Optimization and Engineering
https://doi.org/10.1007/s11081-018-9394-5

RESEARCH ARTICLE

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

L. Adam[1] · M. Hintermüller[1,2] · T. M. Surowiec[3]

Received: 6 February 2017 / Revised: 12 February 2018 / Accepted: 29 May 2018

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 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.

Keywords 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.

✉ M. Hintermüller
hintermueller@wias-berlin.de

Extended author information available on the last page of the article

Published online: 04 July 2018

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 El Kurdi et al. (2010).

One promising approach is to employ lasers to communicate between the individual parts of the microprocessor, see Sun et al. (2010), Camacho-Aguilera et al. (2012), Suess et al. (2013) and Wirths et al. (2015). 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., Suess et al. (2013). A recent study concentrating on modeling these effects on the electronic and optical properties is Dutt et al. (2012). 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: Liu et al. (2010), Camacho-Aguilera et al. (2012), Capellini et al. (2014) and Peschka et al. (2015). One common theme is the presence of a so-called "microbridge" created through standard etching and wetting procedures in photolithograpy. In this work, 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 (Peschka et al. 2015). The question of finding an optimal doping profile will be addressed in future work; see the discussion in Sect. 6 as well as the recent study (Peschka et al. 2015). 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 result from the position of the materials. More specifically, we observe that during the production process, the Ge is laid on an SiO_2 substrate. Upon heating, the lattice structures of the Ge and SiO_2 briefly align near the interface of the materials. However, as this Ge on SiO_2 "stack" cools, the two materials cool at disparate rates. As a result, the previously aligned molecules pull and compress each other, which ultimately leaves a residual force at the interface. This effect can be exacerbated by the addition of a silicon mononitride (SiN) cap or stressor on top of the Ge. We model this fact by using the topology-dependent operator $F(\boldsymbol{\varphi})$ in (3). Multimaterial topology optimization problems that link, e.g., thermoelastic or piezoeletric properties to the underlying topologies have been considered in many works; see

e.g., Sigmund and Torquato (1997, 1999). However, these are fundamentally distinct applications with different objectives.

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

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

$$\textbf{Semiconductors}: \quad -\operatorname{div}(\varepsilon\nabla\phi) = q(C_{dop} + p - n), \tag{1b}$$

$$\dot{n} - \operatorname{div}(D_n\nabla n - \mu_n n\nabla\phi) = -R_{net}(n,p,e(u)), \tag{1c}$$

$$\dot{p} - \operatorname{div}(D_p\nabla p - \mu_p p\nabla\phi) = -R_{net}(n,p,e(u)), \tag{1d}$$

$$\textbf{Optics}: \quad \left[\nabla^2 + \frac{\omega^2}{c^2}\left(n_r + \imath\frac{c}{2\omega}(g-\alpha)\right)^2\right]\Xi_i = \beta_i^2\Xi_i, \tag{1e}$$

$$\dot{S}_i - v_{g,i}(2\mathrm{Im}\beta_i - \alpha_c)S_i - \dot{S}_{\mathrm{sp,i}} = 0. \tag{1f}$$

Note that the spatial domain contains the entire device and is not stated in each component separately. Therefore, it is not necessary to consider implicit or explicit conditions on the (unknown) boundaries within the configuration.

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 van Roosbroeck (1950) and under some assumptions derived in this form in Markowich (1986). This drift-diffusion system also exhibits 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 where the first summand is the diffusion part and the second the drift part. The remaining term $R_{net}(n,p,e(u))$ can be understood as the difference of the 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. Note also that R_{net} typically comprises a sum of several reaction terms. In particular, the so-called radiative (stimulated) recombination mechanism, often denoted by R_{rad}, is of special importance. Moreover, recent studies (e.g., Capellini et al. 2014, 2015; Koprucki et al. 2015; Peschka et al. 2015, 2016a, b) have shown that the stress generated by the displacement has a direct positive influence on R_{net}; see the discussion below. 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., Courant and Hilbert (1924). In future work, we will strive to include the electronics and optics into the topology optimization process.

Some work for model (1) has been already performed in Capellini et al. (2014, 2015), Koprucki et al. (2015) and Peschka et al. (2015, 2016a, b), where two possible designs were studied. It has been found that both of them have a lower lasing threshold than those previously reported in the literature. To be clear, it was observed that the larger the tensile stress, the higher the possibility for radiative recombination at low temperatures. This is important to prevent thermal damage. The idea was to place the materials in such a way that the insulating materials drive the current directly to the middle of the optical cavity.

The idea of considering only (1a) in this paper stems from the observations in Suess et al. (2013) that show a direct relation between the radiative (stimulated) recombination mechanism, R_{rad}, and the biaxial strain within the optical cavity. It should be noted that the optical cavity more or less corresponds to the part occupied by Germanium. The secondary reason for considering only (1a) is of a practical nature and ultimately stems from the fact that the rates derived in Suess et al. (2013) and Virgilio et al. (2015) are not derived in closed form and are therefore unsuitable for our numerical/algorithmic study. It is, however, assumed that it is monotone in $\mathrm{tr}(e(u))$ for small strains (Fig. 1).

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; this 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 and the level set method; see Bendsøe and Sigmund (2003) and the 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

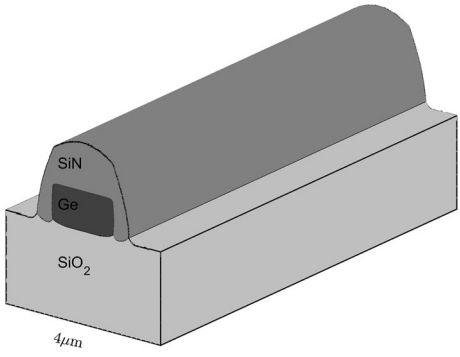

Fig. 1 A possible prototype strained photonic-device based on an optimal configuration determined by our approach; see Sect. 5.3 for details

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 in, e.g., Blank et al. (2014), Zhou and Wang (2006), Burger and Stainko (2006) and Takezawa et al. (2010).

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 Sect. 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. 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 Sect. 4, we present several algorithmic approaches. In particular, we discuss a popular gradient flow approach, projected gradients, and interior methods. The performance and viability of these methods in practice is given in Sect. 5. Ultimately, the methods are able to provide new designs that suggest a 15% increase in strain within the optical cavity, compared to the (empirically determined) benchmarks from Peschka et al. (2015). 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 the design variables. A similar model was considered in Blank et al. (2014), 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 the 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 Blank et al. (2014).

2.1 Forward problem

As stated above, we seek to place 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 $i \in \{1, \ldots, N\}$ we assign a phase-field function φ_i, whose support $\mathrm{supp}\varphi_i$ 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

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

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 upon cooling during the manufacturing process. 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

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

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

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

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

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

Finally, the coincidence sets $\{\varphi_i = 0\}$ should have finite perimeter (Fig. 2).

