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

Preprint ISSN 2198-5855

Optimization of a multiphysics problem in semiconductor laser design

Lukáš Adam[1], Michael Hintermüller[2,3], Dirk Peschka[2], Thomas M. Surowiec[4]

submitted: April 12, 2018

[1] Southern University of Science and Technology
1088 Xueyuan Ave
518055 Shenzhen, China
E-Mail: adam@sustc.edu.cn

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

[3] Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin, Germany

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

No. 2500
Berlin 2018

2010 *Mathematics Subject Classification.* 35J60, 74S05, 35Q93, 49Q10, 90C90, 90C06, 78A60.

Key words and phrases. Optoelectronics, semiconductor laser, strained germanium microbridges, van Roosbroeck, phase field, design optimization, topology optimization, PDE-constrained optimization.

This research is carried out in the framework of MATHEON supported by the Einstein Foundation Berlin.

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

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

Optimization of a multiphysics problem in semiconductor laser design

Lukáš Adam, Michael Hintermüller, Dirk Peschka, Thomas M. Surowiec

Abstract

A multimaterial topology optimization framework is suggested for the simultaneous optimization of mechanical and optical properties to be used in the development of optoelectronic devices. Based on the physical aspects of the underlying device, a nonlinear multiphysics model for the elastic and optical properties is proposed. Rigorous proofs are provided for the sensitivity of the fundamental mode of the device with respect to the changes in the underlying topology. After proving existence and optimality results, numerical experiments leading to an optimal material distribution for maximizing the strain in a Ge-on-Si microbridge are given. The highly favorable electronic properties of this design are demonstrated by steady-state simulations of the corresponding van Roosbroeck (drift-diffusion) system.

1 Introduction

The rapid miniaturization of microprocessors over the last four decades has been matched by a notable increase in computational performance. In particular, these developments have more or less followed Moore's law, which predicts an annual doubling of components per integrated circuit. Nevertheless, there are physical limits to this trend and further improvement requires alternative and innovative approaches. Therefore, silicon photonics integrates optical and electronic components into a single microchip, with the goal of using optical interconnects to provide faster data transfer between and inside microchips and to avoid the limitations of electrical wiring, cf. e.g. [1].

This paper is inspired by the promising approach of using strained germanium (Ge) as the optically active medium for an edge-emitting laser, which serves as the light source for silicon photonics, cf. [2, 3, 4, 5]. The base material used in the production of integrated circuits, silicon (Si), is an indirect-bandgap semiconductor, which implies that stimulated emission is strongly suppressed. The situation is similar for bulk Ge, which is also an indirect-bandgap semiconductor. Seemingly indicating that both materials are disadvantageous to create an integrated light source.

However, the band structure of germanium can be strongly altered using mechanical strains, and with a few percent tensile strain it even becomes a direct bandgap semiconductor, cf. [4]. In [6], the authors focus on modeling the effects of strain and doping on the electronic and optical properties. It is observed that stimulated emission and the resulting lasing threshold usually depend more on the strain than the doping profile.

While an optimal doping profile can be determined by optimizing charge transport using nonlinear drift-diffusion models [8, 9], an optimal material configuration, used to create tensile strain in the Ge, can be found by using techniques from topology optimization applied to linear elastic materials. Since the strain distribution has a larger effect on the gain, we only consider the mechanical properties in our multiphysics forward system. However, we also provide a numerical study of the electronic properties of the optimal device. In both settings we employ material parameter descriptions, which depend on the phase fields encoding the material distribution.

Various engineering studies have focussed on the production of Ge devices, where the light emission is improved by maximizing the strain. This led to a variety of device designs including suspended bridges

DOI 10.20347/WIAS.PREPRINT.2500 Berlin 2018

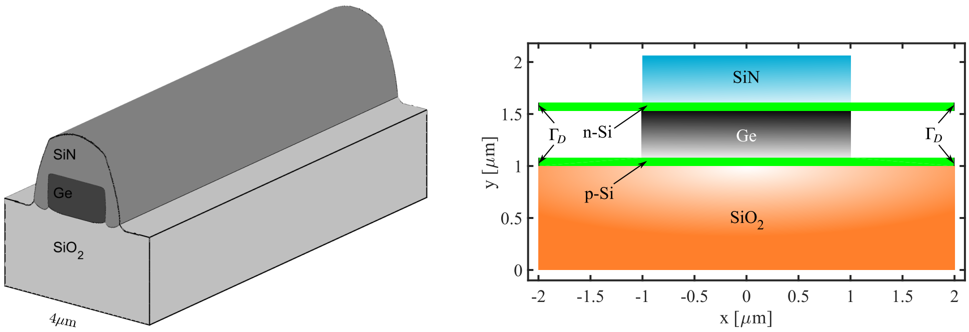

Figure 1: **(left)** A possible prototype strained photonic-device exhibiting the microbridge geometry. This configuration was determined in [7]. **(right)** Cross section of a Ge-on-Si microbridge with material distribution and contacts. This design would exhibit significant current leakage and potentially become damaged due to overheating.

[4, 10] or discs [11]. While the corresponding photoluminescence spectra support the improvement of the optical properties, manufacturability with standard fabrication processes and the incorporation of an optoelectronic loss mechanism are still active research, e.g. [12].

For instance, in general material configurations with large strain in the optically active area are desired. However, an overlap of the optical mode with contacting layers results in losses which lead to further undesirable heating of the device [6]. Furthermore, the improved strain only leads to an improved modal gain when the optical mode and large strain regions coincide, a goal we previously phrased as “overlap engineering” [13]. Summarizing these facts, we arrive at the following:

Goal: *Determine a device topology, which simultaneously ensures that the support of the first fundamental mode, i.e., the optical cavity, is confined inside the germanium and that the strain is maximized within the optical cavity.*

Since the proposed device is static, it is only possible to create a permanent or “built-in” strain field through the shape and topology of the device, where each of the materials supplies a certain amount of strain due to the relaxation process following the manufacturing process. In other words, we need to find an optimal composition and placement of the various necessary materials in order to construct a Ge-on-Si laser. Some ideas for the optimization based on existing empirical, experimental, and analytical studies can be found, e.g., in [14, 3, 12, 13, 15]. We also mention our recent related work [7], in which the optical cavity is assumed to be fixed. The underlying modeling assumption in all of these studies is the usage of a so-called “microbridge” geometry, cf. Figure 1, which in principle can be created by standard manufacturing techniques. As in [13, 7], we again focus on a cross section of an edge-emitter as shown in Figure 1. In the longitudinal direction we assume translation invariance, as it is indicated in Figure 1. We note here that multimaterial and multidisciplinary topology optimization approaches that take into consideration thermoelastic or piezoelectric properties and their relation to the underlying topologies have been considered in many works, see e.g., [16, 17]. However, as we will see below in the modeling section, these are fundamentally different applications with distinct goals.

The rest of the paper is organized as follows. In Section 2, we motivate the usage of a phase-field approach for the topology optimization. In Section 3, we introduce the underlying multiphysics model that appears in the topology optimization problem. In addition, we briefly detail the time-dependent drift-diffusion system, which models the transport of electrons and holes in the device. Afterwards, we introduce the optimization framework in Section 4, which includes a rigorous analysis of the topology-to-eigenmode mapping in Section 4.3. In Section 5, we discuss the numerical solution and necessary structural assumptions. Based on our theoretical results, we provide numerical optimization results in Section 6, which yield an optimal configuration of materials in the microbridge. Using the optimal configuration, we demonstrate the electronic and optical properties of such a design in Section 7.

Finally, our notation is more or less standard for PDE-constrained and topology optimization. Nevertheless, we refer the reader to the well-known monographs [18] for Lebesgue and Sobolev spaces, [19, 20, 21] for PDE-constrained optimization, and [22, 23, 24, 25] for a thorough treatment of topology optimization.

2 A Phase-Field Approach for the Design Parameters

Throughout the entire text, all functions are assumed to be defined on a fixed hold-all-type domain Ω, which represents the cross section of the microbridge. It is assumed that Ω has a sufficiently smooth boundary $\partial\Omega$. Furthermore, $\{\Omega_i\}_{i=1}^N$ denotes a "regular" material distribution, which partitions the domain Ω. Ideally, the partition would be represented by distributed parameters $\{\varphi_i\}_{i=1}^N$, where φ_i serves as the characteristic function for Ω_i. In such a case, we could take $\boldsymbol{\varphi} := (\varphi_1, \ldots, \varphi_N) \in BV(\Omega; \{0,1\}^N)$ (a vector of functions of bounded variation (BV) taking discrete values in $\{0,1\}$) along with the condition that $\sum_{i=1}^N \varphi_i = 1$ for almost every (a.e.) $x \in \Omega$. In order to ensure that the sets $\Omega_i := \{\varphi_i = 1\}$ have finite perimeter $P(\Omega_i, \Omega)$, which is needed to rule out pathological designs and facilitate the mathematical treatment, it suffices that the total variation term $\sum_{i=1}^N TV(\varphi_i, \Omega)$ is finite. In fact, the latter guarantees, by the Fleming-Rishel co-area formula, that $\sum_{i=1}^N P(\Omega_i, \Omega) = \sum_{i=1}^N TV(\varphi_i, \Omega) < +\infty$.

Finding a material distribution for optimizing the device topology, as stated in our goal in Section 1, would lead to a combinatorial problem, which would be computationally intractable. As a remedy, one could relax the integrality condition on each φ_i, as in [26], and attempt to regain the integrality through other means. Such an approach typically depends on structural assumptions. In this paper, as in [7, 27, 28, 29, 30], we use a phase-field approach in which $\boldsymbol{\varphi} \in H^1(\Omega; \mathbb{R}^N) \subset BV(\Omega, \mathbb{R}^N)$. Here, $H^1(\Omega, \mathbb{R}^N)$ is the space of all vector-fields in $\mathbb{R}^N$ with components in the Sobolev space $H^1(\Omega)$, see e.g., [18]. Enforcing approximate integrality of $\boldsymbol{\varphi}$ can then be achieved by considering the following Ginzburg-Landau-type energy functional

$$f_{\mathrm{GL}}(\boldsymbol{\varphi}, \varepsilon) := \int_\Omega \frac{\varepsilon}{2} \nabla\boldsymbol{\varphi} : \nabla\boldsymbol{\varphi} + \frac{1}{2\varepsilon} \boldsymbol{\varphi} \cdot (1 - \boldsymbol{\varphi}) \, \mathrm{d}\mathbf{x} + i_{\mathcal{G}}(\boldsymbol{\varphi}), \tag{1}$$

in the associated topology optimization problem. Here, $i_{\mathcal{G}}$ is the usual indicator functional for the well-known Gibbs simplex (a closed convex set)

$$\mathcal{G} := \left\{ \boldsymbol{\varphi} \in H^1(\Omega; \mathbb{R}^N) \mid \boldsymbol{\varphi} \geq 0, \text{ a.e. } x \in \Omega, \ \varphi_1 + \cdots + \varphi_N = 1, \text{ a.e. in } \Omega \right\}. \tag{2}$$

Note that the non-convex integrand $\frac{1}{2\varepsilon} \boldsymbol{\varphi} \cdot (1 - \boldsymbol{\varphi})$ attempts to force pure phases, whereas the first part in (1) introduces H^1-regularity of $\boldsymbol{\varphi}$ into the problem.

Finally, we note that as $\varepsilon \to 0$, the Ginzburg-Landau energy $f_{\mathrm{GL}}(\cdot, \varepsilon)$ Γ-converges to a set-functional that is related to the phenomenologically derived regularization term $\sum_{i=1}^N TV(\varphi_i, \Omega)$ suggested above, with some modifications. This can be shown by modifying and combining several arguments from [31] and [32]. However, a detailed discussion would go beyond the scope of this paper.

3 The Underlying Multiphysics Problem

In this section, we introduce the underlying multiphysics (forward) problem as well as the drift-diffusion system, which describes the electronic behavior of the device. The forward problem is a nonlinearly coupled system of linear partial-differential equations, which comprises two kinds of physics: elasticity and optics. The elastic properties follow a standard model of linear elasticity that takes into account eigenstrain and thermal pre-stress terms. This represents our mathematical description for the interfacial residual forces discussed in the introduction.

The model is dependent on the distributed material parameter $\boldsymbol{\varphi}$, with components $\varphi_1, \ldots, \varphi_N$ or sometimes φ_{SiN}, φ_{Ge}, φ_{SiO_2}, φ_{air} for emphasis on the actual material components and their effects on the device. These distributed parameters will act as the design/decision variables in our optimization framework. Since it has the largest effect on the lasing properties of the Ge-on-Si microbridge, we focus on optimizing the first/fundamental (eigen)mode of the device. This is done by using a Helmholtz equation, which depends on the material parameters $\boldsymbol{\varphi}$. We provide further physical and mathematical motivations for these models in the subsections below.

3.1 Elasticity

Given $\boldsymbol{\varphi} \in \mathcal{G}$, we consider the following model of elasticity, where the solution is a displacement mapping $\mathbf{u} : \Omega \to \mathbb{R}^2$:

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

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

$$F(\boldsymbol{\varphi}) := e_0(\varphi_{\mathsf{SiO}_2} - 1)\mathbb{C}(\boldsymbol{\varphi})I_{\mathbb{R}^{2\times 2}} - \sigma_0\varphi_{\mathsf{SiN}}I_{\mathbb{R}^{2\times 2}}, \tag{3}$$

incorporates the effect of the eigenstrain generated by thermal relaxation of Ge on SiO_2 and the (pre)stress generated by SiN as discussed in the introduction. This exhibits a slight change in the form of F when compared to the operator in [7]. In fact, (E($\boldsymbol{\varphi}$)) corresponds to solving

$$\begin{aligned} -\operatorname{div}\left[\mathbb{C}(\boldsymbol{\varphi})e(\boldsymbol{u}) - F(\boldsymbol{\varphi}) - e_0\mathbb{C}(\boldsymbol{\varphi})I_{\mathbb{R}^{2\times 2}}\right] &= 0 \qquad \text{in} \quad \Omega, \\ \boldsymbol{u} &= \boldsymbol{g} \qquad \text{on} \quad \partial\Omega, \end{aligned} \tag{E$'$}$$

where $\boldsymbol{g} := e_0\mathbf{x}$ for all $\mathbf{x} \in \Omega$. Indeed, solving first (E($\boldsymbol{\varphi}$)) for $\boldsymbol{u}$, one readily checks that $\boldsymbol{u} + \boldsymbol{g}$ solves (E′). The choice of the particular F is aimed at driving the Ge lattice constant into a tensile region. For small strains from (E($\boldsymbol{\varphi}$)), this lattice constant can be defined by $a(\mathbf{x}) = a_{\mathsf{bulk}}(1 + e(\boldsymbol{u})(\mathbf{x}) - e_0)$, where a_{bulk} is the lattice constant of unstrained Ge and $a(\mathbf{x}) > a_{\mathsf{bulk}}$ is desired, cf. [33]. The Dirichlet boundary condition implies that the device remains fixed and relaxes at $\partial\Omega$.

