Interdisciplinary Information Sciences Vol. 25, No. 2 (2019) 75–146

ISSN 1340-9050 print/1347-6157 online
DOI 10.4036/iis.2019.B.01

The Homogenization Method for Topology Optimization of Structures: Old and New

Grégoire ALLAIRE[1,*], Lorenzo CAVALLINA[2,†], Nobuhito MIYAKE[3,‡], Tomoyuki OKA[3,§] and Toshiaki YACHIMURA[2,¶]

[1]*CMAP, Ecole Polytechnique, 91128 Palaiseau, France*
[2]*RCPAM, Graduate School of Information Sciences, Tohoku University, Sendai 980-8579, Japan*
[3]*Mathematical Institute, Tohoku University, Sendai 980-8578, Japan*

Topology optimization of structures is nowadays a well developed field with many different approaches and a wealth of applications. One of the earliest methods of topology optimization was the so-called homogenization method, introduced in the early eighties. It became extremely popular in its over-simplified version, called SIMP (Solid Isotropic Material with Penalisation), which retains only the notion of material density and forgets about true composite materials with optimal (possibly non isotropic) microstructures. However, the appearance of mature additive manufacturing technologies which are able to build finely graded microstructures (sometimes called lattice materials) drastically change the picture and one can see a resurrection of the homogenization method for such applications. Indeed, homogenization is the right technique to deal with microstructured materials where anisotropy plays a key role, a feature which is absent from SIMP. Homogenization theory allows to replace the microscopic details of the structure (typically a complex networks of bars, trusses and plates) by a simpler effective elasticity tensor describing the mesoscopic properties of the structure. The goal of these lecture notes is to review the necessary mathematical tools of homogenization theory and apply them to topology optimization of mechanical structures. The ultimate application, targeted here, is the topology optimization of structures built with lattice materials. Practical and numerical exercises are given, based on the finite element free software FreeFem++.

KEYWORDS: homogenization, topology optimization, structures, lattice materials

Preface and Acknowledgments

These are the lecture notes of a short course on the homogenization method for topology optimization of structures, given by one of us, Grégoire Allaire, during the "GSIS International Summer School 2018" at Tohoku University (Sendai, Japan). Based on the slides of this course, the four other authors, Lorenzo Cavallina, Nobuhito Miyake, Tomoyuki Oka, Toshiaki Yachimura, have written the present lecture notes, which have been proofread by Grégoire Allaire. Each section of these lecture notes corresponds to one class, except the two first ones which were taught together.

Topology optimization of structures is nowadays a well developed field with many different approaches and a wealth of applications. One of the earliest method of topology optimization was the homogenization method, introduced in the early eighties. It became extremely popular in its over-simplified version, called SIMP (Solid Isotropic Material with Penalization), which retains only the notion of material density and forgets about true composite materials with optimal (possibly non isotropic) microstructures. However, the appearance of mature additive manufacturing technologies which are able to build finely graded microstructures (sometimes called lattice materials) drastically changed the picture and one can see a resurrection of the homogenization method for such applications. Indeed, homogenization is the right technique to deal with microstructured materials where anisotropy plays a key role, a feature which is absent from SIMP. Homogenization theory allows to replace the microscopic details of the structure (typically a complex networks of bars, trusses and plates) by a simpler effective elasticity tensor describing the mesoscopic properties of the structure.

The goal of this course is to review the necessary mathematical tools of homogenization theory and apply them to topology optimization of mechanical structures. The ultimate application, targeted in this course, is the topology optimization of structures built with lattice materials. Practical and numerical exercises are given, based on the finite element free software FreeFem++.

Received January 28, 2019; Accepted May 29, 2019; J-STAGE Advance published September 6, 2019
*E-mail: gregoire.allaire@polytechnique.fr
†E-mail: cava@ims.is.tohoku.ac.jp
‡E-mail: nobuhito.miyake.t2@dc.tohoku.ac.jp
§E-mail: tomoyuki.oka.q3@dc.tohoku.ac.jp
¶E-mail: yachimura@ims.is.tohoku.ac.jp

Finally, the authors would like to express their gratitude to the organizers of the Summer School: Reika Fukuizumi, Kei Funano, Jun Masamune, Jinhae Park, Ruo Li, Shigeru Sakaguchi, Kenjiro Terada, Takayuki Yamada and Lei Zhang. The meeting was partially supported by a grant from the JSPS A3 Foresight Program, JSPS KAKENHI Grant Numbers 26287020 and 26400062, GSIS and RCPAM.

1. Introduction

1.1 Optimal design of structures

A problem of optimal design (material, shape and topology optimization) of structures is defined by three ingredients (see [Al2007-1, BS2003, HM2003, HP2018, KPTZ2000, SK1992]):

(a) a **model** (typically a partial differential equation) to evaluate (or analyze) the mechanical behavior of a structure,
(b) an **objective function** which has to be minimized or maximized, or sometimes several objectives (also called cost functions or criteria),
(c) a **set of admissible designs** which precisely defines the optimization variables, including possible constraints.

The kind of optimal design problems which we focus on in these lecture notes can be roughly divided into three categories, from the "easiest" to the "most difficult" one:

1. **Parametric or sizing optimization**, for which designs are parametrized by a few variables (for example, thickness or member sizes), implying that the set of admissible designs is considerably simplified (see Fig. 1-1, where the variable parameters, the thickness of the two boxes in this case, are symbolized by arrows),
2. **Shape (geometric) optimization**, for which all designs are obtained from an initial guess by moving its boundary without change of its topology due to the generation of new boundaries (see Fig. 1-2, where an admissible shape is drawn with a broken line),
3. **Topology optimization** where both the shape and the topology of the admissible designs can vary without any explicit or implicit restrictions (see Fig. 1-3, where the broken lines show removable holes).

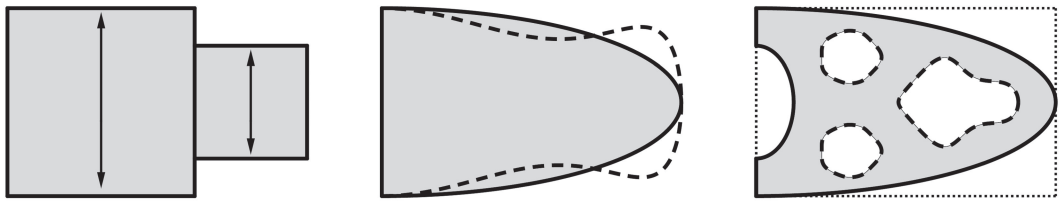


Fig. 1. Three categories of optimal design problems.

The last category in the above is, of course, the most general but also the most difficult. We recall that two shapes share the same topology if there exists a continuous deformation from one to the other. In dimension 2, topology is completely characterized by the number of holes (or, equivalently, of connected components of the boundary). In dimension 3 it is quite more complicated. Indeed, the topology of a set in dimension 3 is not only determined by the number of holes, but it also depends on the number and intricacy of "handles" or "loops."

First of all, one could ask theoretical questions concerning existence, uniqueness, and qualitative properties of the solutions of these shape optimization problems. One could also study the necessary and/or sufficient conditions satisfied by the optimal shapes. Such "optimality conditions" are very important both from a theoretical and a numerical point of view. They are often the basis for numerical algorithms of gradient method type. Furthermore one can investigate the numerical computation of approximate optimal shapes. All these questions will be addressed in the following sections.

1.2 Example of sizing or parametric optimization

First of all, we show some examples of sizing or parametric optimization. Let us consider the thickness optimization of a membrane, where Ω is a mean surface of a (plane) membrane and h is the thickness in the normal direction to the mean surface Ω (see Fig. 2).

In what follows, we consider our membrane to be pre-stressed at its boundary and subject to some vertical force f. Moreover, for small displacements, small deformations and negligible bending effects in the elasticity, the membrane deformation can be modeled by its vertical displacement $u : \Omega \to \mathbb{R}$, solution of the following partial differential equation, the so-called membrane model (see also [Al2007-1, K2016]),

$$\begin{cases} -\mathrm{div}(h\nabla u) = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

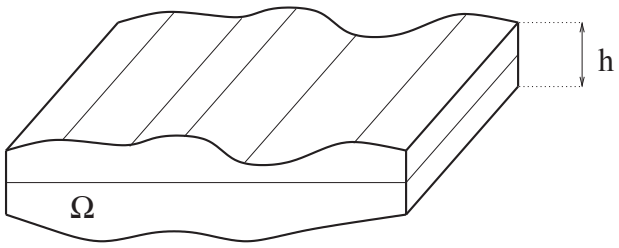


Fig. 2. Membrane with variable thickness h.

where the thickness h is bounded by some given minimum and maximum values:

$$0 < h_{\min} \leq h(x) \leq h_{\max} < \infty.$$

The thickness h is the optimization variable. Notice that we are dealing with a sizing or parametric optimal design problem here, because the computational domain Ω does not change.

Let us define the set of admissible thickness as follows:

$$\mathcal{U}_{\text{ad}} = \left\{ h \in L^\infty(\Omega) : 0 < h_{\min} \leq h(x) \leq h_{\max} \text{ a.e. in } \Omega, \int_\Omega h(x)dx = h_0|\Omega| \right\},$$

where h_0 is an imposed average thickness.

Remark 1.1 (Possible additional "feasibility" constraints)**.** *According to the production process of membranes, the thickness $h(x)$ can be discontinuous, or on the contrary continuous. A uniform bound can be imposed on its first derivative $h'(x)$ (molding-type constraint) or on its second order derivative $h''(x)$, linked to the curvature radius (milling-type constraint).*

The optimization criterion is linked to some mechanical property of the membrane, evaluated through its displacement u, solution of the PDE,

$$J(h) = \int_\Omega j(u)\, dx,$$

where, of course, u depends on h. For example, the global rigidity of a structure is often measured by its compliance, or work done by the load f: the smaller the work, the larger the rigidity (compliance $= -$rigidity). In such a case, we set

$$j(u) = fu.$$

Another example amounts to achieve (at least approximately) a target displacement $u_0(x)$, which is modeled by taking

$$j(u) = |u - u_0|^2.$$

Those two criteria are the typical examples studied in this course. Then, a parametric optimization problem is

$$\inf_{h \in \mathcal{U}_{\text{ad}}} J(h).$$

Other examples of objective functions are the following:

- Introducing the stress vector $\sigma(x) = h(x)\nabla u(x)$, we can minimize the maximum stress norm
$$J(h) = \sup_{x \in \Omega} |\sigma(x)|$$
or more generally, for any $p \geq 1$, the following p-norm
$$J(h) = \left(\int_\Omega |\sigma(x)|^p \, dx \right)^{1/p}.$$
- For a vibrating structure, introducing the first eigenfrequency ω, defined by
$$\begin{cases} -\text{div}(h\nabla u) = \omega^2 u & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$
We consider $J(h) = -\omega$ to maximize it.
- Multiple loads optimization: for n given loads $(f_i)_{1 \leq i \leq n}$ the independent displacements u_i are solutions of
$$\begin{cases} -\text{div}(h\nabla u_i) = f_i & \text{in } \Omega, \\ u_i = 0 & \text{on } \partial\Omega. \end{cases}$$
We then introduce an aggregated criterion

$$J(h) = \sum_{i=1}^{n} c_i \int_\Omega j(u_i)\, dx,$$

with given coefficients c_i, or

$$J(h) = \max_{1 \le i \le n} \left\{ \int_\Omega j(u_i)\, dx \right\}.$$

1.3 Example of shape optimization

In this section, we show two examples of shape optimization. At first let us consider a shape optimization of a membrane's shape. A reference domain for the membrane is denoted by Ω, with a boundary made of three disjoint parts

$$\partial\Omega = \Gamma \cup \Gamma_D \cup \Gamma_N,$$

where Γ is the variable part, Γ_D is the Dirichlet (clamped) part and Γ_N is the Neumann part (loaded by g).

The vertical displacement u is the solution of the following membrane model

$$\begin{cases} -\Delta u = 0 & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_D, \\ \dfrac{\partial u}{\partial n} = g & \text{on } \Gamma_N, \\ \dfrac{\partial u}{\partial n} = 0 & \text{on } \Gamma. \end{cases}$$

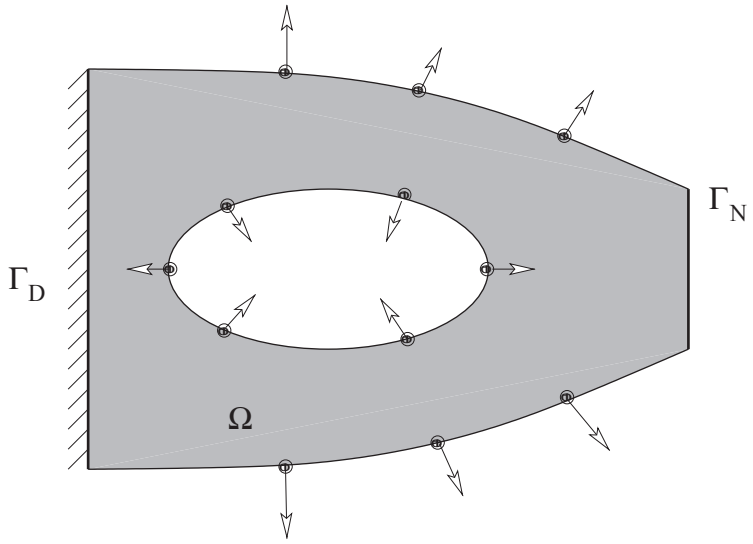


Fig. 3. Shape optimization of a membrane's shape.

From now on the membrane thickness is fixed, equal to 1. Moreover, we consider the parts Γ_D and Γ_N to be given. Thus the set of admissible shapes is

$$\mathcal{U}_{\text{ad}} = \{\Omega \subset \mathbb{R}^N : \Gamma_D \cup \Gamma_N \subset \partial\Omega \text{ and } |\Omega| = V_0\},$$

where $V_0 > 0$ is a given volume. The shape optimization problem reads

$$\inf_{\Omega \in \mathcal{U}_{\text{ad}}} J(\Omega),$$

with, as a criterion, the compliance

$$J(\Omega) = \int_{\Gamma_N} gu\, ds,$$

or a least-square functional to achieve a target displacement $u_0(x)$

$$J(\Omega) = \int_\Omega |u - u_0|^2\, dx.$$

Notice that the true optimization variable is only the free boundary Γ, and therefore the topology of the shape does not change.

Another example is a shape optimization in the elasticity setting. The model of linearized elasticity gives the displacement vector field $u : \Omega \to \mathbb{R}^N$ as the solution of the system of equations

$$\begin{cases} -\mathrm{div}(Ae(u)) = 0 & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_D, \\ (Ae(u)) \cdot n = g & \text{on } \Gamma_N, \\ (Ae(u)) \cdot n = 0 & \text{on } \Gamma, \end{cases}$$

with $e(u) = (\nabla u + (\nabla u)^t)/2$ and $A\xi = 2\mu\xi + \lambda(\mathrm{tr}\,\xi)\mathrm{Id}$, where μ and λ are the Lamé coefficients, and n is the outer unit normal to Ω. The boundary $\partial\Omega$ is again divided into three disjoint parts

$$\partial\Omega = \Gamma \cup \Gamma_D \cup \Gamma_N,$$

where Γ is the free boundary, the true optimization variable. The set of admissible shapes is again

$$\mathcal{U}_{\mathrm{ad}} = \{\Omega \subset \mathbb{R}^N : \Gamma_D \cup \Gamma_N \subset \partial\Omega \text{ and } |\Omega| = V_0\},$$

where V_0 is a given imposed volume. The objective function chosen is either the compliance

$$J(\Omega) = \int_{\Gamma_N} g \cdot u \, ds,$$

or a least-square criterion for the target displacement $u_0(x)$

$$J(\Omega) = \int_{\Omega} |u - u_0|^2 \, dx.$$

As before, the shape optimization problem reads

$$\inf_{\Omega \in \mathcal{U}_{\mathrm{ad}}} J(\Omega).$$

1.4 Topology optimization and the homogenization method

In topology optimization, not only the connected components of the boundary Γ are allowed to move but also new connected components (holes in 2-d) of Γ can appear or disappear. Topology is now optimized too. In order to solve this task, we introduce the homogenization method. The homogenization method is a kind of averaging methods for partial differential equations, and is commonly used to determine the averaged (or effective, or homogenized, or equivalent, or macroscopic) parameters of a heterogeneous medium [Al2002, BLP1978, Ch2000, CD1999, JKO1995, MT1997, TA2000].

How does homogenization apply to optimal design? The homogenization method is based on the concept of "relaxation": it makes ill-posed problems well-posed by enlarging the space of admissible "shapes." It is crucial to introduce "generalized" shapes, that are "not too generalized." In the homogenization method, we think of generalized shapes as "limits" of minimizing sequences of classical shapes. We can then say that homogenization allows, as admissible shapes, composite materials obtained by micro-perforation of the original material (fine mixtures of material and void).

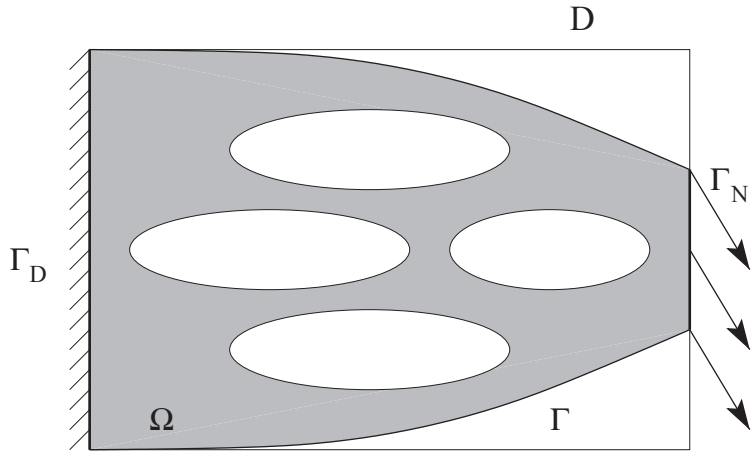


Fig. 4. Topology optimization of a membrane's shape.

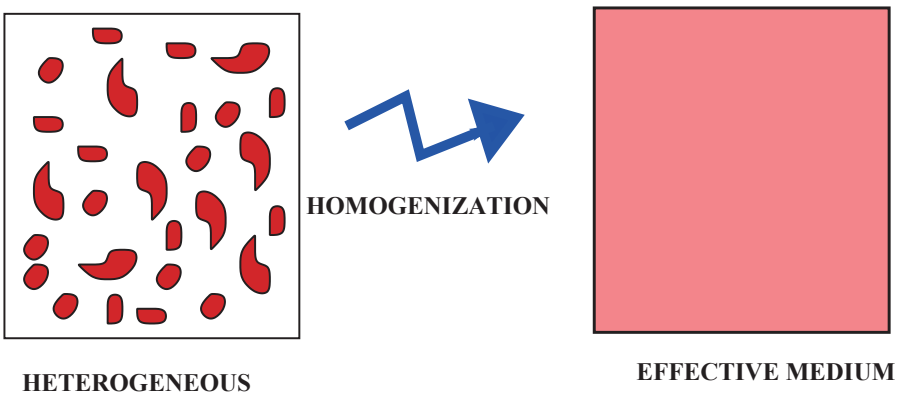


Fig. 5. Homogenization in a nutshell.

1.5 Lattice materials in additive manufacturing

Additive manufacturing, also known as 3D printing, is a process that creates physical structures built layer by layer from a digital design by using metallic powder melted by a laser or an electron beam [GRS2015]. One of the main advantages of additive manufacturing is that, a priori, there are no limitations on the structures that can be built (unfortunately, in practice there are some limitations of manufacturability, like overhangs or the possibility of thermal residual stresses). Moreover, one can even build microstructures or lattice materials.

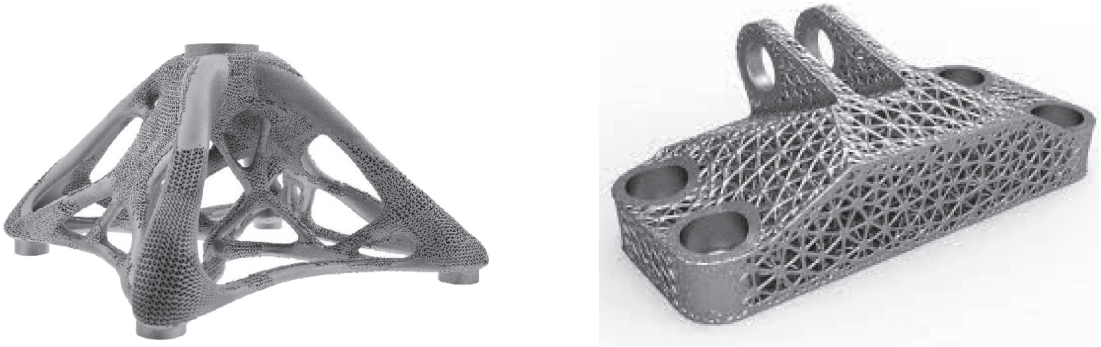


Fig. 6. Some examples of lattice structures. Left: an architectural spider bracket (https://altairenlighten.com/wp-content/uploads/2017/03/architectural-spider-bracket.jpg). Right: crystallon, lattice structures in Rhino and Grasshopper (https://noizear.com/crystallon-lattice-structures-in-rhino-and-grasshopper).

However, it is impossible to describe all the fine details of a lattice structure in a finite element model for optimization purposes. Therefore, homogenization theory is the right tool for dealing with lattice materials and related optimal design problems. In Sect. 6, we will tackle the problem of optimizing lattice structures by using the homogenization method.

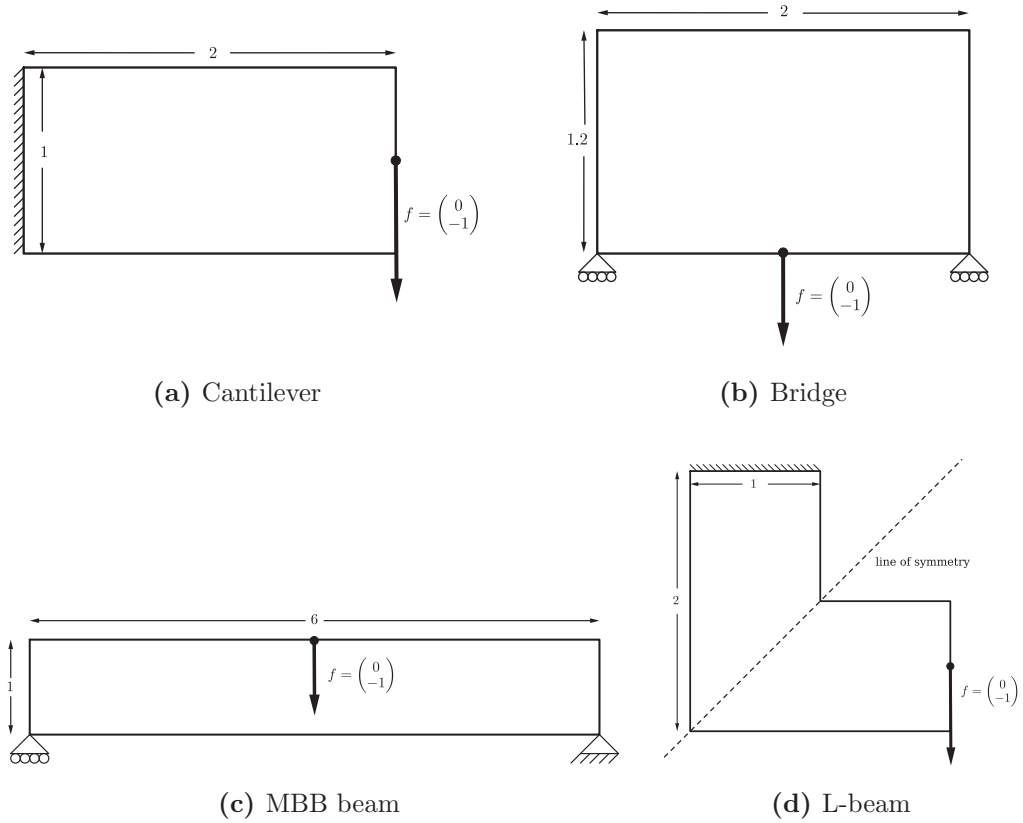
1.6 Goals of these lecture notes

The main goal of these lecture notes is to introduce the homogenization method for topology optimization of structures. The rest of these lecture notes are organized as follows. In Sect. 2, we show some tools in optimization and describe numerical algorithms for computing optimal designs. In Sect. 3, we consider parametric optimization problem and compute gradients of objective functions (by an optimal control approach) for further use in gradient-type algorithms. A representative example of parametric optimization is that of a membrane's thickness. In Sect. 4, we provide a brief survey on homogenization theory. In Sect. 5, we apply the homogenization method to topology optimization. In Sect. 6, we present a resurrection of the homogenization method for the design of lattice materials in additive manufacturing.

Through these lecture notes, numerical exercizes are proposed with the FreeFem++ code (http://www.freefem.org). FreeFem++ is a free software for solving partial differential equations by the finite element method [He2012]. Moreover, you can find some scripts of FreeFem++ for shape optimization in the web site (http://www.cmap.polytechnique.fr/~allaire/freefem_en.html) and in the corresponding educational paper [AP2006].