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

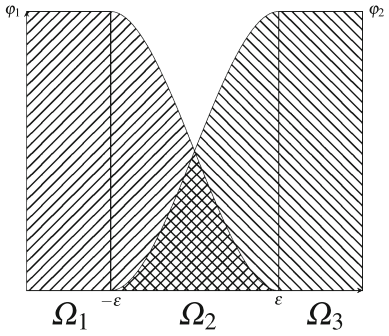

Fig. 2 Phase-field model. On Ω_1 and Ω_3 the phases are pure, on Ω_2 they mix

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

Since there is no term controlling the perimeter of phases, the "optimal solution" may exhibit fractal behavior, i.e., the solution may not exist 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

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

where

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

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 $\mathscr{P}(\{\varphi_i = 1\}; \Omega)$ equals the total variation of φ_i due to the co-area formula Fleming and Rishel (1960, Theorem I). This is a common regularization procedure in topology optimization, cf. Sokolowski and Zolesio (1992) and Bendsøe and Sigmund (2003).

The 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., Mignot (1976), Barbu (1984) or more recent work by Hintermüller and Kopacka (2009), Herzog et al. (2013) and the references therein.

Even though problem (7) admits an optimal solution, as can be seen from Lemma 7 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 Modica (1987). 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:

$$\begin{aligned} &\min J^0(\boldsymbol{u}) + \alpha f_{\mathrm{GL}}(\boldsymbol{\varphi}) \quad \text{over } \boldsymbol{\varphi} \in H^1(\Omega,\mathbb{R}^N),\ \boldsymbol{u} \in H^1_0(\Omega,\mathbb{R}^2) \\ &\text{s.t. } (\boldsymbol{\varphi},\boldsymbol{u}) \text{ satisfies (2)}, \\ &\quad \boldsymbol{\varphi} \in \mathscr{G}_{ad}, \end{aligned} \tag{10}$$

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

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

Due to the presence of $\nabla\boldsymbol{\varphi}$, 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-manifolds (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; and (ii) we are able to work in a Hilbert space setting rather than in a nonreflexive space of functions of bounded variation, which also has advantages 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 options, namely

$$\begin{aligned} &J^0_1(\boldsymbol{u}) := \int_D |\mathrm{tr}(e(\boldsymbol{u})) - e_d|^2 dx, \quad J^0_2(\boldsymbol{u}) := -\int_D |\mathrm{tr}(e(\boldsymbol{u}))|^2 dx, \\ &J^0_3(\boldsymbol{u}) := -\int_D \mathrm{tr}(e(\boldsymbol{u}))dx. \end{aligned} \tag{12}$$

Each J^0_i has advantages and disadvantages. The functional J^0_1 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^0_2 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, the lack of compactness properties prevents 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

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

For existence and uniqueness of solutions for this equation we will make 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 Blank et al. (2014) while ours represent several improvements for the 2D case.

3.1 Existence of an optimal topology

In the sequel, we make the following regularity assumptions, which will hold throughout the remainder of the text:

(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 $\mathrm{cl}\Pi_i \cap \mathrm{cl}\Pi_j = \emptyset$ for all $i \neq j$.

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

(A3) $\mathbb{C}$ 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

$$c_1\|E_1\|^2_{\mathbb{R}^{2\times 2}} \le \widehat{\mathbb{C}}(\phi)E_1 : E_1,$$
$$\widehat{\mathbb{C}}(\phi)E_1 : E_2 \le c_2\|E_1\|_{\mathbb{R}^{2\times 2}}\|E_2\|_{\mathbb{R}^{2\times 2}},$$

where the matrix product is understood as $A : B = \sum_i \sum_j a_{ij}b_{ij}$.

- $\widehat{\mathbb{C}} \in \mathscr{C}^{1,1}(\mathbb{R}^N, \mathbb{R}^{2\times 2\times 2\times 2})$, 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 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 Goldberg et al. (1992, Theorem 7).

Lemma 1 *Under assumptions (A1) and (A3) the mappings* $\mathbb{C} : H^1(\Omega, \mathbb{R}^N) \to L^p(\Omega, \mathbb{R}^{2\times 2\times 2\times 2})$ *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* $p \in [1, \infty)$.

Lemma 1 helps us to prove, and consequently to exploit, more solution regularity of the displacement than originally given by the statement of the elasticity equation. This is particularly helpful later in the text for arguing the validity of the projected gradient method in a Hilbert space setting.

Lemma 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* $\|\boldsymbol{u}\|_{W_0^{1,p}(\Omega,\mathbb{R}^2)} \le M$ *whenever* $(\boldsymbol{\varphi}, \boldsymbol{u})$ *solves* (2) *and* $\boldsymbol{\varphi} \in \mathscr{G}$. *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 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 Bensoussan and Frehse (2002, 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

$$\|\boldsymbol{u}\|_{W_0^{1,q}(\Omega,\mathbb{R}^2)} \le c\|F(\boldsymbol{\varphi})\|_{L^4(\Omega,\mathbb{R}^N)} \tag{14}$$

holds true. Even though this result was presented for the scalar-valued case, it may be generalized to the vector-valued case, which is needed here. Moreover, even though $\boldsymbol{\varphi}$ enters the differential operator, the 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

$$\begin{aligned}&\int_\Omega \mathbb{C}(\boldsymbol{\varphi}^1) e(\boldsymbol{u}^1 - \boldsymbol{u}^2) : e(\boldsymbol{v}) dx \\ &\quad = \int_\Omega (F(\boldsymbol{\varphi}^1) - F(\boldsymbol{\varphi}^2)) : e(\boldsymbol{v}) dx - \int_\Omega (\mathbb{C}(\boldsymbol{\varphi}^1) - \mathbb{C}(\boldsymbol{\varphi}^2)) e(\boldsymbol{u}^2) : e(\boldsymbol{v}) dx.\end{aligned}$$

Using again Bensoussan and Frehse (2002, 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

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

where we have used Lemma 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 *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 the pair $(\boldsymbol{\varphi}, \boldsymbol{u})$ satisfies the elasticity equation (2) follows from Lemma 2. Thus, problem (10) admits a feasible solution. Let $\{(\boldsymbol{\varphi}^k, \boldsymbol{u}^k)\}$ be an infimizing sequence. By Lemma 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, the Ginzburg–Landau energy ensures that $\{\boldsymbol{\varphi}^k\}$ is bounded in $H^1(\Omega, \mathbb{R}^N)$. Thus, there exist $\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$ and $\boldsymbol{u} \in W_0^{1,p}(\Omega, \mathbb{R}^2)$ such that (along a subsequence) $\boldsymbol{\varphi}^k \rightharpoonup \boldsymbol{\varphi}$ in $H^1(\Omega, \mathbb{R}^N)$ and $\boldsymbol{u}^k \rightharpoonup \boldsymbol{u}$ in $W_0^{1,p}(\Omega, \mathbb{R}^2)$. Moreover, we have along the same subsequence $\boldsymbol{\varphi}^k \to \boldsymbol{\varphi}$ in $L^{\frac{2p}{p-2}}(\Omega, R^N)$ due to the Rellich–Kondrachov embedding theorem.

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

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

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

$$e(\boldsymbol{u}^k) \rightharpoonup e(\boldsymbol{u}) \text{ in } L^p(\Omega, \mathbb{R}^2) \text{ and } \mathbb{C}(\boldsymbol{\varphi}^k) \to \mathbb{C}(\boldsymbol{\varphi}) \text{ in } L^{\frac{2p}{p-2}}(\Omega, R^{2\times 2\times 2\times 2}).$$

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 Sect. 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 Blank et al. (2014).