We invoke the following smoothness and ellipticity assumptions throughout:

Assumption (A1). $\mathbb{C}$ is a Nemytskii/superposition operator induced by a tensor-valued mapping $\widehat{\mathbb{C}} : \mathbb{R}^N \to \mathbb{R}^{2\times2\times2\times2}$ such that for some $\boldsymbol{\varphi}$, $\mathbb{C}(\boldsymbol{\varphi})(x) = \widehat{\mathbb{C}}(\boldsymbol{\varphi}(x))$ a.e. on Ω. Moreover, it satisfies: (i) There exist $c_2 > c_1 > 0$ such that for every $\phi \in \mathbb{R}^N$ and $E_1, E_2 \in \mathbb{R}^{2\times2} \setminus \{0\}$ we have

$$c_1\|E_1\|^2_{\mathbb{R}^{2\times2}} \leq \widehat{\mathbb{C}}(\phi)E_1 : E_1, \qquad \widehat{\mathbb{C}}(\phi)E_1 : E_2 \leq c_2\|E_1\|_{\mathbb{R}^{2\times2}}\|E_2\|_{\mathbb{R}^{2\times2}},$$

where the matrix product is understood as $A : B = \sum_i \sum_j a_{ij}b_{ij}$. (ii) $\widehat{\mathbb{C}}$ is globally Lipschitz and continuously differentiable with globally Lipschitz derivative.

Consequently, we have the following regularity and sensitivity result for the topology-to-displacement map $S_u(\boldsymbol{\varphi})$. This is essential for the proof of existence of an optimal solution and derivation of optimality conditions for the associated topology optimization problem. In addition, it is needed for the development of gradient-based numerical optimization methods.

Proposition 3.2 (cf. [7])**.** *Let (A1) hold. Then there exists $p > 2$ such that for every $\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$ the unique solution $\boldsymbol{u}$ of* (E($\boldsymbol{\varphi}$)) *lies in $W_0^{1,p}(\Omega, \mathbb{R}^2)$. Finally, the solution mapping $S_u : H^1(\Omega, \mathbb{R}^N) \to W_0^{1,p}(\Omega, \mathbb{R}^2)$, which maps $\boldsymbol{\varphi} \mapsto \boldsymbol{u}$, is continuously Fréchet differentiable. The directional derivative of S_u at $\boldsymbol{\varphi}$ in direction $\delta\boldsymbol{\varphi}$ is given by $S_u'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi} = \boldsymbol{q}$, where $\boldsymbol{q} \in H_0^1(\Omega, \mathbb{R}^2)$ is a weak solution of the sensitivity equation:*

$$\int_\Omega \mathbb{C}(\boldsymbol{\varphi})e(\boldsymbol{q}) : e(\boldsymbol{v})\mathrm{d}\mathbf{x} = -\int_\Omega [\mathbb{C}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}]e(\boldsymbol{u}) : e(\boldsymbol{v})\mathrm{d}\mathbf{x} + \int_\Omega F'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi} : e(\boldsymbol{v})\mathrm{d}\mathbf{x} \tag{4}$$

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

Here, $W_0^{1,p}(\Omega, \mathbb{R}^2)$ is the Sobolev space of two-dimensional vector fields with components in $W_0^{1,p}(\Omega)$. We utilize this convention throughout the text. In addition, we note that $p > 2$ in Proposition 3.2 ensures via the Sobolev embedding theorem that $\boldsymbol{u}$ is a continuous vector field over $\overline{\Omega}$. This is useful in the existence proof below.

3.3 Optics

As stated above, we focus our attention on finding a topology that confines the bulk of the support of the fundamental mode within the Ge. In this sense, we assume that the governing optical behavior of the device can be modeled by the following $\boldsymbol{\varphi}$-dependent eigenvalue problem in (Θ, λ):

$$\begin{aligned} -\Delta\Theta - g(\boldsymbol{\varphi})\Theta &= \lambda\Theta \quad &&\text{in} \quad \Omega, \\ \Theta &= 0 &&\text{on} \quad \partial\Omega. \end{aligned} \tag{5}$$

Here, we assume that the eigenfunction decays exponentially fast approaching the boundary, so that the homogeneous Dirichlet boundary condition on the outer boundary $\partial\Omega$ does not influence (Θ, λ) significantly. This is justified for certain $g(\boldsymbol{\varphi})$ and the eigenmode corresponding to the smallest eigenvalue and certain, the latter often being the most relevant mode for an edge-emitting laser. This is why we focus our discussion primarily on λ_1, the smallest eigenvalue of $-[\Delta + g(\boldsymbol{\varphi})]$, and Θ_1, the corresponding eigenfunction, with $(\Theta_1, \Theta_1) = 1$. This leads to the following problem:

Find the first eigenvalue λ_1 and corresponding eigenfunction Θ_1 of (5). (H($\boldsymbol{\varphi}$))

We henceforth drop the subscripts, whenever it is clear in context. Note that Ω needs to be connected to ensure that λ_1 has multiplicity one, see [34, Remark 1.2.4].

For this model, we make the standing assumptions throughout:

Assumption (A2). g is a superposition operator induced by $\hat{g} : \mathbb{R}^N \to \mathbb{R}$ such that $\hat{g}(\boldsymbol{\varphi}(x)) = (g(\boldsymbol{\varphi}))(x)$ a.e. on Ω. Moreover, $|\hat{g}|$ is bounded by M, $\hat{g}$ is globally Lipschitz with modulus $L > 0$ and continuously differentiable with globally Lipschitz derivative.

Here, we note that $g(\boldsymbol{\varphi})$ is spatially dependent. Thus, the spectrum of $-[\Delta + g(\boldsymbol{\varphi})]$ is not merely the shifted spectrum of the Laplacian. Nevertheless, $|g(\boldsymbol{\varphi})|$ is uniformly bounded, independently of $\boldsymbol{\varphi}$. Consequently, we may take some fixed $c > M$ and consider the equivalent problem

$$\begin{aligned} -\Delta\Theta + (c - g(\boldsymbol{\varphi}))\Theta &= (c + \lambda)\Theta \quad &&\text{in} \quad \Omega, \\ \Theta &= 0 &&\text{on} \quad \partial\Omega. \end{aligned} \tag{H$_c$}$$

Indeed, the operators $-[\Delta + g(\boldsymbol{\varphi})]$ and $-[\Delta + (g(\boldsymbol{\varphi}) - c)]$ have the same eigenfunctions corresponding to the same eigenvalues shifted by c. Therefore, we may work with the uniformly elliptic bounded linear operator $-[\Delta + (g(\boldsymbol{\varphi}) - c)]$, which allows us to apply elliptic theory. We postpone the sensitivity analysis of the topology-to-eigenmode mapping $\boldsymbol{\varphi} \mapsto \Theta$, denoted by $S_\Theta(\boldsymbol{\varphi})$, until after we state the optimization problem.

3.4 Electronics

In this section, we give a model for the electronic behavior of a given device design. Though we do not consider this as a part of the optimization procedure itself, we provide simulations demonstrating the performance of the optimal designs at the end of this paper and compare them to existing results in the literature. The following drift-diffusion system forms the so-called van Roosbroeck system

$$-\mathrm{div}\,(\varepsilon_0\varepsilon_{\mathsf{r}}\nabla\phi) = q(C_{\mathsf{dop}} + p - n), \tag{6a}$$

$$\dot{n} - q^{-1}\mathrm{div}\,(-q\mu_n n\nabla\phi + qD_n\nabla n) = -R_{\mathsf{net}}, \tag{6b}$$

$$\dot{p} + q^{-1}\mathrm{div}\,(-q\mu_n n\nabla\phi - qD_p\nabla p) = -R_{\mathsf{net}}, \tag{6c}$$

which was introduced for semiconductors in [35] and under several assumptions derived in this form in [36]. Here, ϕ is the electrostatic potential, ε_0 is the vacuum permittivity and ε_{r} is the relative permittivity, n and p are the concentration of electrons and holes, q is the elementary charge, C_{dop} the doping profile. The expression under the divergence are the electron and hole fluxes, where D_n, D_p denote the diffusion constant of electrons and holes and μ_n, μ_p are the corresponding mobilities. Both quantities are related by a generalized Einstein relation, which is $D_\alpha = \mu_\alpha k_B T/q$ for Boltzmann statistics. The remaining function R_{net} is the generation-recombination rate, which vanishes in thermal equilibrium and ensures conservation of charge.

An important feature here is the assumption that the strain of the device plays a role in the model due to its occurrence in the electronic bands and in the optical gain and in electronic recombination rates. We discuss this in more detail along with the numerical experiments sections below.

4 The Optimization Framework

The purpose of this section is to derive an optimization problem for identifying the optimal material distribution for the device as described in Section 1. We start by introducing the objective function. This is followed by a sensitivity study. Finally, we prove existence of a solution and derive first-order optimality conditions.

4.1 Objective Function

Our task is now to identify objective functions that quantify our goal of finding a material distribution, which maximizes the tensile strain inside the optical cavity. In contrast to [7], we do not consider the optical cavity to be fixed. Instead, we assume that the optical cavity is explicitly determined by $\boldsymbol{\varphi}$.

Before providing the mathematical details of the objective function, we further describe the physical motivations leading to its form. When optimizing a laser, the key quantity of interest is the optical gain g, which itself depends on carrier concentrations and photon energy. For an indirect band-gap material such as Ge, the rate of stimulated emission encoded in g is naturally very low and strongly depends on the size of the direct band-gap. Due to the particular band-structure of Ge, tensile strains of $1\text{-}2\,\%$ are sufficient to turn germanium into a direct band-gap material and drastically enhance stimulated emission [37]. Nevertheless, it is believed that much lower strains are sufficient to build a functioning laser, see e.g. [12].

Following this discussion, the main parameter influencing g is the band-gap E_{g}, which itself is a function of the strain e. Using the deformation potential $\mathcal{D}$, see e.g. [38], results in the relation $E_{\text{g}}(e) = E_{\text{g,0}} + \mathcal{D} : e(\mathbf{u})$ Since we do not use information about carrier concentrations or photon energy, we merely exploit the fact that the optical gain in the Ge increases with a decreasing gap and that in Ge this gap decreases with increasing tensile strain. This motivates our approach to minimize the functional

$$-\int_{\Omega} \varphi_{\text{Ge}} \Theta^2 \mathcal{D} : e(\mathbf{u}) \mathrm{d}\mathbf{x} = -\int_{\Omega} j(\boldsymbol{\varphi}, \Theta) \mathrm{tr}\, e(\mathbf{u})\, \mathrm{d}\mathbf{x}$$

where for the moment we assumed that the deformation potential is diagonal, i.e., $\mathcal{D} = D\mathbb{I}_{2\times 2}$ and in our case we have $j(\boldsymbol{\varphi}, \Theta) = \varphi_{\text{Ge}}\Theta^2 D$. Note that $\mathcal{D}$ contains material parameters. However, since $\Theta^2 D$ is scaled by φ_{Ge}, which is a relatively smooth approximation of the indicator function for the subset of Ω corresponding to the Ge concentration, we need only consider the values of D for Ge.

As observed in [13] and discussed in the introduction, we wish to maximize the region of overlap corresponding to the bulk of support for Θ^2 and the region of high tensile strain in Ge. Therefore, at an optimal configuration, we expect the bilinear relationship between $\varphi_{\mathrm{Ge}}\Theta^2$, which is non-negative, and $\mathrm{tr}\; e(\boldsymbol{u})$ to favor large overlap of $\mathrm{supp}\, \varphi_{\mathrm{Ge}}$ and $\mathrm{supp}\, \Theta^2$ along with deformations for which $\mathrm{tr}\, e(\boldsymbol{u})$ is positive on average on $(\mathrm{supp}\, \varphi_{\mathrm{Ge}}) \cap (\mathrm{supp}\, \Theta^2)$. We henceforth denote the objective by

$$J(\boldsymbol{\varphi}, \boldsymbol{u}, \Theta) := -\int_{\Omega} j(\boldsymbol{\varphi}, \Theta) \mathrm{tr}\, e(\boldsymbol{u}) \mathrm{d}\mathbf{x}. \tag{7}$$

For the sake of generality, we allow j to belong to a wide class of functions and make the following assumption:

Assumption (A3). j is a superposition operator induced by a polynomial function $\hat{j} : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ such that $\hat{j}(\boldsymbol{\varphi}(x), \Theta(x)) = (j(\boldsymbol{\varphi}, \Theta))(x)$ a.e. on Ω.

One motivation for admitting higher-order polynomials for $\hat{j}$ is related to the fact that regions where Θ is large are of particular interest and can be emphasized by allowing for exponents significantly larger than 2 in Θ.

4.2 The Optimization Problem

Combining the objectives, constraints, and forward problems from the discussions above, we arrive at the following “full space” formulation of the optimization problem.

$$\begin{aligned} &\min\; -\int_{\Omega} j(\boldsymbol{\varphi}, \Theta) \mathrm{tr}\, e(\boldsymbol{u}) \mathrm{d}\mathbf{x} + \alpha f_{GL}(\boldsymbol{\varphi}, \varepsilon) \text{ over } (\boldsymbol{\varphi}, \boldsymbol{u}, \Theta, \lambda) \in \mathcal{X}, \\ &\quad \text{s.t.} \quad \boldsymbol{u} \text{ solves (E}(\boldsymbol{\varphi})\text{)};\; (\Theta, \lambda) \text{ solves (H}(\boldsymbol{\varphi})\text{)} : (\Theta, \Theta) = 1. \end{aligned} \tag{8}$$

Here, $\alpha > 0$ is a regularization parameter, the space $\mathcal{X}$ represents the Cartesian product $\mathcal{X} := \mathcal{G}_{ad} \times H^1_0(\Omega;\mathbb{R}^2) \times H^1_0(\Omega) \times \mathbb{R}$, and $\mathcal{G}_{ad} := \{\boldsymbol{\varphi} \in \mathcal{G} |\; \varphi_i = 1 \text{ a.e. on } \Pi_i,\; i = 1, \ldots, N\}$ combines the Gibbs simplex (2) and the requirement that material i must be present on $\Pi_i \subset \Omega$. We henceforth impose the following assumptions:

Assumption (A4). $\Omega \subset \mathbb{R}^2$ and $\Pi_i \subset \Omega$ are open, connected and bounded sets with Lipschitz boundary and Π_i are strictly separable, i.e., $\operatorname{cl}\Pi_i \cap \operatorname{cl}\Pi_j = \emptyset$ $(i \neq j)$.

In the next section, we provide a full sensitivity analysis of the parameter-to-state maps, which then motivates our subsequent reformulation of (8) in "reduced space".

4.3 The Topology-to-Eigenmode Mapping S_Θ

In this section, we perform a sensitivity analysis for the Helmholtz equation (H($\boldsymbol{\varphi}$)). The Lipschitz continuity derived in Lemma 4.4 is necessary for the existence result in Proposition 4.8, whereas the differentiability result in Theorem 4.5 is needed for the first-order necessary optimality conditions in Theorem 4.10. The latter are subsequently used for numerical experiments. Obviously any results providing explicit derivative formulae are ultimately useful in adjoint-based solution algorithms.

Though it is possible that larger eigenvalues and eigenfunctions may also be of interest, the nontrivial multiplicity of even the second eigenvalue vastly complicates any differential sensitivity analysis. For higher eigenvalues some path selections as in [39] are necessary. Even with this choice there are still some challenges. For example, it is not possible to write (5) or its equivalent formulation (10) (below) as an equation (H($\boldsymbol{\varphi}$)) because the former allows for all eigenvalues. Later in the proof of Theorem 4.5 we show that this is possible at least locally around the principal eigenvalue.