1.7 Exercises

Problem 1.7.1. *Solve (with FreeFem++) the elasticity equations for the following test cases: cantilever, bridge, MBB beam and L-beam (see Fig. 7).*



(a) Cantilever **(b)** Bridge

(c) MBB beam **(d)** L-beam

Fig. 7. The various boundary conditions of Problem 1.7.1. Here, the loads are to be intended as acting on a very small region of the boundary, around the points represented by the full black dots.

2. Some Tools in Optimization

We review some classical result in optimization theory. More details can be found in textbooks like [Al2007-2, BGLS2006, ET1999, NW1999].

2.1 Generalities

Let V be a Banach space and $K \subset V$ be a non-empty subset. Let $J : V \to \mathbb{R}$. We consider the following minimization problem

$$\inf_{v \in K} J(v).$$

Let us specify some basic definitions.

Definition 2.1. *An element u is called a local minimizer of J on K if*

$$u \in K \text{ and } \exists \delta > 0,\ \forall v \in K,\ \|v - u\| < \delta \Longrightarrow J(v) \geq J(u).$$

Moreover, an element u is called a global minimizer of J on K if

$$u \in K \text{ and } J(v) \geq J(u)\ \forall v \in K.$$

Definition 2.2. *A minimizing sequence of a function J on the set K is a sequence $(u^n)_{n \in \mathbb{N}} \subset K$ such that*

$$\lim_{n \to +\infty} J(u^n) = \inf_{v \in K} J(v).$$

By definition of the infimum value of J on K there always exists at least one minimizing sequence for J on K.

Let us consider the existence of minima for optimization problems in finite dimension. The following result guarantees the existence of a minimum.

Theorem 2.3. *Let K be a non-empty closed subset of $\mathbb{R}^N$ and J a continuous function from K to $\mathbb{R}$ satisfying the so-called "infinite at infinity" property, i.e.,*

$$\forall (u^n)_{n\in\mathbb{N}} \text{ sequence in } K,\ \lim_{n\to+\infty} \|u^n\| = +\infty \Longrightarrow \lim_{n\to+\infty} J(u^n) = +\infty.$$

Then there exists at least one minimizer of J on K. Furthermore, from each minimizing sequence of J over K one can extract a subsequence which converges to a minimum of J on K.

Proof. Let $(u^n)_{n\in\mathbb{N}}$ be a minimizing sequence for J over K. In particular, since J is infinite at infinity and the sequence $(J(u^n))_{n\in\mathbb{N}}$ is bounded, we conclude that $(u^n)_{n\in\mathbb{N}}$ must be bounded as well. Therefore, since closed bounded sets are compact in finite dimension, there exists a subsequence $(u^{n_k})_{k\in\mathbb{N}}$ that converges to a point $u \in \mathbb{R}^N$. Now, $u \in K$ because K is closed, and $J(u^{n_k})$ converges to $J(u)$ by continuity. We conclude that $J(u) = \lim_{k\to\infty} J(u^{n_k}) = \inf_K J$. □

Remark 2.4. *In an infinite dimensional vector space, a continuous function on a closed bounded set does not necessarily attain its minimum. For example, let $H^1(0,1)$ be the usual Sobolev space with the norm $\|v\| = (\int_0^1 (v'(x)^2 + v(x)^2)dx)^{\frac{1}{2}}$. Let*

$$J(v) = \int_0^1 ((|v'(x)| - 1)^2 + v(x)^2)dx.$$

One can check that J is continuous and "infinite at infinity." Nevertheless the minimization problem

$$\inf_{v\in H^1(0,1)} J(v)$$

does not admit a minimizer. Indeed, there exists no $v \in H^1(0,1)$ such that $J(v) = 0$ but, still,

$$\inf_{v\in H^1(0,1)} J(v) = 0.$$

To obtain it, we construct a minimizing sequence $(u^n)_{n\in\mathbb{N}}$ defined for, $n \geq 1$, by

$$u^n(x) = \begin{cases} x - \dfrac{k}{n} & \text{if } \dfrac{k}{n} \leq x \leq \dfrac{2k+1}{2n}, \\ \dfrac{k+1}{n} - x & \text{if } \dfrac{2k+1}{2n} \leq x \leq \dfrac{k+1}{n} \end{cases} \quad \text{for } 0 \leq k \leq n-1,$$

as Fig. 8.

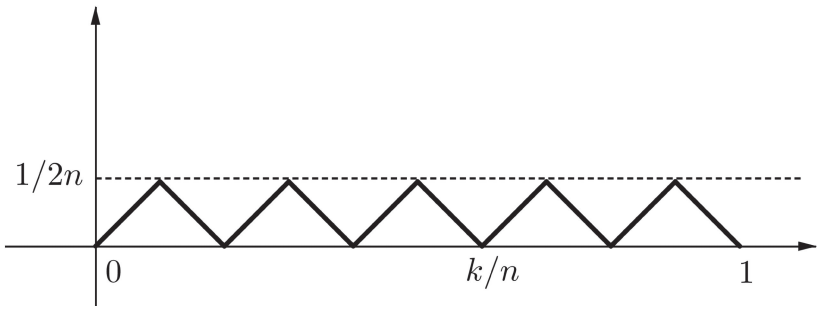


Fig. 8. The function u^n for $n = 5$.

We can easily check that $u^n \in H^1(0,1)$ and $(u^n)' = \pm 1$. Consequently,

$$J(u^n) = \int_0^1 u^n(x)^2 dx = \frac{1}{12n^2} \to 0.$$

We clearly see in this example that the minimizing sequence $(u^n)_{n\in\mathbb{N}}$ is "oscillating" more and more and it is not compact in $H^1(0,1)$ despite being bounded in the same space.

2.2 Convex analysis

As we have seen in Remark 2.4, continuous functions do not necessarily attain their minimum on a bounded closed set. In order to extend the result of Theorem 2.3 to the case of an infinite dimensional Hilbert space, we shall work in a convex framework.

Definition 2.5. *A set $K \subset V$ is said to be convex if, for any $x, y \in K$ and for any $\theta \in [0,1]$, the linear combination $\theta x + (1-\theta)y$ belongs to K.*

Definition 2.6. *A function J, defined from a non-empty convex set $K \subset V$ into $\mathbb{R}$ is convex on K if*

$$J(\theta u + (1-\theta)v) \leq \theta J(u) + (1-\theta)J(v) \quad \forall u, v \in K, \forall \theta \in [0,1]. \tag{2.1}$$

Furthermore, J is said to be strictly convex if the inequality above is strict whenever $u \neq v$ and $\theta \in (0,1)$.

Theorem 2.7. *Let K be a non-empty closed convex set in a reflexive Banach space V (i.e., the dual of V' is V itself), and J be a convex continuous function on K, which is "infinite at infinity," i.e.,*

$$\forall (u^n)_{n\in\mathbb{N}} \text{ sequence in } K, \lim_{n\to+\infty} \|u^n\| = +\infty \implies \lim_{n\to+\infty} J(u^n) = +\infty.$$

Then, there exists a minimizer of J in K.

Remark 2.8. *Theorem 2.7 remains true if V is just the dual of some separable Banach space. In particular it holds true when $V = L^p(\Omega)$ with $1 < p \leq \infty$.*

The proof will follow along the same lines as that of Theorem 2.3. However, the infinite dimensional case is much more delicate, since it relies on *weak convergence* and its relations with convexity. We refer to [Al2007-2, Theorem 9.2.7 and Remark 9.2.9] for a complete proof.

Proposition 2.9. *Under the hypotheses of Theorem 2.7, suppose that J is strictly convex. Then J has at most one minimizer.*

Proof. Suppose by contradiction that $u_1 \neq u_2$ are two distinct minimizers of J over the closed strictly convex set K. If we take $\theta = 1/2$ in (2.1), then we get

$$J\left(\frac{u_1+u_2}{2}\right) < \frac{1}{2}J(u_1) + \frac{1}{2}J(u_2) = \min_{v\in K} J(v),$$

which is contradicts the definition of minimum. □

Proposition 2.10. *If J is convex on the convex set K, then any local minimizer of J on K is a global minimizer.*

Proof. Let u be a local minimizer of J on K. Thus there exists $\delta > 0$ such that $J(v) \geq J(u)$ for any $v \in K \cap B(u,\delta)$. Let $w \in K \setminus B(u,\delta)$. Our aim is to show that $J(w) \geq J(u)$. Let $\theta \in (0,1]$ be such that $u + \theta(w-u) \in B(u,\delta)$. For example we can take $\theta = \frac{\delta}{\|w-u\|_K}$. Since $u + \theta(w-u) \in B(u,\delta)$, it follows that $J(u) \leq J(u+\theta(w-u))$, and by Jensen's inequality

$$J(u) \leq J(u+\theta(w-u)) \leq (1-\theta)J(u) + \theta J(w).$$

Thus, $J(u) \leq J(w)$ follows. □

Convexity is not the only tool to prove existence of minimizers. Another method is, for example, compactness.

2.3 Optimality conditions

In this section, we discuss optimality conditions for objective functions.

Definition 2.11. *Let V be a Banach space. A function J, defined from a neighborhood of $u \in V$ into $\mathbb{R}$, is said to be differentiable in the sense of Fréchet at u if there exists $L \in V'$ such that*

$$J(u+w) = J(u) + L(w) + o(w) \text{ with } \lim_{w\to 0} \frac{|o(w)|}{\|w\|} = 0.$$

We call L the differential (or derivative, or gradient) of J at u and we denote it by

$$L = J'(u), \quad \text{or} \quad L(w) = \langle J'(u), w\rangle_{V',V}. \tag{2.2}$$

Remark 2.12. *If V is a Hilbert space, its dual V' can be identified with V itself thanks to the Riesz representation theorem. Thus, there exists a unique $p \in V$ such that $\langle p, w\rangle = L(w)$. We also write $p = J'(u)$. We use this identification $V = V'$ if $V = \mathbb{R}^n$ or $V = L^2(\Omega)$. In practice, it is often easier to compute the directional derivative $j'(0) = \langle J'(u), w\rangle_{V',V}$ with $j(t) = J(u+tw)$.*

Consider the variational formulation

$$\text{find } u \in V \text{ such that } a(u,w) = L(w) \quad \forall w \in V, \tag{2.3}$$

where a is a symmetric coercive continuous bilinear form and L is a continuous linear form. By the Lax–Milgram theorem we know that the variational formulation (2.3) admits a unique solution. Let us now define the energy

$$J(v) = \frac{1}{2}a(v,v) - L(v).$$

The following lemma tells us the relationship between the energy J and the variational formulation (2.3).

Lemma 2.13. *Let $u \in V$ be the unique solution of the variational formulation (2.3). Then u is the unique minimizer of J, that is,*

$$J(u) = \min_{v \in V} J(v).$$

Conversely, if $u \in V$ is a point giving an energy minimum of $J(v)$, then u is the unique solution of the variational formulation (2.3).

Proof. If u is the solution of the variational formulation (2.3), then thanks to the symmetry of a we have

$$J(u+v) = J(u) + \frac{1}{2}a(v,v) + a(u,v) - L(v) = J(u) + \frac{1}{2}a(v,v) \geq J(u).$$

As $u + v$ is arbitrary in V, u minimizes the energy J in V.

Conversely, let $u \in V$ be such that

$$J(u) = \min_{v \in V} J(v).$$

For $v \in V$ we define $j(t) = J(u + tv)$. Then

$$j(t) = \frac{t^2}{2}a(v,v) + t(a(u,v) - L(v)) + J(u).$$

We differentiate $t \mapsto j(t)$,

$$j'(t) = ta(v,v) + (a(u,v) - L(v)).$$

By definition, $j'(0) = \langle J'(u), v\rangle_{V',V}$, thus

$$\langle J'(u), v\rangle_{V',V} = a(u,v) - L(v).$$

Since $t = 0$ is a minimum point of j, we have $a(u,v) = L(v)$ for all $v \in V$. □

Remark 2.14. *When computing the Fréchet differential of a given functional J at u (see the definition of L and $w \mapsto L(w)$ in (2.2)), there is not always an obvious way to deduce a formula for $J'(u)$, nevertheless most of the time it is enough to know the mapping $w \mapsto \langle J'(u), w\rangle$.*

Example 2.15. 1. *For fixed $f \in L^2(\Omega)$, define*

$$J(v) = \int_\Omega \left(\frac{1}{2}v^2 - fv\right) dx, \quad v \in L^2(\Omega).$$

We have

$$\langle J'(u), w\rangle = \int_\Omega (uw - fw)\, dx.$$

Thus

$$J'(u) = u - f \in L^2(\Omega).$$

Notice that here we identified $L^2(\Omega)$ with its dual.

2. *For fixed $f \in L^2(\Omega)$ define*

$$J(v) = \int_\Omega \left(\frac{1}{2}|\nabla v|^2 - fv\right) dx, \quad v \in H_0^1(\Omega).$$

We have

$$\langle J'(u), w\rangle = \int_\Omega (\nabla u \cdot \nabla w - fw) dx.$$

Therefore, by the usual definition of the duality pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$ (that comes from a formal integration by parts) we get

$$J'(u) = -\Delta u - f \in H^{-1}(\Omega) = (H_0^1(\Omega))'.$$

Notice that here the space $H_0^1(\Omega)$ is **not** *identified with its dual.*

Remark 2.16. *If instead of the "usual" scalar product in L^2 we rather use the H^1 scalar product in the second part of Example 2.15, then we have to identify $J'(u)$ with a different function (in other words, the definition of $J'(u)$ depends on the scalar product used). From the directional derivative*

$$\langle J'(u), w\rangle = \int_\Omega (\nabla u \cdot \nabla w - fw)\, dx,$$

using the H^1 scalar product $\langle \phi, w\rangle = \int_\Omega (\nabla\phi \cdot \nabla w + \phi w)\, dx$, we deduce that

$$-\Delta J'(u) + J'(u) = -\Delta u - f, \quad J'(u) \in H_0^1(\Omega)$$

in the distributional sense. Here we identify $H_0^1(\Omega)$ with its dual.

Theorem 2.17 (Euler inequality)**.** *Let K be a convex Banach space. Take $u \in K$ and let $J : K \to \mathbb{R}$ be differentiable at u. If u is a local minimizer of J in K, then*

$$\langle J'(u), v - u\rangle \geq 0 \quad \forall v \in K. \tag{2.4}$$

On the other hand, if $u \in K$ satisfies this inequality and J is convex, then u is a global minimizer of J in K.

Proof. For $v \in K$ and $\delta \in (0, 1]$, we have $u + \delta(v - u) \in K$. Thus, if δ is sufficiently small, since u is a local minimizer of J in K, we have

$$\frac{J(u + \delta(v - u)) - J(u)}{\delta} \geq 0.$$

We obtain inequality (2.4) by letting $\delta \to 0$ in the above.

We will now prove the second claim of the theorem. Since, by hypothesis J is convex on K, then, the graph of J always lies above its tangent plane at any point $w \in K$. In other words, the following inequality holds true for all $v \in K$:

$$J(v) \geq J(w) + \langle J'(w), v - w\rangle.$$

The conclusion follows by taking $w = u$. □

Remark 2.18. *If u belongs to the interior of K, then we deduce the Euler equation $J'(u) = 0$.*

Remark 2.19. *The Euler inequality is usually just a necessary condition (for instance, it is verified also if u is a local maximizer). It becomes a necessary and sufficient condition under the further assumption that the functional J is also convex.*

2.3.1 Minimization with equality constraints

We consider the following problem

$$\inf_{v \in V, F(v)=0} J(v), \tag{2.5}$$

where $F = (F_1, \ldots, F_M)$ is a differentiable function from V into $\mathbb{R}^M$.

Notice that the set $K = \{v \in V : F(v) = 0\}$ is not necessarily convex. We will therefore need a generalized version of the Euler inequality as stated in Theorem 2.17. To this end we introduce the set of admissible directions for our constrained optimization problem.

Definition 2.20. *At every point $v \in K$, the set*

$$K(v) = \left\{ w \in V : \begin{array}{c} \exists (v^n)_{n\in\mathbb{N}} \subset K,\ \exists (\varepsilon^n)_{n\in\mathbb{N}} \subset (0,\infty), \\ \lim_{n\to\infty} v^n = v,\ \lim_{n\to\infty} \varepsilon^n = 0,\ \lim_{n\to\infty} (v^n - v)/\varepsilon^n = w \end{array} \right\}$$

is called the cone of admissible directions at the point v.

In other words, $K(v)$ is the set of all vectors that are tangent at v to a curve in K that passes through v (hence, if K is a regular manifold, $K(v)$ coincides with the tangent space to K at v). Moreover, notice that, as the name suggests, the set $K(v)$ is a *cone* in the sense of convex analysis: namely, for all $\lambda \geq 0$ and $w \in K(v)$, then also $\lambda w \in K(v)$.

Proposition 2.21 (Euler inequality, general case)**.** *Let u be a local minimum of J over K. Then, if J is differentiable at u, we have*

$$\langle J'(u), w\rangle \geq 0 \quad \forall w \in K(v).$$

Proof. With the same notations of Definition 2.20, set $w^n = (v^n - v)/\varepsilon^n$. By definition, we have that $w \in K(v)$ if and only if there exists a sequence $(w^n)_{n\in\mathbb{N}}$ in V and a sequence of positive real numbers $(\varepsilon^n)_{n\in\mathbb{N}}$ such that

$$\lim_{n\to\infty} w^n = w, \quad \lim_{n\to\infty} \varepsilon^n = 0, \quad \text{and} \quad v + \varepsilon^n w^n \in K \quad \forall n \in \mathbb{N}.$$

Now, since u is a local minimum of J over K, we get

$$\frac{J(u+\varepsilon^n w^n)-J(u)}{\varepsilon^n} \geq 0 \quad \text{for } n \text{ large enough.}$$

Passing to the limit as $n \to \infty$ yields

$$\langle J'(u), w\rangle \geq 0 \quad \forall w \in K$$

as claimed. □

Definition 2.22. *We call Lagrangian of problem (2.5), the function*

$$\mathcal{L}(v,\mu) = J(v) + \sum_{i=1}^{M} \mu_i F_i(v) = J(v) + \mu \cdot F(v) \quad \forall (v,\mu) \in V \times \mathbb{R}^M.$$

The new variable $\mu \in \mathbb{R}^M$ is called Lagrange multiplier for the constraint $F(v) = 0$.

Lemma 2.23. *The constrained minimization problem (2.5) admits the following equivalent formulation using the Lagrangian:*

$$\inf_{v\in V, F(v)=0} J(v) = \inf_{v\in V} \sup_{\mu\in\mathbb{R}^M} \mathcal{L}(v,\mu).$$

Proof. The proof is done by cases. Notice that, if $F(v) = 0$, then $J(v) = \mathcal{L}(v,\mu)$ for all $\mu \in \mathbb{R}^M$. On the other hand, if $F(v) \neq 0$, then $\sup_{\mu\in\mathbb{R}^M} \mathcal{L}(v,\mu) = +\infty$. Putting the two together yields

$$\begin{aligned}\inf_{v\in V} \sup_{\mu\in\mathbb{R}^M} \mathcal{L}(v,\mu) &= \min\left(\inf_{v\in V, F(v)=0} \sup_{\mu\in\mathbb{R}^M} \mathcal{L}(v,\mu),\ \inf_{v\in V, F(v)\neq 0} \sup_{\mu\in\mathbb{R}^M} \mathcal{L}(v,\mu) \right) \\ &= \min\left(\inf_{v\in V, F(v)=0} J(v),\ +\infty \right) = \inf_{v\in V, F(v)=0} J(v).\end{aligned}$$

□

Theorem 2.24 (Stationarity of the Lagrangian)**.** *With the same notation of (2.5), assume that J and F are continuously differentiable in a neighborhood of $u \in V$ such that $F(u) = 0$. If u is a local minimizer and if the vectors $(F_i'(u))_{1\leq i\leq M}$ are linearly independent, then there exist a Lagrange multiplier $\lambda \in \mathbb{R}^M$ such that*

$$\frac{\partial \mathcal{L}}{\partial v}(u,\lambda) = J'(u) + \lambda \cdot F'(u) = 0 \quad \textit{and} \quad \frac{\partial \mathcal{L}}{\partial \mu}(u,\lambda) = F(u) = 0. \tag{2.6}$$

Proof. First define $K = \{v \in V : F(v) = 0\}$ and then the corresponding cone of admissible directions $K(u)$ by Definition 2.20. Now, since the vectors $(F_i'(u))_{1\leq i\leq M}$ are linearly independent by hypothesis, we can use the implicit function theorem in a standard way to deduce that

$$K(u) = \{w \in V : \langle F_i'(u), w\rangle = 0 \text{ for } i = 1,\dots,M\},$$

or equivalently

$$K(u) = \bigcap_{i=1}^{M} [F_i'(u)]^{\perp}.$$

In particular $K(u)$ is a vector space (it is indeed the tangent space to the variety K at the point u). Thus we can successively take w and $-w$ in Proposition 2.21 to get

$$\langle J'(u), w\rangle = 0 \quad \forall w \in \bigcap_{i=1}^{M} [F_i'(u)]^{\perp}.$$

This implies that $J'(u)$ is generated by $(F_i'(u))_{1\leq i\leq M}$ (moreover, since the $F_i'(u)$ are linearly independent, the Lagrange multipliers μ_i are uniquely defined). □

2.3.2 Minimization with inequality constraints

We consider the following minimization problem with inequality constraints

$$\inf_{v\in V, F(v)\leq 0} J(v), \tag{2.7}$$

where $F(v) \leq 0$ here means that $F_i(v) \leq 0$ for $1 \leq i \leq M$, with $F = (F_1,\dots,F_M) : V \to \mathbb{R}^M$ differentiable.

Definition 2.25. *Let u be such that $F(u) \leq 0$. The set*

$$I(u) = \{i \in \{1,\dots,M\} : F_i(u) = 0\}$$

is called the set of active constraints at u. The inequality constraints are said to be qualified at $u \in K$ if the vectors $(F_i'(u))_{i\in I(u)}$ are linearly independent.

There are other (more general) definitions of constraints qualification [BGLS2006].

Definition 2.26. *We call Lagrangian of the previous problem the function*

$$\mathcal{L}(v,\mu) = J(v) + \sum_{i=1}^{M} \mu_i F_i(v) = J(v) + \mu \cdot F(v) \quad \forall (v,\mu) \in V \times (\mathbb{R}_{\geq 0})^M.$$

The new **non negative** *variable $\mu \in (\mathbb{R}_{\geq 0})^M$ is called Lagrange multiplier for the constraint $F(v) \leq 0$.*

The proof of the result below is analogous to that of Lemma 2.23 and thus will be omitted.

Lemma 2.27. *The constrained minimization problem (2.7) is equivalent to*

$$\inf_{v\in V,\ F(v)\leq 0} J(v) = \inf_{v\in V} \sup_{\mu\in(\mathbb{R}_{\geq 0})^M} \mathcal{L}(v,\mu).$$

The existence of (non negative) Lagrange multipliers, analogous to Theorem 2.24, can be proved also for a minimization problem subject to inequality constraints. We refer to [Al2007-2, Theorem 10.2.15] for a proof.

Theorem 2.28 (Stationarity of the Lagrangian for the inequality constraint)**.** *We assume that the constraints are qualified at u satisfying $F(u) \leq 0$. If u is a local minimizer, there exist Lagrange multipliers $\lambda_1, \ldots, \lambda_M \geq 0$ such that*

$$J'(u) + \sum_{i=1}^{M} \lambda_i F_i'(u) = 0, \quad \lambda_i \leq 0, \quad \lambda_i = 0 \text{ if } F_i(u) < 0 \quad \forall i \in \{1,\ldots,M\}. \tag{2.8}$$

The condition (2.8) is indeed the stationarity of the Lagrangian since

$$\frac{\partial \mathcal{L}}{\partial v}(u,\lambda) = J'(u) + \lambda \cdot F'(u) = 0,$$

and the condition $F(u) \leq 0$ and $\lambda \cdot F(u) = 0$ for $\lambda \geq 0$, is equivalent to the Euler inequality (Theorem 2.17) associated to the maximization problem $\sup \mathcal{L}(u,\mu)$ with respect to the variable μ in the closed convex set $(\mathbb{R}_{\geq 0})^M$. Indeed

$$\frac{\partial \mathcal{L}}{\partial \mu}(u,\lambda) \cdot (\mu - \lambda) = F(u) \cdot (\mu - \lambda) \leq 0 \quad \forall \mu \in (\mathbb{R}_{\geq 0})^M,$$

and thus $F(u) \cdot \mu \leq F(u) \cdot \lambda = 0$ for all $\mu \in (\mathbb{R}_{\geq 0})^M$ as claimed.

2.3.3 Interpreting the Lagrange multipliers

Define the Lagrangian for the minimization of $J(v)$ under the constraint $F(v) = c$ as follows:

$$\mathcal{L}(v,\mu,c) = J(v) + \mu \cdot (F(v) - c).$$

We claim that the value of the Lagrange multiplier represents the sensitivity of the minimal value with respect to variations of the constraint c. To this end, let $u(c)$ and $\lambda(c)$ denote the minimizer and the optimal Lagrange multiplier respectively. Moreover, we assume that they are differentiable with respect to c. Then

$$\nabla_c J(u(c)) = -\lambda(c).$$

In other words, λ gives the derivative of the minimal value with respect to c without any further calculation. Indeed

$$\nabla_c J(u(c)) = \nabla_c \mathcal{L}(u(c),\lambda(c),c) = \frac{\partial \mathcal{L}}{\partial v} \nabla_c u(c) + \frac{\partial \mathcal{L}}{\partial \mu} \nabla_c \mu(c) + \frac{\partial \mathcal{L}}{\partial c} = -\lambda(c),$$

where, in the last equality we used

$$\frac{\partial \mathcal{L}}{\partial v}(u(c),\lambda(c),c) = 0 \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial \mu}(u(c),\lambda(c),c) = 0,$$

which are a consequence of Theorem 2.24 and the constraint $F(u(c)) = c$ respectively.