Lemma 4 *Assume that (A1)–(A3) hold. 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* $S'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi} = \boldsymbol{q}$, *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

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

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 equalities and rearranging the terms results in

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

Now we set $\boldsymbol{v} = \boldsymbol{u}^2 - \boldsymbol{u}^1 - \boldsymbol{q}$ and apply Korn's lemma (Zeidler 1988, Corollary 62.13) coupled with the ellipticity assumption from (A3) on the left-hand side to obtain

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

where $p > 2$ is the exponent from Lemma 3 and $X := L^q(\Omega, \mathbb{R}^{2\times 2\times 2\times 2})$ with $q := \frac{2p}{p-2}$ (stemming from the Rellich–Kondrachov embedding theorem). 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 1 and 2. Thus, we have shown that S is Fréchet differentiable. The continuity of the derivative may be shown as in Lemma 3. □

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

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

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

Lemma 5 *Assume that (A1)–(A3) hold and consider $\boldsymbol{\varphi} \in \mathscr{G}$. If $J^0 : H_0^1(\Omega, \mathbb{R}^2) \to \mathbb{R}$ is (continuously) Fréchet differentiable at $\boldsymbol{\varphi}$, then $\mathscr{J} : H^1(\Omega, \mathbb{R}^N) \to \mathbb{R}$ 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*

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

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

$$\begin{aligned}-\mathrm{div}\mathbb{C}(\boldsymbol{\varphi})e(\boldsymbol{p}) &= -(J_u^0)'(\boldsymbol{\varphi}, \boldsymbol{u}) \\ \boldsymbol{p} &= 0\end{aligned} \tag{19}$$

Proof The (continuous) differentiability of $\mathscr{J}$ at $\boldsymbol{\varphi}$ follows from the chain rule. By the standard technique, see Hinze et al. (2009, Section 1.6.2), we obtain

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

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,

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

from which (18) follows by substitution. □

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

Theorem 1 *Assume (A1)–(A3) hold and let $(\boldsymbol{\varphi}, \boldsymbol{u})$ be an optimal solution to (10). Then the following first-order optimality condition holds:*

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

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

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

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

However, the usual method of deriving the existence of such multipliers via the constraint qualification (cf. Bonnans and Shapiro 2000; Robinson 1976; Zowe and Kurcyusz 1979),

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

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 Sect. 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. With this goal, we first show that the control-to-state mapping is twice continuously differentiable. The proof mimics that of Lemma 4. In this section, we need to strengthen assumption (A3). We thus augment (A3):

(A3a) In addition to assumption (A3), we assume that $\widehat{\mathbb{C}}$ is twice continuously differentiable with the second derivative being globally Lipschitz.

Moreover, we mention that our method is inspired by the general approach presented in Bonnans and Shapiro (2000, Chapters 3.2, 3.3). However, there are some differences, which we detail as they appear. Although there may be more general conditions, the results in Theorems 2 and 3 capture the basic forms.

Lemma 6 *Assume that (A1)–(A3a) hold. 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* $[S''(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^1]\delta\boldsymbol{\varphi}^2 = \boldsymbol{r}$, *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 $S' \in \mathscr{L}(H^1(\Omega, \mathbb{R}^N), \mathscr{L}(H^1(\Omega, \mathbb{R}^N), H_0^1(\Omega, \mathbb{R}^2)))$, we need to show that

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

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

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

as well as

$$-\int_\Omega \mathbb{C}(\boldsymbol{\varphi})e(\boldsymbol{q}^2) : e(\boldsymbol{v})dx = \int_\Omega \mathbb{C}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^2 e(\boldsymbol{u}) : e(\boldsymbol{v})dx - \int_\Omega F'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}^2 : e(\boldsymbol{v})dx$$

and

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

Summing these three equalities and rearranging the terms results in

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

Now set $\boldsymbol{v} = \hat{\boldsymbol{q}} - \boldsymbol{q}^2 - \boldsymbol{r}$ and apply Korn's lemma (Zeidler 1988, Corollary 62.13) coupled with the ellipticity assumption from (A3) on the left-hand side and the estimates on the right-hand side to obtain

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

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 the linear functional

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

Before stating the second-order conditions, we recall some results from 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 other than $\mathscr{J}'(\boldsymbol{\varphi})(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) \geq 0$ for all $\hat{\boldsymbol{\varphi}} \in \mathscr{G}$, due to to the convexity of $\mathscr{G}$, we can write (21) equivalently as

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

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 $\boldsymbol{\lambda} = -\mathscr{J}'(\boldsymbol{\varphi})$. Further, we recall the definition of the radial cone

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

and the annihilator $\{\cdot\}^\perp$ using the dual pairing on $H^1(\Omega, \mathbb{R}^N)$ and $H^1(\Omega, \mathbb{R}^N)^*$, defined by

$$\{\boldsymbol{\lambda}\}^\perp := \{\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N) | \langle \boldsymbol{\lambda}, \boldsymbol{\varphi} \rangle = 0\}.$$

Finally we define the linear operator $A : H^1(\Omega, \mathbb{R}^N) \to H^1(\Omega, \mathbb{R}^N)^*$ by

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

Theorem 2 *Let (A1)–(A3a) be satisfied and let $\boldsymbol{\varphi} \in \mathscr{G}$ 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 critical directions $\boldsymbol{d} \in \overline{R_{\mathscr{G}}(\boldsymbol{\varphi}) \cap \{-\mathscr{J}'(\boldsymbol{\varphi})\}^\perp}$.

Remark 1 In other words, if $\boldsymbol{\varphi}$ is a local minimum, then the "curvature" of the reduced objective functional (in the sense of second-order epiderivatives, cf. Bonnans and Shapiro (2000, Chap. 3) is non-negative in all critical directions. This is therefore an infinite-dimensional analogue of the standard result in $\mathbb{R}^n$.

Proof Set $\boldsymbol{\lambda} := -\mathscr{J}'(\boldsymbol{\varphi})$ and fix any $\boldsymbol{d} \in R_{\mathscr{G}}(\boldsymbol{\varphi}) \cap \{\boldsymbol{\lambda}\}^\perp$. 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 $\boldsymbol{r} := (S'(\boldsymbol{\varphi}))^* \tau_D$. Note that all terms (excluding $\boldsymbol{\lambda}$) arise directly from the definition of $\mathscr{J}$.

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

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

We now group like terms 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:

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

Due to Lemma 6, 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) to

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

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

It is known that no "curvature" should appear in either second-order necessary or sufficient optimality conditions if the underlying constraint set M 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 arguments similar to those in Mignot (1976). The general case of N phases goes beyond the scope and purpose of this text. We observe that the results of Wachsmuth (2016) cannot be used here and that if M is not polyhedric, then $\overline{R_{\mathscr{G}}(\boldsymbol{\varphi}) \cap \{-\mathscr{J}'(\boldsymbol{\varphi})\}^\perp}$ might be too small, i.e., $\{0\}$.