We recall that due to $|g(\boldsymbol{\varphi})(x)| \leq M$ for a.e. $x \in \Omega$ independent of $\boldsymbol{\varphi}$, it is possible to shift the operators and obtain a simpler but equivalent eigenvalue problem. Choosing $c > M$, we make the operator on the left-hand side of (9) below elliptic.

$$(-\Delta - g(\boldsymbol{\varphi}) + cI)\Theta = (\lambda + c)\Theta, \tag{9}$$

It then readily follows from [40, Theorem 8.6.1, Remark 8.6.1] that all eigenvalues of $[-\Delta - [g(\boldsymbol{\varphi}) + c]]$ are real and that λ_1 may be computed as the Lagrange multiplier for the normalization constraint in the (nonconvex) Courant-Fisher optimization problem

$$\min\left\{(\nabla\Theta, \nabla\Theta) - (g(\boldsymbol{\varphi})\Theta, \Theta) \text{ over } \Theta \in H_0^1(\Omega) \;\mid (\Theta, \Theta) = 1\right\} \tag{10}$$

Here and below, $(\cdot, \cdot)$ represents the usual $L^2(\Omega)$-inner product. Moreover, the above problem admits an optimal solution and all minimizers are the eigenfunctions corresponding to the smallest eigenvalue.

Before showing the first result, we comment on some direct consequences of assumption (A2). First, the smallest eigenvalue in (H($\boldsymbol{\varphi}$)) has multiplicity one and the corresponding eigenfunction can be chosen to be positive almost everywhere, see [41, Theorem 8.38] or [34, Theorem 1.2.5]. Assumption (A2) implies that $g : L^p(\Omega, \mathbb{R}^N) \to L^p(\Omega)$ is globally Lipschitz with modulus L and $g : L^{2p}(\Omega, \mathbb{R}^N) \to L^p(\Omega)$ is continuously differentiable with global Lipschitz derivative for all $p \in [1, \infty]$, see [42]. Moreover, $\|g(\boldsymbol{\varphi})\|_{L^\infty(\Omega)} \leq M$ for all $\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$.

For notational simplicity, we define the solution mappings $S_u : \boldsymbol{\varphi} \mapsto \boldsymbol{u}$, $S_\lambda : \boldsymbol{\varphi} \mapsto \lambda$ and $S_\Theta : \boldsymbol{\varphi} \mapsto \Theta$ as solutions to (E($\boldsymbol{\varphi}$)) and (H($\boldsymbol{\varphi}$)), respectively. We start with derivation of the Lipschitz continuity of S_λ.

Lemma 4.4. *Assume (A2) and (A4). Then the following holds true:*

(i) There exists $\tilde{M} > 0$ such that for all $\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$ and the corresponding eigenfunction we have $\|S_\Theta(\boldsymbol{\varphi})\|_{H_0^1(\Omega)} \leq \tilde{M}$.

(ii) *The mapping S_λ is globally Lipschitz from $L^\infty(\Omega,\mathbb{R}^N) \to \mathbb{R}$ with modulus L and globally Lipschitz from $L^2(\Omega,\mathbb{R}^N) \to \mathbb{R}$.*

Proof. Let Θ_0 by feasible for (10). Then for any $\boldsymbol{\varphi} \in H^1(\Omega,\mathbb{R}^N)$ and $\Theta = S_\Theta(\boldsymbol{\varphi}) \in H_0^1(\Omega)$ we have $(\nabla\Theta,\nabla\Theta) - (g(\boldsymbol{\varphi})\Theta,\Theta) \leq (\nabla\Theta_0,\nabla\Theta_0) - (g(\boldsymbol{\varphi})\Theta_0,\Theta_0)$ which due to the normalization condition implies

$$\|\Theta\|^2_{H_0^1(\Omega)} \leq (2M + \|\Theta_0\|^2_{H_0^1(\Omega)}) =: \tilde{M}^2, \tag{11}$$

This yields (i). Next, fix $\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}} \in H^1(\Omega,\mathbb{R}^N)$ and let $\Theta, \hat{\Theta} \in H_0^1(\Omega)$ be the corresponding eigenfunctions. Since $\Theta, \hat{\Theta}$ are minimizers of (10), we have

$$\begin{aligned}(\nabla\Theta,\nabla\Theta) - (g(\boldsymbol{\varphi})\Theta,\Theta) &\leq (\nabla\hat{\Theta},\nabla\hat{\Theta}) - (g(\hat{\boldsymbol{\varphi}})\hat{\Theta},\hat{\Theta}) + ((g(\hat{\boldsymbol{\varphi}}) - g(\boldsymbol{\varphi}))\hat{\Theta},\hat{\Theta}) \\ &\leq (\nabla\hat{\Theta},\nabla\hat{\Theta}) - (g(\hat{\boldsymbol{\varphi}})\hat{\Theta},\hat{\Theta}) + \|g(\hat{\boldsymbol{\varphi}}) - g(\boldsymbol{\varphi})\|_{L^2}\|\hat{\Theta}\|^2_{L^4} \\ &\leq (\nabla\hat{\Theta},\nabla\hat{\Theta}) - (g(\hat{\boldsymbol{\varphi}})\hat{\Theta},\hat{\Theta}) + \tilde{L}\|\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_{L^2(\Omega,\mathbb{R}^N)}.\end{aligned} \tag{12}$$

Here, $\tilde{L}$ combines the Lipschitz modulus L, the embedding constant from $H^1(\Omega)$ into $L^4(\Omega)$ and $\tilde{M}$. Since we may switch the roles of $\boldsymbol{\varphi}$ and $\hat{\boldsymbol{\varphi}}$, we obtain

$$|(\nabla\Theta,\nabla\Theta) - (g(\boldsymbol{\varphi})\Theta,\Theta) - (\nabla\hat{\Theta},\nabla\hat{\Theta}) + (g(\hat{\boldsymbol{\varphi}})\hat{\Theta},\hat{\Theta})| \leq \tilde{L}\|\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_{L^2(\Omega,\mathbb{R}^N)}. \tag{13}$$

As the eigenvalue is the Lagrange multiplier associated with the constraint in (10), we obtain from the first-order optimality conditions for (10):

$$\begin{aligned}(\nabla\Theta,\nabla\Theta) - (g(\boldsymbol{\varphi})\Theta,\Theta) &= \lambda(\Theta,\Theta) = \lambda, \\ (\nabla\hat{\Theta},\nabla\hat{\Theta}) - (g(\hat{\boldsymbol{\varphi}})\hat{\Theta},\hat{\Theta}) &= \hat{\lambda}(\hat{\Theta},\hat{\Theta}) = \hat{\lambda},\end{aligned}$$

where λ and $\hat{\lambda}$ are the corresponding eigenvalues. Plugging this into (13), we see that $|\lambda - \hat{\lambda}| \leq \tilde{L}\|\boldsymbol{\varphi} - \hat{\boldsymbol{\varphi}}\|_{L^2(\Omega,\mathbb{R}^N)}$, which proves the second statement in (ii). The proof of the first statement is analogous and can be obtained from (12) using the upper bound:

$$\begin{aligned}((g(\hat{\boldsymbol{\varphi}}) - g(\boldsymbol{\varphi}))\hat{\Theta},\hat{\Theta}) \leq \|g(\hat{\boldsymbol{\varphi}}) - g(\boldsymbol{\varphi})\|_{L^\infty(\Omega)}\|\hat{\Theta}\|^2_{L^2} \\ = \|g(\hat{\boldsymbol{\varphi}}) - g(\boldsymbol{\varphi})\|_{L^\infty(\Omega)} \leq L\|\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_{L^\infty(\Omega,\mathbb{R}^N)}.\end{aligned}$$

□

Theorem 4.5. *Under (A2) the solution mapping $S := (S_\lambda, S_\Theta)$ is Fréchet differentiable at any $\boldsymbol{\varphi} \in H^1(\Omega,\mathbb{R}^N)$ and, given a direction $\delta\varphi \in H^1(\Omega,\mathbb{R}^N)$, its directional derivative $S'(\boldsymbol{\varphi})(\delta\boldsymbol{\varphi}) = (\delta\lambda,\delta\Theta)$ can be computed as the unique solution $(\delta\lambda,\delta\Theta) \in \mathbb{R} \times H_0^1(\Omega)$ of the system*

$$\begin{aligned}-[\Delta + g(\boldsymbol{\varphi}) + \lambda]\delta\Theta &= \delta\lambda\Theta + [g'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}]\Theta, \\ (\Theta,\delta\Theta) &= 0.\end{aligned} \tag{14}$$

Proof. Based on (5) and (H($\boldsymbol{\varphi}$)) we consider the following system of equations in strong form:

$$\begin{aligned}(-\Delta - g(\boldsymbol{\varphi}))\Theta - \lambda\Theta &= 0 \qquad \text{in} \quad \Omega, \\ \Theta &= 0 \qquad \text{on} \quad \partial\Omega, \\ (\Theta,\Theta) - 1 &= 0.\end{aligned} \tag{15}$$

Multiplying $(-\Delta - g(\boldsymbol{\varphi}))\Theta$ by $\psi \in H_0^1(\Omega)$ and integrating over Ω, it follows from Green's theorem that:

$$\int_\Omega (-\Delta - g(\boldsymbol{\varphi}))\Theta\psi \mathrm{d}\mathbf{x} = (\nabla\Theta, \nabla\psi) - (g(\boldsymbol{\varphi})\Theta, \psi).$$

Therefore, there exists a unique coercive bounded linear operator $A : H_0^1(\Omega) \to H^{-1}(\Omega)$ such that $\langle A\Theta, \psi\rangle = (\nabla\Theta, \nabla\psi)$. Nevertheless, we allow a slight abuse of notation and denote A by $-\Delta$. The boundary condition in (15) is therefore "absorbed" by the operator.

Continuing, we denote the solution mapping of (15) by $\hat{S} : \boldsymbol{\varphi} \mapsto (\lambda, \Theta)$. Note that $\hat{S}$ is in fact multivalued (for every $\boldsymbol{\varphi}$, $\hat{S}(\boldsymbol{\varphi})$ is the set of all eigenpairs). Nevertheless, since S_λ is single-valued, the Lipschitz continuity of S_λ from Lemma 4.4 implies that there exists an open ball around $\boldsymbol{\varphi}$ such that $\hat{S}$ coincides locally with S. To derive differentiability of S it suffices then to apply the implicit function theorem [43, Theorem 4.B] to (15).

Denote the function on the left-hand side of (15) by $G(\boldsymbol{\varphi}; \lambda, \Theta)$. By formally differentiating this mapping in direction $(\delta\boldsymbol{\varphi}, \delta\lambda, \delta\Theta)$, we obtain the formula:

$$G'(\boldsymbol{\varphi}, \lambda, \Theta)(\delta\boldsymbol{\varphi}, \delta\lambda, \delta\Theta) = \begin{pmatrix} -\Delta\delta\Theta - [g'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}]\Theta - g(\boldsymbol{\varphi})\delta\Theta - \delta\lambda\Theta - \lambda\delta\Theta \\ 2(\Theta, \delta\Theta) \end{pmatrix}.$$

Furthermore, by substituting this formula into the usual difference quotient, it is not difficult to verify that $G : H^1(\Omega, \mathbb{R}^N) \times \mathbb{R} \times H_0^1(\Omega) \to H^{-1}(\Omega) \times \mathbb{R}$ is in fact continuously Fréchet differentiable. Clearly, G is also continuous. Finally, we show that the partial derivative $G'_{\lambda,\Theta}(\boldsymbol{\varphi}, \lambda, \Theta)$ is bijective. We have

$$G'_{\lambda,\Theta}(\boldsymbol{\varphi}, \lambda, \Theta)(\delta\lambda, \delta\Theta) = \begin{pmatrix} -\Delta\delta\Theta - g(\boldsymbol{\varphi})\delta\Theta - \delta\lambda\Theta - \lambda\delta\Theta \\ 2(\Theta, \delta\Theta) \end{pmatrix}. \tag{16}$$

To demonstrate injectivity, we need to show that

$$\begin{aligned} -\Delta\delta\Theta - g(\boldsymbol{\varphi})\delta\Theta - \delta\lambda\Theta - \lambda\delta\Theta = 0, \\ (\Theta, \delta\Theta) = 0, \end{aligned} \tag{17}$$

admits only the trivial solution $(\delta\lambda, \delta\Theta) = (0, 0) \in \mathbb{R} \times H_0^1(\Omega)$. To this aim, suppose $(\delta\lambda, \delta\Theta)$ is some solution pair. Using Θ as a test function in the first equation in (17) we obtain

$$(\nabla\delta\Theta, \nabla\Theta) - (g(\boldsymbol{\varphi})\delta\Theta, \Theta) - \delta\lambda(\Theta, \Theta) - \lambda(\delta\Theta, \Theta) = 0. \tag{18}$$

Realizing that $(\nabla\delta\Theta, \nabla\Theta) - (g(\boldsymbol{\varphi})\delta\Theta, \Theta) - \lambda(\delta\Theta, \Theta) = 0$ due to symmetry and the definition of the eigenvalue, relation (18) reduces to $0 = \delta\lambda(\Theta, \Theta) = \delta\lambda$. Plugging this back into (17) we see that $(\lambda, \delta\Theta)$ is an eigenpair. But this implies $\delta\Theta = 0$ because the multiplicity of λ is one and $\delta\Theta$ is orthogonal to Θ. Thus, we have shown injectivity.

For surjectivity, we need to show that for any $v \in H^{-1}(\Omega)$ and $\mu \in \mathbb{R}$ system

$$\begin{aligned} -\Delta\delta\Theta - g(\boldsymbol{\varphi})\delta\Theta - \delta\lambda\Theta - \lambda\delta\Theta = v \\ (\Theta, \delta\Theta) = \mu \end{aligned} \tag{19}$$

has a solution $(\delta\lambda, \delta\Theta)$. In what follows, we will construct a solution pair $(\delta\Theta, \delta\lambda)$ associated with (v, μ). We use aspects of the proof of [44, Section 6.2, Theorem 4]. Fix some $\gamma > M + \lambda$ and define the mappping $\mathcal{L}_\gamma := -\Delta - g(\boldsymbol{\varphi}) - \lambda I + \gamma I$. Since $\gamma > M + \lambda$, the operator $\mathcal{L}_\gamma$ is $H_0^1(\Omega)$-coercive, bounded, and linear. In what follows, we let $\mathcal{L} := \mathcal{L}_0$. Hence, $\mathcal{L}_\gamma^{-1}$ exists. Moreover, since the canonical embedding $E_{1,-1}$ of $H_0^1(\Omega)$ into $H^{-1}(\Omega)$ is compact, the operator $K := (E_{1,-1} \circ \mathcal{L}_\gamma^{-1})$ is a compact linear operator from $H^{-1}(\Omega)$ into itself.

Note that for the sake of making the compact embedding of $H_0^1(\Omega)$ into $H^{-1}(\Omega)$ explicit, we include the embedding operator $E_{1,-1}$. However, we have left this out of the notation on many other occasions for the sake of readability, e.g. in the definition of $\mathcal{L}_\gamma$.