2.4 Dual energy

In this section, we shall associate to a minimizing problem with a maximizing problem, so called dual problem. To simplify the argument, we will assume that V and Y are two Banach spaces. Let V' and Y' be the corresponding dual spaces. The following argument is according to [ET1999] and see the book for the more general setting. For $J: V \to \mathbb{R} \cup \{\infty\}$, we consider the following minimizing problem:

$$\inf_{v\in V} J(v). \tag{2.9}$$

For given problem (2.9), we are now able to define a dual problem. We shall consider a function $\Phi: V \times Y \to \mathbb{R} \cup \{\infty\}$ such that

$$\Phi(v, 0) = J(v), \quad v \in V.$$

We define the conjugate function $\Phi^*: V' \times Y' \to \mathbb{R} \cup \{\infty\}$ as

$$\Phi^*(v^*, p^*) := \sup_{(v,p)\in V\times Y} \{\langle v^*, v\rangle + \langle p^*, p\rangle - \Phi(v, p)\}, \quad (v^*, p^*) \in V' \times Y'.$$

We call the problem

$$\sup_{p^*\in Y'} \{-\Phi^*(0, p^*)\} \tag{2.10}$$

the dual problem of (2.9).

In the following, we will mention the relationship between (2.9) and (2.10) in a special case. Let $\Lambda: V \to Y$ be a continuous linear operator. Assume that J can be rewritten as

$$J(v) = \tilde{J}(v, \Lambda v), \quad v \in V,$$

where $\tilde{J}$ is a function of $V \times Y$ into $\mathbb{R} \cup \{\infty\}$. In this case, the function Φ will be

$$\Phi(v, p) := \tilde{J}(v, \Lambda v - p), \quad (v, p) \in V \times Y.$$

Then the conjugate function Φ^* becomes

$$\begin{aligned}\Phi^*(0, p^*) &= \sup_{(v,p)\in V\times Y} \{\langle p^*, p\rangle - \tilde{J}(v, \Lambda v - p)\} \\ &= \sup_{v\in V} \sup_{q\in Y} \{\langle p^*, \Lambda v\rangle - \langle p^*, q\rangle - \tilde{J}(v, q)\} \\ &= \sup_{(v,q)\in V\times Y} \{\langle \Lambda^* p^*, v\rangle - \langle p^*, q\rangle - \tilde{J}(v, q)\}.\end{aligned}$$

For this case, we can see the following relationship.

Theorem 2.29. *Assume that $\tilde{J}$ is convex and (2.9) is finite. We also assume that there exists $v_0 \in V$ such that $\tilde{J}(v_0, \Lambda v_0) < \infty$ and the function $p \mapsto \tilde{J}(v_0, p)$ is continuous at Λv_0. Then*

$$\inf_{v\in V} J(v) = \sup_{p^*\in Y'} \{-\Phi^*(0, p^*)\}$$

and maximizing problem (2.10) has at least one solution.

To show Theorem 2.29, we will use convex analysis. For the details of the proof, see [ET1999, Sect. III, Theorem 4.1].

Example 2.30. *We show an application of Theorem 2.29. Let $\Omega \subset \mathbb{R}^N$ be a smooth domain. We consider the Dirichlet problem*

$$\begin{cases} -\operatorname{div}(h\nabla u) = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where $f \in L^2(\Omega)$ and $h: \Omega \to \mathbb{R}$ is a positive given function. The solution u of the problem above is the minimizer of

$$\frac{1}{2}\int_\Omega h|\nabla v|^2\,dx - \int_\Omega fv\,dx, \quad v \in H_0^1(\Omega).$$

We can apply Theorem 2.29 with

$$V = H_0^1(\Omega), \quad Y = L^2(\Omega)^N, \quad \Lambda = \nabla, \quad \tilde{J}(v, p) = \frac{1}{2}\int_\Omega h|p|^2\,dx - \int_\Omega fv\,dx.$$

In the case, we see that

$$\begin{aligned}\Phi^*(0, p^*) &= \sup_{v\in H_0^1(\Omega)} \sup_{q\in L^2(\Omega)^N} \left\{\int_\Omega \left(p^*\cdot\nabla v + fv - \frac{1}{2}h|q|^2 - p^*\cdot q\right)dx\right\} \\ &= \begin{cases} \dfrac{1}{2}\displaystyle\int_\Omega h^{-1}|p^*|^2\,dx & \text{if } -\operatorname{div} p^* = f, \\ \infty & \text{otherwise} \end{cases}\end{aligned}$$

and hence

$$\sup_{p^* \in Y'} \{-\Phi^*(0, p^*)\} = - \inf_{\substack{p^* \in L^2(\Omega)^N \\ -\operatorname{div} p^* = f}} \int_\Omega h^{-1} |p^*|^2 \, dx.$$

2.5 Numerical algorithms

In this section we present some numerical algorithms in order to solve the kind of minimization problems that were treated in this section. All these algorithms are of iterative nature: starting from a give initial value u^0, we construct a sequence $(u^n)_{n \in \mathbb{N}}$, which can be shown to converge to the solution u of the given minimization problem under some hypotheses.

2.5.1 A gradient-type algorithm (non-constrained case)

Suppose that $V = \mathbb{R}^N$ (or, more generally, a Hilbert space, that we will identify with its dual V'). We consider the following minimization problem without constraints:

$$\inf_{v \in V} J(v). \tag{2.11}$$

We initialize the algorithm by choosing some initial value $u^0 \in V$ and iteratively update it as follows:

$$u^{n+1} = u^n - \mu J'(u^n), \tag{2.12}$$

where μ is a positive parameter that we choose in advance (a more sophisticate algorithm involving the optimal choice of $\mu = \mu^n$ for each iteration is discussed in [Al2007-1, Theorem 3.38]).

Theorem 2.31. *Let V be a Hilbert space and suppose that the functional $J : V \to \mathbb{R}$ is strongly convex, that is, for some $\alpha > 0$*

$$\langle J'(u) - J'(v), u - v \rangle \geq \alpha \|u - v\|^2 \quad \forall u, v \in V.$$

Moreover, assume that J is differentiable with Lipschitz continuous derivative J'. Then, if μ is small enough (depending on α and on the Lipschitz constant of J'), the gradient-type algorithm described above converges. In other words, for all u^0, the sequence $(u^n)_{n \in \mathbb{N}}$ defined in (2.12) converges to the solution u of (2.11).

For a proof, see [Al2007-2].

Remark 2.32. *Choosing the right step length is not an easy task. Let us use the line search strategy as follows: start with a given step $\mu^0 > 0$. Now, at each iteration, increase the current step, $\mu_{n+1} = 1.1 \times \mu_n$, if J decreases, and reduce it, $\mu_{n+1} = 0.5 \times \mu_n$ if J increases.*

2.5.2 A gradient-type algorithm (constrained case)

Suppose that J is a real valued strictly convex differentiable functional defined on a nonempty closed convex subset K of the Hilbert space V. The set K represents the imposed constraints. We consider the following minimization problem

$$\inf_{v \in K} J(v). \tag{2.13}$$

Theorem 2.3 ensures the existence of a minimizer u for (2.13) (which is unique by Proposition (2.9)). Moreover, according to Theorem 2.17, the minimizer u is characterized by the condition

$$\langle J'(u), v - u \rangle \geq 0 \quad \forall v \in K.$$

Notice that the condition above can be rephrased as follows. For all $\mu > 0$

$$\langle u - (u - \mu J'(u)), v - u \rangle \geq 0 \quad \forall v \in K. \tag{2.14}$$

Let $P_K : V \to K$ denote the projection operator onto the convex subset K. Then (2.14) just states that u is the orthogonal projection of $u - \mu J'(u)$ onto K. In other words

$$u = P_K(u - \mu J'(u)) \quad \forall \mu > 0.$$

Therefore we devise a (projected) gradient-type algorithm, defined by the following iteration

$$u^{n+1} = P_K(u^n - \mu J'(u^n)), \tag{2.15}$$

where μ is a fixed positive parameter.

Theorem 2.33. *Let J be a differentiable strongly convex functional, with derivative J' Lipschitz continuous on V. Then, if μ is small enough, the projected gradient algorithm with fixed step defined above converges. In other words, for all initial values $u^0 \in K$, the sequence $(u^n)_{n \in \mathbb{N}}$ defined by (2.15) converges to the solution u of (2.13).*

We refer to [Al2007-2, Theorem 10.5.8] for a proof.

Remark 2.34. *Another possibility is to penalize the constraints, i.e., for small ε we replace the problem*

$$\inf_{v\in V,\, F(v)\le 0} J(v) \quad \text{by} \quad \inf_{v\in V}\left\{ J(v) + \frac{1}{\varepsilon}\sum_{i=1}^{M} (\max(F_i(v),0)^2 \right\}.$$

Example 2.35 (Some projection operators $\boldsymbol{P_K}$)**.** *Here we present some projection operators that can be computed explicitly.*

- *If $V = \mathbb{R}^M$ and $K = \prod_{i=1}^M [a_i, b_i]$, then for $x = (x_1, \ldots, x_M) \in \mathbb{R}^M$ we have*
$$P_K(x) = y \quad \text{with} \quad y_i = \min(\max(a_i, x_i), b_i) \quad \text{for } 1 \le i \le M.$$
- *If $V = \mathbb{R}^M$ and $K = \{x \in \mathbb{R}^M : \sum_{i=1}^M x_i = c_0\}$, then*
$$P_K(x) = y \quad \text{with} \quad y_i = x_i - \lambda, \quad \lambda = \frac{1}{M}\left(-c_0 + \sum_{i=1}^{M} x_i\right).$$
- *Similarly, if $V = L^2(\Omega)$ and $K = \{\phi \in V :\ a(x) \le \phi(x) \le b(x)\}$ or $K = \{\phi \in V : \int_\Omega \phi\, dx = c_0\}$ the corresponding projection operators P_K can be obtained by replacing finite sums with integrals in the two examples above.*

For more general closed convex sets K, the corresponding projection operator P_K can be very hard to determine. In such cases one can use the so called Uzawa algorithm [Al2007-2] which looks for a saddle point of the Lagrangian.

2.6 Exercises

Problem 2.6.1. *For a given $f \in L^2(\Omega)$, Ω being a rectangle in 2D, solve the following optimization problem numerically under the constraints $0 \le u(x) \le 1$ and $\int_\Omega u\, dx = |\Omega|/2$:*

$$\min_{u\in L^2(\Omega)} \int_\Omega |u - f|^2\, dx.$$

Problem 2.6.2. *For a given $f \in L^2(\Omega)$, Ω being a rectangle in 2D, and $\varepsilon > 0$, solve the following optimization problem numerically under the constraints $0 \le u(x) \le 1$:*

$$\min_{u\in L^2(\Omega)} \int_\Omega (|u - f|^2 + \varepsilon^2 |\nabla u|^2)\, dx.$$

3. Parametric Optimal Design

3.1 Optimization of a membrane's thickness

In this section, we consider a parametric optimal design problem of a membrane. Let Ω be a bounded domain in $\mathbb{R}^N$ ($N \ge 2$) and $f \in L^2(\Omega)$ be external forces. Let us consider the displacement $u \in H_0^1(\Omega)$, defined as the solution of

$$\begin{cases} -\mathrm{div}(h\nabla u) = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{3.1}$$

where $h = h(x)$ is the thickness of membrane. Note that Lax–Milgram theorem ensures that there exists a unique solution $u \in H_0^1(\Omega)$ of (3.1) if $f \in L^2(\Omega)$. For some given constants $0 < h_{\min} < h_0 < h_{\max}$, we seek to optimize the membrane by varying its thickness $h(x)$ in the admissible set defined by

$$\mathcal{U}_{\mathrm{ad}} = \left\{ h \in L^\infty(\Omega) : 0 < h_{\min} \le h(x) \le h_{\max} \text{ a.e. in } \Omega,\ \int_\Omega h(x)dx = h_0|\Omega| \right\}.$$

We consider the following parametric shape optimization problem:

$$\inf_{h\in\mathcal{U}_{\mathrm{ad}}} \left\{ J(h) = \int_\Omega j(u)\, dx \right\},$$

where u depends on h through the state equation (3.1), and j is a C^1 function from $\mathbb{R}$ to $\mathbb{R}$ such that $|j(u)| \le C(u^2 + 1)$ and $|j'(u)| \le C(|u| + 1)$. As examples of function j, we can take $j(u) = fu$ if we want to minimize the compliance (maximize the rigidity of the membrane), or $j(u) = |u - u_0|^2$ if we want to minimize the least-square criterion to reach a target displacement $u_0 \in L^2(\Omega)$.

Before studying the existence of an optimal thickness, we show the continuity of the cost function.

Proposition 3.1. *The application*

$$h \mapsto J(h) = \int_\Omega j(u)\,dx$$

is a continuous mapping from $\mathcal{U}_{\text{ad}}$ *into* $\mathbb{R}$.

Proof. The result follows immediately by composition of the two continuous functions that appear in the following lemmas: Lemma 3.2 and Lemma 3.3. □

Lemma 3.2. *The map* $v \mapsto \int_\Omega j(v)\,dx$ *is continuous from* $L^2(\Omega)$ *into* $\mathbb{R}$.

Proof. The result follows by the Lebesgue dominated convergence theorem. □

Lemma 3.3. *The map* $h \mapsto u$, *where* $u \in H_0^1(\Omega)$ *is the solution of (3.1), is a continuous function from* $\mathcal{U}_{\text{ab}}$ *into* $H_0^1(\Omega)$.

Proof. Let $(h_n)_{n\in\mathbb{N}} \subset \mathcal{U}_{\text{ab}}$ be a sequence converging in the L^∞-norm to some $h_\infty \in L^\infty(\Omega)$. Let $u_n \in H_0^1(\Omega)$ denote the unique solution of the membrane equation with associated thickness h_n:

$$\begin{cases} -\text{div}(h_n \nabla u_n) = f & \text{in } \Omega, \\ u_n = 0 & \text{on } \partial\Omega, \end{cases}$$

or the equivalent weak formulation

$$\int_\Omega h_n \nabla u_n \cdot \nabla\phi\,dx = \int_\Omega f\phi\,dx \quad \forall \phi \in H_0^1(\Omega). \tag{3.2}$$

We will prove that u_n is a Cauchy sequence in $H_0^1(\Omega)$ and thus it converges. Take $n, m \in \mathbb{N}$ and subtract the variational formulation for u_n (3.2) from that of u_m for fixed $\phi \in H_0^1(\Omega)$ to be chosen later. We get

$$\int_\Omega h_m \nabla(u_m - u_n) \cdot \nabla\phi\,dx = \int_\Omega (h_n - h_m)\nabla u_n \cdot \nabla\phi\,dx \quad \forall \phi \in H_0^1(\Omega).$$

Choosing $\phi = u_m - u_n$ we deduce

$$\|\nabla(u_m - u_n)\|_{L^2(\Omega)} \le \frac{C}{h_{\min}^2} \|f\|_{L^2(\Omega)} \|h_m - h_n\|_{L^\infty(\Omega)},$$

which proves the claim. □

3.2 Existence theories

The question of the existence of optimal shapes is far from simple. We cannot apply the results of Sect. 2 directly since $J(h)$ is not generally convex function. In fact, there exists no optimal shape in general. General counter-examples have been found by Murat [Mu1977]. It is an important issue because this non-existence phenomenon has dramatic consequences for the numerical computations. Thus the definition of the set $\mathcal{U}_{\text{ab}}$ of admissible designs has to be modified in order to obtain existence of optimal shapes. The main strategies employed to gain the existence of optimal shapes are discretization (when the admissible set is made finite dimensional), regularization (when the admissible set is made compact), and sometimes a miracle (when the given optimization problem happens to be convex).

3.2.1 Definition of a counter-example

First, let us show a counter-example to the existence of optimal design for the membrane problem. For simplicity, let $N = 2$ and $\Omega = (0, 1) \times (0, 1)$.

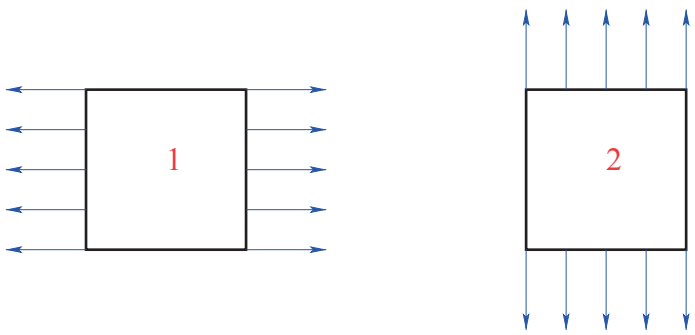


Fig. 9. The setting of the counter-example: we seek a membrane that is strong for horizontal loading (1) and weak for vertical loading (2).

We want to minimize the following objective function for $h \in \mathcal{U}_{\text{ad}}$:

$$J(h) = \int_{\partial\Omega} e_1 \cdot nu_1 \, ds - \int_{\partial\Omega} e_2 \cdot nu_2 \, ds, \tag{3.3}$$

where e_1, e_2 are the horizontal and vertical directions $(1,0)$, $(0,1)$ respectively and u_1, u_2 are the solutions of the following membrane problems:

$$\begin{cases} -\mathrm{div}(h\nabla u_1) = 0 & \text{in } \Omega, \\ h\nabla u_1 \cdot n = e_1 \cdot n & \text{on } \partial\Omega, \end{cases} \quad \begin{cases} -\mathrm{div}(h\nabla u_2) = 0 & \text{in } \Omega, \\ h\nabla u_2 \cdot n = e_2 \cdot n & \text{on } \partial\Omega. \end{cases}$$

When we minimize (3.3), we want the membrane to be strong for horizontal loading (we minimize compliance in the e_1 direction), and at the same time weak for vertical loading (we maximize the compliance in the direction e_2). This property of the objective function makes the problem ill-posed in the following sense.

Theorem 3.4. *The infimum of (3.3) is not attained by any $h \in \mathcal{U}_{\mathrm{ad}}$.*

Since the rigorous proof of Theorem 3.4 is a little bit technical, here we will only explain the main ideas by means of a "hand-waving argument." First of all, notice that if h is uniform (i.e., h is a constant function), then by definition the membrane is isotropic. Therefore, also the domain Ω is isotropic, that is to say that it shows the same mechanical behavior in all direction. However, it is better to build horizontal layers of alternating small and large thicknesses in order to minimize the objective function (3.3) (see Fig. 10). In other words, we are building a laminated structure that is horizontally strong but vertically weak. In order to intuitively justify this statement consider the following. Vertically, the lines of forces must cross the layers of minimal thickness: this means that the structure is thus weak with respect to vertical stress. On the other hand, horizontally, the lines of forces follow the layers of maximal thickness: this means that the structure is thus strong with respect to horizontal stress. However, since the boundary conditions are uniform, the membrane is horizontally stronger if the layers are finer, as the lines of forces are deviating from the horizontal to a lesser extent. If h oscillates at a small scale, we obtain an anisotropic composite material. To reach the minimum, the oscillation scale must go to 0. Therefore, there does not exist any real optimal design that does not involve a microstructure at an infinitely small scale. We refer the interested reader to Sect. 5.2 in [Al2007-1] for the details.

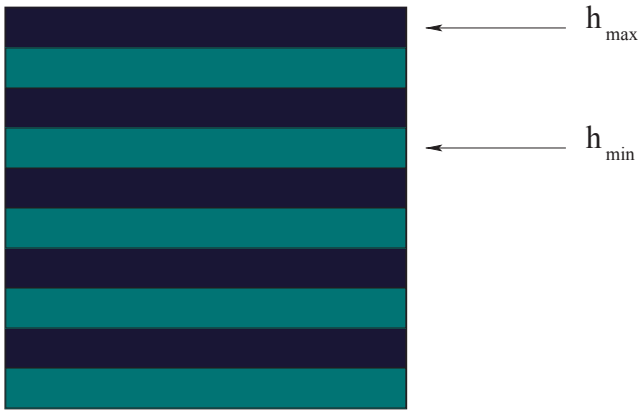


Fig. 10. Horizontal layers of alternating small and large thicknesses.

3.2.2 Existence for a discretized model

One way to avoid non-existence due to a loss of compactness consists in working with a discretized (and hence finite-dimensional) model. Let $(\omega_i)_{1\le i\le n}$ be a partition of Ω such that

$$\overline{\Omega} = \bigcup_{i=1}^{n} \overline{\omega}_i, \quad \omega_i \cap \omega_j = \emptyset \quad \text{for } i \neq j.$$

We introduce the subset $\mathcal{U}_{\mathrm{ad}}^n$ of $\mathcal{U}_{\mathrm{ab}}$ defined by

$$\mathcal{U}_{\mathrm{ab}}^n = \{h \in \mathcal{U}_{\mathrm{ab}} : h(x) \equiv h_i \in \mathbb{R} \text{ in } \omega_i, \quad 1 \le i \le n\}.$$

In other words, any function $h \in \mathcal{U}_{\mathrm{ab}}^n$ is uniquely determined by the choice of the vector $(h_i)_{1\le i\le n} \in \mathbb{R}^n$ and thus $\mathcal{U}_{\mathrm{ad}}^n$ is identified with a closed subset of $\mathbb{R}^n$.

Theorem 3.5 (Existence in finite dimension)**.** *The discretized optimization problem*

$$\inf_{h\in\mathcal{U}_{\mathrm{ad}}^n} J(h)$$

admits at least one minimizer.

Proof. Since $\mathcal{U}_{\text{ab}}^n$ is a compact subset of $\mathbb{R}^N$ and $J(h)$ is a continuous function on $\mathcal{U}_{\text{ab}}^n$, the existence of a minimizer of J in $\mathcal{U}_{\text{ab}}^n$ follows from Theorem 2.3. □

3.2.3 Existence with a regularity constraint

Another classical way of ensuring the existence of minimizers relies in imposing additional regularity. For example, consider the space $C^1(\overline{\Omega})$ which is a Banach space with the norm

$$\|\phi\|_{C^1(\overline{\Omega})} = \max_{x\in\overline{\Omega}}(|\phi(x)| + |\nabla\phi(x)|).$$

Take a given constant $R > 0$ and introduce the subspace $\mathcal{U}_{\text{ad}}^{\text{reg}}$:

$$\mathcal{U}_{\text{ad}}^{\text{reg}} = \{h \in \mathcal{U}_{\text{ad}} \cap C^1(\overline{\Omega}) : \|h\|_{C^1(\overline{\Omega})} \leq R\}.$$

The upper bound on the C^1-norm of h in the definition above can be interpreted as a "feasibility" (or "manufacturability") constraint, as, in practice, the thickness cannot vary too rapidly. Then the following theorem holds:

Theorem 3.6. *The regularized optimization problem*

$$\inf_{h\in\mathcal{U}_{\text{ad}}^{\text{reg}}} J(h)$$

admits at least one minimizer.

Proof. Consider a minimizing sequence $(h_n)_{n\in\mathbb{N}} \subset \mathcal{U}_{\text{ad}}^{\text{reg}}$ such that

$$\lim_{n\to\infty} J(h_n) = \inf_{h\in\mathcal{U}_{\text{ad}}^{\text{reg}}} J(h).$$

By definition, the sequence $(h_n)_{n\in\mathbb{N}}$ is bounded uniformly in n in the space $C^1(\overline{\Omega})$. We then apply a variant of Rellich theorem which states that one can extract a subsequence (still denoted by h_n for simplicity) that converges in $C^0(\overline{\Omega})$ to a limit function h_∞ (furthermore, we know that $h_\infty \in C^1(\overline{\Omega})$). We already know that $h \mapsto J(h)$ is a continuous mapping from $\mathcal{U}_{\text{ad}}$ into $\mathbb{R}$ by Proposition 3.1, therefore

$$\lim_{n\to\infty} J(h_n) = J(h_\infty),$$

which proves that h_∞ is a global minimizer of J in $\mathcal{U}_{\text{ad}}^{\text{reg}}$ as claimed. □

Remark 3.7. *Theorem 3.6 is actually a theorem of limited practical interest for the following reasons.*

- *In the practical cases, it is not clear how to choose the upper bound R in the definition of* $\mathcal{U}_{\text{ad}}^{\text{reg}}$.
- *Usually we do not have convergence as R goes to infinity.*
- *It is not clear whether, numerically, we have global or local minimizers.*
- *Numerically, an upper bound on the* H^1*-norm is preferred instead:*

$$\|h\|_{H^1(\Omega)} \leq R.$$

3.3 Computation of a continuous gradient

In this section, we will calculate the gradient of the objective function $J(h)$. This tells us the necessary conditions for optimality of the optimal shape and allows us to establish a numerical algorithm for calculating the optimal shape.