For the sufficient second-order conditions, we define the cone

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

This is strongly reminiscent of the *approximate critical cone* used in Bonnans and Shapiro (2000). However, there are several key differences:

1. Here, $\boldsymbol{\varphi}$ is assumed to be a stationary point. For Bonnans and Shapiro (2000), the approximate critical cone is defined for any feasible point.
2. We make direct use of dual information, i.e., $\boldsymbol{\lambda}\in N_{\mathscr{G}}(\boldsymbol{\varphi})$, in the definition of $\mathscr{K}_\eta(\boldsymbol{\varphi},\boldsymbol{\lambda})$.
3. Here, $\boldsymbol{d}\in\mathscr{K}_\eta(\boldsymbol{\varphi},\boldsymbol{\lambda})$ is required to be in the radial cone $R_{\mathscr{G}}(\boldsymbol{\varphi})$, whereas in Bonnans and Shapiro (2000) $\boldsymbol{d}$ is taken to be "close" to the linearization cone.

In particular, 3. means that $\mathscr{K}_\eta(\boldsymbol{\varphi},\boldsymbol{\lambda})$ is potentially smaller than the approximate critical cone used in Bonnans and Shapiro (2000).

Theorem 3 *Assume that (A1)–(A3a) holds, let $\boldsymbol{\varphi}$ be a stationary point, and assume that the following growth condition holds: there exist $\eta>0$ and $\beta>0$ such that*

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

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

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

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

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

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 $\boldsymbol{\lambda}:=-\mathscr{J}'(\boldsymbol{\varphi})$. Since $\boldsymbol{\varphi}$ is a stationary point, we have $\boldsymbol{\lambda}\in N_{\mathscr{G}}(\boldsymbol{\varphi})$ and

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

where we used the differentiability of $\mathscr{J}$ and (32). Moreover, $\boldsymbol{\lambda}\in N_{\mathscr{G}}(\boldsymbol{\varphi})$ and $\boldsymbol{\varphi}_n\in\mathscr{G}$ imply that $0\ge\langle\boldsymbol{\lambda},\boldsymbol{\varphi}_n-\boldsymbol{\varphi}\rangle=t_n\langle\boldsymbol{\lambda},\boldsymbol{d}_n\rangle$. This yields the inequality $\langle\boldsymbol{\lambda},\boldsymbol{d}_n\rangle\le 0$, which together with (33) implies that for large enough n we have $\boldsymbol{d}_n\in\mathscr{K}_\eta(\boldsymbol{\varphi},\boldsymbol{\lambda})$. But then by arguments similar to those in the proof of Theorem 2 we obtain

$$\frac{\beta}{2} \geq \frac{\mathscr{J}(\boldsymbol{\varphi}_n) - \mathscr{J}(\boldsymbol{\varphi})}{\frac{1}{2}t_n^2} = \frac{\mathscr{J}(\boldsymbol{\varphi}_n) - \mathscr{J}(\boldsymbol{\varphi}) - t_n\langle 0, \boldsymbol{d}_n\rangle}{\frac{1}{2}t_n^2}$$

$$= \alpha\varepsilon \int_\Omega |\nabla \boldsymbol{d}_n|^2 dx - \frac{\alpha}{\varepsilon}\int_\Omega \sum_{i=1}^N (d_n)_i^2 dx - \tau_D([u''(\boldsymbol{\varphi})\boldsymbol{d}_n]\boldsymbol{d}_n) + \frac{o(t_n^2)}{t_n^2} - \frac{t_n\langle \boldsymbol{\lambda}, \boldsymbol{d}_n\rangle}{\frac{1}{2}t_n^2}$$

$$\geq \beta + \frac{o(t_n^2)}{t_n^2} - \frac{t_n\langle \boldsymbol{\lambda}, \boldsymbol{d}_n\rangle}{\frac{1}{2}t_n^2}.$$

Since $\boldsymbol{\lambda} \in N_{\mathscr{G}}(\boldsymbol{\varphi})$, 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 taking 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 briefly describe these methods in this section. For 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 (Behrman 1998; Blowey and Elliott 1993) in which one introduces an artificial dependence of $\boldsymbol{\varphi}$ on time. In other words, suppose we are given an energy functional $E : \mathscr{U} \to \overline{\mathbb{R}}$, where $\mathscr{U}$ is a real Hilbert space, $E(u) := E_1(u) + E_2(u)$ such that E_1 is continuously Fréchet differentiable, and E_2 is proper, convex, lower-semicontinuous and subdifferentiable. Then the first-order optimality conditions for minimizing E over $\mathscr{U}$ are given by the generalized equation

$$0 \in \nabla E_1(u) + \partial E_2(u).$$

We can then define a (non-smooth) gradient flow-type problem by considering the (generalized) evolution equation: Find $u : [0, \infty) \to \mathscr{U}$ such that

$$\frac{\partial u}{\partial t}(t) \in \nabla E_1(u(t)) + \partial E_2(u(t)), \quad t \in (0, \infty), \text{ and } u(0) := u_0.$$

Using an implicit-Euler scheme to discretize in time yields the iteration:

$$u^{t+1} \in u^t + \delta t \nabla E_1(u^{t+1}) + \delta t \partial E_2(u^{t+1}), \quad t = 0, 1, \ldots; \ \delta t > 0.$$

In our setting, $E_2 := i_{\mathscr{G}_{ad}}$, the indicator functional for the H^1-Gibbs simplex. In such cases, the previous inclusion can be rewritten:

$$\langle u^{t+1} - u^t - \delta t \nabla E_1(u(t^{+1})), v - u^{t+1} \rangle_{\mathscr{U}^*,\mathscr{U}} \geq 0, \ \forall v \in \mathscr{G}_{ad}, \quad t = 0, 1, \ldots.$$

Since in our application $\nabla E_1(u^{t+1})$ has a rather complicated form, we use u^{t+1} in only part of the gradient; elsewhere we use u^t.

More specifically, we embed (21) into a gradient flow scheme by introducing $\alpha\varepsilon\frac{\partial\boldsymbol{\varphi}}{\partial t}$. Then, 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:

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

where $\boldsymbol{u}^t$ and $\boldsymbol{p}^t$ are solutions of the elasticity and adjoint equation, respectively, obtained by setting $\boldsymbol{\varphi} = \boldsymbol{\varphi}^t$. 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 $\mathscr{G}$, for which there exist efficient function-space-based numerical approaches; see Adam et al. (2018a). The algorithm should stop when $\|\boldsymbol{\varphi}^{t+1} - \boldsymbol{\varphi}^t\|_{H^1}$ is sufficiently small, 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. Therefore, we stop after a fixed number of iterations.