The dual operator of K, denoted by K', is given by $K' = \mathcal{L}_\gamma^{-1} E_{1,-1}$. This is a mapping from $H_0^1(\Omega)$ into itself. The latter follows from the fact that $E_{1,-1} : H_0^1(\Omega) \to H^{-1}(\Omega)$ is defined by $E_{1,-1}\psi = (\psi, \cdot)_{L^2}$, where $\psi \in H_0^1(\Omega)$. Therefore, for any $\xi \in H_0^1(\Omega)$, we have $\langle E_{1,-1}\psi, \xi\rangle = (\psi, \xi)_{L^2} = \langle \psi, E_{1,-1}\xi\rangle$. Hence, $E_{1,-1}$ coincides with its dual operator. Similarly, for some $h \in H^{-1}(\Omega)$, there exists a unique $z_h := \mathcal{L}_\gamma^{-1} h \in H_0^1(\Omega)$. Then given an arbitrary $k \in H^{-1}(\Omega)$ we have

$$\langle \mathcal{L}_\gamma^{-1} h, k\rangle = \langle z_h, k\rangle = \langle z_h, \mathcal{L}_\gamma \mathcal{L}_\gamma^{-1} k\rangle = \langle \mathcal{L}_\gamma' z_h, \mathcal{L}_\gamma^{-1} k\rangle = \langle \mathcal{L}_\gamma z_h, \mathcal{L}_\gamma^{-1} k\rangle = \langle h, \mathcal{L}_\gamma^{-1} k\rangle.$$

The second-to-last equality follows from the specific form of $\mathcal{L}_\gamma$. Hence, $\mathcal{L}_\gamma^{-1}$ also coincides with its dual operator.

Next, using $R : H_0^1(\Omega) \to H^{-1}(\Omega)$ with $R = -\Delta$ as the Riesz isometry, we define the adjoint $K^* : H^{-1}(\Omega) \to H^{-1}(\Omega)$ of K by $K^* = RK'R^{-1} = -\Delta \mathcal{L}_\gamma^{-1} E_{1,-1}(-\Delta)^{-1}$. In addition, we observe that $K'\Theta = \gamma^{-1}\Theta$, since $z = K'\Theta = \mathcal{L}_\gamma^{-1} E_{1,-1}\Theta$ means

$$\mathcal{L}_\gamma z = E_{1,-1}\Theta \Leftrightarrow [\mathcal{L} + \gamma]z = E_{1,-1}\Theta \Rightarrow z = \gamma^{-1}\Theta. \tag{20}$$

This property carries over to the adjoint as well since $K^*R\Theta = RK'R^{-1}R\Theta = RK'\Theta = \gamma^{-1}R\Theta$, i.e., $K^*R\Theta = \gamma^{-1}R\Theta$.

Continuing, we use the Fredholm alternative, see e.g., [44, Appendix D, Theorem 5], which implies

$$\mathrm{Rng}(\gamma^{-1}I - K) = \mathrm{Ker}(\gamma^{-1}I - K^*)^\perp. \tag{21}$$

Here, I is the identity on $H^{-1}(\Omega)$ and the orthogonal complement is defined using the inner product on $H^{-1}(\Omega)$.

Next consider that for $w \in \mathrm{Ker}(\gamma^{-1}I - K^*)$, $w \in H^{-1}(\Omega)$, we have

$$\begin{aligned} \gamma^{-1}w - K^*w = 0 &\Leftrightarrow \gamma^{-1}w - RK'R^{-1}w = 0 \\ &\Leftrightarrow \gamma^{-1}R^{-1}w - \mathcal{L}_\gamma^{-1}E_{1,-1}R^{-1}w = 0 \end{aligned}$$

But then $\mathcal{L}_\gamma R^{-1}w = \gamma E_{1,-1}R^{-1}w \Rightarrow [\mathcal{L} + \gamma]R^{-1}w = \gamma E_{1,-1}R^{-1}w$, which futhermore implies, $\mathcal{L}R^{-1}w = 0 \Rightarrow R^{-1}w = \Theta$. Hence, $w = R\Theta$.

Conversely, for any $t \in \mathbb{R}$, we can show that $R\Theta \in \mathrm{Ker}(\gamma^{-1}I - K^*)$ using an analogous argument. Hence, it follows from this and (21) that

$$\mathrm{span}(R\Theta)^\perp = \mathrm{Rng}(\gamma^{-1}I - K). \tag{22}$$

In fact, for any for any $h \in \mathrm{span}(R\Theta)^\perp$ it follows from (20) that

$$(Kh, R\Theta) = (h, K^*R\Theta) = \gamma^{-1}(h, R\Theta) = 0, \tag{23}$$

where $(\cdot, \cdot)$ represents the inner product on $H^{-1}(\Omega)$. Hence,

$$Kh \in \mathrm{span}(R\Theta)^\perp, \tag{24}$$

as well. Then, taking v from (19), we observe that

$$\begin{aligned} \langle v - (v, R\Theta)_{H^{-1}} E_{1,-1}\Theta, \Theta\rangle_{H^{-1},H_0^1} &= \langle v, \Theta\rangle_{H^{-1},H_0^1} - (v, R\Theta)_{H^{-1}}(\Theta, \Theta)_{L^2} \\ &= (v, R\Theta)_{H^{-1}} - (v, R\Theta)_{H^{-1}} \cdot 1 = 0, \end{aligned}$$

where we once again make use of the Riesz representation theorem. It follows that $(v-(v,R\Theta)E_{1,-1}\Theta) \in \operatorname{span}(R\Theta)^{\perp}$. Furthermore, by (24), we also have $K(v-(v,R\Theta)E_{1,-1}\Theta) \in \operatorname{span}(R\Theta)^{\perp}$. Then by the Fredholm alternative theorem, in particular due to (22), there exists a $h \in H^{-1}(\Omega)$ such that

$$\gamma^{-1}h - Kh = K(v-(v,R\Theta)E_{1,-1}\Theta).$$

In fact, as the above equality implies $h = E_{1,-1}(\gamma\mathcal{L}_{\gamma}^{-1}(h+v-(v,R\Theta)E_{1,-1}\Theta))$ we readily infer that $h \in H_0^1(\Omega)$. Furthermore, it follows that for any $\psi \in H_0^1(\Omega)$

$$\langle \gamma^{-1}\mathcal{L}h, \psi\rangle = \langle v, \psi\rangle - \langle v, \Theta\rangle(\Theta, \psi). \tag{25}$$

Now define $\delta\Theta := \gamma^{-1}h + (\mu - (\gamma^{-1}h, \Theta))\Theta,\ \delta\lambda := -\langle v, \Theta\rangle$. Then we have for any $\psi \in H_0^1(\Omega)$ that

$$\langle \mathcal{L}\delta\Theta - \delta\lambda\Theta, \psi\rangle = \langle \gamma^{-1}\mathcal{L}h, \psi\rangle + (\mu - (\gamma^{-1}h, \Theta))\langle \mathcal{L}\Theta, \psi\rangle + \langle v, \Theta\rangle\langle \Theta, \psi\rangle = \langle v, \psi\rangle.$$

due to (25) and $\mathcal{L}\Theta = 0$ due to the definition of eigenfunction. But this means that $(\delta\Theta, \delta\lambda)$ solves the first equation in (19). Since obviously $(\Theta, \delta\Theta) = \mu$, the second equality holds true as well. Thus, we have verified the assumptions of the implicit function theorem. □

4.6 Existence of an Optimal Topology

For notational simplicity, we define here the reduced objective $\mathcal{J}$ by

$$\mathcal{J}(\boldsymbol{\varphi}) := -\int_{\Omega} j(\boldsymbol{\varphi}, S_{\Theta}(\boldsymbol{\varphi}))\operatorname{tr} e(\boldsymbol{S}_u(\boldsymbol{\varphi}))\mathrm{d}\mathbf{x}. \tag{26}$$

In order prove existence of a solution, we require the following technical lemma.

Lemma 4.7. *Under assumptions (A1)-(A4), $\mathcal{J}$ is bounded on $\mathcal{G}_{ad}$.*

Proof. From [7, Lemma 3.2] we obtain that S_u is bounded in $H_0^1(\Omega, \mathbb{R}^2)$ on $\mathcal{G}_{ad}$. Since $\Theta := S_{\Theta}(\boldsymbol{\varphi})$ solves (9), we obtain for $\lambda := S_{\lambda}(\boldsymbol{\varphi})$ the estimate

$$\|\Theta\|_{H_0^1(\Omega)} \leq C(\lambda + c)\|\Theta\|_{L^2(\Omega)} = C(\lambda + c)$$

for some $C > 0$. Then the boundedness of S_{Θ} in $H_0^1(\Omega)$ on $\mathcal{G}_{ad}$ follows. With $\boldsymbol{u} := S_u(\boldsymbol{\varphi})$ and $\Theta := S_{\Theta}(\boldsymbol{\varphi})$ we infer

$$|\mathcal{J}(\boldsymbol{\varphi})| = |J(\boldsymbol{\varphi}, \boldsymbol{u}, \Theta)| \leq \int_{\Omega} |j(\boldsymbol{\varphi}, \Theta)\operatorname{tr} e(\boldsymbol{u})|\mathrm{d}\mathbf{x} \leq \|j(\boldsymbol{\varphi}, \Theta)\|_{L^2(\Omega)}\|\boldsymbol{u}\|_{H_0^1(\Omega,\mathbb{R}^2)}.$$

By (A3)-(A4), $j(\boldsymbol{\varphi}, \Theta)$ is bounded in the $L^2(\Omega)$-norm. The assertion follows. □

Using the previous lemma, we now prove the existence result.

Proposition 4.8. *Under assumptions (A1)-(A4),* (8) *has an optimal solution.*

Proof. Due to Lemma 4.7 we may consider an infimizing sequence $\{\boldsymbol{\varphi}^k\}$ of (8). Given the form of $\mathcal{G}$, this sequence is bounded in $L^\infty(\Omega, \mathbb{R}^N)$. From Lemma 4.7 and from the form of the Ginzburg-Landau energy we see that $\boldsymbol{\varphi}^k$ is bounded in $H^1(\Omega, \mathbb{R}^N)$. This allows to select subsequences, denoted by the same indices, such that $\boldsymbol{\varphi}^k \rightharpoonup \boldsymbol{\varphi}$ in $H^1(\Omega, \mathbb{R}^N), \boldsymbol{u}^k := S_u(\boldsymbol{\varphi}^k) \rightharpoonup \boldsymbol{u}$ in $H^1_0(\Omega, \mathbb{R}^2), \Theta^k := S_\Theta(\boldsymbol{\varphi}^k) \rightharpoonup \Theta$ in $H^1_0(\Omega)$for some $\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$, $\boldsymbol{u} \in H^1_0(\Omega, \mathbb{R}^2)$ and $\Theta \in H^1_0(\Omega)$. From [7, Lemma 3.2] we obtain $\boldsymbol{\varphi} \in \mathcal{G}$ and $\boldsymbol{u} = S_u(\boldsymbol{\varphi})$. Moreover, for any $\hat{\Theta} \in H^1_0(\Omega)$ with $(\hat{\Theta}, \hat{\Theta}) = 1$ we have

$$\begin{aligned}(\nabla\Theta, \nabla\Theta) - (g(\boldsymbol{\varphi})\Theta, \Theta) &\le \operatorname{liminf}_k \left[(\nabla\Theta^k, \nabla\Theta^k) - (g(\boldsymbol{\varphi}^k)\Theta^k, \Theta^k)\right] \\ &\le \operatorname{liminf}_k \left[(\nabla\hat{\Theta}, \nabla\hat{\Theta}) - (g(\boldsymbol{\varphi}^k)\hat{\Theta}, \hat{\Theta})\right] \\ &= (\nabla\hat{\Theta}, \nabla\hat{\Theta}) - (g(\boldsymbol{\varphi})\hat{\Theta}, \hat{\Theta}),\end{aligned}$$

where in the second inequality we have used that Θ^k globally minimizes (10) for $\boldsymbol{\varphi}^k$. Since the minimizers of (10) are the vectors corresponding to the smallest eigenvalue, we have $\Theta = S_\Theta(\boldsymbol{\varphi})$. Finally, we obtain

$$\lim_k \int_\Omega j(\boldsymbol{\varphi}^k, \Theta^k)\operatorname{tr} e(\boldsymbol{S}_u(\boldsymbol{\varphi}^k)) = \int_\Omega j(\boldsymbol{\varphi}, \Theta)\operatorname{tr} e(\boldsymbol{u}), \qquad \operatorname{liminf}_k f_{GL}(\boldsymbol{\varphi}^k) \ge f_{GL}(\boldsymbol{\varphi}).$$

Since $\{\boldsymbol{\varphi}^k\}$ is a minimizing sequence and $\boldsymbol{\varphi}$ is feasible, $\boldsymbol{\varphi}$ is optimal for (8). □

Before concluding this section, we mention that it is possible to obtain variational convergence arguments for the multiphase Ginzburg-Landau-type function f_{GL} based on arguments in [31] and [32]. From a mathematical standpoint, this is necessary to show that accumulation points of ε-dependent solutions converge to solution of a related sharp-interface model. However, this would go beyond the scope of this paper and will therefore be the focus of a future study.

4.9 First-Order Optimality Conditions

We now derive first-order necessary optimality conditions. As a byproduct of this result, we obtain useful adjoint formulae for the elasticity and eigenvalue problems.

Theorem 4.10. *Assume that (A1)-(A4) are satisfied. If $\boldsymbol{\varphi}$, with the corresponding $\boldsymbol{u} = S_u(\boldsymbol{\varphi})$ and $\Theta = S_\Theta(\boldsymbol{\varphi})$, is an optimal solution to* (8)*, then the following first-order necessary optimality conditions are satisfied:*

$$\begin{aligned}&\alpha\varepsilon(\nabla\boldsymbol{\varphi}, \nabla(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})) + \frac{\alpha}{2\varepsilon}(1 - 2\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) + \int_\Omega [\mathbb{C}'(\boldsymbol{\varphi})(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})]e(\boldsymbol{u}) : e(\boldsymbol{p})\mathrm{d}\mathbf{x} \\ &- \int_\Omega F'(\boldsymbol{\varphi})(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) : e(\boldsymbol{p})\mathrm{d}\mathbf{x} - \int_\Omega [g'(\boldsymbol{\varphi})(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})]\Theta q\mathrm{d}\mathbf{x} \ge 0, \ \forall\hat{\boldsymbol{\varphi}} \in \mathcal{G}_{ad},\end{aligned} \tag{27}$$

where $\boldsymbol{p} \in H^1_0(\Omega, \mathbb{R}^2)$ is the adjoint state associated with the elasticity equation

$$-\operatorname{div} \mathbb{C}(\boldsymbol{\varphi})e(\boldsymbol{p}) = -J'_u(\boldsymbol{\varphi}, \boldsymbol{u}, \Theta) \qquad \text{in} \quad \Omega, \tag{28}$$

and $q \in H^1_0(\Omega)$ is the adjoint state associated with the Helmholtz equation

$$\begin{aligned}&-\Delta q - g(\boldsymbol{\varphi})q - \lambda q = \langle J'_\Theta(\boldsymbol{\varphi}, \boldsymbol{u}, \Theta), \Theta\rangle\Theta - J'_\Theta(\boldsymbol{\varphi}, \boldsymbol{u}, \Theta) \qquad \text{in} \quad \Omega, \\ &(q, \Theta) = 0.\end{aligned} \tag{29}$$

Proof. For the first part of the objective of (8) we have $\mathcal{J} = J_2 \circ J_1$, where $J_1 : H^1(\Omega, \mathbb{R}^N) \to L^q(\Omega, \mathbb{R}^N) \times L^2(\Omega) \times L^q(\Omega)$ and $J_2 : L^q(\Omega, \mathbb{R}^N) \times L^2(\Omega) \times L^q(\Omega) \to \mathbb{R}$ are defined by

$$J_1(\boldsymbol{\varphi}) := (\boldsymbol{\varphi}, \operatorname{tr} e(S_u(\boldsymbol{\varphi})), S_\Theta(\boldsymbol{\varphi})), \quad J_2(\boldsymbol{\varphi}, v, \Theta) := -\int_\Omega \hat{j}(\boldsymbol{\varphi}(\cdot), \Theta(\cdot))v(\cdot)\mathrm{d}\mathbf{x}.$$

Then J_1 is differentiable for all $q \in [1, \infty)$ due to Lemma 3.2 and Theorem 4.5. Since j is a polynomial due to (A3), by direct computation it can be shown that J_2 is differentiable, as well. Consequently, the reduced objective of (8) is differentiable.