First, we consider the boundary value problem

$$\begin{cases} -\text{div}(h\nabla u) = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{3.4}$$

where h belong to the following convex set which is larger than $\mathcal{U}_{\text{ad}}$:

$$\mathcal{U} = \{h \in L^\infty(\Omega) : \exists h_0 > 0 \text{ such that } h(x) \geq h_0 \text{ a.e. in } \Omega\}.$$

Lemma 3.8. *The application* $h \mapsto u(h)$*, which gives the solution* $u(h) \in H_0^1(\Omega)$ *of (3.4) for* $h \in \mathcal{U}$*, is differentiable and its directional derivative at h in the direction* $k \in L^\infty(\Omega)$ *is given by*

$$\langle u'(h), k\rangle = v,$$

where v is the unique solution in $H_0^1(\Omega)$ *of*

$$\begin{cases} -\text{div}(h\nabla v) = \text{div}(k\nabla u) & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{3.5}$$

Proof. Formally, one simply differentiates equation (3.4) with respect to h. However, to be mathematically rigorous one should rather work at the level of the variational formulation. To compute the directional derivative with respect to

$k \in L^\infty(\Omega)$, we define $h(t) = h + tk$ for $t > 0$. For $t > 0$, let $u(t)$ be the solution for the thickness $h(t)$. Differentiating with respect to t leads to

$$\begin{cases} -\mathrm{div}(h(t)\nabla u'(t)) = \mathrm{div}(h'(t)\nabla u(t)) & \text{in } \Omega, \\ u'(t) = 0 & \text{on } \partial\Omega, \end{cases}$$

and, since $h'(0) = k$, we deduce $u'(0) = v$.

Let us justify the above calculation by showing that the map $h \mapsto u(h)$ is differentiable in the sense of Fréchet. First, there exists a unique solution v of (3.5) in $H_0^1(\Omega)$ thanks to the Lax–Milgram Theorem applied to the variational formulation

$$\int_\Omega h\nabla v \cdot \nabla\phi \, dx = -\int_\Omega k\nabla u \cdot \nabla\phi \, dx \quad \forall \phi \in H_0^1(\Omega). \tag{3.6}$$

We combine (3.6) with the following variational formulation for $u(t)$

$$\int_\Omega h(t)\nabla u(t) \cdot \nabla\phi \, dx = \int_\Omega f\phi \, dx \quad \forall \phi \in H_0^1(\Omega). \tag{3.7}$$

Since $u(1) = u(h + k)$ and $u(0) = u(h)$, we obtain by difference

$$\int_\Omega h\nabla(u(h+k) - u(h) - v) \cdot \nabla\phi \, dx = -\int_\Omega k\nabla(u(h+k) - u(h)) \cdot \nabla\phi \, dx.$$

Taking $\phi = u(h + k) - u(h) - v$ as a test function in the above yields

$$\begin{aligned} &\|\nabla(u(h+k) - u(h) - v)\|^2_{L^2(\Omega)} \\ &\quad = -\int_\Omega k\nabla(u(h+k) - u(h)) \cdot \nabla(u(h+k) - u(h) - v) \, dx \end{aligned} \tag{3.8}$$

which implies

$$\|\nabla(u(h+k) - u(h) - v)\|_{L^2(\Omega)} \le C\|k\|_{L^\infty(\Omega)}\|\nabla(u(h+k) - u(h))\|_{L^2(\Omega)}, \tag{3.9}$$

where we used Cauchy–Schwarz's inequality and the H_0^1 boundedness of v. Furthermore, by (3.7) we have

$$\int_\Omega (h+k)\nabla(u(h+k) - u(h)) \cdot \nabla\phi \, dx = -\int_\Omega k\nabla u(h) \cdot \nabla\phi \, dx. \tag{3.10}$$

Taking the test function as $\phi = u(h + k) - u(h)$ in (3.10), we obtain the following estimate:

$$\|\nabla(u(h+k) - u(h))\|_{L^2(\Omega)} \le C\|k\|_{L^\infty(\Omega)}. \tag{3.11}$$

Combining (3.8) with (3.11), we have

$$\|\nabla(u(h+k) - u(h) - v)\|_{L^2(\Omega)} \le C\|k\|^2_{L^\infty(\Omega)}.$$

Therefore we obtain $u(h + k) = u(h) + v + o(k)$ as $\|k\|_{L^\infty(\Omega)} \to 0$, which proves the claim. □

Lemma 3.9. *For $h \in \mathcal{U}$, let $u(h) \in H_0^1(\Omega)$ be the solution to (3.4) and*

$$J(h) = \int_\Omega j(u(h)) \, dx,$$

where j is a C^1 function from $\mathbb{R}$ into $\mathbb{R}$ such that $|j(u)| \le C(u^2 + 1)$ and $|j'(u)| \le C(|u| + 1)$ for any $u \in \mathbb{R}$. The application $J(h)$, from $\mathcal{U}$ into $\mathbb{R}$, is differentiable and its directional derivative at h in the direction $k \in L^\infty(\Omega)$ is given by

$$\langle J'(h), k\rangle = \int_\Omega j'(u(h))v \, dx,$$

where $v = \langle u'(h), k\rangle$ is the unique solution in $H_0^1(\Omega)$ of

$$\begin{cases} -\mathrm{div}(h\nabla v) = \mathrm{div}(k\nabla u) & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

Proof. By simple composition of differentiable applications. To justify it, one only has to check that all the terms are well defined. We omit the details of the proof. □

3.3.1 Adjoint state

In order to treat the derivative of the objective function $J(h)$, we introduce the adjoint state p, defined as the unique solution in $H_0^1(\Omega)$ of

$$\begin{cases} -\mathrm{div}(h\nabla p) = -j'(u) & \text{in } \Omega, \\ p = 0 & \text{on } \partial\Omega. \end{cases} \tag{3.12}$$

Theorem 3.10. *The cost function $J(h)$ is differentiable on $\mathcal{U}$ and*

$$J'(h) = \nabla u \cdot \nabla p.$$

If $h \in \mathcal{U}_{\mathrm{ad}}$ is a local minimizer of J in $\mathcal{U}_{\mathrm{ad}}$, then it satisfies the necessary optimality condition

$$\int_\Omega \nabla u \cdot \nabla p(k-h)\,dx \geq 0$$

for any $k \in \mathcal{U}_{\mathrm{ad}}$.

Proof. To make explicit $J'(h)$ from Lemma 3.9, we must eliminate $v = \langle u'(h), k\rangle$. To this end, we employ the use of the adjoint state, solution of (3.12). Multiplying the equation for v by p and that for p by v, we integrate by parts

$$\int_\Omega h\nabla p \cdot \nabla v\,dx = -\int_\Omega j'(u)v\,dx,$$

$$\int_\Omega h\nabla v \cdot \nabla p\,dx = -\int_\Omega k\nabla u \cdot \nabla p\,dx.$$

Comparing these two equalities we deduce

$$\langle J'(h), k\rangle = \int_\Omega j'(u)v\,dx = \int_\Omega k\nabla u \cdot \nabla p\,dx$$

for any $k \in L^\infty(\Omega)$. Since $\nabla u \cdot \nabla p$ belongs to $L^1(\Omega)$, we check that $J'(h)$ is continuous on $L^\infty(\Omega)$. To obtain the condition of optimality, it suffices to apply Theorem 2.17 since $\mathcal{U}_{\mathrm{ad}}$ is a closed non-empty convex subset of $L^\infty(\Omega)$. □

Remark 3.11 (How to find the adjoint state)**.** *For independent variable $(\hat{h}, \hat{u}, \hat{p}) \in L^\infty(\Omega) \times H_0^1(\Omega) \times H_0^1(\Omega)$, we introduce the Lagrangian*

$$\mathcal{L}(\hat{h}, \hat{u}, \hat{p}) = \int_\Omega j(\hat{u})\,dx + \int_\Omega \hat{p}(-\mathrm{div}(\hat{h}\nabla\hat{u}) - f)\,dx,$$

where $\hat{p}$ is a Lagrange multiplier (a function) for the constraint which connects u to h. By integration by parts we get

$$\mathcal{L}(\hat{h}, \hat{u}, \hat{p}) = \int_\Omega j(\hat{u})\,dx + \int_\Omega \left(\hat{h}\nabla\hat{p} \cdot \nabla\hat{u} - f\hat{p}\right)dx.$$

The partial derivative of $\mathcal{L}$ at $\hat{u} = u$ in the direction $\phi \in H_0^1(\Omega)$ is given by

$$\left\langle \frac{\partial\mathcal{L}}{\partial\hat{u}}(\hat{h}, u, \hat{p}), \phi \right\rangle = \int_\Omega j'(u)\phi\,dx + \int_\Omega (\hat{h}\nabla p \cdot \nabla\phi)\,dx.$$

Notice that, requiring that $\langle \frac{\partial\mathcal{L}}{\partial\hat{u}}(h, u, p), \phi\rangle = 0$ for all directions ϕ is nothing else than the variational formulation of the adjoint equation (3.12).

3.3.2 A simple formula for the derivative

It is possible to compute the derivative of J by means of the Lagrangian in the following way:

$$J'(h) = \frac{\partial\mathcal{L}}{\partial h}(h, u, p),$$

where u is the state function (solution to (3.4)) and p is the adjoint state (solution to problem (3.12)). Indeed, we have

$$J(h) = \mathcal{L}(h, u, \hat{p}) \quad \forall \hat{p} \in H_0^1(\Omega)$$

by definition of the state function u. Thus, if the map $h \mapsto u(h)$ is differentiable, we get for $k \in L^\infty(\Omega)$

$$\langle J'(h), k\rangle = \left\langle \frac{\partial\mathcal{L}}{\partial h}(h, u, \hat{p}), k \right\rangle + \left\langle \frac{\partial\mathcal{L}}{\partial u}(h, u, \hat{p}), \frac{\partial u}{\partial h}(k) \right\rangle.$$

Then, taking $\hat{p} = p$, the adjoint we obtain

$$\langle J'(h), k\rangle = \left\langle \frac{\partial\mathcal{L}}{\partial h}(h, u, p), k \right\rangle.$$

By the above discussion, we obtain the following theorem.

Theorem 3.12. *Let $\mathcal{L}(\hat{h}, \hat{u}, \hat{p})$ be the Lagrangian defined as the sum of the objective function and the variational formulation of the state equation, i.e.,*

$$\mathcal{L}(\hat{h}, \hat{u}, \hat{p}) = \int_\Omega j(\hat{u})\,dx + \int_\Omega (\hat{h}\nabla\hat{p}\cdot\nabla\hat{u} - f\hat{p})\,dx.$$

Let p be the solution of the adjoint equation

$$\left\langle \frac{\partial \mathcal{L}}{\partial u}(h, u, p), \phi \right\rangle = 0 \quad \forall \phi \in H_0^1(\Omega).$$

Assume that the solution $u = u(h)$ of the state equation (3.4) is differentiable with respect to h. Then the objective function J is differentiable and

$$J'(h) = \frac{\partial \mathcal{L}}{\partial h}(h, u, p).$$

This theorem is the practical method for computing $J'(h)$. Once the gradient of the cost function has been obtained, it is natural and quite easy to implement a gradient method to minimize $J(h)$ numerically. In Sect. 3.5, we provide numerical algorithms to compute the optimal thickness.

3.4 A discrete approach

One can wonder whether the such optimal design problems get simpler after discretization. Unfortunately, the answer is "no." In this section, we consider a discrete approach to the problems. Applying a finite element method, the equation becomes a linear system of order n

$$K(h)y(h) = b,$$

where $K(h)$ is the rigidity matrix of the membrane (which depends on h), b is a vector representing the forces f, and $y(h)$ the vector of the coordinates of the solution u in the finite element basis (of dimension n). We also discretize the admissible set as follows:

$$\mathcal{U}_{\text{ad}}^{\text{disc}} = \left\{ h \in \mathbb{R}^N : h_{\max} \geq h_i \geq h_{\min} > 0,\ \sum_{i=1}^n c_i h_i = h_0|\Omega| \right\},$$

where the finite sum

$$\sum_{i=1}^n c_i h_i$$

is an approximation of

$$\int_\Omega h(x)\,dx.$$

Approximating the cost function, the discrete problem becomes

$$\inf_{h\in\mathcal{U}_{\text{ad}}^{\text{disc}}} \{J^{\text{disc}}(h) = j^{\text{disc}}(y(h))\},$$

where j^{disc} is a smooth approximation of j from $\mathbb{R}^N$ into $\mathbb{R}$. In the case of the compliance we have:

$$j^{\text{disc}}(y(h)) = b\cdot y(h) = K(h)^{-1}b\cdot b.$$

In the case of a least-square criterion for a target displacement we have:

$$j^{\text{disc}}(y(h)) = B(y(h) - y_0)\cdot(y(h) - y_0),$$

where B is a mass matrix. In practice, we need a way to compute the gradient of $J^{\text{disc}}(h)$. This can be applied to both finding the optimality condition and the implementation of a numerical method of minimization.

First, we consider the following "naive idea." Since $y(h) = K(h)^{-1}b$, we have

$$(J^{\text{disc}})'(h) = y'(h)(j^{\text{disc}})'(y(h)) \quad \text{with} \quad y'(h) = -K(h)^{-1}K(h)'K(h)^{-1}b, \tag{3.13}$$

where we used the notation $f'(h) = (\partial f(h)/\partial h_i)_{1\leq i\leq n}$ and the second identity in (3.13) is a direct application of the formula for the derivative of a matrix. We remark that this method is not practically useful because one must solve $n+1$ linear systems with respect to the matrix $K(h)$ in order to obtain all components of $y'(h)$. Recall that $K(h)$ is a very large matrix (of size $n\times n$) and its inverse is never explicitly computed as it would take too long. As a consequence, we do not use the explicit formula $y(h) = K(h)^{-1}b$. We rather use an adjoint method.

3.4.1 Adjoint state

Definition 3.13. *We define the adjoint state $p \in \mathbb{R}^N$ as the solution of*

$$K(h)p(h) = -(j^{\text{disc}})'(y(h)). \tag{3.14}$$

By rearranging the second equality of (3.13) we get

$$K(h)y'(h) = -K'(h)y(h). \tag{3.15}$$

Now, taking the scalar product of (3.15) with $p(h)$ and that of (3.14) with $y'(h)$, we obtain, for each component $i = 1, \dots, n$:

$$K(h)p(h) \cdot \frac{\partial y}{\partial h_i}(h) = -\frac{\partial K}{\partial h_i}(h)y(h) \cdot p(h) = -(j^{\text{disc}})'(y(h)) \cdot \frac{\partial y}{\partial h_i}(h),$$

from which we deduce

$$(J^{\text{disc}})'(h) = K'(h)y(h) \cdot p(h) = \left(\frac{\partial K}{\partial h_i}(h)y(h) \cdot p(h)\right)_{1 \le i \le n}.$$

In practice, this is the very formula that we use for evaluating the gradient $(J^{\text{disc}})'(h)$ since it requires only to solve two linear systems.

There is no simplification in using a discrete approach rather than a continuous one. Some authors prefer to discretize first and optimize afterwards. This approach guarantees a perfect compatibility between the gradient and the cost function, but it requires a deep knowledge of the numerical solver. Here, we follow another philosophy, "first optimize in a continuous framework, then discretize." It is much simpler, and no precision is lost if the finite element spaces are adequately chosen.

3.5 Numerical algorithms

In this section, we show numerical algorithms to seek the optimal thickness of h. First, we consider the following projected gradient algorithm.

Algorithm 1 Projected gradient algorithm

1. Initialization of the thickness $h_0 \in \mathcal{U}_{\text{ad}}$ (for example, a constant function which satisfies the constraints);
2. Iterations until convergence, for $n \ge 0$ set
$$h_{n+1} = P_{\mathcal{U}_{\text{ad}}}(h_n - \mu J'(h_n)),$$
where $\mu > 0$ is a small descent step, $P_{\mathcal{U}_{\text{ad}}}$ is the projection operator on the closed convex set $\mathcal{U}_{\text{ad}}$ and the derivative of J is given by
$$J'(h_n) = \nabla u_n \cdot \nabla p_n$$
with state u_n and adjoint p_n (both defined with respect to the thickness h_n).

To make the algorithm fully explicit, we have to specify how to compute the projection operator $P_{\mathcal{U}_{\text{ad}}}$.

We define the projection operator $P_{\mathcal{U}_{\text{ad}}}$ as follows:

$$(P_{\mathcal{U}_{\text{ad}}}(h))(x) = \max(h_{\min}, \min(h_{\max}, h(x) + \ell)), \quad x \in \Omega,$$

where ℓ is the unique Lagrange multiplier such that

$$\int_\Omega P_{\mathcal{U}_{\text{ad}}}(h)\, dx = h_0|\Omega|.$$

The determination of the constant ℓ is not explicit but based on an iterative algorithm. First, notice that the function

$$h \longmapsto F(\ell) = \int_\Omega \max(h_{\min}, \min(h_{\max}, h(x) + \ell))\, dx$$

is strictly increasing on the interval $[\ell^-, \ell^+]$, the inverse image of the closed interval $[h_{\min}|\Omega|, h_{\max}|\Omega|]$. Thanks to this monotonicity property, we propose a simple iterative algorithm: we first bracket the root by an interval $[\ell^1, \ell^2]$ such that

$$F(\ell^1) \le h_0|\Omega| \le F(\ell^2),$$

then we proceed by dichotomy to find the root ℓ.

Remark 3.14.

1. *In practice, we rather use a projected gradient algorithm with a variable step (not optimal) which guarantees the decrease of the functional $J(h_{n+1}) < J(h_n)$.*
2. *The algorithm is rather slow. A possible acceleration is based on the quasi-Newton algorithm.*
3. *The overhead generated by the adjoint computation is very modest: one has to build a new right-hand-side (using the state) and solve the corresponding linear system (with the same rigidity matrix).*

4. *Convergence is detected when the optimality condition is satisfied with a threshold* $\varepsilon > 0$

$$|h_n - \max\big(h_{\min}, \min(h_{\max}, h_n - \mu_n J'(h_n) + \ell_n)\big)| \leq \varepsilon \mu_n h_{\max}.$$

3.5.1 Another numerical algorithm for the compliance

When $j(u) = fu$, we find $p = -u$ since $j'(u) = f$. This particular case is said to be self-adjoint. We use the dual or complementary energy (see Sect. 2.4)

$$\int_\Omega fu\,dx = \min_{\substack{\tau \in L^2(\Omega)^N, \\ -\operatorname{div}\tau = f \text{ in } \Omega}} \int_\Omega h^{-1}|\tau|^2\,dx$$

in order to rewrite the original optimization problem as a double minimization problem:

$$\inf_{h \in \mathcal{U}_{\rm ad}} \min_{\substack{\tau \in L^2(\Omega)^N, \\ -\operatorname{div}\tau = f \text{ in } \Omega}} \int_\Omega h^{-1}|\tau|^2\,dx,$$

and the order of minimization is irrelevant. This problem is convex and therefore it admits a minimizer.

By elementary calculation, we can show that the following lemma holds.

Lemma 3.15. *The function* $\phi(a,\sigma) = a^{-1}|\sigma|^2$, *defined from* $\mathbb{R}_{\geq 0} \times \mathbb{R}^N$ *into* $\mathbb{R}$, *satisfies*

$$\phi(a,\sigma) = \phi(a_0,\sigma_0) + \phi'(a_0,\sigma_0)\cdot(a - a_0, \sigma - \sigma_0) + \phi\left(a, \sigma - \frac{a}{a_0}\sigma_0\right), \tag{3.16}$$

where the derivative is given by

$$\phi'(a_0,\sigma_0)\cdot(b,\tau) = -\frac{b}{a_0^2}|\sigma_0|^2 + \frac{2}{a_0}\sigma_0\cdot\tau.$$

In particular, since by (3.16), the graph of $\phi(a,\sigma)$ *lies above its linear approximation at each point* (a_0,σ_0), *then* ϕ *is convex.*

As a result, we obtain the following.

Lemma 3.16 (Optimality conditions)**.** *For a given* $\tau \in L^2(\Omega)^N$, *the problem*

$$\min_{h \in \mathcal{U}_{\rm ad}} \int_\Omega h^{-1}|\tau|^2\,dx$$

admits a minimizer $h(\tau)$ *in* $\mathcal{U}_{\rm ad}$ *given by*

$$h(\tau)(x) = \begin{cases} h^*(x) & \text{if } h_{\min} < h^*(x) < h_{\max}, \\ h_{\min} & \text{if } h^*(x) \leq h_{\min}, \\ h_{\max} & \text{if } h^*(x) \geq h_{\max} \end{cases} \qquad \text{with } h^*(x) = \frac{|\tau(x)|}{\sqrt{\ell}}, \tag{3.17}$$

where ℓ *is the Lagrange multiplier such that*

$$\int_\Omega h(\tau)(x)\,dx = h_0|\Omega|.$$

Sketch of the proof. By Lemma 3.15 we obtain that the map $h \mapsto \int_\Omega h^{-1}|\tau|^2 dx$ is convex in $\mathcal{U}_{\rm ad}$. Therefore, Theorem 2.7 ensures the existence of a minimum point h. This point is then characterized by the optimality condition given by Theorem 2.17. We refer to [Al2007-1, Lemma 5.2.25] for more details. □

Lemma 3.16 tells us the following numerical algorithm for the compliance:

Algorithm 2 Optimality criteria method

1. Initialization of the thickness $h_0 \in \mathcal{U}_{\rm ad}$.
2. Iterations until convergence, for $n \geq 0$,
   (a) Computation of the state τ_n, unique solution of

$$\min_{\substack{\tau \in L^2(\Omega)^N, \\ -\operatorname{div}\tau = f \text{ in } \Omega}} \int_\Omega h_n^{-1}|\tau|^2\,dx. \tag{3.18}$$

   (b) Update of the thickness:

$$h_{n+1} = h(\tau_n),$$

   where $h(\tau)$ is the minimizer defined by (3.17). Finally, the Lagrange multiplier ℓ is computed by dichotomy.

Remark that, by the dual energy approach introduced in Sect. 2.4, minimizing (3.18) in τ is equivalent to solving the equation

$$\begin{cases} -\mathrm{div}(h_n \nabla u_n) = f & \text{in } \Omega, \\ u_n = 0 & \text{on } \partial\Omega, \end{cases}$$

and then recovering τ_n by the formula

$$\tau_n = h_n \nabla u_n.$$

The algorithm can be interpreted as an alternate minimization in τ and h of the objective function. In particular, we deduce that the objective function always decreases through the iterations. Indeed, for all $n \geq 0$,

$$J(h_{n+1}) = \int_\Omega h_{n+1}^{-1} |\tau_{n+1}|^2 \, dx \leq \int_\Omega h_{n+1}^{-1} |\tau_n|^2 \, dx \leq \int_\Omega h_n^{-1} |\tau_n|^2 \, dx = J(h_n),$$

where, for fixed h_{n+1} we minimized in τ and then, for fixed τ_n we minimized in h. This algorithm can also be interpreted as an optimality criteria method.

3.6 Thickness optimization of an elastic plate

We consider the following elasticity problem for an elastic plate Ω

$$\begin{cases} -\mathrm{div}\, \sigma = f & \text{in } \Omega, \\ \sigma = 2\mu h e(u) + \lambda h \, \mathrm{tr}(e(u))\mathrm{Id} & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_D, \\ \sigma \cdot n = g & \text{on } \Gamma_N \end{cases}$$

with strain tensor $e(u) = (\nabla u + (\nabla u)^t)/2$. The set of admissible thicknesses is

$$\mathcal{U}_{\mathrm{ad}} = \left\{ h \in L^\infty(\Omega) : h_{\max} \geq h(x) \geq h_{\min} > 0 \text{ a.e. in } \Omega, \int_\Omega h(x) \, dx = h_0 |\Omega| \right\}.$$

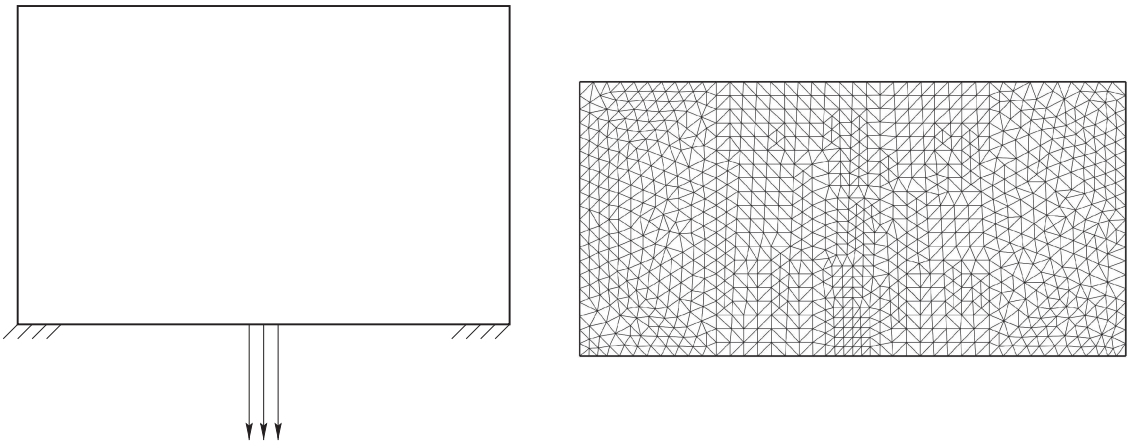


Fig. 11. Boundary conditions and mesh for an elastic plate.

The compliance optimization reads

$$\inf_{h \in \mathcal{U}_{\mathrm{ad}}} \left\{ J(h) = \int_\Omega f \cdot u \, dx + \int_{\Gamma_N} g \cdot u \, ds \right\}. \tag{3.19}$$

The theoretical results are the same of previous sections. We apply the optimality criteria method for the compliance optimization (3.19). In order to compute (3.19), we use FreeFem++. You can see its scripts on the web page http://www.cmap.polytechnique.fr/~allaire/freefem_en.html.