Algorithm 1 Gradient flow

Input: initial point $\varphi^0 \in \mathscr{G}$, $k \leftarrow 0$
1: **repeat**
2: Solve (2) and (19) for u^k and p^k, respectively, with $\varphi = \varphi^k$
3: Solve (34) for φ^{k+1}; set $k \leftarrow k+1$
4: **until** stopping criterion is satisfied
5: **return** φ^k

4.2 Projected gradients

The idea of projected gradients goes back to Goldstein (1964), Levitin and Polyak (1966). At each step, we compute the following update

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

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

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

In an implementation, we make use of the generalized Armijo step size rule (Bertsekas 1976), where we choose some $t^k > 0$ such that the following inequality

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

is satisfied. Here, $\sigma > 0$ is a given parameter, and we observe that if $\mathscr{G}$ were the whole space, then (37) would reduce to the classical Armijo rule. For the stopping criterion, we check the residual in H^1 of the optimality condition

$$\boldsymbol{\varphi} - Proj_{\mathscr{G}}\big(\boldsymbol{\varphi} - c\mathscr{J}'(\boldsymbol{\varphi})_{Riesz}\big), \tag{38}$$

where $c > 0$ is a fixed constant. Since $\mathscr{G}$ is a convex set, we obtain that once $\boldsymbol{\varphi}^k \in \mathscr{G}$ 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 2. A convergence proof was given for the first time in Bertsekas (1976) for finite dimensions. Recently, it has been generalized in Blank and Rupprecht (2015) to an intersection of a Hilbert with a Banach space satisfying certain properties. This paper was motivated by Blank et al. (2014), where the authors worked with the space $H^1(\Omega,\mathbb{R}^N) \cap L^\infty(\Omega,\mathbb{R}^N)$ and used a 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. Nevertheless, as noted above, there exist fast numerical methods to do this; see Adam et al. (2018a). In particular, after several iterations, the initial estimations of active and inactive are fine enough that the solvers suggested in Adam et al. (2018a) need only two or three iterations to obtain the projection. Each iteration requires the solution of a sparse linear system.

Algorithm 2 Projected gradients

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

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

4.3 Interior point method

Although the projected gradient method is largely successful for solving (10), it can in some instances require many steps to obtain a reasonable tolerance for the residual (38). 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 instead consider interior point methods, which ensure feasibility of $\boldsymbol{\varphi}^k$ throughout. For an excellent review with many references, see Forsgren et al. (2002). For its applications to optimal control with PDE constraints see Schiela and Weiser (2008), Ulbrich and Ulbrich (2009).

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. The remaining system to be solved at each iteration is as follows:

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

where $\beta = 0$. The last condition, together with $\boldsymbol{\lambda} \geq 0$ and $\boldsymbol{\varphi} \geq 0$, is nothing more than the complementarity condition for the remaining inequality constraint. At each "inner" loop (i.e., for each fixed $\gamma > 0$) 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 $t^k \leftarrow \min\{c_f \cdot \min\{t_\varphi^l, t_\lambda^l\}, 1\}$. 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 $c_f \in (0,1)$. The parameter β is then decreased and the process is repeated. We give the interior point method in Algorithm 3. By G we denote the left-hand side of (22) or (39). Moreover, we 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 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 $\beta_{\min}$, $k \leftarrow 0$

repeat
  based on (22) or (39) compute $d^k \leftarrow -G'(y^k)^{-1}G(y^k)$ ▷ direction
  based on (40) compute $t^k \leftarrow \min\{c_f \cdot \min\{t^l_\varphi, t^l_\lambda\}, 1\}$ ▷ step size
  $y^{k+1} \leftarrow y^k + t^k d^k$ ▷ new iterate
  $\beta^{k+1} \leftarrow \max(c_\beta \beta^k, \beta_{\min})$
  $k \leftarrow k+1$
until stopping criterion is satisfied
return φ^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 Bendsøe and Sigmund (2003), Haslinger and Neittaanmäki (1988). In both applications, the elasticity tensor has the form

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

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

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

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

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

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

For the projection onto the Gibbs simplex $\mathcal{G}$, we discretized the problem and used the semismooth Newton's method (Hintermüller et al. 2002), which is equivalent to a primal-dual active set strategy. Another possibility would be to use the path-following method from Adam et al. (2018a). We use the former in all the experiments.

5.1 Updating the parameters

The general model contains a number of parameters, which 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 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. To refine the mesh, we employ the standard red refinement strategy; see, e.g., Brenner and Carstensen (2004). 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. The meshes were refined three times for the first application and four times for the second one.

Concerning the parameters, α was chosen as small as possible; see Sect. 5.4. The cutoff parameters δ_1 and δ_2 were chosen so that the cutoff has a negligible effect on the interval (0, 1). The fraction to the boundary c_f was chosen close to 1 and c_β

Table 1 List of parameters

	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}

close to 0 to promote rapid convergence. Since the second application is more demanding, we decrease c_f, increase c_β and set a minimal value $\beta_{\min}$. For γ we chose a relatively high value to obtain only a small violation of constraints (4). As the convergence for the interior point was rapid 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 the residual development see Fig. 6. Finally, δt was chosen small to ensure small steps for the gradient flow method. Even though $h_{\min} = \frac{1}{128}$ may seem too large, the mesh is rather fine because Ω is not the unit square.

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 subsequently omitted it. For the second application, we ran the projected gradients and interior point applied on both (22) and (39). They performed comparably, and 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 since the multipliers need not exist. Second, for the gradient flow and the projected gradients one iteration involves solving the elasticity and adjoint equations and performing the line search, whereas for the interior point method one iteration involves solving one large system. 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 50% 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 $\boldsymbol{g} = (0, -5000)$. The problem has the form

$$\min \int_{\Gamma_N} \boldsymbol{g}\cdot\boldsymbol{u}\, dS + \frac{\alpha}{2}\int_\Omega \left(\varepsilon|\nabla\boldsymbol{\varphi}|^2 + \frac{1}{\varepsilon}\boldsymbol{\varphi}\cdot(1-\boldsymbol{\varphi})\right)dx \text{ over } \boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N),\ \boldsymbol{u} \in H_0^1(\Omega, \mathbb{R}^2)$$

$$\text{s.t. } \hat{E}(\boldsymbol{\varphi}, \boldsymbol{u}) = 0, \quad \boldsymbol{\varphi} \in \mathscr{G}_{ad}, \quad \int_\Omega \varphi_1 dx = \frac{1}{2}|\Omega|.$$

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

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

For more details we refer to Blank et al. (2014, Section 6.1) or to our codes available online (Adam et al. 2018b).

The resulting bridge shape is presented in Fig. 3. 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 of 1000 iterations was not imposed, it did converge to the same solution. Further evidence is summarized in Table 2. Every row describes one method. The first column gives the total objective $\mathscr{J}(\boldsymbol{\varphi})$ and the second column gives the compliance $\int_{\Gamma_N} \boldsymbol{g}\cdot\boldsymbol{u}\, dS$. The next four columns show the numbers of iterations and the last four columns the numbers 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.

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 are $\lambda_{Ge} = 44.279$, $\lambda_{SiN} = 110.369$, $\lambda_{SiO_2} = 16.071$, $\mu_{Ge} = 27.249$, $\mu_{SiN} = 57.813$, $\mu_{SiO_2} = 20.798$, $\epsilon_0 = 2.5{\cdot}10^{-3}$ and $\sigma_0 = -2.5$; see Lu (2007), Vlassak and Nix (1992), Wortman and Evans (1965). Concerning the objective function, the weak lower semicontinuity of J^0 from (A2) is satisfied only for J_1^0 and

Fig. 3 Optimal design for the bridge construction problem for the interior point method and projected gradients (left) and gradient flow (right). There are differences at the corners and at the width of the interfacial region

Table 2 Numerical evidence for the application described in Sect. 5.2. The rows correspond to the interior point method (IP), projected gradients (PG) and gradient flow (GF). The 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 are the numbers of nodes

	Objective		# Iterations per (M)esh				# Nodes per (M)esh			
	$\mathscr{J}(\boldsymbol{\varphi})$	$\int_{\Gamma_N} \boldsymbol{g}\cdot\boldsymbol{u}\, dS$	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

J_3^0 but not for J_2^0, all of them being defined in (12). Since we do not have an estimate for e_d, we have decided to work with J_0^3.