By a standard technique, see e.g., [45, Section 1.6.2], we obtain

$$\mathcal{J}'(\boldsymbol{\varphi}) = J'_\varphi(\boldsymbol{\varphi}, \boldsymbol{u}) + E'_\varphi(\boldsymbol{\varphi}, \boldsymbol{u})^*\boldsymbol{p} + G'_\varphi(\boldsymbol{\varphi}, \boldsymbol{u})^*(q_\Theta, q_\lambda), \tag{30}$$

where E denotes the operator on the left-hand side of the elasticity equation (E($\boldsymbol{\varphi}$)) and G is defined as in the proof of Theorem 4.5. Here $\boldsymbol{p} \in H^1_0(\Omega, \mathbb{R}^2)$ is the solution of the adjoint equation $E'_u(\boldsymbol{\varphi}, \boldsymbol{u})^*\boldsymbol{p} = -J'_u(\boldsymbol{\varphi}, \boldsymbol{u}, \Theta)$ and similarly $(q_\lambda, q_\Theta) \in \mathbb{R} \times H^1_0(\Omega)$ solves the second adjoint equation $G'_{\lambda,\Theta}(\boldsymbol{\varphi}, \boldsymbol{u})^*\boldsymbol{p} = -J'_{\lambda,\Theta}(\boldsymbol{\varphi}, \boldsymbol{u}, \Theta)$. While the first adjoint equation amounts to (28), the second adjoint equation is given by

$$\begin{aligned} -\Delta q_\Theta - g(\boldsymbol{\varphi})q_\Theta &= \lambda q_\Theta + q_\lambda \Theta - J'_\Theta(\boldsymbol{\varphi}, \boldsymbol{u}, \Theta) \qquad \text{in} \quad \Omega, \\ (q_\Theta, \Theta) &= 0. \end{aligned} \tag{31}$$

Using Θ as a test function in the first equation, the boundary condition and along with the fact that (λ, Θ) is an eigenpair implies $q_\lambda = \langle J'_\Theta(\boldsymbol{\varphi}, \boldsymbol{u}, \Theta), \Theta\rangle$. Plugging this back to (31) and setting $q := q_\Theta$, we obtain (29). The rest of the proof follows from standard optimality theory, see e.g., [46]. □

Remark 4.11. *As in [7], we cannot guarantee the existence of Lagrange multipliers for the constraints defining $\mathcal{G}_{\mathrm{ad}}$. Consequently, we only have the variational optimality conditions as opposed to a multiplier-based system.*

5 Solution Method for the Optimization Problem

5.1 Data Assumptions

As explained in the introduction, the choice of integrand $j(\boldsymbol{\varphi}, \Theta)$ is crucial for forcing an overlap of the first eigenmode with the Ge experiencing the highest levels of strain. For our numerical experiments, we choose

$$j(\boldsymbol{\varphi}, \Theta) := \varphi_{\mathsf{Ge}}\Theta^2. \tag{32}$$

This is justified since, on the one hand, $\varphi_{\mathrm{Ge}} \in [0, 1]$ is in effect a smoothed characteristic function for the region Ω_{Ge} occupied by Ge. On the other hand, since we are optimizing for tensile strain, we expect $\operatorname{tr} e(\mathbf{u}) \geq 0$ (at least on average over the domain Ω). Therefore, a minimization procedure should force Θ^2 to be as large as possible on Ω_{Ge}. Since the eigenvalue problem includes a normalization constraint: $(\Theta, \Theta) = 1$, this ultimately confines the bulk of $\operatorname{supp}\Theta$ to Ω_{Ge}.

Concerning $\mathcal{G}_{\mathrm{ad}}$, we need to fix certain regions $\Pi_i \subset \Omega$. This is due to the fact that the actual physical values associated with SiN and Ge, see Table 1, are somewhat counterproductive to the aims of the choice of objective function. Since the strain $|e(\mathbf{u})(x)|$ is largest for $x \in \Omega_{\mathrm{SiN}}$ and $|\Theta(x)|$ is largest for $x \in \Omega_{\mathrm{Ge}}$ (followed by Ω_{SiN}), it is more advantageous to maximize the strain when Ge is omitted altogether.

There are several possibilities to circumvent this issue in the model: Prescribe a part of the domain; add a volume constraint; or add an appropriate term to the objective as in (32). Although the first possibility restricts the freedom of the design space, the latter two options are not appropriate for coarse meshes as they are bypassed whenever φ_{Ge} has fractional values. As we employ a warm-start/coarse-to-fine mesh refinement strategy in our numerical experiments, we choose the first option.

5.2 Optimization Algorithm

For the actual optimization algorithm, we use a standard projected gradient step, as in [47, 48]. Denoting the entire reduced objective by $\hat{\mathcal{J}} := \mathcal{J} + \alpha f_{GL}$, we thus obtain at each step

$$\boldsymbol{\varphi}^{k+1}(t) = \mathrm{Proj}_{\mathcal{G}}\big(\boldsymbol{\varphi}^k - tR_{\mathrm{Riesz}}^{-1}(\nabla\hat{\mathcal{J}}(\boldsymbol{\varphi}^k))\big), \quad t > 0. \tag{33}$$

Since each $\boldsymbol{\varphi}^k \in H^1(\Omega;\mathbb{R}^N)$, the gradient $\nabla\hat{\mathcal{J}}(\boldsymbol{\varphi}^k) \in (H^1(\Omega;\mathbb{R}^N))^*$, which is not identified with $H^1(\Omega;\mathbb{R}^N)$. Therefore, we need to obtain the Riesz representation $R_{\mathrm{Riesz}}^{-1}(\nabla\hat{\mathcal{J}}(\boldsymbol{\varphi}^k)) \in H^1(\Omega;\mathbb{R}^N)$. Failure to do so may result in a theoretical inconsistency on the continuous level as well as a drastically reduced convergence rate or even lack of convergence in the discrete setting (asymptotically, assuming conforming discretizations). Fortunately, the Riesz representation $\boldsymbol{\xi} = R_{\mathrm{Riesz}}^{-1}(\nabla\hat{\mathcal{J}}(\boldsymbol{\varphi}^k))$ can be easily calculated by solving a linear elliptic PDE:

$$\begin{aligned} -\Delta\boldsymbol{\xi} + \boldsymbol{\xi} &= \nabla\hat{\mathcal{J}}(\boldsymbol{\varphi}^k) && \text{in} \quad \Omega, \\ \partial_n\boldsymbol{\xi} &= 0 && \text{on} \quad \partial\Omega. \end{aligned} \tag{34}$$

As suggested in [49], we make use of the generalized Armijo step size rule in order to select the step size $t^k = t$ in (33) and set $\boldsymbol{\varphi}^{k+1} := \boldsymbol{\varphi}^{k+1}(t^k)$. Here, for a given $\sigma > 0$ we use a simple backtracking strategy to find the largest $t^k > 0$ such that

$$\hat{\mathcal{J}}(\boldsymbol{\varphi}^k) - \hat{\mathcal{J}}(\boldsymbol{\varphi}^{k+1}) \geq \sigma(t^k)^{-1}\|\boldsymbol{\varphi}^k - \boldsymbol{\varphi}^{k+1}\|^2_{H^1(\Omega,\mathbb{R}^N)} \tag{35}$$

is satisfied. We then iterate until

$$\|\boldsymbol{\varphi}^k - \mathrm{Proj}_{\mathcal{G}}\left(\boldsymbol{\varphi}^k - R_{\mathrm{Riesz}}^{-1}(\nabla\hat{\mathcal{J}}(\boldsymbol{\varphi}^k)\right)\|_{H^1(\Omega,\mathbb{R}^N)} \leq \mathrm{tol}_{PG} \tag{36}$$

is satisfied for some tolerance $\mathrm{tol}_{PG} > 0$.

It may at first seem odd that we employ this first-order numerical method. In particular, as noted in [49], the projected gradient method is best used when the set $\mathcal{G}$ is simple enough that $\mathrm{Proj}_{\mathcal{G}}$ is a trivial calculation; in contrast to the current situation that requires the solution of an elliptic variational inequality for each projection. Nevertheless, a direct application of second-order optimization techniques as in [7] is not possible here. Indeed, if we were to write the Helmholtz equation in the form of optimality conditions (5), there would be no guarantee that the steps generated by a second-order method are related to the smallest eigenvalue.

Finally, in order to calculate $\nabla\hat{\mathcal{J}}$ we need to solve both forward equations and both adjoint equations (see (4.10)). For the solution of the Helmholtz equation (H($\boldsymbol{\varphi}$)), we first apply the shift as described in Section 5.3 below, then (for the discretized problem) we solve the resulting eigenvalue problem via MATLAB function `eigs` (which is built on top of the ARPACK library by [50]) and finally apply a shift back. Since the directional derivative from (14) has a unique solution, it can be simply solved as a system of linear equations.

5.3 Estimating the Shift Parameter c

The choice of the shift parameter c in (9) is a delicate matter as it has a major impact on the computation of the smallest eigenvalue. Several methods such as the inverse method [51] find the eigenvalue closest to zero and the rate of convergence equals to the ratio of the two eigenvalues closest to zero. Thus, if the shift is too small, a different eigenvalue may be found while if the shift is too big, the convergence will be slow.

To keep positivity of the smallest eigenvalue, it is always possibly to choose $c = M$, where M is the bounding constant from (A2). However, this choice may be suboptimal. Here, we present two possibilities for a shift which ensures positivity of the smallest eigenvalue.

Lemma 5.4. *Set $\Omega = (0,a) \times (0,b)$ and assume that (A2) and (A4) hold. Consider the shift*

$$c := L(M + \|g(0)\|_{L^\infty(\Omega)}) - 2\pi^2/ab\,. \tag{37}$$

Then the smallest eigenvalue of $-\Delta - g(\boldsymbol{\varphi}) + cI$ is nonnegative for all $\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$.

Proof. Denote by λ_1 the smallest eigenvalue of the operator $-\Delta - g(\boldsymbol{\varphi})$ and by $\lambda_1(\Omega)$ the smallest eigenvalue of the operator $-\Delta$. From Lemma 4.4 we infer

$$|\lambda_1 - \lambda_1(\Omega)| \le L\|g(\boldsymbol{\varphi}) - g(0)\|_{L^\infty(\Omega)} \le L(M + \|g(0)\|_{L^\infty(\Omega)}).$$

From [40, Proposition 8.5.2] we obtain $\lambda_1((0,a)\times(0,b)) = \frac{2\pi^2}{ab}$. The assertion follows. □

Let $\boldsymbol{\varphi}$ be the current iterate of a procedure for solving our optimization problem. If c is the shift and the computed smallest eigenvalue of operator $-\Delta - g(\boldsymbol{\varphi}) + cI$ equals λ_1, then the optimal shift is $c - \lambda_1$. Even though we cannot use this information for determining $\boldsymbol{\varphi}$, it is of use for determining the next iterate. In fact, in this case we may use the shift in (38). In what follows, C_P denotes the Poincaré constant, i.e., for every $\Theta \in H_0^1(\Omega)$ one has $\|\Theta\|_{L^2(\Omega)} \le C_P\|\nabla\Theta\|_{L^2(\Omega)}$.

Lemma 5.5. *Assume that (A2) and (A4) hold and that for $\boldsymbol{\varphi} \in H^1(\Omega,\mathbb{R}^N)$ and for some shift c we know the eigenvalue λ_1 of operator $-\Delta + g(\boldsymbol{\varphi}) + cI$. Consider $\delta\boldsymbol{\varphi} \in H^1(\Omega,\mathbb{R}^N)$, define*

$$\hat{c} := c - \lambda_1 + 2^{-3/4}C_P^{-2}(C_P^2+1)(2MC_P^2+1)L\|\delta\boldsymbol{\varphi}\|_{L^2(\Omega)} \tag{38}$$

and denote by $\hat{\lambda}_1$ the smallest eigenvalue of operator $-\Delta + g(\boldsymbol{\varphi}+\delta\boldsymbol{\varphi}) + \hat{c}I$. Then $\hat{\lambda}_1 \ge 0$. Moreover, denote the second smallest eigenvalues of the previous two operators by λ_2 and $\hat{\lambda}_2$, respectively. Let $\kappa := 2^{-3/4}C_P^{-2}(C_P^2+1)(2MC_P^2+1)L$. If

$$\|\delta\boldsymbol{\varphi}\|_{L^2(\Omega)} < (\lambda_2 - \lambda_1)/(2\kappa), \tag{39}$$

then

$$0 \le \hat{\lambda}_1\hat{\lambda}_2^{-1} \le 2\kappa(\lambda_2-\lambda_1)^{-1}\|\delta\boldsymbol{\varphi}\|_{L^2(\Omega)}. \tag{40}$$

Proof. Due to [52, Chapter 3, Lemma 3.3] we for any $\Theta \in H_0^1(\Omega)$ have

$$\begin{aligned}\|\Theta\|_{L^4(\Omega)}^2 \le 2^{\frac{1}{4}}\|\nabla\Theta\|_{L^2(\Omega)}\|\Theta\|_{L^2(\Omega)} \le 2^{-\frac{3}{4}}\big(\|\nabla\Theta\|_{L^2(\Omega)}^2 + \|\Theta\|_{L^2(\Omega)}^2\big)\\ \le 2^{-\frac{3}{4}}(1+C_P^2)\|\nabla\Theta\|_{L^2(\Omega)}^2\end{aligned}$$

By definition, $C_P^{-1} = \inf\{\|\nabla u\|_{L^2(\Omega)}/\|u\|_{L^2(\Omega)} : u \in H_0^1(\Omega)\}$. The latter optimization problem can be reformulated as $\inf\{\|\nabla u\|_{L^2(\Omega)} : u \in H_0^1(\Omega),\ \|u\|_{L^2(\Omega)} = 1\}$, which has a solution based on our analysis of (10). Therefore, there exists some $\hat{\Theta}_0 \in H_0^1(\Omega)$ with $1 = \|\hat{\Theta}_0\|_{L^2(\Omega)} = C_P\|\nabla\hat{\Theta}_0\|_{L^2(\Omega)}$. Now we provide an estimate for constant $\tilde{L}$ in (12). Fix any $\hat{\boldsymbol{\varphi}} \in H^1(\Omega, \mathbb{R}^N)$ and let $\hat{\Theta} \in H_0^1(\Omega)$ be the corresponding minimizer of (10). Then we have

$$\begin{aligned}\|g(\hat{\boldsymbol{\varphi}}) - g(\boldsymbol{\varphi})\|_{L^2(\Omega)}\|\hat{\Theta}\|^2_{L^4(\Omega)} &\le L\|\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_{L^2(\Omega)}\|\hat{\Theta}\|^2_{L^4(\Omega)}\\ &\le 2^{-\frac{3}{4}}L(C_P^2+1)\|\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_{L^2(\Omega)}\|\nabla\hat{\Theta}\|^2_{L^2(\Omega)}\\ &\le 2^{-\frac{3}{4}}L(C_P^2+1)\|\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_{L^2(\Omega)}(2M + \|\nabla\hat{\Theta}_0\|^2_{L^2(\Omega)})\\ &= 2^{-\frac{3}{4}}C_P^{-2}(C_P^2+1)(2MC_P^2+1)L\|\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_{L^2(\Omega)},\end{aligned}$$

where the third inequality is due to (11) and L is the Lipschitz constant of $\hat{g}$. Thus, we have $\tilde{L} = 2^{-\frac{3}{4}}C_P^{-2}(C_P^2+1)(2MC_P^2+1)L$. Then the first two eigenvalues of the operator $-\Delta + g(\boldsymbol{\varphi}) + \hat{c}I$ are equal to $\tilde{\lambda}_1 := \tilde{L}\|\delta\boldsymbol{\varphi}\|_{L^2(\Omega)}$ and $\tilde{\lambda}_2 := \tilde{L}\|\delta\boldsymbol{\varphi}\|_{L^2(\Omega)} + \lambda_2 - \lambda_1$, respectively. From Lemma 4.4 we then obtain $0 \le \hat{\lambda}_1 \le 2\tilde{L}\|\delta\boldsymbol{\varphi}\|_{L^2(\Omega)},\ \lambda_2 - \lambda_1 \le \hat{\lambda}_2$. □

Note that the shift $c = M$ and the shift from Lemma 5.4 are independent of $\boldsymbol{\varphi}$, where the one from Lemma 5.5 depends on the perturbation. Observe, furthermore, that an iterative scheme for solving the optimization problem (8) in reduced form yields $\delta\boldsymbol{\varphi} \to 0$ in $L^2(\Omega, \mathbb{R}^N)$. Hence, the shift in (38) converges and the ratio in (40) will tend to zero.