In Fig. 12, we used finite elements $P2$ for u and $P0$ for h. However, numerical instabilities like checkerboards occur if we use finite elements $P1$ for u and $P0$ for h (see Fig. 14). Therefore we consider a "regularization" in order to avoid the instabilities.

3.6.1 Regularization

In what follows, let us consider the "regularized" framework to avoid numerical instabilities. The main idea, similar to that introduced in Remark 2.16 is as follows.

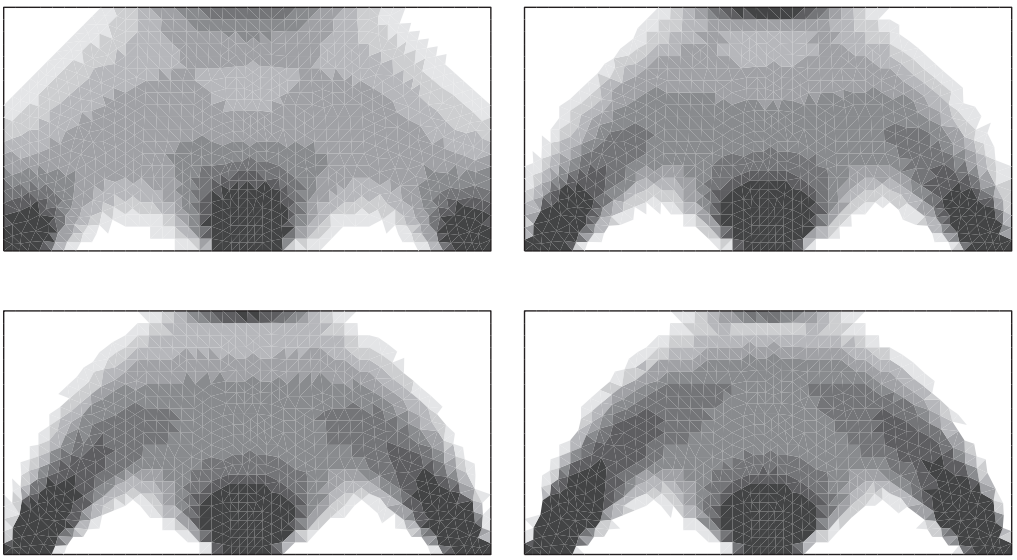


Fig. 12. Thickness at iterations 1, 5, 10, 30 (uniform initialization), where $h_{\min} = 0.1, h_{\max} = 1.0, h_0 = 0.5$ (increasing thickness from white to black).

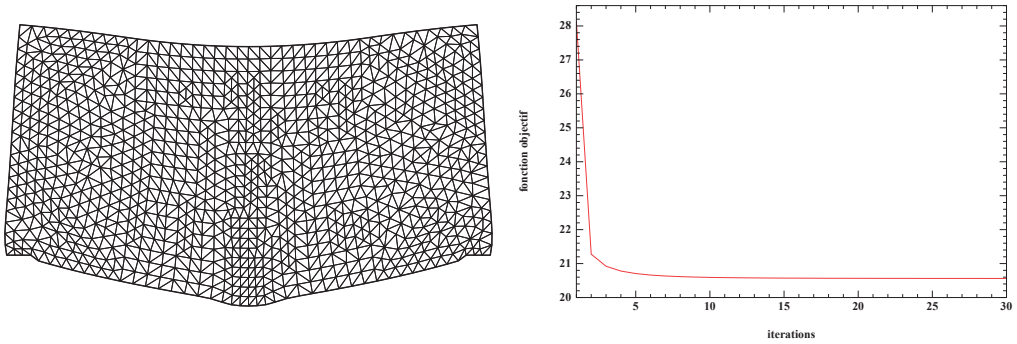


Fig. 13. Final deformed shapes and convergence history.

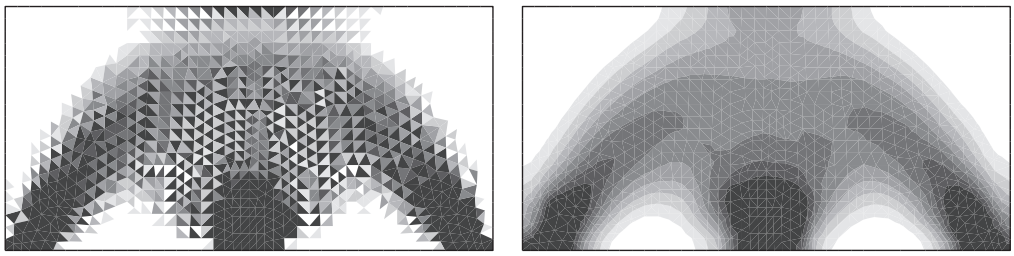


Fig. 14. Left: numerical instabilities (checkerboards), right: regularized optimal shape.

We are going to replace the scalar product

$$\langle J'(h), k\rangle = \int_\Omega k\nabla u \cdot \nabla p \, dx, \quad k \in \mathcal{U}_{\text{ad}}$$

with a different one. Previously we identified $\mathcal{U}_{\text{ad}}$ with a subspace of $L^2(\Omega)$, thus

$$\langle J'(h), k\rangle = \int_\Omega J'(h)k \, dx \Longrightarrow J'(h) = \nabla u \cdot \nabla p.$$

Now, we identify a "regularized" admissible set $\mathcal{U}_{\text{ad}}^{\text{reg}}$ to a subspace $H^1(\Omega)$, thus

$$\langle J'(h), k\rangle = \int_\Omega (\varepsilon^2 \nabla J'(h) \cdot \nabla k + J'(h)k) \, dx,$$

where $\varepsilon > 0$ is a regularization parameter (which can be interpreted as a length scale). Therefore, we deduce a new formula for the gradient

$$\begin{cases} -\varepsilon^2 \Delta J'(h) + J'(h) = \nabla u \cdot \nabla p & \text{in } \Omega, \\ \dfrac{\partial J'(h)}{\partial n} = 0 & \text{on } \partial\Omega. \end{cases} \tag{3.20}$$

Solving (3.20) and using a gradient algorithm such as projected gradient method, we obtain regularized optimal shape (see Fig. 14).

3.7 Exercises

Problem 3.7.1. *Check the numerical instabilities (Fig. 14, left) by using FreeFem++.*

Problem 3.7.2. *Solve (3.20) and see the regularized optimal shape (Fig. 14, right) by using FreeFem++.*

4. Homogenization Theory

In this section, we explain the homogenization method in order to apply shape optimization problems in Sect. 5. Homogenization method is one of the averaging methods for partial differential equations. It is often concerned with the derivation of (macroscopic) equations whose solutions are defined as limits of solutions to (microscopic) equations with rapidly varying coefficients. A particular case of homogenization is obtained when the coefficients of the partial differential equation are periodically and rapidly oscillating. Indeed, in many fields of science and technology one has to solve boundary value problems in periodic media. In such a case, homogenization is simpler and can be achieved, at least formally, by using asymptotic expansions. This section is devoted to an elementary introduction of periodic homogenization, without providing a fully rigorous justification. Of course, homogenization methods using functional analysis method were considered for mathematical justification. The interested reader is referred to the classical books [BLP1978], [CD1999], [H1996], [JKO1995], for further details.

Note that, for applications in shape optimization, one should rely, in full rigor, on a more general homogenization method, called H-convergence, introduced in [MT1997] (see the textbook [Al2002] for more details). For simplicity, we restrict ourselves to the setting of periodic homogenization which is enough for a formal understanding.

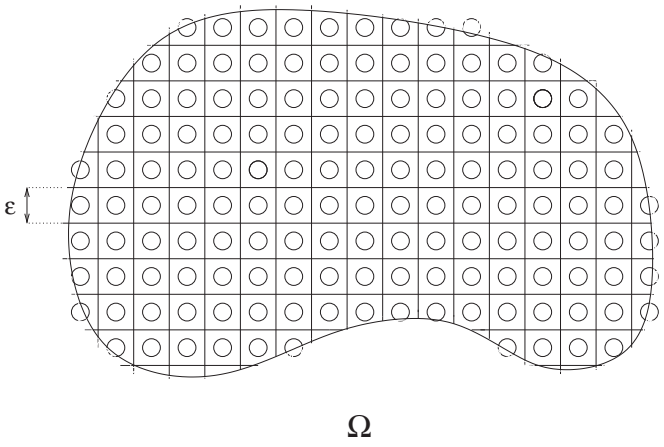


Fig. 15. A periodic heterogeneous medium.

4.1 Homogenization based on two-scale asymptotic expansions

In what follows, we consider an elastic membrane made of a composite material with a fine periodic structure and apply the periodic homogenization method. We assume that ratio between the period and the characteristic size of the structure equals to $\varepsilon \ll 1$. We will find the "true" problem by the limit problem obtained as $\varepsilon \to 0$.

Let Ω be a bounded domain in $\mathbb{R}^N$ ($N \geq 1$) and $f = f(x)$ be a load. Then we consider the displacement u_ε, which is defined as the solution of the following boundary value problem:

$$\begin{cases} -\mathrm{div}\left(A\left(\dfrac{x}{\varepsilon}\right) \nabla u_\varepsilon \right) = f & \text{in } \Omega, \\ u_\varepsilon = 0 & \text{on } \partial\Omega, \end{cases} \tag{4.1}$$

where the coefficient $A(y)$ satisfies the variable Hooke's law, that is, $A(y)$ is a Y-periodic function with $Y = (0,1)^N$. Thus for any i-th vector of the canonical basis e_i, the coefficient $A(y)$ satisfies

$$A(y + e_i) = A(y).$$

If we replace y by x/ε, then we obtain that the map $x \mapsto A(x/\varepsilon)$ is a periodic of period ε in all the coordinate directions $e_1, \ldots, e_N$. A direct computation of u_ε can be very expensive (since the mesh size h should satisfy $h \ll \varepsilon$), thus we seek only the averaged values of u_ε. We assume that the solution u_ε can be expanded as follows:

$$u_\varepsilon(x) = \sum_{i=0}^{+\infty} \varepsilon^i u_i\left(x, \frac{x}{\varepsilon}\right), \tag{4.2}$$

with $u_i(x, y)$ function of the two variables x and y, periodic in y, with periodicity cell given by $Y = (0, 1)^N$. Plugging the series (4.2) in the Eq. (4.1), we use the derivation rule

$$\nabla\left(u_i\left(x, \frac{x}{\varepsilon}\right)\right) = (\varepsilon^{-1}\nabla_y u_i + \nabla_x u_i)\left(x, \frac{x}{\varepsilon}\right). \tag{4.3}$$

Then we get

$$\nabla u_\varepsilon(x) = \varepsilon^{-1}\nabla_y u_0\left(x, \frac{x}{\varepsilon}\right) + \sum_{i=0}^{\infty} \varepsilon^i(\nabla_y u_{i+1} + \nabla_x u_i)\left(x, \frac{x}{\varepsilon}\right). \tag{4.4}$$

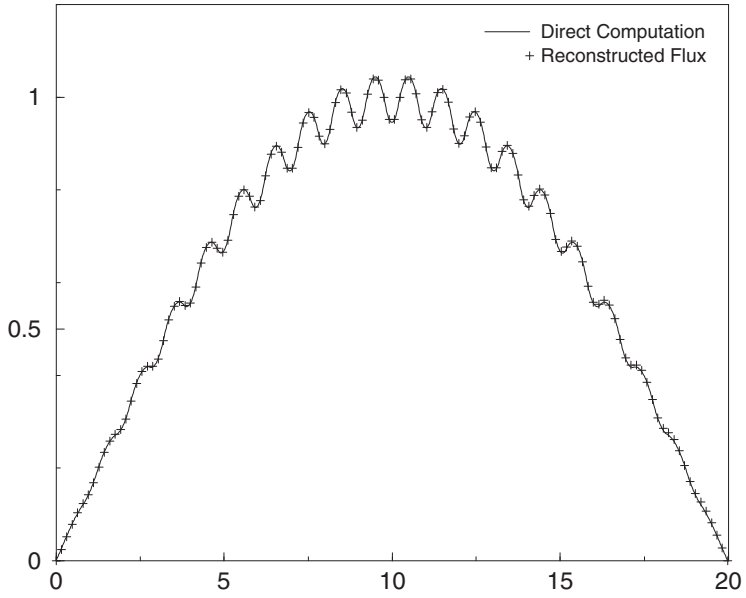


Fig. 16. Typical oscillating behavior of $x \mapsto u_i\left(x, \frac{x}{\varepsilon}\right)$.

Substituting (4.4) into (4.2), the equation becomes a series in ε

$$\begin{aligned} &-\varepsilon^{-2}[\mathrm{div}_y(A(\nabla_y u_0))]\left(x, \frac{x}{\varepsilon}\right) - \varepsilon^{-1}[\mathrm{div}_y(A(\nabla_x u_0 + \nabla_y u_1)) + \mathrm{div}_x(A\nabla_y u_0)]\left(x, \frac{x}{\varepsilon}\right) \\ &- \sum_{i=0}^{+\infty} \varepsilon^i[\mathrm{div}_x(A(\nabla_x u_i + \nabla_y u_{i+1})) + \mathrm{div}_y(A(\nabla_x u_{i+1} + \nabla_y u_{i+2}))]\left(x, \frac{x}{\varepsilon}\right) = f(x). \end{aligned} \tag{4.5}$$

In order to find the solution of the limit equation as $\varepsilon \to 0$, we identify each power of ε. The most important terms are only the first three terms of the series. We start by a technical lemma:

Lemma 4.1. *Let $g \in L^2(Y)$ and suppose that $A(y)$ is a Y-periodic $N \times N$ matrices satisfying*

$$A(y)\xi \cdot \xi \geq \lambda|\xi|^2 \quad \forall \xi \in \mathbb{R}^N,$$

for some $\lambda > 0$. Moreover, let $H^1_\#(Y)/\mathbb{R}$ denote the quotient space, $H^1_\#(Y)$ up to an additive constant, equipped with the norm $\|\nabla \cdot\|_{L^2(Y)}$. Then the problem

$$\begin{cases} -\mathrm{div}_y(A(y)\nabla_y v(y)) = g(y) & \text{in } Y, \\ y \mapsto v(y) & Y\text{-periodic} \end{cases}$$

admits a unique solution $v \in H^1_\#(Y)/\mathbb{R}$ if and only if

$$\int_Y g(y)\, dy = 0.$$

Proof. Let us check that g being of zero mean over Y is a necessary condition for existence. As a matter of fact, integrating the equation over Y, we get

$$\int_Y \operatorname{div}_y(A(y)\nabla_y v(y))\,dy = \int_{\partial Y} A(y)\nabla_y v(y)\cdot n\,ds = 0$$

because of the periodic boundary condition. Indeed $A(y)\nabla_y v(y)$ is periodic, but the normal n changes its sign on opposite faces of Y.

The sufficient condition is obtained by applying Lax–Milgram theorem with respect to $H^1_\#(Y)/\mathbb{R}$. Indeed, $a(u,v) = \int_Y A(y)\nabla u(y)\cdot\nabla v(y)\,dy$ is a coercive continuous bilinear form on $H^1_\#(Y)/\mathbb{R}$ by uniform ellipticity. Furthermore, the map $F : H^1_\#(Y)/\mathbb{R} \to \mathbb{R}$, defined by $F(\phi) = \int_Y g(y)\phi(y)\,dy$, is a well defined bounded linear functional on $H^1_\#(Y)/\mathbb{R}$ because g is a function of zero mean over Y. Indeed, for all $\phi \in H^1_\#(Y)$, if we let $\bar\phi = \fint_Y \phi(y)\,dy$ denote the mean value of ϕ over Y, then, we get

$$\begin{aligned}\int_Y g(y)\phi(y)\,dy &= \int_Y g(y)(\phi(y)-\bar\phi)\,dy\\ &\le \|g\|_{L^2(Y)}\|\phi-\bar\phi\|_{L^2(Y)} \le \|g\|_{L^2(Y)}\|\nabla\phi\|_{L^2(Y)},\end{aligned}$$

where we used the Poincaré–Wirtinger inequality in the last inequality. This implies that the map $F : H^1_\#(Y)/\mathbb{R} \to \mathbb{R}$ defined above is bounded in the norm $\|\nabla\cdot\|_{L^2(Y)}$ as claimed. Hence, by Lax–Milgram's theorem, there exists a unique solution $v \in H^1_\#(Y)/\mathbb{R}$ such that

$$\int_Y A(y)\nabla v(y)\cdot\nabla\phi(y)\,dy = \int_Y g(y)\phi(y)\,dy \quad \forall\phi \in H^1_\#(Y)/\mathbb{R}.$$

□

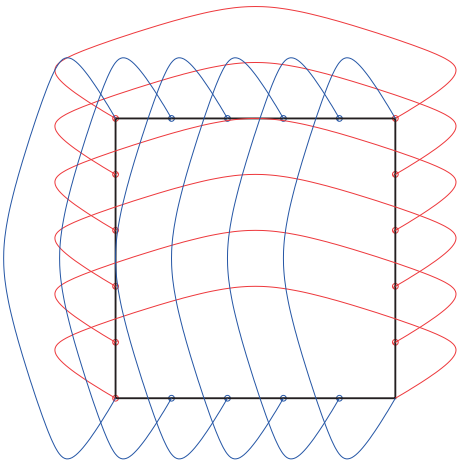


Fig. 17. Periodic boundary conditions in $H^1_\#(Y)$.

By using Lemma 4.1, we can find the solution of the limit equation. Let us consider the equations that arise when we consider the first three terms of the series in ().

ε^{-2}:

$$\begin{cases} -\operatorname{div}_y(A(y)\nabla_y u_0(x,y)) = 0 & \text{in } Y,\\ y \mapsto u_0(x,y) & Y\text{-periodic.}\end{cases} \tag{4.6}$$

It is a partial differential equation with respect to y in Y (here x is just a parameter). By the uniqueness of the solution up to an additive constant, we deduce that

$$u_0(x,y) \equiv u(x). \tag{4.7}$$

ε^{-1}:

$$\begin{cases} -\operatorname{div}_y(a(y)\nabla_y u_1(x,y)) = \operatorname{div}_y(a(y)\nabla_x u_0(x,y)) & \text{in } Y,\\ y \mapsto u_1(x,y) & Y\text{-periodic.}\end{cases} \tag{4.8}$$

The necessary and sufficient condition of existence is satisfied. Thus, by (4.7), u_1 (seen as an element of $H^1_\#(Y)/\mathbb{R}$) depends linearly on $\nabla_x u(x)$. In particular, if we let $(e_i)_{1\le i\le N}$ denote the canonical basis of $\mathbb{R}^N$, then it is easy to check that

$$u_1(x,y) = \sum_{i=1}^{N} \frac{\partial u}{\partial x_i}(x) w_i(y), \tag{4.9}$$

where w_i is the solutions of the following auxiliary problems (cell problems) for $i = 1, \ldots, N$:

$$\begin{cases} -\mathrm{div}_y(A(y)(\nabla_y w_i(y) + e_i)) = 0 & \text{in } Y, \\ y \mapsto w_i(y) & Y\text{-periodic}. \end{cases} \tag{4.10}$$

The functions w_i are usually called the correctors.

ε^0:

$$\begin{cases} -\mathrm{div}_y(A(y)\nabla_y u_2(x,y)) = f(x) + \mathrm{div}_y(a(y)\nabla_x v_1) + \mathrm{div}_x(a(y)(\nabla_y v_1 + \nabla_x u)) & \text{in } Y, \\ y \mapsto u_2(x,y) & Y\text{-periodic}. \end{cases} \tag{4.11}$$

By using Lemma 4.1, the necessary and sufficient condition of existence of the solution u_2 is

$$\int_Y (\mathrm{div}_y(A(y)\nabla_x u_1) + \mathrm{div}_x(A(y)(\nabla_y u_1 + \nabla_x u)) + f(x))dy = 0.$$

By employing the use of the representation formula (4.9), we can rewrite u_1 in terms of $\nabla_x u(x)$:

$$\mathrm{div}_x \int_Y A(y)\left(\sum_{i=1}^{N} \frac{\partial u}{\partial x_i}(x)\nabla_y w_i(y) + \nabla_x u(x)\right)dy + f(x) = 0.$$

In other words, we have succeeded in identifying the the homogenized problem

$$\begin{cases} -\mathrm{div}_x(A^*\nabla_x u(x)) = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{4.12}$$

where the homogenized tensor A^* is defined by

$$A^*_{ji} = \int_Y A(y)(e_i + \nabla_y w_i) \cdot e_j \, dy, \tag{4.13}$$

or, integrating by parts

$$A^*_{ji} = \int_Y A(y)(e_i + \nabla_y w_i(y)) \cdot (e_j + \nabla_y w_j(y)) \, dy.$$

Indeed, the cell problems (4.10) yield

$$\int_Y A(y)(e_i + \nabla_y w_i(y)) \cdot \nabla_y w_j(y) \, dy = 0.$$

Remark 4.2. *The formula for A^* is not fully explicit because cell problems (4.10) must be solved. However A^* does not depend on Ω, nor f, nor the boundary conditions. It only characterizes the microstructure. Later, we shall compute explicitly some examples of A^*.*

Under mild smoothness assumptions on the data, one can justify the expansion in $H^1(\Omega)$ [BLP1978, JKO1995].

Theorem 4.3. *Assume that the homogenized solution u is smooth. Then the following expansion holds in $H^1(\Omega)$:*

$$u_\epsilon(x) = u(x) + \varepsilon u_1\left(x, \frac{x}{\varepsilon}\right) + r_\varepsilon \text{ with } \|r_\varepsilon\|_{H^1} \leq C\varepsilon^{1/2}.$$

In particular

$$\|u_\varepsilon - u\|_{L^2(\Omega)} \leq C\varepsilon^{1/2}.$$

Remark 4.4 (Rigorous justification)**.** *Employing a formal asymptotic expansion is a very useful method. However we don't know a priori whether the solution of the microscopic equation can be expanded as (4.2). We refer the interested reader to Tartar's method [MT1997] and the two-scale convergence method [Ng1989, Al1992] for a rigorous mathematical justification.*

Remark 4.5 (Homogenized coefficients A^*)**.** *In dimension $N = 1$, the explicit formula for A^* is the so-called harmonic mean. In dimension $N \geq 2$, there is no explicit formula for A^*, which has to be computed numerically. Nevertheless, one can obtain explicit bounds on A^*.*

Remark 4.6. *Homogenization works for non-periodic media too (H-convergence or G-convergence).*

Remark 4.7 (Asymptotic expansions for the stress)**.** *We assume that*

$$u_\varepsilon(x) = \sum_{i=0}^{+\infty} \varepsilon^i u_i\left(x, \frac{x}{\varepsilon}\right),\ \sigma_\varepsilon(x) = A\left(\frac{x}{\varepsilon}\right)\nabla u_\varepsilon(x) = \sum_{i=0}^{+\infty} \varepsilon^i \sigma_i\left(x, \frac{x}{\varepsilon}\right),$$

where $\sigma_i(x, y)$ is a function of the two variables x and y, periodic in y with period $Y = (0, 1)^N$. Plugging this series in the Eq. (4.1), we find

$$-\mathrm{div}_y\sigma_0 = 0,\ \mathrm{div}_x\sigma_0 - \mathrm{div}_y\sigma_1 = f.$$

On the other hand,

$$\sigma_0(x, y) = A(y)(\nabla_x u(x) + \nabla_y u_1(x, y))$$

and

$$\sigma_0(x, y) = A^*\nabla_x u(x) + \tau(x, y) \text{ with } \int_Y \tau\, dy = 0.$$

One can prove that τ is the solution of the dual cell problem.

4.2 Composite materials

Composite materials are ubiquitous in engineering, mechanics and physics and their effective properties can be understood through homogenization theory [Al2002, Ch2000, MI2001]. In what follows, we identify a composite material by its homogenized tensor A^*. We restrict ourselves to two-phase composites. We mix two isotropic constituents $A(y) = \alpha\chi(y) + \beta(1 - \chi(y))$, where $\chi : Y \to \{0, 1\}$ is a characteristic function. Let $\theta = \int_Y \chi(y)dy$ be the volume fraction of phase α and $(1 - \theta)$ be that of phase β.

We focus on the characterization of G_θ defined as follows:

Definition 4.8 (The set of all homogenized tensors G_θ). *Let G_θ be the set of all homogenized tensors A^* obtained by homogenization of the two phases α and β in proportions θ and $(1 - \theta)$.*

Remark 4.9. *Of course, we have $G_0 = \{\beta\mathrm{Id}\}$ and $G_1 = \{\alpha\mathrm{Id}\}$. However, G_θ is usually a (very) large set of tensors (corresponding to different choices of $\chi(y)$).*

4.2.1 Lamination for two phase composites

For two phase composites, the density $\theta(x)$, as well as the homogenized tensor $A^*(x)$, depends on the position x. For two-phase mixtures, an explicit characterization of G_θ is possible by the variational principle of Hashin and Shtrikman [HS1963]. We make the following assumptions:

(i) Linear model of conduction or membrane stiffness (it is more delicate for linearized elasticity and very few results are known in the non-linear case).
(ii) Perfect interfaces between the phases (continuity of both displacement and normal stress), no possible effects of delamination or debonding.