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, potentially, 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 Peschka et al. (2015) which we refer to as the original configuration. We show the design differences in Fig. 4. We see that the difference between the two designs is significant; in particular, the SiN stressor encapsulates the entire section of Ge. The biaxial strain is depicted in Fig. 5.

The results are summarized in Table 3 which is very similar to Table 2. Here, the objective value reached by the projected gradient method is lower than that obtained using the interior point method, but the difference is essentially negligible. On the other hand, the strain profile for the interior point method is slightly better. In both cases we obtained an improvement in the strain of approximately 15% compared to the original configuration. The next five columns give the numbers of iterations on individual meshes, which are approximately constant for the interior point and double on the last mesh. For the projected gradients, the number of iterations

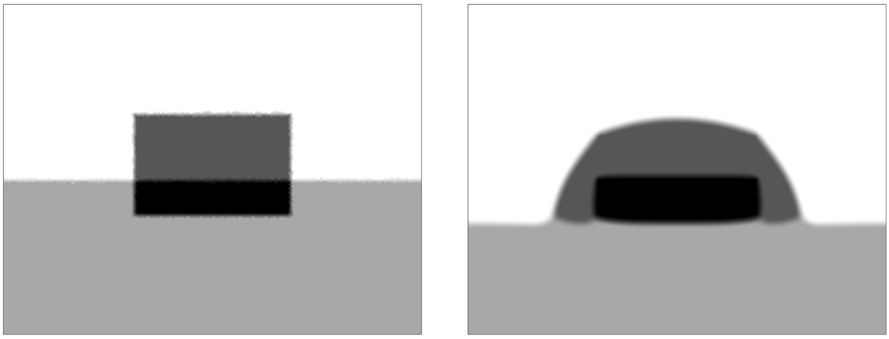

Fig. 4 The original configuration (left) and optimal design (right) for the application presented in Sect. 5.3. The gray scales are as follows: black (Ge), dark gray (SiN) and light gray (SiO_2)

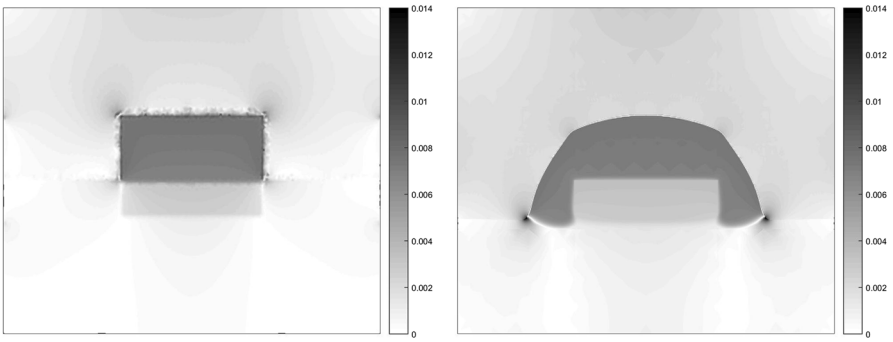

Fig. 5 The biaxial strain for the original configuration (left) and optimal design (right) for the application presented in Sect. 5.3

doubles every mesh refinement. For the residual development on the next-to-last mesh, see Fig. 6. Concerning the precise meaning of the iteration numbers, please refer to the end of the previous subsection. The last five columns give the numbers of nodes. Observe that the number of variables is much higher: for example the resulting matrix in system (22) has dimension 929113×929113 on the finest mesh.

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 Hintermüller and Kunisch (2006a, b) 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 the objective

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

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 Fig. 7.

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 Fig. 8. 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 surprising as the regularization term disappears for $\alpha \to 0$.

Finally, in Fig. 9 we perform a similar analysis for the dependence of the optimal design on the eigenstrain parameter ε_0; see (3). In accordance with results in Fig. 8 we fix $\alpha = 5 \times 10^{-5}$. The top left figure depicts the value of the Ginzburg–Landau energy. It develops in a continuous way but a jump occurs around $\varepsilon_0 = 0.003185$, which means that the perimeter of the optimal design changed dramatically. The

Table 3 Numerical evidence for the application described in Sect. 5.2. The rows correspond to the interior point method (IP) and projected gradients (PG). The first two columns are the values of the objective function, the third gives the improvement over the initial configuration and the remaining columns are the numbers of iterations and nodes

	Objective			# Iterations per (M)esh					# Nodes per (M)esh				
	$\mathscr{J}(\boldsymbol{\varphi})$	$-J_3^0(\boldsymbol{\varphi})$	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

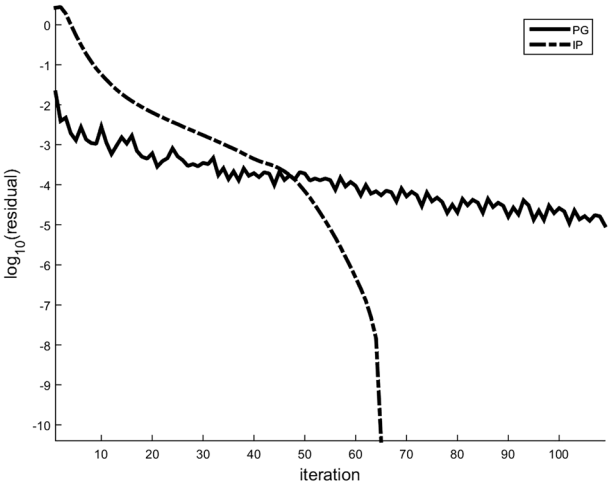

Fig. 6 Residual development for the interior point method (IP) and the projected gradients (PG) for the application presented in Sect. 5.3

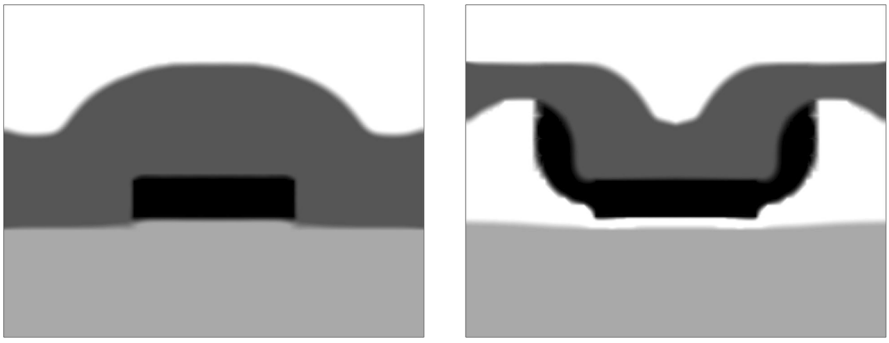

Fig. 7 Optimal designs for weighted objective J_0^4 with weights $a = 1$, $b = 10$ (left) and $a = 10$, $b = 1$ (right). The gray scales are as follows: black (Ge), dark gray (SiN) and light gray (SiO_2)

