization and Applications*, 39(3):369–393, 2008.

[43] M. J. Suess, R. Geiger, R. A. Minamisawa, G. Schiefler, J. Frigerio, D. Chrastina, G. Isella, R. Spolenak, J. Faist, and H. Sigg. Analysis of enhanced light emission from highly strained germanium microbridges. *Nat Photon*, 7(6):466–472, 2013.

[44] X. Sun, L. Jifeng, L. Kimerling, and J. Michel. Toward a germanium laser for integrated silicon photonics. *Selected Topics in Quantum Electronics, IEEE Journal of*, 16(1):124–131, 2010.

[45] A. Takezawa, S. Nishiwaki, and M. Kitamura. Shape and topology optimization based on the phase field method and sensitivity analysis. *Journal of Computational Physics*, 229(7):2697 – 2718, 2010.

[46] M. Ulbrich and S. Ulbrich. Primal-dual interior-point methods for PDE-constrained optimization. *Mathematical Programming*, 117(1):435–485, 2009.

[47] W. van Roosbroeck. Theory of the flow of electrons and holes in germanium and other semiconductors. *Bell. Syst. Tech. J*, 29(4):560–607, 1950.

[48] M. Virgilio, T. Schroeder, Y. Yamamoto, and G. Capellini. Radiative and non-radiative recombinations in tensile strained ge microstrips: Photoluminescence experiments and modeling. *Journal of Applied Physics*, 118(23), 2015.

[49] J. J. Vlassak and W. D. Nix. A new bulge test technique for the determination of Young's modulus and Poisson's ratio of thin films. *Journal of Materials Research*, 7:3242–3249, 1992.

[50] G. Wachsmuth. A guided tour of polyhedric sets: Basic properties, new results on intersections and applications. TU Chemnitz, 2016.

[51] S. Wirths, R. Geiger, N. von den Driesch, G. Mussler, T. Stoica, S. Mantl, Z. Ikonic, M. Luysberg, S. Chiussi, J. M. Hartmann, H. Sigg, J. Faist, D. Buca, and D. Grutzmacher. Lasing in direct-bandgap GeSn alloy grown on Si. *Nat Photon*, 9(2):88–92, 2015.

[52] J. J. Wortman and R. A. Evans. Young's modulus, shear modulus, and Poisson's ratio in silicon and germanium. *Journal of Applied Physics*, 36(1):153–156, 1965.

[53] E. Zeidler. *Nonlinear Functional Analysis and its Applications IV: Applications to Mathematical Physics*. Springer, 1988.

[54] J. Zowe and S. Kurcyusz. Regularity and stability for the mathematical programming problem in Banach spaces. *Applied Mathematics and Optimization*, 5(1):49–62, 1979.

A Appendix

In the next lemma, we show that problem (7) admits an optimal solution. We will use shortened notation $\mathcal{P}(E) := \mathcal{P}(E; \mathbb{R}^N)$ for the perimeter of E, see its definition in (8).

Lemma A.1. *Assume that (A1)-(A3) hold true and that Ω has finite perimeter. Then problem* (7) *admits an optimal solution.*

Proof. Due to assumption (A1) problem (7) admits a feasible point. Consider now $\{(\boldsymbol{\varphi}^k, \boldsymbol{u}^k)\}$ to be a minimizing sequence of problem (7). Due to (4), we deduce that $\{\boldsymbol{\varphi}^k\}$ is uniformly bounded in $L^\infty(\Omega, \mathbb{R}^N)$. From Lemma 3.2 we obtain that $\{\boldsymbol{u}^k\}$ is uniformly bounded in $W_0^{1,p}(\Omega, \mathbb{R}^2)$ for some $p > 2$. Due to the constraints (4) and (5) we obtain that φ_i has binary values. Due to assumption (A2), $\{J^0(\boldsymbol{u}^k)\}$ is bounded below, which from the definition of a minimizing sequence implies that $\mathcal{P}(\{\varphi_i^k = 1\}; \Omega)$ is uniformly bounded above. Since

$$\mathcal{P}(\{\varphi_i^k = 1\}) \leq \mathcal{P}(\{\varphi_i^k = 1\}; \Omega) + \mathcal{P}(\Omega)$$

due to [34], Equation (12.26), we may invoke [34], Theorem 12.26 to obtain the existence of some sets $A_i \subset \Omega$ such that, upon possibly passing to a subsequence, $\{\varphi_i^k = 1\} \to A_i$ for all i, which means that

$$a_i^k := |(\{\varphi_i^k = 1\} \setminus A_i) \cup (A_i \setminus \{\varphi_i^k = 1\})| \to 0. \tag{42}$$

Now define $\boldsymbol{\varphi}$ with components $\varphi_i := \chi_{A_i}$ for $i = 1, \ldots, N$, where χ_{A_i} is the characteristic function to A_i. Fixing any $q \in [1, \infty)$, due to (42) and the binarity of φ_i^k we have

$$\|\varphi_i - \varphi_i^k\|_{L^q(\Omega)}^q = \|\chi_{A_i} - \varphi_i^k\|_{L^q(\Omega)}^q = \int_\Omega |\chi_{A_i} - \varphi_i^k|^q dx = a_i^k \to 0$$

and thus $\boldsymbol{\varphi}^k \to \boldsymbol{\varphi}$ in $L^q(\Omega, \mathbb{R}^N)$ for all $q \in [1, \infty)$. Thus, we may possibly pass to another subsequence to obtain that the above sequence converges pointwise almost everywhere as well. Thus, $\boldsymbol{\varphi}$ satisfies (4), (5) and (6). Denoting $\boldsymbol{u} = S(\boldsymbol{\varphi})$ and $\boldsymbol{u}^k = S(\boldsymbol{\varphi}^k)$ due to a slight modification of the last part of the proof of Lemma 3.2 we have $\boldsymbol{u}^k \to \boldsymbol{u}$ in $H_0^1(\Omega, \mathbb{R}^2)$, which due to assumption (A2) further implies

$$J^0(\boldsymbol{u}) \leq \liminf_k J^0(\boldsymbol{u}^k). \tag{43}$$

Due to [34], Proposition 12.15 we also have

$$\mathcal{P}(\{\varphi_i = 1\}; \Omega) = \mathcal{P}(A_i; \Omega) \leq \liminf_k \mathcal{P}(\{\varphi_i^k = 1\}; \Omega). \tag{44}$$

Since $\boldsymbol{\varphi}^k$ is a minimizing sequence, from (43) and (44) we obtain that $\boldsymbol{\varphi}$ is an optimal solution, which proves the assertion. □