In dimension one, the cell problem (4.10) reads:

$$\begin{cases} -(A(y)(1 + w'(y)))' = 0 & \text{in } [0, 1), \\ y \mapsto w(y) & 1\text{-periodic.} \end{cases}$$

The solution computed explicitly as follows:

$$w(y) = -y + \int_0^y \frac{C_1}{A(t)}\, dt + C_2 \text{ with } C_1 = \left(\int_0^1 \frac{1}{A(y)}\, dy\right)^{-1}.$$

By (4.13), we know that $A^* = \int_0^1 A(y)(1 + w'(y))^2\, dy$, which yields the harmonic mean of $A(y)$:

$$A^* = \left(\int_0^1 \frac{1}{A(y)}\, dy\right)^{-1}.$$

Therefore, if we choose $A(y) = \alpha\chi(y) + \beta(1 - \chi(y))$, then homogenized tensor of any two-phase material is just

$$A^* = \left(\frac{\theta}{\alpha} + \frac{1 - \theta}{\beta}\right)^{-1}.$$

This formula tells us that, in one dimension, the homogenized tensor depends on the characteristic function χ by means of its volume fraction θ only.

In dimension $N \geq 2$, we cannot express A^* explicitly in general as mentioned in Remark 4.5. However it is possible under the following special case. We consider parallel layers of two isotropic phases α and β, orthogonal to the direction e_1. Assume that A^ε depends only on y_1. Let

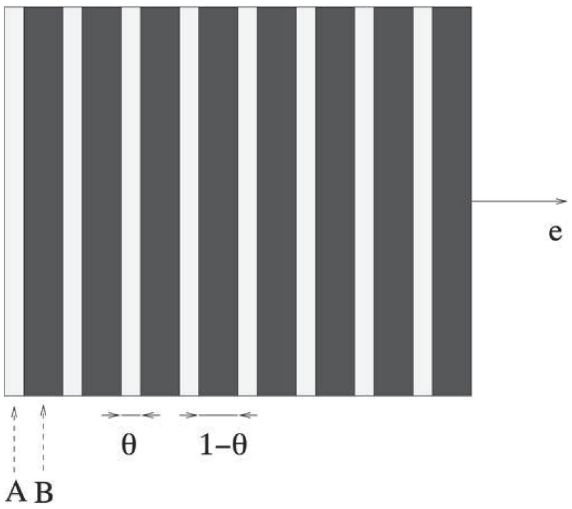


Fig. 18. Simple laminated composites.

$$\chi(y_1) = \begin{cases} 1 & \text{if } 0 < y_1 < \theta \\ 0 & \text{if } \theta < y_1 < 1 \end{cases} \quad \text{with } \theta = \int_Y \chi \, dy.$$

We denote by A^* the homogenized tensor of $A(y) = (\alpha\chi(y_1) + \beta(1 - \chi(y_1)))I$. Then we obtain the following lemma. This lemma is a simple case of the more general Lemma 4.12.

Lemma 4.10. *Define* $\lambda_\theta^- = (\frac{\theta}{\alpha} + \frac{1-\theta}{\beta})^{-1}$ *and* $\lambda_\theta^+ = \theta\alpha + (1-\theta)\beta$. *Then we have*

$$A^* = \begin{pmatrix} \lambda_\theta^- & & & 0 \\ & \lambda_\theta^+ & & \\ & & \ddots & \\ 0 & & & \lambda_\theta^+ \end{pmatrix}. \tag{4.14}$$

Remark 4.11 (Interpretation (resistance = inverse of conductivity))**.** *In the context of electrical conductivity, the harmonic mean is the effective conductivity of a mixture of conductors placed in series (in the direction* e_1*), while the arithmetic mean is the effective conductivity of a mixture of conductors placed in parallel (in any direction orthogonal to* e_1*).*

Lemma 4.12 (Simple laminate of two non-isotropic phases)**.** *The homogenized tensor* A^* *of a simple laminate made of* A *and* B *in proportions* θ *and* $(1-\theta)$ *in the direction* e_1 *is*

$$A^* = \theta A + (1-\theta)B - \frac{\theta(1-\theta)(A-B)e_1 \otimes (A-B)^t e_1}{(1-\theta)Ae_1 \cdot e_1 + \theta Be_1 \cdot e_1}. \tag{4.15}$$

Moreover, if we assume that $(A - B)$ *is invertible, then this formula is equivalent to*

$$\theta(A^* - B)^{-1} = (A-B)^{-1} + \frac{(1-\theta)}{Be_1 \cdot e_1} e_1 \otimes e_1. \tag{4.16}$$

Proof. Recall that by definition (4.13)

$$A^*_{ji} = \int_Y A(y)(e_i + \nabla_y w_i) \cdot e_j \, dy = \int_Y A(y)(e_i + \nabla_y w_i(y)) \cdot (e_j + \nabla_y w_j(y)) \, dy,$$

namely

$$A^* e_i = \int_Y A(y)(e_i + \nabla_y w_i) \, dy.$$

Consequently, for any $\xi \in \mathbb{R}^N$, we have

$$A^* \xi = \int_Y A(y)(\xi + \nabla_y w_\xi) \, dy, \tag{4.17}$$

where $w_\xi(y) = \sum_{i=1}^N \xi_i w_i(y)$ is the solution of

$$\begin{cases} -\mathrm{div}_y(A(y)(\xi + \nabla w_\xi(y))) = 0 & \text{in } Y, \\ y \mapsto w_\xi(y) & Y\text{-periodic}. \end{cases}$$

Defining $u(y) = \xi \cdot y + w_\xi(y)$, we seek a solution u such that the gradient of u is constant in each phase,

$$\nabla u(y) = a\chi(y_1) + b(1 - \chi(y_1)).$$

Thus, we have

$$u(y) = \chi(y_1)(c_a + a \cdot y) + (1 - \chi(y_1))(c_b + b \cdot y), \tag{4.18}$$

where c_a and c_b are constant vectors.

Let Γ be the interface between the two phases. By continuity of (4.18) through the interface Γ, we have

$$c_a + a \cdot y = c_b + b \cdot y. \tag{4.19}$$

Since c_a and c_b are constant vectors, by (4.19) we have

$$(a - b) \cdot x = (a - b) \cdot y \quad \forall x, y \in \Gamma.$$

Since $(x - y)$ is orthogonal to e_1, there exists a real number $t \in \mathbb{R}$ such that $b - a = te_1$.

Moreover, by continuity of the flux $A(y)\nabla u \cdot n$ through the interface Γ, we have

$$Aa \cdot e_1 = Bb \cdot e_1. \tag{4.20}$$

In particular, it implies $-\mathrm{div}(A(y)\nabla u) = 0$ in the weak sense.

Since $b - a = te_1$, (4.20) yields the following value for t:

$$t = \frac{(A - B)a \cdot e_1}{Be_1 \cdot e_1}.$$

Since w_ξ is periodic, it satisfies $\int_Y \nabla w_\xi \, dy = 0$, thus by the definition of u we have

$$\int_Y \nabla u \, dy = \theta a + (1 - \theta)b = \xi.$$

On the other hand, by (4.17) and the definition of u we have

$$A^*\xi = \int_Y A(y)(\xi + \nabla w_\xi) \, dy = \int_Y A(y)\nabla u \, dy = \theta Aa + (1 - \theta)Bb.$$

Thus we obtain

$$A^*(\theta a + (1 - \theta)b) = \theta Aa + (1 - \theta)Bb.$$

Since $b = a + te_1$ with $t = \frac{(A-B)a \cdot e_1}{Be_1 \cdot e_1}$, we find

$$a = \xi - (1 - \theta)\frac{(A - B)\xi \cdot e_1}{(1 - \theta)Ae_1 \cdot e_1 + \theta Be_1 \cdot e_1} e_1.$$

Then, a simple computation gives

$$A^*\xi = \theta A\xi + (1 - \theta)B\xi - \frac{\theta(1 - \theta)(A - B)\xi \cdot e_1}{(1 - \theta)Ae_1 \cdot e_1 + \theta Be_1 \cdot e_1}(A - B)e_1.$$

The other formula is a consequence of the following fact: if M is invertible, then

$$(M + c(Me) \otimes (M^t e))^{-1} = M^{-1} - \frac{c}{1 + c(Me \cdot e)} e \otimes e,$$

where $c \in \mathbb{R}$ and e is a unit vector in $\mathbb{R}^N$ which determines the direction of the lamination. □

The composite A^* is said to be a single lamination in the direction e_1 of the two phases A and B in proportions θ and $(1 - \theta)$ (see Fig. 18). By varying the proportion θ and the direction e_1, we obtain a whole family of composite materials. This family can still be enlarged by laminating again these simple laminates. Then we laminate again the preceding composite with always the same phase B.

A sequential laminate is obtained by an iterative process of lamination where the previous laminate is laminated again with a single pure phase (always the same one). By using the special form of (4.16) (which does not deliver directly the value of A^*, contrary to (4.15)), the iterative or sequential laminate can be explicitly characterized. Let $(e_i)_{1 \le i \le p}$ be a collection of unit vectors and $(\theta_i)_{1 \le i \le p}$ be proportions in $[0, 1]$. By (4.16) a simple laminate A_1^* of A and B in proportions θ, $(1 - \theta)$ is

$$\theta_1(A_1^* - B)^{-1} = (A - B)^{-1} + \frac{(1 - \theta_1)}{Be_1 \cdot e_1} e_1 \otimes e_1.$$

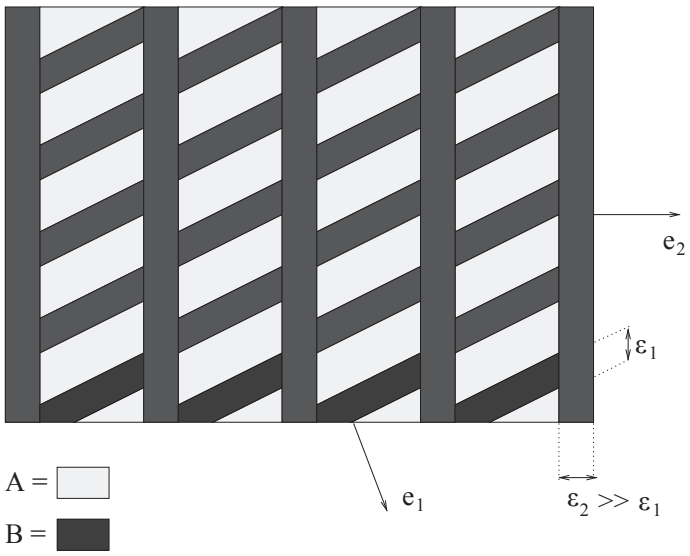


Fig. 19. A sequential laminate composite.

This simple laminate A_1^* can again be laminated with phase B, in direction e_2 and in proportions θ_2, $(1-\theta_2)$ respectively, to obtain a new laminate denoted by A_2^*. By induction, we obtain A_p^* by lamination of A_{p-1}^* and B, in direction e_p and in proportions θ_p, $(1-\theta_p)$, respectively. Then the homogenized tensor A_p^* is

$$\theta_p(A_p^* - B)^{-1} = (A_{p-1} - B)^{-1} + \frac{(1-\theta_p)}{Be_p \cdot e_p} e_p \otimes e_p. \tag{4.21}$$

Replacing $(A_{p-1}^* - B)^{-1}$ in (4.21) by the similar formula defining $(A_{p-2}^* - B)^{-1}$, and so on up to $A_0^* \equiv A$, we obtain a formula of the same type as (4.16), namely,

$$\left(\prod_{j=1}^{p} \theta_j\right)(A_p^* - B)^{-1} = (A - B)^{-1} + \sum_{i=1}^{p}\left((1-\theta_i)\prod_{j=1}^{i-1}\theta_j\right)\frac{e_i \otimes e_i}{Be_i \cdot e_i}. \tag{4.22}$$

We remark that we always laminate an intermediate laminate with the same phase B. In other words, the other phase A is coated by several layers of B. One can say that B plays the role of a matrix phase, and A plays the role of a core phase. Globally, A^* can be seen as a mixture of A and B in different layers having a large separation of scales (see Fig. 19).

Let us define rank-p sequential laminate with matrix B and inclusion A.

Lemma 4.13 (rank-p sequential laminate)**.** *If we laminate p times with B, we obtain a rank-p sequential laminate with matrix B and inclusion A, in proportions $(1-\theta)$ and θ, is defined by*

$$\theta(A_p^* - B)^{-1} = (A - B)^{-1} + (1-\theta)\sum_{i=1}^{p} m_i \frac{e_i \otimes e_i}{Be_i \cdot e_i}$$

with $\sum_{i=1}^{p} m_i = 1$ and $m_i \geq 0$, $1 \leq i \leq p$.

Proof. By (4.22) we already have

$$\left(\prod_{j=1}^{p} \theta_j\right)(A_p^* - B)^{-1} = (A - B)^{-1} + \sum_{i=1}^{p}\left((1-\theta_i)\prod_{j=1}^{i-1}\theta_j\right)\frac{e_i \otimes e_i}{Be_i \cdot e_i}.$$

We make the change of variables

$$\theta = \prod_{i=1}^{p}\theta_i \quad \text{and} \quad (1-\theta)m_i = (1-\theta_i)\prod_{j=1}^{i-1}\theta_j,\ 1 \leq i \leq p$$

which is indeed one-to-one with the constraints on the m_i's and the θ_i's. □

Of course the same can be done when exchanging the roles of A and B.

Lemma 4.14. *A rank-p sequential laminate with matrix A and inclusion B, in proportions θ and $(1-\theta)$, is defined by*

$$(1-\theta)(A_p^* - A)^{-1} = (B - A)^{-1} + \theta\sum_{i=1}^{p} m_i \frac{e_i \otimes e_i}{Ae_i \cdot e_i}$$

with $\sum_{i=1}^{p} m_i = 1$ *and* $m_i \geq 0$, $1 \leq i \leq p$.

Remark 4.15. *Sequential laminates form a very rich and explicit class of composite materials which, as we shall see, completely describes the boundaries of the set* G_θ.

4.2.2 Characterization of G_θ

From now on, we assume that the microscopic tensor $A(y)$ is symmetric. Then A^* is also symmetric. Furthermore, A^* is characterized by the following variational principle:

$$A^*\xi \cdot \xi = \min_{w \in H^1_\#(Y)/\mathbb{R}} \int_Y A(y)(\xi + \nabla w) \cdot (\xi + \nabla w)\, dy \quad \forall \xi \in \mathbb{R}^N. \tag{4.23}$$

Indeed, if w_ξ is a minimizer of (4.23), then it satisfies the Euler optimality condition

$$\begin{cases} -\mathrm{div}(A(y)(\xi + \nabla w_\xi(y))) = 0 & \text{in } Y, \\ y \mapsto w_\xi(y) & Y\text{-periodic}. \end{cases}$$

By linearity, we have $w_\xi = \sum_{i=1}^{N} \xi_i w_i$, where w_i $(i = 1, \ldots, N)$ denotes the solution of (4.10), and thus, by (4.13) we get

$$\int_Y A(y)(\xi + \nabla w_\xi) \cdot (\xi + \nabla w_\xi)\, dy = \sum_{i,j=1}^{N} \xi_i \xi_j A^*_{ij} = A^*\xi \cdot \xi.$$

By using the variational principle of A^* (4.23), we can obtain arithmetic and harmonic mean bounds for A^*.

Lemma 4.16 (Arithmetic and harmonic mean bounds)**.** *Any homogenized tensor* A^* *satisfies the arithmetic mean bound*

$$A^*\xi \cdot \xi \leq \left(\int_Y A(y)\, dy \right) \xi \cdot \xi$$

and the harmonic mean bound

$$\left(\int_Y A^{-1}(y)\, dy \right)^{-1} \xi \cdot \xi \leq A^*\xi \cdot \xi.$$

Proof. Taking $w = 0$ in the variational principle (4.23), we deduce the arithmetic mean bound. For the harmonic mean bound we enlarge the minimization space as follows. Indeed, since $\int_Y \nabla w\, dy = 0$, we replace ∇w with any vector field $\zeta(y)$ with zero-average on Y

$$A^*\xi \cdot \xi \geq \min_{\substack{\zeta \in L^2_\#(Y)^N, \\ \int_Y \zeta\, dy = 0}} \int_Y A(y)(\xi + \zeta(y)) \cdot (\xi + \zeta(y))\, dy.$$

The Euler equation for the minimizer $\zeta_\xi(y)$ of this convex problem is

$$A(y)(\xi + \zeta_\xi(y)) = \lambda,$$

where $\lambda \in \mathbb{R}$ is the Lagrange multiplier for the constraint $\int_Y \zeta\, dy = 0$. Thus

$$\xi = \left(\int_Y A(y)^{-1}\, dy \right) \lambda$$

and

$$\int_Y A(y)(\xi + \zeta_\xi(y)) \cdot (\xi + \zeta_\xi(y))\, dy = \left(\int_Y A(y)^{-1}\, dy \right)^{-1} \xi \cdot \xi.$$

□

Lemma 4.16 can be improved for two-phase composites. Next, we consider two isotropic phases $A = \alpha \mathrm{Id}$ and $B = \beta \mathrm{Id}$ with $0 < \alpha < \beta$.

Theorem 4.17 (Hashin and Shtrikman bounds [HS1963, TA2000])**.** *The set* G_θ *of all homogenized tensors obtained by mixing* α *and* β *in proportions* θ *and* $(1 - \theta)$ *is the set of all symmetric matrices* A^* *with eigenvalues* $\lambda_1, \ldots, \lambda_N$ *such that*

$$\left(\frac{\theta}{\alpha} + \frac{1 - \theta}{\beta} \right)^{-1} = \lambda_\theta^- \leq \lambda_i \leq \lambda_\theta^+ = \theta\alpha + (1 - \theta)\beta, \ 1 \leq i \leq N, \tag{4.24}$$

$$\sum_{i=1}^{N} \frac{1}{\lambda_i - \alpha} \leq \frac{1}{\lambda_\theta^- - \alpha} + \frac{N-1}{\lambda_\theta^+ - \alpha}, \tag{4.25}$$

$$\sum_{i=1}^{N} \frac{1}{\beta - \lambda_i} \leq \frac{1}{\beta - \lambda_\theta^-} + \frac{N-1}{\beta - \lambda_\theta^+}. \tag{4.26}$$

Furthermore, these so-called Hashin and Shtrikman bounds are optimal and attained by rank-N sequential laminates.

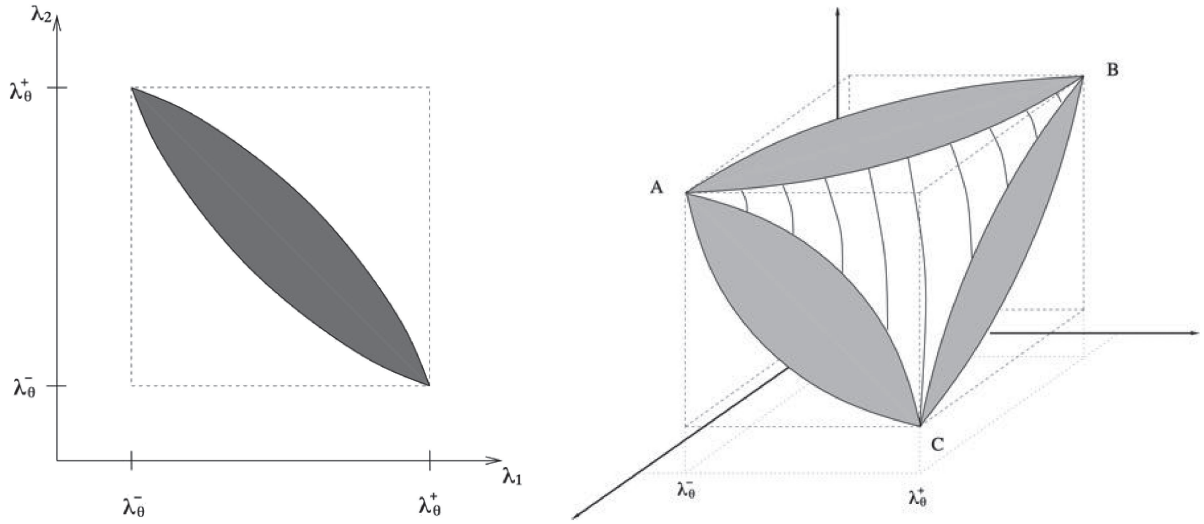


Fig. 20. The set G_θ in dimension $N = 2$ and 3 respectively.

Proof. We first show that all matrices satisfying these inequalities (Hashin–Shtrikman bounds) belong to G_θ. Let us start by showing that the upper bound (4.26) is attained by sequential laminates. Take a matrix A^* such that

$$\sum_{i=1}^{N} \frac{1}{\beta - \lambda_i} = \frac{1}{\beta - \lambda_\theta^-} + \frac{N-1}{\beta - \lambda_\theta^+}.$$

Define a rank-N sequential laminate A_L^* of matrix β and inclusion α, with lamination directions being the (orthogonal) eigenvectors of A^*. By Lemma 4.13 we have

$$\theta(A_L^* - \beta \mathrm{Id})^{-1} = \frac{1}{\alpha - \beta}\mathrm{Id} + (1-\theta)\sum_{i=1}^{N} m_i \frac{e_i \otimes e_i}{\beta} \text{ with } m_i \geq 0, \ \sum_{i=1}^{N} m_i = 1.$$

We obtain $A^* = A_L^*$ if we can choose the m_i's such that

$$\frac{\theta}{\lambda_i - \beta} = \frac{1}{\alpha - \beta} + \frac{m_i(1-\theta)}{\beta},$$

that is,

$$m_i = \frac{\beta(\lambda_\theta^+ - \lambda_i)}{(1-\theta)(\beta - \alpha)(\beta - \lambda_i)}.$$

We check that $0 < m_i < 1$ is equivalent to $\lambda_\theta^- < \lambda < \lambda_\theta^+$ and that

$$\sum_{N=1}^{N} m_i = 1 \Longleftrightarrow \sum_{i=1}^{N} \frac{1}{\beta - \lambda_i} = \frac{1}{\beta - \lambda_\theta^-} + \frac{N-1}{\beta - \lambda_\theta^+}.$$

Thus any matrix on the upper bound (4.26) is a rank-N sequential laminate with matrix β and inclusion α. The same proof works for the lower bound (4.25) upon exchanging the role of α (now the matrix) and β (now the inclusion).

A simple but lengthy computation shows that all the matrices satisfying the inequalities (4.24), (4.25), and (4.26) can be obtained as a rank-N sequential laminate of two suitable matrices, one realizing the equality in the upper bound (4.26) and the other realizing the equality in the lower bound (4.25) (see the full proof of [Al2002, Theorem 2.2.13] for the details). It remains to prove that the lower and upper Hashin–Shtrikman bounds hold true. To establish the lower bound (4.25) we introduce the so-called Hashin and Shtrikman variational principle. Main idea is to use Fourier analysis and Plancherel theorem.

By definition of A^*, for $\xi \in \mathbb{R}^N$, we have

$$A^*\xi \cdot \xi = \min_{w \in H^1_\#(Y)} \int_Y (\chi(y)\alpha + (1-\chi(y))\beta)(\xi + \nabla w) \cdot (\xi + \nabla w)\, dy.$$

Subtracting a reference material α,

$$\int_Y (\chi\alpha + (1-\chi)\beta)|\xi + \nabla w|^2\, dy = \int_Y (1-\chi)(\beta-\alpha)|\xi + \nabla w|^2\, dy + \int_Y \alpha|\xi + \nabla w|^2\, dy.$$

We use convex duality (or Legendre transform): for any symmetric positive definite matrix K, the following holds

$$K\zeta \cdot \zeta = \max_{\eta \in \mathbb{R}^N} (2\zeta \cdot \eta - K^{-1}\eta \cdot \eta) \quad \forall \zeta \in \mathbb{R}^N. \tag{4.27}$$

Since $0 < \alpha < \beta$, we apply the formula (4.27) at each point in Y. Then we get

$$\begin{aligned} &\int_Y (1-\chi)(\beta-\alpha)|\xi + \nabla w|^2\, dy \\ &\quad = \max_{\eta \in L^2_\#(Y)^N} \int_Y (1-\chi)(2(\xi + \nabla w) \cdot \eta - (\beta-\alpha)^{-1}|\eta|^2)\, dy, \end{aligned}$$

which becomes an inequality if we restrict the minimization to constant η in Y

$$\begin{aligned} \int_Y (1-\chi)(\beta-\alpha)|\xi + \nabla w|^2\, dy &\geq \max_{\eta \in \mathbb{R}^N} \int_Y (1-\chi)(2(\xi + \nabla w) \cdot \eta - (\beta-\alpha)^{-1}|\eta|^2)\, dy \\ &\geq (2\xi \cdot \eta - (\beta-\alpha)^{-1}|\eta|^2) - 2\int_Y \chi \nabla w \cdot \eta\, dy. \end{aligned}$$

On the other hand, because of periodicity, $\int_Y \nabla w\, dy = 0$ which implies

$$\int_Y \alpha|\xi + \nabla w|^2\, dy = \alpha|\xi|^2 + \int_Y \alpha|\nabla w|^2\, dy.$$

Overall, we obtain that, for any $\eta \in \mathbb{R}^N$,

$$A^*\xi \cdot \xi \geq \alpha|\xi|^2 + (1-\theta)(2\xi \cdot \eta - (\beta-\alpha)^{-1}|\eta|^2) - g(\chi, \eta), \tag{4.28}$$

where $g(\chi, \eta)$ is a so-called non-local term, defined by

$$g(\chi, \eta) = \min_{w \in H^1_\#(Y)} \int_Y (\alpha|\nabla w|^2 - 2\chi \nabla w \cdot \eta)\, dy.$$

We can now use Fourier analysis to compute $g(\chi, \eta)$. By periodicity, both χ and the test function w can be written as Fourier series:

$$\chi(y) = \sum_{k \in \mathbb{Z}^N} \hat{\chi}(k) e^{2i\pi k \cdot y}, \quad w(y) = \sum_{k \in \mathbb{Z}^N} \hat{w}(k) e^{2i\pi k \cdot y}.$$

Since χ and w are real-valued, their Fourier coefficients satisfy

$$\overline{\hat{\chi}(k)} = \hat{\chi}(-k) \text{ and } \overline{\hat{w}(k)} = \hat{w}(-k) \quad \forall k \in \mathbb{Z}^N.$$

The gradient of w at $y \in Y$ is given by

$$\nabla w(y) = \sum_{k \in \mathbb{Z}^N} 2i\pi e^{2i\pi k \cdot y} \hat{w}(k) k.$$

Then, Plancherel formula yields

$$\begin{aligned} \int_Y (\alpha|\nabla w|^2 - 2\chi \nabla w \cdot \eta)\, dy &= \sum_{k \in \mathbb{Z}^N} (4\pi^2 \alpha |\hat{w}(k)k|^2 - 4i\pi \overline{\hat{\chi}(k)} \hat{w}(k) k \cdot \eta) \\ &= \sum_{k \in \mathbb{Z}^N} (4\pi^2 \alpha |k|^2 |\hat{w}(k)|^2 + 4\pi \mathbb{I}m(\overline{\hat{\chi}(k)} \hat{w}(k)) \eta \cdot k). \end{aligned}$$

Notice that minimizing in $w(y) \in H^1_\#(Y)$ is equivalent to minimizing in $\hat{w}(k) \in \mathbb{C}$. For $k \neq 0$ the minimum is achieved by

$$\hat{w}(k) = -\frac{i\hat{\chi}(k)}{2\pi\alpha|k|^2} \eta \cdot k,$$

and we deduce that

$$g(\chi,\eta)=\left(\alpha^{-1}\sum_{k\in\mathbb{Z}^N,k\neq 0}|\hat{\chi}(k)|^2\,\frac{k}{|k|}\otimes\frac{k}{|k|}\right)\eta\cdot\eta=\alpha^{-1}\theta(1-\theta)M\eta\cdot\eta, \tag{4.29}$$

where M is a symmetric non-negative matrix defined by

$$M=\frac{1}{\theta(1-\theta)}\sum_{k\in\mathbb{Z}^N,k\neq 0}|\hat{\chi}(k)|^2\,\frac{k}{|k|}\otimes\frac{k}{|k|}\,.$$

Since, by Plancherel theorem, we have

$$\sum_{k\in\mathbb{Z}^N,k\neq 0}|\hat{\chi}(k)|^2=\int_Y|\chi(y)-\theta|^2\,dy=\theta(1-\theta),$$

we deduce that the trace of M is equal to 1.