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 Fig. 8, 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 converges 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 include only the elasticity equation, this represents an important first step in the design process. Following a recent paper on multi-material topology optimization (Blank et al. 2014), we formulated the

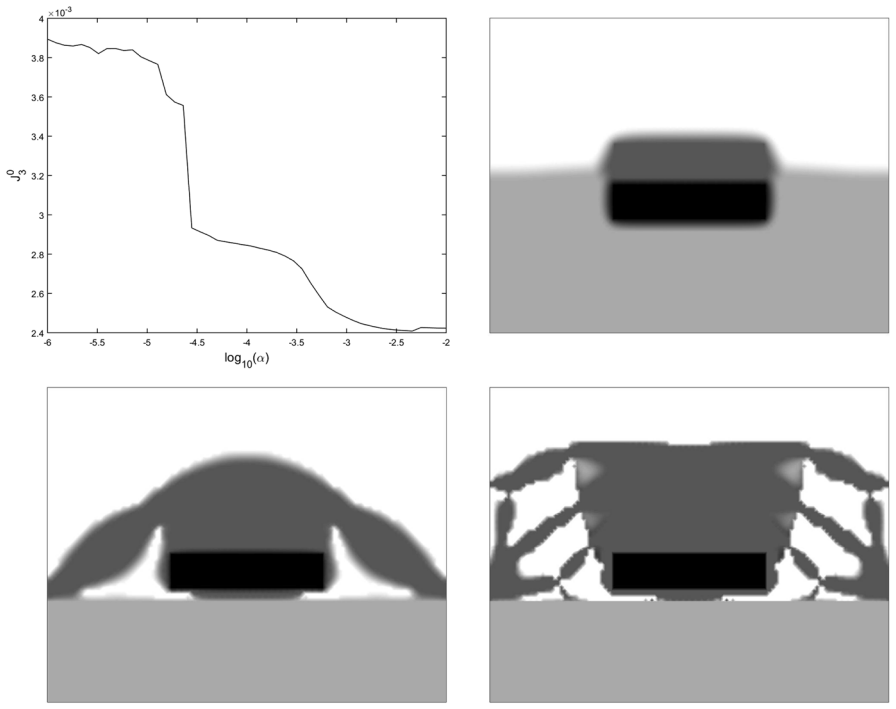

Fig. 8 Dependence of the strain on α (top left) and optimal designs for $\alpha = 10^{-2}$ (top right), $\alpha = 10^{-5}$ (bottom left) and $\alpha = 10^{-6}$ (bottom right). The gray scales are as follows: black (Ge), dark gray (SiN) and light gray (SiO_2)

problem using the phase-field approach and derived first- and second-order optimality conditions. On this basis, we compared the performance of several (popular) algorithms from nonlinear optimization, namely gradient flow, projected gradients and the 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.

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.

Appendix

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

Lemma 7 *Assume that (A1)–(A3) hold and that Ω has a finite perimeter. Then problem (7) admits an optimal solution.*

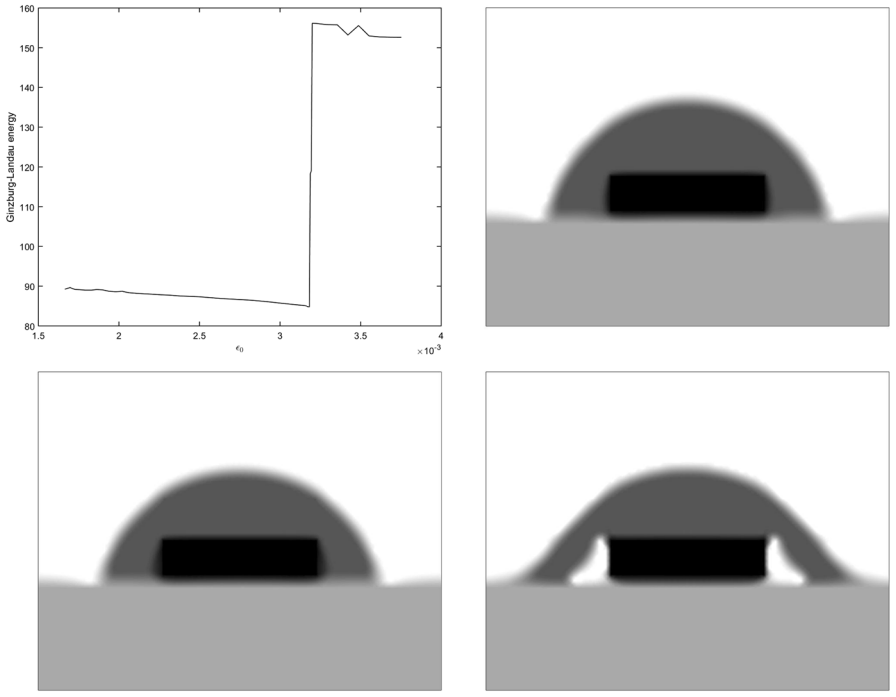

Fig. 9 Dependence of the Ginzburg–Landau energy on the eigenstrain parameter ε_0 (top left) and optimal designs for $\varepsilon_0 = 0.001667$ (top right), $\varepsilon_0 = 0.003182$ (bottom left) and $\varepsilon_0 = 0.003188$ (bottom right). The gray scales are as follows: black (Ge), dark gray (SiN) and light gray (SiO_2)

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 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 $\mathscr{P}(\{\varphi_i^k = 1\}; \Omega)$ is uniformly bounded above. Since

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

due to Maggi (2012, Equation (12.26)), we may invoke Maggi (2012, 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 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 Maggi (2012, Proposition 12.15) we also have

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

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. □

References

Adam L, Hintermüller M, Surowiec TM (2018a) A semismooth Newton method with analytical path-following for the H[1]-projection onto the Gibbs simplex. IMA J Numer Anal. https://doi.org/10.1093/imanum/dry034

Adam L, Hintermüller M, Surowiec TM (2018b) Matlab source code. http://staff.utia.cas.cz/adam/research.html

Barbu V (1984) Optimal control of variational inequalities. Research notes in mathematics. Pitman Advanced Publishing Program, Boston

Behrman W (1998) An efficient gradient flow method for unconstrained optimization. Ph.D. thesis, Stanford University

Bendsøe M, Sigmund O (2003) Topology optimization: theory, methods, and applications. Springer, Berlin

Bensoussan A, Frehse J (2002) Regularity results for nonlinear elliptic systems and applications. Springer, Berlin

Bertsekas DP (1976) On the Goldstein–Levitin–Polyak gradient projection method. IEEE Trans Autom Control 21(2):174–184

Blank L, Rupprecht C (2015) 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

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

Blowey JF, Elliott CM (1993) Curvature dependent phase boundary motion and parabolic double obstacle problems. In: Ni WM, Peletier LA, Vazquez JL (eds) Degenerate diffusions. Springer, New York, pp 19–60

Bonnans JF, Shapiro A (2000) Perturbation analysis of optimization problems. Springer, Berlin

Brenner SC, Carstensen C (2004) Finite element methods. In: Encyclopedia of computational mechanics, Chap. 4. Wiley Online Library