6 Numerical Experiments

In this section, we present the results of numerical optimization experiments. The optimal solution is then used in the final section below to demonstrate the electronic properties of the associated microbridge design.

6.1 Structural Assumptions: Elasticity and Optics

For the elasticity equation, we primarily follow the setting in [7]. The $\boldsymbol{\varphi}$-dependent elasticity tensor is of the form

$$\mathbb{C}(\boldsymbol{\varphi}) := \widehat{\mathrm{cut}}(\varphi_1)\mathbb{C}_1 + \cdots + \widehat{\mathrm{cut}}(\varphi_N)\mathbb{C}_N,$$

where $\mathbb{C}_i$ is a standard elasticity tensor associated with material i. Thus, for $E_1, E_2 \in \mathbb{R}^{2\times 2}$ we have $\mathbb{C}_iE_1{:}E_2 = \lambda_i \mathrm{tr}\, E_1 \mathrm{tr}\, E_2 + 2\mu_i E_1{:}E_2$, where λ_i and μ_i are Lamé constants of individual materials and $\widehat{\mathrm{cut}} : \mathbb{R} \to \mathbb{R}$ is the cutoff function

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

for some small $\delta_1 > 0$, large $\delta_2 > 0$, $a = \delta_1^4/\pi$, and $b = (1 - 2\delta_1^3)\pi/\delta_1^4$. Note that the cutoff function is a twice continuously differentiable increasing function with the property $\widehat{\mathrm{cut}}(x) \ge \delta_1^4/2$ for all $x \in \mathbb{R}$ and thus (A2) is satisfied. As in [7], where we employed second-order optimization methods,

$\widehat{\mathrm{cut}}$ is chosen to ensure that $\mathbb{C}$ is sufficiently smooth and the resulting differentiable operator remains elliptic. Note that as $\delta_1 \to 0$ and $\delta_2 \to 1$, $\widehat{\mathrm{cut}}$ approaches the identity on $[0,1]$.

Concerning the Helmholtz equation, we define g by

$$g(\boldsymbol{\varphi}) := 2\pi^2\lambda^{-2}(\varepsilon_1 \,\mathrm{cut}(\varphi_1) + \cdots + \varepsilon_N \,\mathrm{cut}(\varphi_N)). \tag{42}$$

Here, $\lambda > 0$ is the desired wavelength and $\varepsilon_i > 0$, $i = 1, \ldots, N$, are the relative permittivities of the individual materials. For some small $\delta_3 > 0$, the cutoff function $\mathrm{cut} : \mathbb{R} \to \mathbb{R}$

$$\mathrm{cut}(x) := \begin{cases} \delta_3 \,\mathrm{arctg}(\frac{x}{\delta_3}) & \text{if } x < 0, \\ x & \text{if } x \in [0,1], \\ 1 + \delta_3 \,\mathrm{arctg}(\frac{x-1}{\delta_3}) & \text{if } x > 1 \end{cases} \tag{43}$$

is necessary for assumption (A2) to hold true. Since the requirements on $\mathbb{C}$ and g are different, we work with different cutoff functions.

6.2 Discretization and Refinement Strategies

For the numerical implementation, we discretize the underlying function spaces using P1-finite elements. All numerical experiments are carried out using MATLAB. In order to increase the computational efficiency of the scheme, we use an adaptivity heuristic to generate new meshes following the "red" refinement strategy, cf. [53], which is implemented in the package P1-AFEM, see [54]. The marking heuristic is as follows: After solving (8) on a given mesh, every element on which the phases are not pure or where there is a transition between two materials is refined. Otherwise, we coarsen or leave the element unchanged if there exist pure phases and no transition.

In addition to the role of the various phases in the refinement strategy, we need to take into account the interfacial thickness parameter ε, which appears in the Ginzburg-Landau-term f_{GL}. Since ε corresponds to the interfacial thickness, the initial ε is chosen to be twice the length of the largest element. Subsequently, we divide ε by 2 upon every mesh refinement. We refine the mesh in our experiments five times. For the projection onto the Gibbs simplex $\mathcal{G}$, we use the potentially mesh-dependent semismooth Newton method as suggested in [55], where it is shown to be equivalent to a primal-dual active set strategy with warm start. This strategy is efficient provided the active sets are stable over mesh refinements. Another possibility would be to use the path-following method from [56], as it is mesh-independent.

6.3 Parameters and Starting Values

As mentioned above, we consider three possible materials: Ge, SiN, SiO$_2$ as well as air. In Table 1 we summarize their physical properties, see [57, 58, 59] and the fixed domains Π_i are given by: $\Pi_{\mathrm{Ge}} := [-0.125, 0.125] \times [1, 1.49],\ \Pi_{\mathrm{SiN}} := [-0.75, 0.75] \times [1.5, 1.75],\ \Pi_{\mathrm{SiO_2}} := [-2, 2] \times [0, 0.99],\ \Pi_{\mathrm{air}} := [-2, 2] \times [2.5, 3]$ (in μm) Since the general model contains a number of parameters, we list them here for convenience:

- N: Number of phases
- α: Weights in the objective for the Ginzburg-Landau energy f_{GL}
- ε: Parameter corresponding to interfacial thickness

	λ [GPa]	μ [GPa]	ε [-]	σ_0 [GPa]	ε_0 [-]
Ge	44.279	27.249	17.64	$\cdot$	$\cdot$
SiN	110.369	57.813	4	-3.8	$\cdot$
SiO_2	16.071	20.798	2.25	$\cdot$	$2.6{\cdot}10^{-3}$

Table 1: List of material properties for elasticity.

α	N	$h_{\min}$	$\varepsilon_{\min}$	δ_1	δ_2	δ_3	tol_{PG}	σ
$4{\cdot}10^{-4}$	4	2^{-8}	2^{-7}	10^{-3}	10^{16}	10^{-3}	10^{-6}	10^{-4}

Table 2: List of parameters.

- $\delta_1, \delta_2, \delta_3$: Cutoff parameters from (41) and (43)
- ϵ_0, δ_0: Constants for the eigenstrain generated by SiO_2 and the thermal (pre-)stress generated by SiN, see (3)
- λ, ε_i: The wavelength and the relative permittivities of materials, see (42)
- tol_{PG}: Stopping tolerance for first-order system (36)
- $h_{\min}, \varepsilon_{\min}$: Width of the smallest triangle and value of ε on the finest mesh

The parameter values are summarized in Table 2. The cutoff parameters δ_1, δ_2, and δ_3 were chosen so that the cutoff has a negligible effect on the interval $(0, 1)$. Since $\Omega = (-2, 2) \times (0, 3)$ (in μm), the values $h_{\min} = \frac{1}{256}\mu$m and $\varepsilon_{\min} = \frac{1}{128}\mu$m give rise to a rather fine mesh along the interface. For the wavelength we choose $\lambda = 1.64\mu$m.

6.4 Numerical results

In Figure 2 we depict the optimal φ (left) and the corresponding strain field (right). To keep the discrete instances small, we employ the previously described mesh refinement strategy. Furthermore, we coarsened all elements where there was only one pure phase. The meshes after the first, third and final fifth refinement are shown in Figure 3. Since we are able to drive ε to a rather small value, the final design has a rather sharp interface, see Figure 2.

The number of active nodes (where no material is prescribed) is depicted in the left-hand side of Table 3. The small increase from the penultimate to the final mesh is caused by the disapprearance of an artefact above the structure, whose presence can be inferred from the structure of the refined mesh in Figure 3. The refined region above the structure in the final mesh is a remanent of this artefact, which disappears in the final phase field φ, see Figure 2. The number of iterations is shown on the right-hand side of Table 3. Note that on the intermediate meshes 3, 4, and 5, we accepted a suboptimal, i.e., substationary, solution after reaching 500 iterations. Nevertheless, on Mesh 6, the algorithm only needs 223 iterations to reach a tolerance below 10^{-7}.

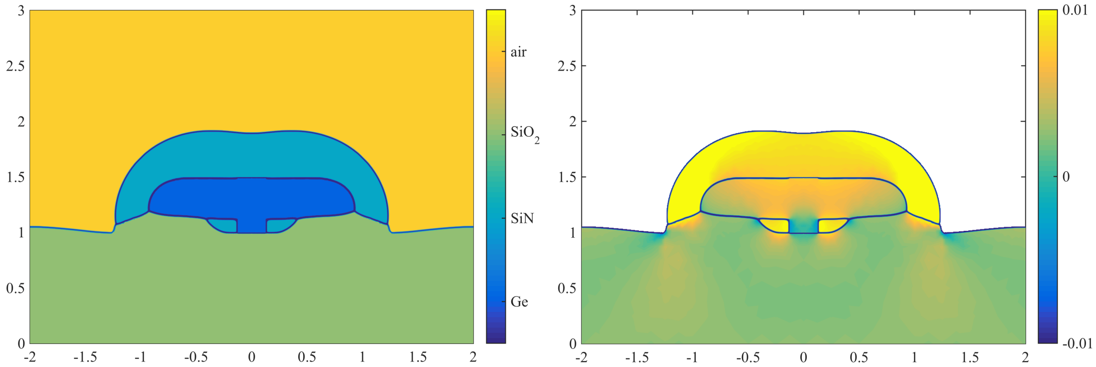

Figure 2: Optimal φ (left) and its corresponding strain field (right).

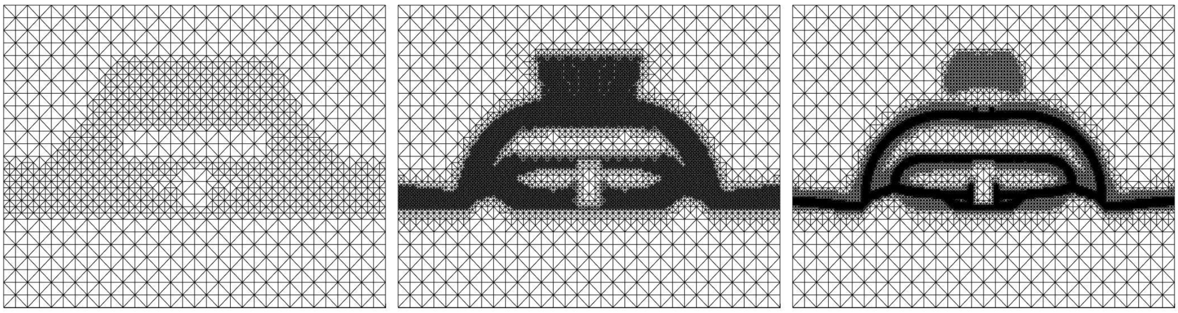

Figure 3: Adaptively updated mesh (both refined and coarsened) after first, third and fifth refinement.

	Mesh 1	Mesh 2	Mesh 3	Mesh 4	Mesh 5	Mesh 6
active	345	1043	3027	9902	25824	36273
iter	37	427	500+	500+	500+	223
res	$9.33\cdot10^{-8}$	$8.95\cdot10^{-8}$	$8.42\cdot10^{-5}$	$9.95\cdot10^{-5}$	$3.26\cdot10^{-7}$	$7.94\cdot10^{-8}$
GL	9.298	8.821	8.659	9.154	8.630	8.580

Table 3: `active`: Number of active nodes (with no material prescribed), `iter`: number of iterations, `res`: the best residual (36) and `GL`: the value of the Ginzburg-Landau energy for all meshes.

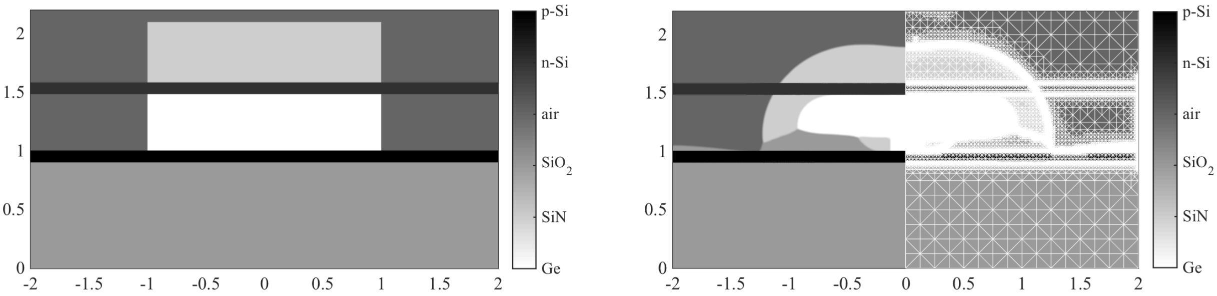

Figure 4: **(left)** Original phase fields φ_i with inserted Si contacts and **(right)** optimized phase fields φ_i with inserted Si contacts and mesh with shading indicating $i \in \{$Ge,SiN,SiO$_2$,air,n-Si,p-Si$\}$.

7 Simulation of Carrier Transport for the Optimal Design

In order to highlight the improvement by the optimization procedure, we conclude the paper with a detailed simulation of stationary solutions for the associated drift-diffusion equations for the initial and optimized material configuration. In fact, we will see below that the strain field generated by the optimal configuration $\boldsymbol{\varphi}$ leads to a positive net gain as compared to the initial design with negative net gain.