Substituting (4.29) to (4.28), for any $\xi,\eta\in\mathbb{R}^N$,

$$A^*\xi\cdot\xi\geq\alpha|\xi|^2+(1-\theta)(2\xi\cdot\eta-(\beta-\alpha)^{-1}|\eta|^2)-\alpha^{-1}\theta(1-\theta)M\eta\cdot\eta. \tag{4.30}$$

The minimum (in ξ) of the inequality (4.30) is obtained when

$$\xi=(1-\theta)(A^*-\alpha)^{-1}\eta.$$

Then we deduce

$$(1-\theta)(A^*-\alpha)^{-1}\eta\cdot\eta\leq(\beta-\alpha)^{-1}|\eta|^2+\alpha^{-1}\theta M\eta\cdot\eta\quad\forall\eta\in\mathbb{R}^N.$$

Thus, we have

$$(1-\theta)(A^*-\alpha)^{-1}\leq(\beta-\alpha)^{-1}I+\alpha^{-1}\theta M. \tag{4.31}$$

Taking the trace of this matrix inequality (4.31), and recalling that $\operatorname{tr}M=1$, we obtain the lower Hashin–Shtrikman bound. The proof of the upper bound is similar. □

4.3 The elasticity setting

In what follows, let us consider the elasticity setting. The homogenization method can be generalized to the elasticity setting. However, an explicit characterization of G_θ is still lacking in the elasticity setting.

We set

$$\begin{aligned}A\xi&=2\mu_A\xi+\lambda_A(\operatorname{tr}\xi)I_2,\\ B\xi&=2\mu_B\xi+\lambda_B(\operatorname{tr}\xi)I_2,\end{aligned} \tag{4.32}$$

with the identity matrix I_2, and $\kappa_{A,B}=\lambda_{A,B}+2\mu_{A,B}/N$. We assume B to be weaker than A:

$$0\leq\mu_B<\mu_A,\quad 0\leq\kappa_B<\kappa_A.$$

We work with stresses rather than strains, thus we use inverse elasticity tensors. The similar results of the two-phase composites in the elasticity setting as follows (in details, see [Al2002, Sect. 2.3]):

Lemma 4.18 (Sequential laminates in elasticity)**.** *The Hooke's law of a simple laminate of A and B, in proportions θ and $(1-\theta)$, respectively, in the direction e, is*

$$(1-\theta)(A^{*^{-1}}-A^{-1})^{-1}=(B^{-1}-A^{-1})^{-1}+\theta f_A^c(e),$$

where $f_A^c(e)$ is the tensor, defined, for any symmetric matrix ξ, by

$$f_A^c(e_i)\xi\cdot\xi=A\xi\cdot\xi-\frac{1}{\mu_A}|A\xi e_i|^2+\frac{\mu_A+\lambda_A}{\mu_A(2\mu_A+\lambda_A)}((A\xi)e_i\cdot e_i)^2.$$

Proposition 4.19 (Reiterated lamination formula)**.** *A rank-p sequential laminate with matrix A and inclusions B, in proportions θ and $(1-\theta)$, respectively, in the directions $(e_i)_{1\leq i\leq p}$ with parameter $(m_i)_{1\leq i\leq p}$ such that $0\leq m_i\leq 1$ and $\sum_{i=1}^p m_i=1$, is given by*

$$(1-\theta)(A^{*^{-1}}-A^{-1})^{-1}=(B^{-1}-A^{-1})^{-1}+\theta\sum_{i=1}^p m_i f_A^c(e_i).$$

Theorem 4.20 (Hashin–Shtrikman bounds in elasticity)**.** *Let A^* be a homogenized elasticity tensor in G_θ which is assumed isotropic*

$$A^* = 2\mu_* I_4 + \left(\kappa_* - \frac{2\mu_*}{N}\right) I_2 \otimes I_2.$$

Its bulk κ_* and shear μ_* moduli satisfy

$$\frac{1-\theta}{\kappa_A - \kappa_*} \leq \frac{1}{\kappa_A - \kappa_B} + \frac{\theta}{2\mu_A + \lambda_A} \text{ and } \frac{\theta}{\kappa_* - \kappa_B} \leq \frac{1}{\kappa_A - \kappa_B} + \frac{1-\theta}{2\mu_B + \lambda_B},$$

$$\frac{1-\theta}{2(\mu_A - \mu_*)} \leq \frac{1}{2(\mu_A - \mu_B)} + \frac{\theta(N-1)(\kappa_A + 2\mu_A)}{(N^2+N-2)\mu_A(2\mu_A + \lambda_A)},$$

$$\frac{\theta}{2(\mu_* - \mu_B)} \leq \frac{1}{2(\mu_A - \mu_B)} - \frac{(1-\theta)(N-1)(\kappa_B + 2\mu_B)}{(N^2+N-2)\mu_B(2\mu_B + \lambda_B)}.$$

Furthermore, the two lower bounds, as well as the two upper bounds are simultaneously attained by a rank-p sequential laminate with $p = 3$ if $N = 2$, and $p = 6$ if $N = 3$.

Proof. We refer to [Al2002, Theorem 2.3.13] □

Remark 4.21. *These bounds do not characterize all possible isotropic homogenized tensors A^* in G_θ. In other words, there exist isotropic elasticity tensors with moduli satisfying these bounds that are not composite materials obtained by mixing phases A and B in proportions θ, $(1-\theta)$, respectively.*

Proposition 4.22 (Hashin–Shtrikman optimal energy bound)**.** *Let G_θ be the set of all homogenized elasticity tensors obtained by mixing the two phases A and B in proportions θ and $(1-\theta)$. Let L_θ be the subset of G_θ made of sequential laminated composites. For any stress σ,*

$$HS(\sigma) = \min_{A^* \in G_\theta} A^{*^{-1}}\sigma \cdot \sigma = \min_{A^* \in L_\theta} A^{*^{-1}}\sigma \cdot \sigma.$$

Furthermore, the minimum is attained by a rank-N sequential laminate with lamination directions given by the eigendirections of σ.

Remark 4.23. *An optimal tensor A^* can be interpreted as the most rigid composite material in G_θ able to sustain the stress σ. $HS(\sigma)$ is called Hashin–Shtrikman optimal energy bound. In practical conclusion, G_θ can be replaced by L_θ for compliance minimization.*

4.4 Numerical applications

Let us consider the case of parametrized periodicity cells. For example, the square cell with a rectangular hole (as used in the seminal work of Bendsøe and Kikuchi [BK1988]), parametrized by m_1, m_2, and denoted by $Y(m)$ (see Fig. 21).

We compute the so-called correctors or cell solutions:

$$\begin{cases} \operatorname{div}(A(e_{ij} + e(w_{ij}))) = 0 & \text{in } Y(m), \\ A(e_{ij} + e(w_{ij})) \cdot n = 0 & \text{on } \Gamma_{\text{int}}, \\ y \mapsto w_{ij}(y) & (0,1)^2\text{-periodic}, \end{cases}$$

where $e_{ij} = (e_i \otimes e_j + e_j \otimes e_i)/2$ is a basis of the symmetric tensors of order 2, and n is the normal to the hole's boundary Γ_{int} in $Y(m)$. Hence we find a unique solution (up to an additive translation) $w_{ij} \in H^1_\#(Y(m), \mathbb{R}^2)$ to the variational formulation:

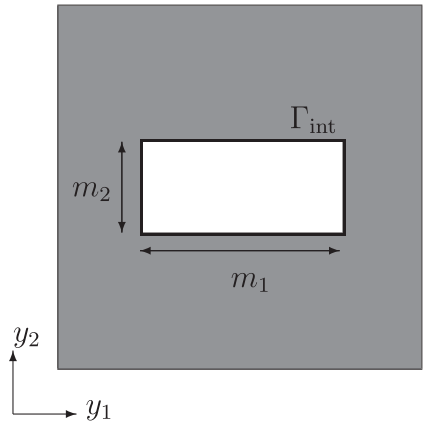


Fig. 21. Square cell with a rectangular hole.

$$\int_{Y(m)} Ae(w_{ij}) \cdot e(\phi)\, dy + \int_{Y(m)} Ae_{ij} \cdot e(\phi)\, dy = 0 \quad \forall \phi \in H^1_{\#}(Y(m), \mathbb{R}^2). \tag{4.33}$$

The tensor A^* is then given by

$$A^*_{ijkl} = \int_{Y(m)} A(e_{ij} + e(w_{ij})) \cdot (e_{kl} + e(w_{kl}))\, dy, \quad i, j, k, l \in \{1, 2\}. \tag{4.34}$$

4.5 Exercises

Problem 4.5.1. *Compute numerically A^* for various parameters m_1, m_2.*

Problem 4.5.2. *Orthotropic composite: check numerically that $A^*_{1112} = A^*_{2212} = 0$, then prove it theoretically.*

Problem 4.5.3. *Check that if $m_1 \to 1$, then A^* is close to the formula of a rank-1 laminate.*

Problem 4.5.4. *What happens if $m_1 \to 0$? Is A^* close to A?*

Problem 4.5.5. *If $m_1 = m_2$, is A^* isotropic?*

Problem 4.5.6. *Check numerically that A^* is isotropic for a honeycomb structure with hexagonal holes.*

5. Topology Optimization by the Homogenization Method

5.1 Why topology optimization?

Shape optimization consists in "shape tracking" algorithms, hence the method cannot change the topology, such as the number of holes in the case of 2-dimension. On the other hand, topology optimization consists in "shape capturing" algorithms, that allow us to consider the optimization in a wider class which includes the different topological properties. There are several methods of topology optimization but we focus on just one, called the homogenization method (see [Al2002, BS2003] and references therein). In the following of this section, we introduce a model problem for topology optimization with a constraint on the volume of holes and study it using the method of homogenization as mentioned in Sect. 4.

5.2 Homogenization method in the conductivity setting

In this section, we apply the homogenization method in the conductivity setting. As we mentioned in Sect. 3.2, there is no minimizer for the corresponding minimizing problem (we will explain the detail below) in general. To solve the problem, we introduce a set of generalized shapes as the limit of classical shapes. More precisely, goals of the homogenization method for topology optimization are following:

- To introduce the notion of generalized shapes made of composite material,
- To show that those generalized shapes are limits of sequences of classical shapes (in the sense of homogenization),
- To compute the generalized objective function and its gradient,
- To prove an existence theorem for optimal generalized shapes,
- To deduce new numerical algorithms for topology optimization.

In order to consider the limit of classical shapes, we recall one of the main results of homogenization theory. For the details of the proof, see [Al2002, Theorem 1.2.16 and 2.1.2].

Theorem 5.1. *Let Ω be a bounded domain in $\mathbb{R}^N$ and $(\chi_\varepsilon(x))_{\varepsilon>0}$ be a sequence of characteristic functions in Ω. Set $A_\varepsilon(x) = \alpha\chi_\varepsilon(x) + \beta(1 - \chi_\varepsilon(x))$ for $x \in \Omega$. Then there exists a subsequence, still denoted by χ_ε, a density $0 \le \theta(x) \le 1$ and a homogenized tensor A^* such that*

$$\chi_\varepsilon \rightharpoonup \theta \quad \text{weakly * in } L^\infty(\Omega; [0, 1])$$

and A_ε converges in the sense of homogenization to A^*, i.e., for all $f \in L^2(\Omega)$, the solution u_ε of the problem

$$\begin{cases} -\mathrm{div}(A_\varepsilon(x)\nabla u_\varepsilon) = f & \text{in } \Omega, \\ u_\varepsilon = 0 & \text{on } \partial\Omega \end{cases}$$

converges strongly in $L^2(\Omega)$ to the solution u of the homogenized problem

$$\begin{cases} -\mathrm{div}(A^*(x)\nabla u) = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{5.1}$$

Furthermore, for almost all $x \in \Omega$, $A^*(x)$ belongs to the set $G_{\theta(x)}$ defined in Definition 4.8.

5.2.1 Relaxed problem for the conductivity model

In this subsection, we introduce the conductivity model and derive generalized shapes as the limits of classical

shapes by the homogenization method. We impose a simplifying assumption: the "holes" (with a Neumann, free boundary condition) are filled with a weak ("ersatz") material α, while the other space is filled with a material β, that is, $\alpha < \beta$. We consider a membrane we introduced in Sect. 3 with two thicknesses α and β, whose distribution is give by $h_\chi = \alpha\chi + \beta(1-\chi)$, where χ is a characteristic function which denotes the position of the holes. If $f \in L^2(\Omega)$ is the applied load, the displacement u_χ satisfies

$$\begin{cases} -\mathrm{div}(h_\chi \nabla u_\chi) = f & \text{in } \Omega, \\ u_\chi = 0 & \text{on } \partial\Omega. \end{cases}$$

We will optimize the membrane's shape amounts by considering the minimizing problem

$$\inf_{\chi \in \mathcal{U}_{\mathrm{ad}}} J(\chi), \tag{5.2}$$

with

$$\mathcal{U}_{\mathrm{ad}} = \left\{ \chi \in L^\infty(\Omega; \{0,1\}) : \int_\Omega \chi(x)\,dx = V_\alpha \right\} \quad \text{and} \quad J(\chi) = \int_\Omega j(u_\chi)\,dx,$$

where V_α is a given positive constant, denoting the volume of the holes and

$$j(u) = fu, \text{ or } |u - u_0|^2. \tag{5.3}$$

Since there is no minimizer of (5.2) in general, we apply Theorem 5.1 to introduce a set of generalized shapes. Let χ_ε be a sequence (minimizing sequence of (5.2) or not) of characteristic functions. Applying Theorem 5.1, we see that there exist θ and A^* such that

$$\begin{aligned} &\chi_\varepsilon \rightharpoonup \theta && \text{weakly * in } L^\infty(\Omega; [0,1]), \\ &A_\varepsilon \to A^* && \text{in the sense of homogenization} \end{aligned}$$

and

$$J(\chi_\varepsilon) = \int_\Omega j(u_\varepsilon)\,dx \to \int_\Omega j(u)\,dx =: J(\theta, A^*),$$

where u is the solution of (5.1) and j is defined by (5.3). From this convergence result, we define the set of admissible homogenized shapes

$$\mathcal{U}^*_{\mathrm{ad}} := \left\{ (\theta, A^*) \in L^\infty(\Omega; [0,1] \times \mathbb{R}^{N^2}) \ : \ A^*(x) \in G_{\theta(x)} \text{ a.e. in } \Omega, \ \int_\Omega \theta(x)\,dx = V_\alpha \right\}$$

and consider the following relaxed or homogenized optimization problem:

$$\inf_{(\theta, A^*) \in \mathcal{U}^*_{\mathrm{ad}}} J(\theta, A^*). \tag{5.4}$$

We easily check that $\mathcal{U}_{\mathrm{ad}} \subset \mathcal{U}^*_{\mathrm{ad}}$ if we identify $\chi \in \mathcal{U}_{\mathrm{ad}}$ with the pair $(\chi, \alpha\chi \mathrm{Id} + \beta(1-\chi)\mathrm{Id}) \in \mathcal{U}^*_{\mathrm{ad}}$. The inclusion implies that we have enlarged the set of admissible shapes. Moreover, one can prove that the relaxed problem (5.4) always admits an optimal solution, and the homogenized formulation is a relaxation of the original topology optimization problem as follows.

Theorem 5.2. *The homogenized formulation is a relaxation of the original topology optimization problem in the sense that:*

- *there exists, at least, one optimal composite shape $(\theta, A^*) \in \mathcal{U}^*_{\mathrm{ad}}$, i.e., a minimizer of (5.4),*
- *for any minimizing sequence $(\chi_n)_{n\in\mathbb{N}}$ of (5.2), there exists a minimizer $(\theta, A^*) \in \mathcal{U}^*_{\mathrm{ad}}$ of (5.4) such that, up to subsequence, χ_n converges weakly * in $L^\infty(\Omega; [0,1])$ to θ and $A_n := \alpha\chi_n + \beta(1-\chi_n)$ converges to A^* in the sense of homogenization,*
- *any composite optimal solution $(\theta, A^*) \in \mathcal{U}^*_{\mathrm{ad}}$ of (5.4) is the limit of a minimizing sequence of (5.2).*

Moreover, the infima of the original and homogenized objective functions coincide

$$\inf_{\chi \in \mathcal{U}_{\mathrm{ad}}} J(\chi) = \min_{(\theta, A^*) \in \mathcal{U}^*_{\mathrm{ad}}} J(\theta, A^*).$$

For the proof of Theorem 5.2, see [Al2002, Theorem 3.2.1]. Theorem 5.2 means that the topology optimization problem is not changed by relaxation. Moreover, close to any optimal composite shape, we are sure to find a quasi-optimal classical shape. This theorem is at the root of new numerical algorithms.

Remark 5.3. *The homogenized formulation is similar to a parametric or sizing optimization problem. This is the main reason why the homogenization method is computationally cheap and works like a shape capturing algorithm. Moreover, computing gradients or optimality conditions are thus very simple. On the other hand, the design parameters (θ, A^*) are quite complicated. Another further (drastic and unjustified) simplification is to suppress the parameter A^**

and to keep only the material density θ. This is the main idea of the SIMP method [BS2003], we will mention in Sect. 5.4.

5.2.2 Optimality conditions

In Sect. 5.2.1, we introduced the relaxed formulation (5.4) of the original optimization problem (5.2). One of the advantages of the relaxed problem is that we always have the existence of a minimizer of (5.4). There is also another advantage: we can get the optimality condition for the relaxed problem since it is possible to perform variations of composite designs. In this subsection, we will consider the optimality condition for problem (5.4).

We now compute the gradient of the following objective function

$$J(\theta, A^*) = \int_\Omega |u - u_0|^2 \, dx,$$

where u is the solution to (5.1) and u_0 is a given function in $L^2(\Omega)$. We introduce the adjoint state p of u as the unique solution in $H_0^1(\Omega)$ of

$$\begin{cases} -\mathrm{div}(A^* \nabla p) = -2(u - u_0) & \text{in } \Omega, \\ p = 0 & \text{on } \partial\Omega. \end{cases} \tag{5.5}$$

By the use of the adjoint state p, we can obtain the derivative of the functional J.

Proposition 5.4. *Let $\alpha > 0$ and $\mathcal{M}_\alpha$ be the set of symmetric positive definite matrices M such that $M \geq \alpha \mathrm{Id}$. The functional J is differentiable with respect to A^* in $L^\infty(\Omega; \mathcal{M}_\alpha)$, and its derivative is*

$$\nabla_{A^*} J(\theta, A^*) = \nabla u \otimes \nabla p,$$

i.e.,

$$\langle \nabla_{A^*} J(\theta, A^*), B^* \rangle = \int_\Omega B^* \nabla u \cdot \nabla p \, dx, \quad B^* \in L^\infty(\Omega; \mathcal{M}_\alpha)$$

where $u \in H_0^1(\Omega)$ is the unique solution of (5.1) and p is the adjoint state (5.5) of u.

Proof. We can prove the proposition by the same strategy that we used in Lemma 3.8, 3.9 and Theorem 3.10. Thus we omit the proof. □

Remark that the partial derivative with respect to θ vanishes because θ appears only in the constraint of A^*. Moreover, we can also consider the Lagrangian

$$\mathcal{L}(A^*, v, q) = \int_\Omega |v - v_0|^2 \, dx + \int_\Omega A^* \nabla v \cdot \nabla q \, dx - \int_\Omega fq \, dx,$$

where $(A^*, v, q) \in L^\infty(\Omega; \mathcal{M}_\alpha) \times H_0^1(\Omega) \times H_0^1(\Omega)$. The partial derivatives of $\mathcal{L}$ with respect to q and v yield the state and adjoint state respectively. Furthermore, the functional J is also differentiable with respect to A^* with derivative

$$\nabla_{A^*} J(\theta, A^*) = \frac{\partial \mathcal{L}}{\partial A^*}(A^*, u, p) = \nabla u \otimes \nabla p.$$

The essential consequence of this section is the following optimality condition.

Theorem 5.5. *Let (θ, A^*) be a global minimizer of J in $\mathcal{U}_{\mathrm{ad}}^*$ which admits u and p as state and adjoint. Then there exists $(\tilde{\theta}, \tilde{A}^*)$, another global minimizer of J in $\mathcal{U}_{\mathrm{ad}}^*$, which admits the same state and adjoint u and p, and such that $\tilde{A}^*$ is a rank-1 simple laminate.*

Proof. We fix $\theta \in L^\infty(\Omega; [0, 1])$ with

$$\int_\Omega \theta(x) \, dx = V_\alpha.$$

We remark that by Theorem 4.17, the set

$$\mathcal{G}_\theta := \{A^0 \in L^\infty(\Omega; \mathbb{R}^{N^2}) : A^0(x) \in G_{\theta(x)} \text{ a.e. } x \in \Omega\}$$

is a convex set. Then, by Theorem 2.17 and Proposition 5.4, we have

$$\int_\Omega (A^0 - A^*) \nabla u \cdot \nabla p \, dx \geq 0 \quad \forall A^0 \in \mathcal{G}_\theta.$$

We easily check that the above inequality is equivalent to the point-wise constraint

$$A^*(x) \nabla u(x) \cdot \nabla p(x) = \min_{A^0 \in G_{\theta(x)}} (A^0 \nabla u(x) \cdot \nabla p(x)) \tag{5.6}$$

for almost all $x \in \Omega$. Fix $x \in \Omega$ which satisfies (5.6). If $\nabla u(x)$ or $\nabla p(x)$ vanishes, then any $A^* \in G_{\theta(x)}$ is optimal in (5.6)

(the fact that u and p does not change at these points is more delicate to establish and we refer to [Al2002, TA2000] for details). Otherwise, we define two unit vectors

$$e = \frac{\nabla u}{|\nabla u|} \text{ and } e' = \frac{\nabla p}{|\nabla p|}.$$

Then (5.6) is equivalent to finding a minimizer $A^*(x)$ of

$$4A^0 e \cdot e' = A^0(e+e') \cdot (e+e') - A^0(e-e') \cdot (e-e').$$

A lower bound is easily seen to be

$$\begin{aligned} \min_{A^0 \in G_\theta} 4A^0 e \cdot e' &\geq \min_{A^0 \in G_\theta} A^0(e+e') \cdot (e+e') - \max_{A^0 \in G_\theta} A^0(e-e') \cdot (e-e') \\ &= \lambda_\theta^- |e+e'|^2 - \lambda_\theta^+ |e-e'|^2, \end{aligned} \tag{5.7}$$

where $\lambda_\theta^\pm$ is defined in Theorem 4.17. We can see that the lower bound is attained by a matrix A^1 corresponding to a rank-1 laminate (see [Al2002, Remark 2.2.14]). More precisely, A^1 satisfies

$$A^1(e+e') = \lambda_\theta^-(e+e'), \quad A^1(e-e') = \lambda_\theta^+(e-e'). \tag{5.8}$$

Thus we obtain

$$\min_{A^0 \in G_\theta} 4A^0 e \cdot e' = \lambda_\theta^- |e+e'|^2 - \lambda_\theta^+ |e-e'|^2.$$

Moreover, if A^* is any optimal tensor, then, A^* must also satisfy

$$A^*(e+e') = \lambda_\theta^-(e+e') \text{ and } A^*(e-e') = \lambda_\theta^+(e-e'). \tag{5.9}$$

Indeed, if (5.9) does not hold true, the bounds on eigenvalues in Lemma 4.17 imply the strict inequality

$$4A^* e \cdot e' = A^*(e+e') \cdot (e+e') - A^*(e-e') \cdot (e-e') > \lambda_\theta^- |e+e'|^2 - \lambda_\theta^+ |e-e'|^2,$$

which is a contradiction with (5.6) and (5.7). By (5.8) and (5.9), we see that

$$\begin{aligned} 2A^* \nabla u &= (\lambda_\theta^+ + \lambda_\theta^-)\nabla u + (\lambda_\theta^+ - \lambda_\theta^-)\frac{|\nabla u|}{|\nabla p|}\nabla p = 2A^1 \nabla u, \\ 2A^* \nabla p &= (\lambda_\theta^+ + \lambda_\theta^-)\nabla p + (\lambda_\theta^+ - \lambda_\theta^-)\frac{|\nabla p|}{|\nabla u|}\nabla u = 2A^1 \nabla p \end{aligned}$$

for almost all $x \in \Omega$. Therefore any optimal tensor A^* can be replaced by the rank-1 simple laminate A^1 without changing u and p. □

Remark 5.6. *Theorem 5.5 implies that in the definition of $\mathcal{U}_{\text{ad}}^*$, the set G_θ can be replaced by its simpler subset of rank-1 simple laminates. We actually use this simplification in the numerical algorithms. We remark that this simplification holds true for other objective functions as well. However, it does not hold for multiple loads optimization in general. For the details concerning the conditions that the objective function J must satisfy in order to obtain the optimality conditions, see* [Al2002, Sect. 3.2.2].