Burger M, Stainko R (2006) Phase-field relaxation of topology optimization with local stress constraints. SIAM J Control Optim 45(4):1447–1466

Camacho-Aguilera RE, Cai Y, Patel N, Bessette JT, Romagnoli M, Kimerling LC, Michel J (2012) An electrically pumped germanium laser. Opt Express 20(10):11,316–11,320

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

Capellini G, Virgilio M, Yamamoto Y, Zimmermann L, Tillack B, Peschka D, Thomas M, Glitzky A, Nürnberg R, Gärtner K, Koprucki T, Schroeder T (2015) Modeling of an edge-emitting strained-Ge laser. In: Advanced solid state lasers

Courant R, Hilbert D (1924) Methoden der mathematischen Physik. Springer, Berlin

Dutt B, Sukhdeo DS, Nam D, Vulovic BM, Yuan Z, Saraswat KC (2012) Roadmap to an efficient germanium-on-silicon laser: strain vs. n-type doping. IEEE Photon J 4(5):2002–2009

El Kurdi M, Fishman G, Sauvage S, Boucaud P (2010) Band structure and optical gain of tensile-strained germanium based on a 30 band $k\cdot p$ formalism. J Appl Phys 107(1):013710

Fleming WH, Rishel R (1960) An integral formula for total gradient variation. Archiv der Mathematik 11(1):218–222

Forsgren A, Gill PE, Wright MH (2002) Interior methods for nonlinear optimization. SIAM Rev 44(4):525–597

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

Goldstein AA (1964) Convex programming in Hilbert space. Bull Am Math Soc 70(5):709–710

Haslinger J, Neittaanmäki P (1988) Finite element approximation for optimal shape design: theory and applications. Wiley, New York

Herzog R, Meyer C, Wachsmuth G (2013) B- and strong stationarity for optimal control of static plasticity with hardening. SIAM J Optim 23(1):321–352

Hintermüller M, Kunisch K (2006a) Feasible and noninterior path-following in constrained minimization with low multiplier regularity. SIAM J Control Optim 45(4):1198–1221

Hintermüller M, Kunisch K (2006b) Path-following methods for a class of constrained minimization problems in function space. SIAM J Optim 17(1):159–187

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

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

Hinze M, Pinnau R, Ulbrich M, Ulbrich S (2009) Optimization with PDE constraints. Springer, Berlin

Koprucki T, Peschka D, Thomas M (2015) Towards the optimization of on-chip germanium lasers. In: WIAS annual research report 2015

Levitin ES, Polyak BT (1966) Constrained minimization methods. Zh Vychisl Mat Mat Fiz 6(5):787–823

Liu J, Sun X, Camacho-Aguilera R, Kimerling LC, Michel J (2010) Ge-on-si laser operating at room temperature. Opt Lett 35(5):679–681

Lu Z (2007) Dynamics of wing cracks and nanoscale damage in silica glass. Ph.D. thesis, University of Southern California

Maggi F (2012) Sets of finite perimeter and geometric variational problems: an introduction to geometric measure theory. Cambridge University Press, Cambridge

Markowich PA (1986) The stationary semiconductor device equations. Springer, New York

Mignot F (1976) Contrôle dans les inéquations variationelles elliptiques. J Funct Anal 22(2):130–185

Modica L (1987) The gradient theory of phase transitions and the minimal interface criterion. Arch Ration Mech Anal 98(2):123–142

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

Peschka D, Rotundo N, Thomas M (2016a) Towards doping optimization of semiconductor lasers. J Comput Theor Transp 45(5):410–423

Peschka D, Thomas M, Glitzky A, Nürnberg R, Virgilio M, Guha S, Schroeder T, Capellini G, Koprucki T (2016b) Robustness analysis of a device concept for edge-emitting lasers based on strained germanium. Opt Quantum Electron 48:156

Robinson SM (1976) First order conditions for general nonlinear optimization. SIAM J Appl Math 30(4):597–607

Schiela A, Weiser M (2008) Superlinear convergence of the control reduced interior point method for PDE constrained optimization. Comput Optim Appl 39(3):369–393

Sigmund O, Torquato S (1997) Design of materials with extreme thermal expansion using a three-phase topology optimization method. J Mech Phys Solids 45(6):1037–1067

Sigmund O, Torquato S (1999) Design of smart composite materials using topology optimization. Smart Mater Struct 8(3):365
Sokolowski J, Zolesio JP (1992) Introduction to shape optimization: shape sensitivity analysis. Springer, Berlin
Suess MJ, Geiger R, Minamisawa RA, Schiefler G, Frigerio J, Chrastina D, Isella G, Spolenak R, Faist J, Sigg H (2013) Analysis of enhanced light emission from highly strained germanium microbridges. Nat Photon 7(6):466–472
Sun X, Jifeng L, Kimerling L, Michel J (2010) Toward a germanium laser for integrated silicon photonics. IEEE J Sel Top Quantum Electron 16(1):124–131
Takezawa A, Nishiwaki S, Kitamura M (2010) Shape and topology optimization based on the phase field method and sensitivity analysis. J Comput Phys 229(7):2697–2718
Ulbrich M, Ulbrich S (2009) Primal-dual interior-point methods for PDE-constrained optimization. Math Program 117(1):435–485
van Roosbroeck W (1950) Theory of the flow of electrons and holes in germanium and other semiconductors. Bell Syst Tech J 29(4):560–607
Virgilio M, Schroeder T, Yamamoto Y, Capellini G (2015) Radiative and non-radiative recombinations in tensile strained Ge microstrips: photoluminescence experiments and modeling. J Appl Phys 118(23):233110
Vlassak JJ, Nix WD (1992) A new bulge test technique for the determination of Young's modulus and Poisson's ratio of thin films. J Mater Res 7:3242–3249
Wachsmuth G (2016) A guided tour of polyhedric sets: basic properties, new results on intersections and applications. TU Chemnitz, Chemnitz
Wirths S, Geiger R, von den Driesch N, Mussler G, Stoica T, Mantl S, Ikonic Z, Luysberg M, Chiussi S, Hartmann JM, Sigg H, Faist J, Buca D, Grutzmacher D (2015) Lasing in direct-bandgap GeSn alloy grown on Si. Nat Photon 9(2):88–92
Wortman JJ, Evans RA (1965) Young's modulus, shear modulus, and Poisson's ratio in silicon and germanium. J Appl Phys 36(1):153–156
Zeidler E (1988) Nonlinear functional analysis and its applications IV: applications to mathematical physics. Springer, Berlin
Zhou S, Wang MY (2006) Multimaterial structural topology optimization with a generalized Cahn–Hilliard model of multiphase transition. Struct Multidiscip Optim 33(2):89
Zowe J, Kurcyusz S (1979) Regularity and stability for the mathematical programming problem in Banach spaces. Appl Math Optim 5(1):49–62

Affiliations

L. Adam[1] · M. Hintermüller[1,2] · T. M. Surowiec[3]

[1] Institut für Mathematik, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

[2] Weierstrass Institute, Mohrenstr. 39, 10117 Berlin, Germany

[3] FB12 Mathematik und Informatik, Philipps-Universität Marburg, Hans-Meerwein Straße 6, 35032 Marburg, Germany