However, in order to make both material distributions technically feasible, we first manually add two phase fields φ_i with $i = n$-Si or $i = p$-Si representing two thin, highly n and p-doped silicon layers above and below the germanium, c.f. Figure 4.

Then, to solve for stationary solutions of the van Roosbroeck system (6), we transform the equation from charge carrier densities (n, p) to the so-called quasi-Fermi potentials (ϕ_n, ϕ_p), defined by

$$n = N_c F\left(\frac{q(\phi - \phi_n) - E_\mathsf{c}}{k_\mathsf{B} T}\right), \qquad p = N_v F\left(\frac{q(\phi_p - \phi) + E_\mathsf{v}}{k_\mathsf{B} T}\right), \tag{44}$$

where $F(\eta) = \exp(-\eta)$ for Boltzmann distributions or $F(\eta) = F_{3/2}(\eta)$ the complete Fermi-Dirac integral with index $3/2$ for Fermi-Dirac distributions. By N_c, N_v we denote the material dependent effective density of states, $E_\mathsf{c}, E_\mathsf{v}$ are the conduction and valence band-edges, and q is, as before, the elementary charge. Rewriting the van Roosbroeck system in these terms has the advantage of guaranteeing positivity of (n, p) and it will formally ensure continuity of (ϕ_n, ϕ_p) at heterojunctions, where (n, p) are usually discontinuous. We may now write the stationary form of the van Roosbroeck system (6) in its weak form, where we seek $(\phi, \phi_n, \phi_p) \in H^1_D(\Omega)^3$ such that the nonlinear system of equations

$$\int_\Omega \varepsilon_0 \varepsilon_\mathsf{r} \nabla\phi \cdot \nabla v \,\mathrm{d}\mathbf{x} = q \int_\Omega (C_\mathsf{dop} + p - n) v \,\mathrm{d}\mathbf{x}, \tag{45a}$$

$$\int_\Omega q n \mu_n \nabla\phi_n \cdot \nabla v_n \,\mathrm{d}\mathbf{x} = \int_\Omega q R_\mathsf{net} v_n \,\mathrm{d}\mathbf{x}, \tag{45b}$$

$$\int_\Omega q p \mu_p \nabla\phi_p \cdot \nabla v_p \,\mathrm{d}\mathbf{x} = -\int_\Omega q R_\mathsf{net} v_p \,\mathrm{d}\mathbf{x}, \tag{45c}$$

holds for all $(v, v_n, v_p) \in H^1_0(\Omega)^3$. The material data $\mu_n, \mu_p, N_c, N_v, E_c, E_v, \varepsilon_r, C_\mathsf{dop}$ depend on space through the phase fields via interpolated material parameters as introduced in Tab. 4, e.g., $\mu_n(x) = \sum_i \mu_n^i \varphi_i(x)$.

At Ohmic contacts Γ_D we enforce inhomogeneous Dirichlet boundary conditions for the potentials, otherwise we have natural boundary conditions for (45). In order to rewrite R_net for general distribution functions F one has to ensure $R_{\text{net}} = 0$ in thermal equilibrium, where $\phi_n = \phi_p = 0$, which we can satisfy by using

$$R_{\text{net}} = \Big(1 - \exp\big(\tfrac{q}{k_{\text{B}}T}(\phi_n - \phi_p)\big)\Big)\,\frac{np}{\tau_p(n+n_i) + \tau_p(p+n_i)},$$

for Shockley-Read-Hall recombination terms, see also [60]. In the Boltzmann situation this expression reduces to the well-known form. We use a closed-form approximation for $F_{3/2}$ from [61] and discretize the weak form (45) using triangular P1 finite elements, where all the integration of nonlinearities is performed using a standard 7-point Gauss quadrature. Boundary conditions are enforced using Lagrange multipliers. The resulting discretized system of equations is solved by using Newton's method. In order ensure its convergence for large applied biases V_{ext} we employ the following strategy:

(i) First, find pointwise $\bar{\phi}$ so that the right-hand-side of (45a) vanishes with $\phi_n = 0$ and $\phi_p = 0$, i.e., solve $C_{\text{dop}} + p - n = 0$.

(ii) Solve (45a) for ϕ with $\phi_n = \phi_p = 0$, i.e., solve the nonlinear Poisson equation

$$-\nabla\cdot(\varepsilon_0\varepsilon_{\text{r}}\nabla\phi_{\text{eq}}) = q(C + N_vF(\eta_p) - N_cF(\eta_n))$$

where the previously determined $\bar{\phi}$ serves as initial data for Newton's method with $V_{\text{ext}} = 0$. The resulting solution $(\phi_{\text{eq}} = \phi, \phi_n = 0, \phi_p = 0)$ represents the solution at thermodynamic equilibrium.

(iii) Employ a path-following method to solve the entire system (45) for (ϕ, ϕ_n, ϕ_p) with $V_{\text{ext}} > 0$. This is done by using Newton's method with the solution with a smaller bias as initialization. The Dirichlet conditions at the two Ohmic contacts Γ_{D_i} for $i = 1, 2$, are $\phi = \bar{\phi} + V^i_{\text{ext}},\ \phi_n = V^i_{\text{ext}},\ \phi_p = V^i_{\text{ext}},$ where $V^1_{\text{ext}} = 0$, $V^2_{\text{ext}} = V_{\text{ext}}$ and $\Gamma_D = \Gamma_{D_1} \cup \Gamma_{D_2}$ and $\Gamma_{D_1} \cap \Gamma_{D_2} = \emptyset$.

Even though standard approaches to van Roosbroeck systems employ Scharfetter-Gummel type discretizations using finite volume methods, [62], we decided to solve the multiphysics problem using finite elements similar to [63]. This will in general lead to boundary layers for ϕ_n, ϕ_p at Ohmic contacts. We address this by using a heuristic refinement strategy near contacts. Furthermore, in order to compute the total current $\mathbf{j} = \mathbf{j}_p + \mathbf{j}_n$ with $\mathbf{j}_n = -qn\mu_n\nabla\phi_n$ and $\mathbf{j}_p = -qp\mu_p\nabla\phi_p$ at a contact Γ_{D_1} we use the identity

$$J = \int_{\Gamma_{D_1}} \mathbf{j}\cdot\mathbf{n}\,\mathrm{d}a = -\int_{\Omega} q(n\mu_n\nabla\phi_n\cdot\nabla u + p\mu_p\nabla\phi_p\cdot\nabla u)\,\mathrm{d}\mathbf{x}.$$

Here, we exploit the fact that $\nabla\cdot\mathbf{j} = 0, \mathbf{j}\cdot\mathbf{n} = 0$ on $\partial\Omega\setminus\Gamma_D$, using a constructed test function u with $u = 1$ on Γ_{D_1} and $u = 0$ on Γ_{D_2}. This definition of J is useful as it redistributes the evaluation of $\nabla\phi_n$, $\nabla\phi_p$ from the boundary layer to an evaluation in the volume Ω. For simplicity we choose u to be harmonic with the above mentioned Dirichlet data. In order to evaluate the optoelectronic performance of the optimized design, we compute and show the currents, current densities, and the net-gains, where the latter is defined (pointwise) by subtracting optical losses from optical gain and scaling the result by the optical mode. The resulting gain model is the same as published in [9], and generally higher net-gains for given current is desired.

7.1 Discussion and conclusion

As desired, the topology delivers a rather smooth material distribution, which increases the in-plane biaxial strain in the Ge phase for the initial design from an average $\bar{e}_{xx} = 2\cdot 10^{-4}$ to an average strain of $\bar{e}_{xx} = 9 \cdot 10^{-4}$ for the improved design, see Fig. 2. While loss mechanisms due to low confinement or recombination are not included in the optimization, the cost functional in (8) is designed to optimize the overlap of the optical mode and regions of large tensile strain. Therefore, the optimal designs exhibit overall improvements for the integrated strain (on average) versus the maximal/peak in-plane strains. For the latter, we see here that the maximal (pointwise) in-plane strain in the Ge cavity only features an increase by a factor of $\times 1.2$.

Another interesting feature of the optimal designs is that the Ge phase is surrounded by an SiN stressor. This is very similar to the all-around stressor designs considered for germanium microdiscs in [64].

The optimized design also features an aperture, which, as we showed previously, can be highly beneficial for lowering the threshold current of an edge-emitting laser. The main idea of the aperture is visible in the hole-currents in Fig. 6, where the currents in the optimized microbridge (right) are guided efficiently into the optical mode to recombine without creating a shortcut pathway around the center of the optical mode, as it is the case for the initial microbridge (left). For better interpretation we also indicate the material boundaries between the phases by plotting regions where $\varphi_i\varphi_j{>}0$ between material i and j in white. However, while a doping optimization can produce such an aperture geometry from a suitably defined cost functional, the aperture of the optimized design is more likely created artificially due to the location of the highly doped Si contacts above the cavity.

Nevertheless, due to the improved strain the Ge phase also features much higher modal gain at the prescribed external bias, see Fig. 8. Also, the characteristic curve in Fig. 5 features a lower current, certainly due to higher Ohmic resistance based on the implementation of the aperture. Most noticeable, however, is that the modal gain as well as the net-gain show significant improvement of the optimized design as compared to the initial design. For a recent study containing a thorough explanation of the calculation of the gain curves, we refer the interested reader to [9].

This allows us to conclude that, even though not yet fully coupled, topology optimization for optoelectronic devices can improve device designs significantly. The optimized designs are similar to what is considered by engineers. The optoelectronic simulations prove the feasibility of the optimization strategy. Nevertheless, since optoelectronic devices also suffer from loss mechanisms due to recombination, future optimization studies might even consider the fully coupled optoelectronic system.

Acknowledgement

The authors would like to cordially thank Marita Thomas for the many helpful discussions and comments. This research of DP is carried out in the framework of MATHEON supported by the Einstein Foundation Berlin in the MATHEON project OT8.

param.	phys. unit	Ge	SiN	SiO$_2$	air	Si$^{\text{top}}$	Si$^{\text{bottom}}$
ε_{r}	[1]	16.2	7.5	3.8	1	11.9	11.9
μ_{n}	[m$^2V^{-1}s^{-1}$]	0.39	10^{-4}	10^{-4}	10^{-4}	0.14	0.14
μ_{p}	[m$^2V^{-1}s^{-1}$]	0.19	10^{-4}	10^{-4}	10^{-4}	0.045	0.045
N_{c}	[10^{19}cm^{-3}]	1.256	10^{-2}	10^{-2}	10^{-2}	3.2	3.2
N_{v}	[10^{19}cm^{-3}]	0.118	10^{-2}	10^{-2}	10^{-2}	1.8	1.8
C_{dop}	[10^{19}cm^{-3}]	5	0	0	0	$+20$	-20
E_{c}	[eV]	0.76*	1	1	1	1.169	1.169
E_{v}	[eV]	0.09*	0	0	0	1.169	1.169
$\mathcal{D}^{\text{c}}$	[eV]	-3.5	0	0	0	0	0
$\mathcal{D}^{\text{v}}$	[eV]	$+1.4$	0	0	0	0	0

Table 4: Spatial interpolation $\pi(x) = \sum_i \varphi_i(x)\pi_i$ for electronic simulation given phase fields $\varphi_i(x)$ and pure phase material parameters π_i. For the topology optimization we have $i = 1\dots4$, whereas for the electronic simulation we introduce additional Si contact layers with $i = 5, 6$. The global parameters $\tau_n = \tau_p = 10\,\text{ns}$ and $n_i = 10^6\,\text{cm}^{-3}$ are used for the recombination. ($\star$) Given a strain distribution $e(\mathbf{u})$, the bandgaps are modified by deformation potentials $\mathcal{D}^\alpha_{kl} = \mathcal{D}^\alpha\delta_{kl,xx}$ via $E_\alpha(x) = \sum_i(E_i + \mathcal{D}^\alpha e_{xx})\varphi_i(x)$ with $\alpha \in \{$c,v$\}$ and in-plane biaxial strain e_{xx}. The electronic parameters and deformation potentials are from [13], the values for $\mu_n, \mu_p, N_{\text{c}}, N_{\text{v}}, E_{\text{c}}, E_{\text{v}}$ for SiN, SiO$_2$ and air are chosen to prevent existence and transport of carriers.

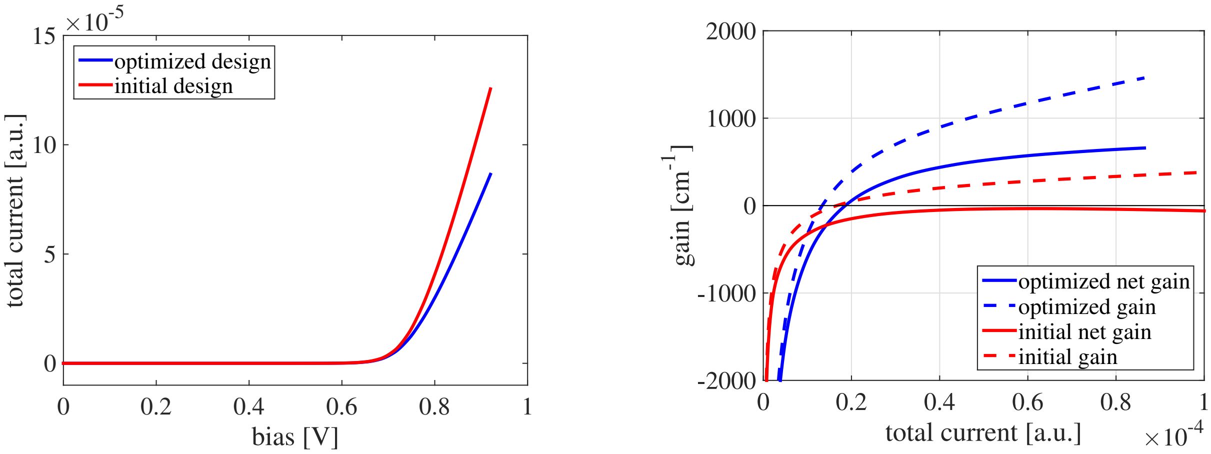

Figure 5: **(left)** Current-voltage characteristic of initial (red) and optimized (blue) device **(right)** current-gain (solid) and current-net gain (dashed) characteristics of initial and optimized device showing that the optimized configuration yields considerably higher gain and net gain compared to the initial design.

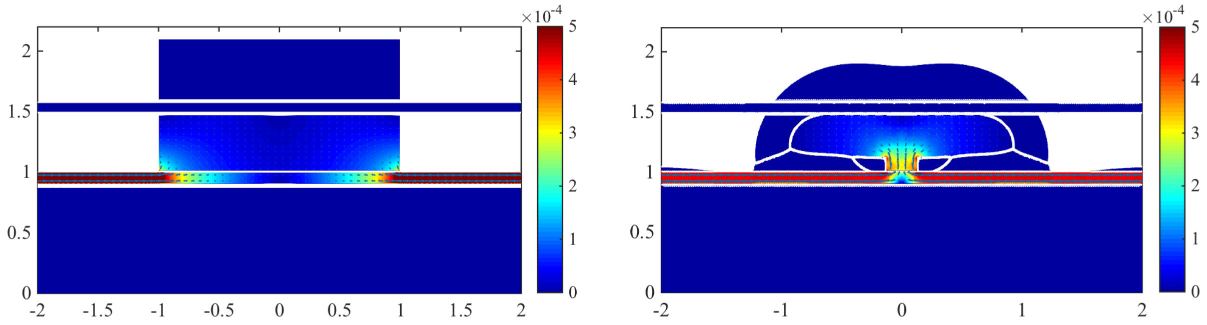

Figure 6: Hole currents for **(left)** initial design and **(right)** and for optimized design. Material boundaries are indicated in white.

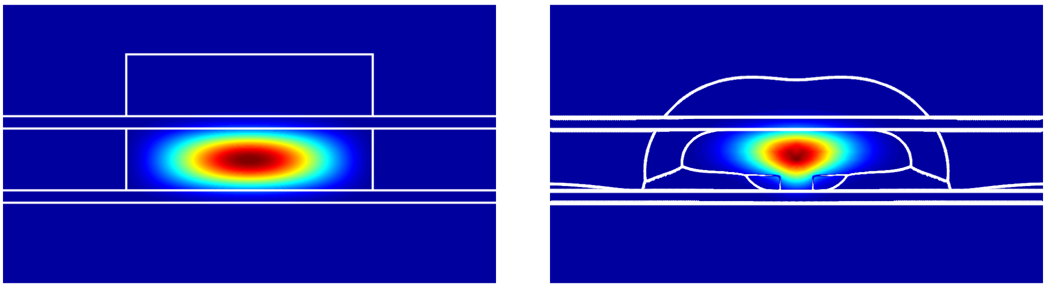

Figure 7: Optical mode $|\Theta|^2$ (shading) and material boundaries indicated in white **(left)** for initial design and **(right)** for optimized design.

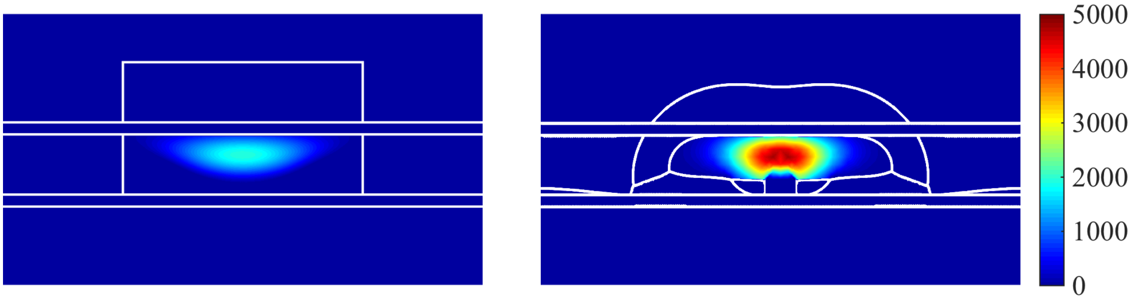

Figure 8: Modal gain $g|\Theta|^2$ [cm^{-1}] for **(left)** initial design and **(right)** optimized design shows almost threefold increase in gain due to optimized design. Material boundaries are indicated in white.

References

[1] M. El Kurdi, G. Fishman, S. Sauvage, and P. Boucaud, “Band structure and optical gain of tensile-strained germanium based on a 30 band $k \cdot p$ formalism,” *Journal of Applied Physics*, vol. 107, no. 1, 2010.

[2] X. Sun, L. Jifeng, L. Kimerling, and J. Michel, “Toward a germanium laser for integrated silicon photonics,” *Selected Topics in Quantum Electronics, IEEE Journal of*, vol. 16, no. 1, pp. 124–131, 2010.

[3] R. E. Camacho-Aguilera, Y. Cai, N. Patel, J. T. Bessette, M. Romagnoli, L. C. Kimerling, and J. Michel, “An electrically pumped germanium laser,” *Opt. Express*, vol. 20, no. 10, pp. 11316–11320, 2012.

[4] M. J. Suess, R. Geiger, R. A. Minamisawa, G. Schiefler, J. Frigerio, D. Chrastina, G. Isella, R. Spolenak, J. Faist, and H. Sigg, “Analysis of enhanced light emission from highly strained germanium microbridges,” *Nat Photon*, vol. 7, no. 6, pp. 466–472, 2013.

[5] S. Wirths, R. Geiger, N. von den Driesch, G. Mussler, T. Stoica, S. Mantl, Z. Ikonic, M. Luysberg, S. Chiussi, J. M. Hartmann, H. Sigg, J. Faist, D. Buca, and D. Grutzmacher, “Lasing in direct-bandgap GeSn alloy grown on Si,” *Nat Photon*, vol. 9, no. 2, pp. 88–92, 2015.

[6] B. Dutt, D. S. Sukhdeo, D. Nam, B. M. Vulovic, Z. Yuan, and K. C. Saraswat, “Roadmap to an efficient germanium-on-silicon laser: Strain vs. n-type doping,” *IEEE Photonics Journal*, vol. 4, no. 5, pp. 2002–2009, 2012.

[7] L. Adam, M. Hintermüller, and T. M. Surowiec, “A PDE-Constrained Optimization Approach for Topology Optimization of Strained Photonic Devices,” *Submitted, Preprint available at: WIAS Preprint Series, DOI 10.20347/WIAS.PREPRINT.2377*, 2017.

[8] M. Hinze and R. Pinnau, “An optimal control approach to semiconductor design,” *Mathematical Models and Methods in Applied Sciences*, vol. 12, no. 01, pp. 89–107, 2002.

[9] D. Peschka, N. Rotundo, and M. Thomas, “Doping optimization for optoelectronic devices,” *Optical and Quantum Electronics*, vol. 50, no. 3, p. 125, 2018.

[10] J. R. Jain, A. Hryciw, T. M. Baer, D. A. Miller, M. L. Brongersma, and R. T. Howe, “A micromachining-based technology for enhancing germanium light emission via tensile strain,” *Nature Photonics*, vol. 6, no. 6, p. 398, 2012.

[11] A. Ghrib, M. El Kurdi, M. De Kersauson, M. Prost, S. Sauvage, X. Checoury, G. Beaudoin, I. Sagnes, and P. Boucaud, “Tensile-strained germanium microdisks,” *Applied Physics Letters*, vol. 102, no. 22, p. 221112, 2013.

[12] G. Capellini, C. Reich, S. Guha, Y. Yamamoto, M. Lisker, M. Virgilio, A. Ghrib, M. E. Kurdi, P. Boucaud, B. Tillack, and T. Schroeder, “Tensile Ge microstructures for lasing fabricated by means of a silicon complementary metal-oxide-semiconductor process,” *Opt. Express*, vol. 22, no. 1, pp. 399–410, 2014.

[13] D. Peschka, M. Thomas, A. Glitzky, R. Nurnberg, K. Gartner, M. Virgilio, S. Guha, T. Schroeder, G. Capellini, and T. Koprucki, “Modeling of edge-emitting lasers based on tensile strained germanium microstrips,” *IEEE Photonics J.*, vol. 7, no. 3, 2015.

[14] J. Liu, X. Sun, R. Camacho-Aguilera, L. C. Kimerling, and J. Michel, “Ge-on-si laser operating at room temperature,” *Opt. Lett.*, vol. 35, no. 5, pp. 679–681, 2010.

[15] D. Peschka, M. Thomas, A. Glitzky, R. Nürnberg, M. Virgilio, S. Guha, T. Schroeder, G. Capellini, and T. Koprucki, “Robustness analysis of a device concept for edge-emitting lasers based on strained germanium,” *Optical and Quantum Electronics*, vol. 48, no. 156, 2016.

[16] O. Sigmund and S. Torquato, “Design of materials with extreme thermal expansion using a three-phase topology optimization method,” *Journal of the Mechanics and Physics of Solids*, vol. 45, no. 6, pp. 1037 – 1067, 1997.

[17] O. Sigmund and S. Torquato, “Design of smart composite materials using topology optimization,” *Smart Materials and Structures*, vol. 8, no. 3, p. 365, 1999.

[18] R. A. Adams and J. J.-F. Fournier, *Sobolev Spaces*. Elsevier, Amsterdam, second ed., 2008.

[19] J.-L. Lions, *Optimal control of systems governed by partial differential equations*. Translated from the French by S. K. Mitter. Die Grundlehren der mathematischen Wissenschaften, Band 170, Springer-Verlag, New York-Berlin, 1971.

[20] F. Tröltzsch, *Optimal control of partial differential equations*, vol. 112 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2010.

[21] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE constraints*, vol. 23 of *Mathematical Modelling: Theory and Applications*. Springer, New York, 2009.

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

[23] G. Allaire, *Shape optimization by the homogenization method*, vol. 146 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2002.

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

[25] A. A. Novotny and J. Sokołowski, *Topological derivatives in shape optimization*. Interaction of Mechanics and Mathematics, Springer, Heidelberg, 2013.

[26] M. Burger and M. Hintermüller, “Projected Gradient Flows for BV/Level Set Relaxation,” *PAMM*, vol. 5, no. 1, pp. 11–14, 2005.

[27] L. Blank, H. Garcke, M. H. Farshbaf-Shaker, and V. Styles, “Relating phase field and sharp interface approaches to structural topology optimization,” *ESAIM Control. Optim. Calc. Var.*, vol. 20, no. 02, pp. 1025–1058, 2014.

[28] S. Zhou and M. Y. Wang, “Multimaterial structural topology optimization with a generalized cahn–hilliard model of multiphase transition,” *Structural and Multidisciplinary Optimization*, vol. 33, p. 89, Jul 2006.

[29] M. Burger and R. Stainko, “Phase-field relaxation of topology optimization with local stress constraints,” *SIAM Journal on Control and Optimization*, vol. 45, no. 4, pp. 1447–1466, 2006.

[30] A. Takezawa, S. Nishiwaki, and M. Kitamura, “Shape and topology optimization based on the phase field method and sensitivity analysis,” *Journal of Computational Physics*, vol. 229, no. 7, pp. 2697 – 2718, 2010.

[31] L. Modica, “The gradient theory of phase transitions and the minimal interface criterion,” *Archive for Rational Mechanics and Analysis*, vol. 98, no. 2, pp. 123–142, 1987.

[32] S. Baldo, “Minimal interface criterion for phase transitions in mixtures of Cahn-Hilliard fluids,” *Annales de l’I.H.P. Analyse non linéaire*, vol. 7, no. 2, pp. 67–90, 1990.

[33] G. Capellini, M. De Seta, P. Zaumseil, G. Kozlowski, and T. Schroeder, “High temperature x ray diffraction measurements on ge/si (001) heterostructures: A study on the residual tensile strain,” *Journal of Applied Physics*, vol. 111, no. 7, p. 073518, 2012.

[34] A. Henrot, *Extremum Problems for Eigenvalues of Elliptic Operators*. Frontiers in Mathematics, Birkhäuser Basel, 2006.

[35] W. van Roosbroeck, “Theory of the flow of electrons and holes in germanium and other semiconductors,” *Bell. Syst. Tech. J*, vol. 29, no. 4, pp. 560–607, 1950.

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

[37] B. Dutt, D. S. Sukhdeo, D. Nam, B. M. Vulovic, Z. Yuan, and K. C. Saraswat, “Roadmap to an efficient germanium-on-silicon laser: strain vs. n-type doping,” *IEEE Photonics Journal*, vol. 4, no. 5, pp. 2002–2009, 2012.

[38] S. Chuang, *Physics of Optoelectronic Devices*. Wiley Series in Pure and Applied Optics, Wiley, 1995.

[39] D. Klindworth and K. Schmidt, “An efficient calculation of photonic crystal band structures using Taylor expansions,” *Commun. Comput. Phys*, vol. 16, no. 5, p. 1355ï£¡1388, 2014.

[40] H. Attouch, G. Buttazzo, and G. Michaille, *Variational Analysis in Sobolev and BV Spaces: Applications to PDEs and Optimization*. SIAM Philadelphia, 2006.

[41] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*. Springer-Verlag Berlin Heidelberg, third ed., 2001.

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

[43] E. Zeidler, *Nonlinear Functional Analysis and its Applications I: Fixed-Point Theorems*. Springer, 1986.

[44] L. C. Evans, *Partial Differential Equations*. American Mathematical Society, 1994.

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

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

[47] A. A. Goldstein, “Convex programming in Hilbert space,” *Bulletin of the American Mathematical Society*, vol. 70, no. 5, pp. 709–710, 1964.

[48] E. S. Levitin and B. T. Polyak, “Constrained minimization methods,” *Zh. Vychisl. Mat. Mat. Fiz.*, vol. 6, no. 5, pp. 787–823, 1966.

[49] D. P. Bertsekas, “On the Goldstein-Levitin-Polyak gradient projection method,” *IEEE Trans. Automat. Contr.*, vol. 21, no. 2, pp. 174–184, 1976.

[50] R. Lehoucq, D. Sorensen, and C. Yang, *ARPACK Users' Guide*. Society for Industrial and Applied Mathematics, 1998.

[51] B. N. Parlett, *The Symmetric Eigenvalue Problem*. Prentice-Hall, Inc., 1980.

[52] R. Temam, *Navier-Stokes equations : Theory and numerical analysis*. North-Holland Publishing Company, 1977.

[53] S. C. Brenner and C. Carstensen, “Finite element methods,” in *Encyclopedia of Computational Mechanics*, ch. 4, Wiley Online Library, 2004.

[54] S. Funken, D. Praetorius, and P. Wissgott, “Efficient implementation of adaptive P1-FEM in Matlab,” *Computational Methods in Applied Mathematics Comput. Methods Appl. Math.*, vol. 11, no. 4, pp. 460–490, 2011.

[55] M. Hintermüller, K. Ito, and K. Kunisch, “The primal-dual active set strategy as a semismooth Newton method,” *SIAM J. Optim.*, vol. 13, no. 3, pp. 865–888, 2003.

[56] L. Adam, M. Hintermüller, and T. M. Surowiec, “A semismooth Newton method with analytical path-following for the H^1-projection onto the Gibbs simplex,” *Submitted, Preprint available at: WIAS Preprint Series, DOI 10.20347/WIAS.PREPRINT.2340*, 2017.

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

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

[59] J. J. Wortman and R. A. Evans, “Young's modulus, shear modulus, and Poisson's ratio in silicon and germanium,” *Journal of Applied Physics*, vol. 36, no. 1, pp. 153–156, 1965.

[60] J. Piprek, *Nitride semiconductor devices: principles and simulation*. John Wiley & Sons, 2007.

[61] D. Bednarczyk and J. Bednarczyk, “The approximation of the Fermi-Dirac integral $F_{1/2}(\eta)$,” *Physics Letters A*, vol. 64, no. 4, pp. 409 – 410, 1978.

[62] D. L. Scharfetter and H. K. Gummel, “Large-signal analysis of a silicon read diode oscillator,” *IEEE Transactions on electron devices*, vol. 16, no. 1, pp. 64–77, 1969.

[63] M. A. der Maur, G. Penazzi, G. Romano, F. Sacconi, A. Pecchia, and A. Di Carlo, “The multiscale paradigm in electronic device simulation,” *IEEE Transactions on electron devices*, vol. 58, no. 5, pp. 1425–1432, 2011.

[64] A. Ghrib, M. El Kurdi, M. Prost, S. Sauvage, X. Checoury, G. Beaudoin, M. Chaigneau, R. Ossikovski, I. Sagnes, and P. Boucaud, “All-around sin stressor for high and homogeneous tensile strain in germanium microdisk cavities,” *Advanced Optical Materials*, vol. 3, no. 3, pp. 353–358, 2015.