As stated in Remark 5.6, we can simplify the admissible set of the minimization problem (5.4). We consider the parametrization of rank-1 laminates. For simplicity, we consider the case $N = 2$. A rank-1 laminate is defined by

$$A^*(\theta, \phi) = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} \lambda_\theta^+ & 0 \\ 0 & \lambda_\theta^- \end{pmatrix} \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix},$$

where the angle $\phi \in [0, \pi]$ determines the orientation of the unit cell. Hence the admissible set is rewritten as

$$\mathcal{U}_{\text{ad}}^L := \left\{ (\theta, \phi) \in L^\infty(\Omega; [0,1] \times [0,\pi]) : \int_\Omega \theta(x)\, dx = V_\alpha \right\}.$$

If we set $J(\theta, \phi) := J(\theta, A^*(\theta, \phi))$, then the derivative of the objective function $J(\theta, \phi)$ follows by Proposition 5.4 immediately.

Proposition 5.7. *The objective function $J(\theta, \phi)$ is differentiable with respect to (θ, ϕ) in $\mathcal{U}_{\text{ad}}^L$, and its partial derivatives are*

$$\nabla_\phi J(\theta, \phi) = \frac{\partial A^*}{\partial \phi} \nabla u \cdot \nabla p \text{ and } \nabla_\theta J(\theta, \phi) = \frac{\partial A^*}{\partial \theta} \nabla u \cdot \nabla p.$$

5.2.3 Numerical algorithm

In this subsection, we show the numerical algorithm to seek the optimal shape θ of (5.4). As stated in Sect. 5.2.2, we can treat $J(\theta, \phi)$ instead of $J(\theta, A^*)$ if $N = 2$. We explain the projected gradient algorithm for the minimization of

$J(\theta, \phi)$ by the use of Proposition 5.7.

Algorithm 3 Projected gradient algorithm for (5.4)

1. We initialize the design parameters θ_0 and ϕ_0 (for example, equal to constants).
2. Until convergence, for $k \geq 0$ we iterate by computing the state u_k and adjoint p_k, solutions of (5.1) and (5.5) respectively with respect to the previous design parameters (θ_k, ϕ_k), then we update these parameters by

$$\theta_{k+1} = \max\left(0, \min\left(1, \theta_k - t_k\left(\ell_k + \frac{\partial A^*}{\partial \theta}(\theta_k, \phi_k)\nabla u_k \cdot \nabla p_k\right)\right)\right),$$

$$\phi_{k+1} = \phi_k - t_k \frac{\partial A^*}{\partial \theta}(\theta_k, \phi_k)\nabla u_k \cdot \nabla p_k,$$

where ℓ_k a Lagrange multiplier for the volume constraint, and $t_k > 0$ a descent step such that $J(\theta_{k+1}, \phi_{k+1}) < J(\theta_k, \phi_k)$.

For the details about the multiplier ℓ_k, we refer to Sect. 3.5. In the following, we will consider two simpler self-adjoint cases. Finally, we will show the algorithm to obtain the optimal shape which is close to the classical shapes.

First example of a self-adjoint case

A first example is the minimization of the torsional rigidity (maximization of compliance)

$$\min_{(\theta, A^*) \in \mathcal{U}_{\text{ad}}^L} \left\{ J(\theta, A^*) = -\int_\Omega u(x)\, dx \right\}, \tag{5.10}$$

where u is the solution of

$$\begin{cases} -\text{div}(A^*\nabla u) = 1 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{5.11}$$

In this case, the adjoint state p is the solution of

$$\begin{cases} -\text{div}(A^*\nabla p) = 1 & \text{in } \Omega, \\ p = 0 & \text{on } \partial\Omega, \end{cases}$$

i.e., the adjoint state is just $p = u$.

Before applying a numerical algorithm, we simplify the minimization problem by using the argument in Sect. 5.2.2. By the similar argument as in Proposition 5.4, we get

$$\langle \nabla_{A^*} J(\theta, A^*), B^* \rangle = \int_\Omega B^* \nabla u \cdot \nabla u\, dx \geq 0, \quad B^* \in L^\infty(\Omega; \mathcal{M}_\alpha).$$

Hence we have to decrease A^* to minimize J. Indeed, by (5.9) any minimizer (θ, A^*) satisfies

$$A^*\nabla u = \lambda_\theta^- \nabla u \quad \text{a.e. in } \Omega,$$

i.e., the optimal composite is the worst possible conductor. This condition allows us to eliminate the angle ϕ and it remains to optimize with respect to θ only, i.e., instead of (5.11) we have to consider the state of the problem

$$\begin{cases} -\text{div}(\lambda_\theta^- \nabla u) = 1 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{5.12}$$

Moreover, we recall that the state u, which is the solution of (5.12), is characterized as the minimizer of the corresponding energy, i.e.,

$$\int_\Omega \lambda_\theta^- |\nabla u|^2\, dx - 2\int_\Omega u\, dx = \min_{v \in H_0^1(\Omega)} \left\{ \int_\Omega \lambda_\theta^- |\nabla v|^2\, dx - 2\int_\Omega v\, dx \right\}.$$

On the other hand, by the weak form of the state u, we have

$$\int_\Omega \lambda_\theta^- |\nabla u|^2\, dx = \int_\Omega u\, dx$$

and thus

$$-\int_\Omega u\, dx = \min_{v \in H_0^1(\Omega)} \left\{ \int_\Omega \lambda_\theta^- |\nabla v|^2\, dx - 2\int_\Omega v\, dx \right\}.$$

Therefore, we can rewrite the minimizing problem (5.10) as

$$\min_{(\theta,v)}\left\{\int_\Omega \lambda_\theta^-|\nabla v|^2\,dx - 2\int_\Omega v\,dx\right\},$$

where the minimum in the right hand side of the above equation is taken over the set

$$\left\{(\theta,v)\in L^\infty(\Omega;[0,1])\times H_0^1(\Omega):\int_\Omega \theta\,dx = V_\alpha\right\}.$$

Furthermore, since the function $(\theta,v)\mapsto\lambda_\theta^-|\nabla v|^2$ is convex, there are only global minima by Proposition 2.10.

Numerically, we use an algorithm based on alternate direction minimization (see Sect. 3.5). We solve in the domain $\Omega=(0,1)^2$ with phases $\alpha=1$ and $\beta=2$ respectively. We work with a volume constraint 50% of phase α. We initialize with a constant value of $\theta=0.5$ and a constant zero lamination angle $\phi=0$. We perform 30 iterations. We show the numerical result how the objective function J convergent to some value (Fig. 22) and the volume fraction θ at some iteration numbers (Fig. 23). Here, the horizontal axis in Fig. 22 means the iteration number.

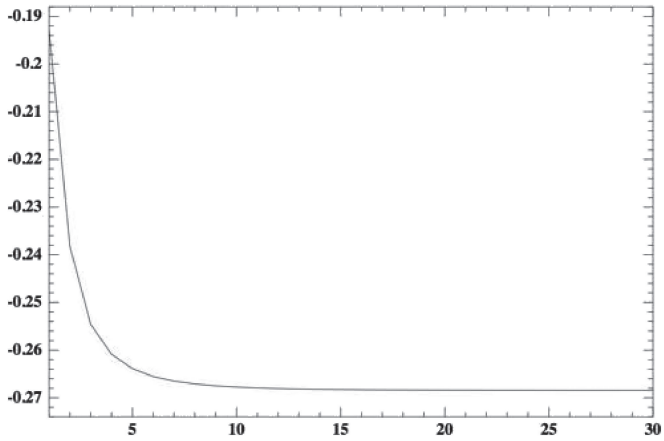


Fig. 22. Convergence history of J.

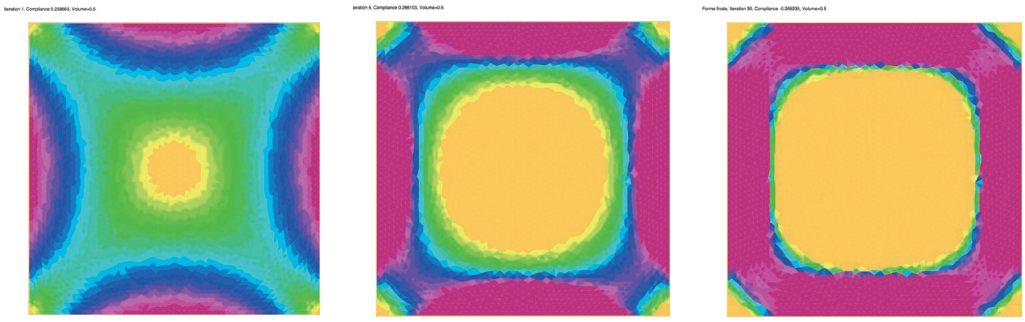


Fig. 23. Volume fraction θ (iteration number 1, 5, and 30 respectively) under the following color convention: red = 1, yellow = 0.

Second example of a self-adjoint case

A second self-adjoint example is a compliance minimization

$$\min_{(\theta,A^*)\in\mathfrak{U}_{\rm ad}^L}\left\{J(\theta,A^*)=\int_\Omega u(x)\,dx\right\},\tag{5.13}$$

where u is the solution of (5.11). In this case, the adjoint state is $p=-u$.

In order to apply the numerical algorithm, we will simplify the minimizing problem as we did in the first example. By the same argument as in the first example, we see that

$$\langle\nabla_{A^*}J(\theta,A^*),B^*\rangle = -\int_\Omega B^*\nabla u\cdot\nabla u\,dx\le 0,\quad B^*\in L^\infty(\Omega;\mathcal{M}_\alpha)$$

and

$$A^*\nabla u=\lambda_\theta^+\nabla u$$

if (θ,A^*) is a minimizer of (5.13). Hence the optimal composite is the best possible conductor and we have only to

consider the following problem instead of (5.11):

$$\begin{cases} -\mathrm{div}(\lambda_\theta^+ \nabla u) = 1 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

Therefore, as in the previous section, we can eliminate the dependency on the angle ϕ and then optimize with respect to θ only. We rewrite the optimization problem thanks to the dual energy

$$\int_\Omega u\,dx = \min_{\substack{\tau\in L^2(\Omega)^N, \\ -\mathrm{div}\,\tau=1 \text{ in } \Omega}} \int_\Omega (\lambda_\theta^+)^{-1}|\tau|^2 dx.$$

We can obtain the dual energy by the similar calculation in Example 2.30. Thus instead of (5.13) we obtain a double minimization

$$\min_{(\theta,\tau)} \int_\Omega (\lambda_\theta^+)^{-1}|\tau|^2 dx,$$

where the minimum is taken over the set

$$\left\{(\theta,\tau) \in L^\infty(\Omega;[0,1]) \times (L^2(\Omega))^N : \int_\Omega \theta\,dx = V_\alpha,\ -\mathrm{div}\,\tau = 1 \text{ in } \Omega\right\}.$$

Furthermore, since the function $(\theta,\tau) \mapsto (\lambda_\theta^+)^{-1}|\tau|^2$ is convex, there are only global minima by Proposition 2.10.

We apply the same algorithm as in the first example. The setting of the problem is also same as in the first example. Fig. 24 shows the numerical result of the volume fraction θ at some iteration numbers.

Fig. 24. Volume fraction θ (iteration number 1, 10 and 30 respectively) under the following color convention: red = 1, yellow = 0.

Remark 5.8. *Thanks to the convexity properties of the functionals, the convergence to a global minimum is guaranteed. In practice, it can be checked by numerical experiments with various initializations converging to the same solution.*

If one is interested by shape optimization rather than two-phase optimization, then, in numerical practice, holes can be mimicked by a very weak phase α, such as $10^{-3}\beta$. Mathematically, when $\alpha \to 0$ we obtain Neumann boundary conditions on the holes boundaries.

Penalization

By the algorithm stated in the two examples, we obtain optimal shapes in the wider class of composite shapes. Since, in practice, we are rather interested in classical shapes, we choose to use a penalization process to force the density to take values close to 0 or 1.

Algorithm 4 Penalization process for (5.4)

Apply either of the following algorithms after convergence to a composite shape by the previous algorithm.

1. We add a penalization term to the objective function
$$J(\theta, A^*) + c_{\mathrm{pen}} \int_\Omega \theta(1-\theta)\,dx,$$
where c_{pen} is a constant for penalization.
2. We continue the previous algorithm with a modified "penalized" density
$$\theta_{\mathrm{pen}} := \frac{1 - \cos(\pi\theta_{\mathrm{opt}})}{2},$$
where θ_{opt} is the optimal density obtained by the previous algorithm.

In the second algorithm, we note that if $0 < \theta_{\text{opt}} < 1/2$, then $\theta_{\text{pen}} < \theta_{\text{opt}}$, while $1/2 < \theta_{\text{opt}} < 1$, then $\theta_{\text{pen}} > \theta_{\text{opt}}$. Hence we see that the density θ goes to 0 or 1 when we apply the algorithm.

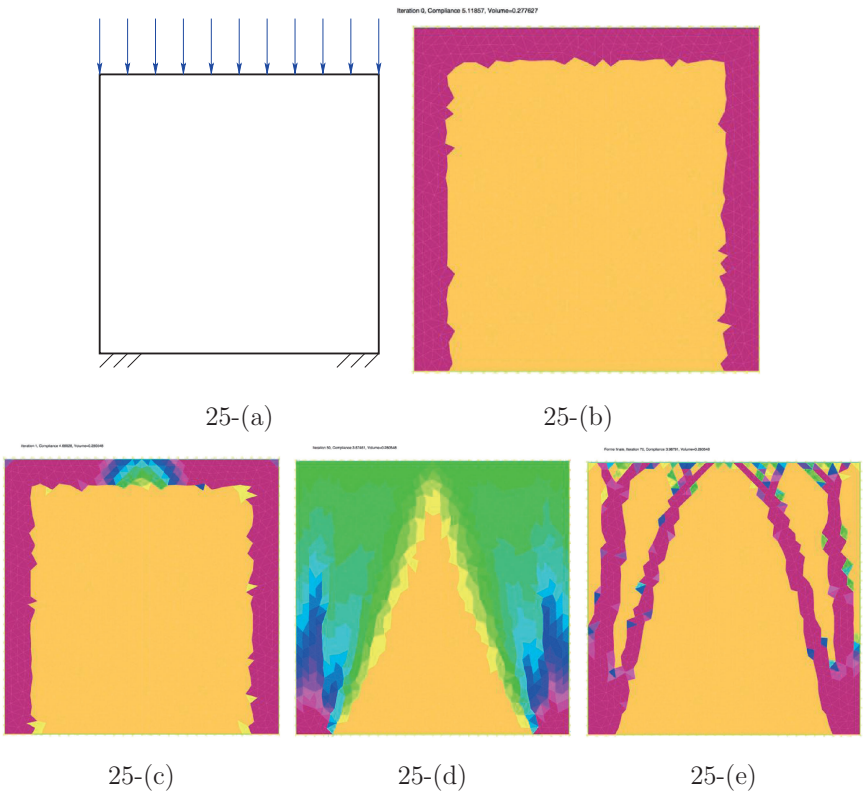
Example 5.9. *We consider the optimal radiator (Fig. 25-(a))*

$$\begin{cases} -\text{div}(A^*\nabla u) = 0 & \text{in } \Omega, \\ A^*\nabla u \cdot n = 1 & \text{on } \Gamma_N, \\ A^*\nabla u \cdot n = 0 & \text{on } \Gamma, \\ u = 0 & \text{on } \Gamma_D. \end{cases}$$

We minimize the temperature where heating takes place

$$\min_{(\theta,A^*)\in\mathcal{U}_{\text{ad}}^L} \left\{ J(\theta, A^*) = \int_{\Gamma_N} u\, ds \right\}.$$

This is another case of compliance minimization. Thus, the problem is self-adjoint and $p = -u$. We solve in the domain $\Omega = (0,1)^2$ with phases $\alpha = 0.01$ and $\beta = 1$. We work with a volume constraint 50% of phase α. We initialize with a value θ as Fig. 25-(b). Fig. 25-(c), (d), and (e) show the numerical results of the volume density at some iteration numbers under the above penalization algorithm.



25-(a) 25-(b)

25-(c) 25-(d) 25-(e)

Fig. 25. Optimal radiator: (a) setting of the problem, (b) initial condition, (c)–(e) volume fraction θ (iteration number 1, 50 and 70 respectively). Here we used the following color convention: red = 1, yellow = 0.

5.3 Homogenization method in the elasticity setting

In this section, we will apply the homogenization method in the elasticity setting. We remark that it is very similar to the conductivity setting but there are some additional hurdles. We shall review the results without proofs, however, the basic ingredients of the homogenization method which we will consider in this section are the sames:

- Introduction of composite designs characterized by (θ, A^*),
- Hashin–Shtrikman bounds for composites,
- Sequential laminates are optimal microstructures for compliance minimization, which we will consider the following.

We remark that, unfortunately, the full set of composites G_θ is unknown as stated in Sect. 4.3, unlike the case we considered in the Sect. 5.2.

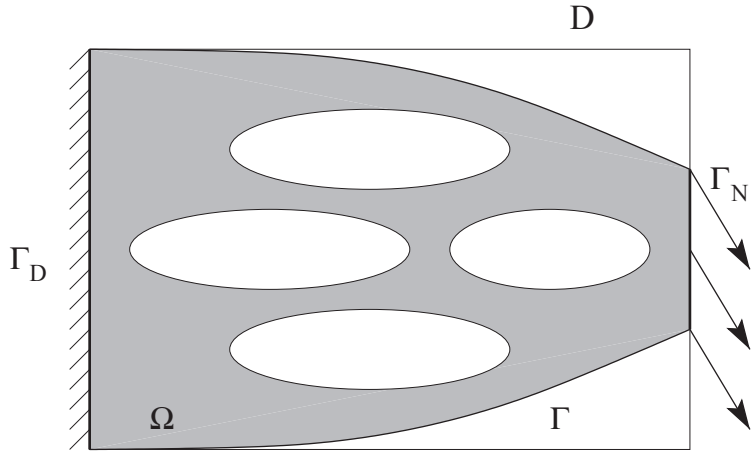


Fig. 26. Setting of the problem (5.14).

5.3.1 Introduction of the model of the elasticity and relaxed problem

We introduce the model compliance minimization problem (Fig. 26). Let $N = 2$ or 3 and $D \subset \mathbb{R}^N$ be a bounded domain. If the Hooke's law A is isotropic, with positive bulk and shear moduli κ and μ, we have

$$A = \left(\kappa - \frac{2\mu}{N}\right) I_2 \otimes I_2 + 2\mu I_4.$$

Let $\Gamma_D \subset \partial D$ be the Dirichlet part and $\Gamma_N \subset \partial D$ be the Neumann part loaded by g. For any domain $\Omega \subset D$ with Γ_D, $\Gamma_N \subset \partial\Omega$, the displacement vector field $u\colon \Omega \to \mathbb{R}^N$ is defined as the solution of the problem

$$\begin{cases} \operatorname{div}\sigma = 0 & \text{in } \Omega, \\ \sigma = 2\mu e(u) + \lambda \operatorname{tr}(e(u))\mathrm{Id} & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_D, \\ \sigma \cdot n = g & \text{on } \Gamma_N, \\ \sigma \cdot n = 0 & \text{on } \Gamma, \end{cases} \tag{5.14}$$

where $\Gamma := \partial\Omega \setminus (\Gamma_D \cup \Gamma_N)$, $e(u) := (\nabla u + (\nabla u)^t)/2$ and $\lambda := \kappa - 2\mu/N$. We consider the following minimization problem to obtain the optimal shape such that the weight is minimized and the rigidity is maximized:

$$\inf_{\Omega \subset D} \left\{ J(\Omega) = \int_{\Gamma_N} g \cdot u \, ds + \ell \int_{\Omega} dx \right\}, \tag{5.15}$$

where ℓ is a Lagrange multiplier and the infimum is taken over the all subset Ω of D with Γ_D, $\Gamma_N \subset \partial\Omega$.

The shape optimization problem (5.15) can be approximated by a two-phase optimization problem: the original material A and the holes of rigidity $B \approx 0$. Then the Hooke's law of the mixture in D is rewritten as

$$\chi_\Omega(x)A + (1 - \chi_\Omega(x))B \approx \chi_\Omega(x)A, \quad x \in D.$$

Hence the admissible set becomes

$$\mathcal{U}_{\text{ad}} = \{\chi \in L^\infty(D; \{0, 1\})\}.$$

As in conductivity (membrane) case, we can apply the relaxation approach based on homogenization theory.

We introduce composite structures characterized by a local volume fraction $\theta(x)$ of the phase A and a homogenized tensor $A^*(x)$, corresponding to its microstructure. The set of admissible homogenized designs is

$$\mathcal{U}^*_{\text{ad}} = \{(\theta, A^*) \in L^\infty(D; [0, 1] \times \mathbb{R}^{N^4}) : A^*(x) \in G_{\theta(x)} \text{ for a.e. } x \in D\},$$

where, for fixed x in D, $G_\theta(x)$ denotes the set of all possible two-phase composite materials at fixed volume fraction $\theta(x)$. In this case, the homogenized state equation is

$$\begin{cases} \operatorname{div}\sigma = 0 & \text{in } D, \\ \sigma = A^* e(u) & \text{in } D, \\ u = 0 & \text{on } \Gamma_D, \\ \sigma \cdot n = g & \text{on } \Gamma_N, \\ \sigma \cdot n = 0 & \text{on } \partial D \setminus (\Gamma_D \cup \Gamma_N) \end{cases} \tag{5.16}$$

and the homogenized compliance is defined by

$$c(\theta, A^*) = \int_{\Gamma_N} g \cdot u \, ds.$$

By the above setting, the relaxed or homogenized optimization problem is derived as follows:

$$\min_{(\theta,A^*)\in\mathcal{U}_{\text{ad}}^*} \left\{ J(\theta, A^*) = c(\theta, A^*) + \ell \int_D \theta(x)\, dx \right\}. \tag{5.17}$$

In the elasticity setting, an explicit characterization of G_θ is still lacking, it is a major inconvenience of the problem (5.17). Fortunately, for compliance one can replace G_θ by its explicit subset of laminated composites. The key argument to avoid the knowledge of G_θ is that, thanks to the complementary energy minimization, the compliance can be rewritten as

$$c(\theta, A^*) = \int_{\Gamma_N} g \cdot u \, ds = \min_{\substack{\operatorname{div}\sigma=0 \text{ in } D,\\ \sigma\cdot n=g \text{ on } \Gamma_N,\\ \sigma\cdot n=0 \text{ on } \partial D\setminus(\Gamma_N\cup\Gamma_D)}} \int_D (A^*)^{-1}\sigma\cdot\sigma \, dx.$$

The complementary energy is followed by a similar argument in Example 2.30. The shape optimization problem (5.17) thus becomes a double minimization problem

$$\min_{(\theta,A^*)\in\mathcal{U}_{\text{ad}}^*} \left\{ \min_{\substack{\operatorname{div}\sigma=0 \text{ in } D,\\ \sigma\cdot n=g \text{ on } \Gamma_N,\\ \sigma\cdot n=0 \text{ on } \partial D\setminus(\Gamma_N\cup\Gamma_D)}} \int_D (A^*)^{-1}\sigma\cdot\sigma \, dx + \ell \int_D \theta(x)\, dx \right\}. \tag{5.18}$$

We will exchange the order of the minimization (5.18). Since the order of minimization is irrelevant, (5.18) can be rewritten as

$$\min_{\substack{\operatorname{div}\sigma=0 \text{ in } D,\\ \sigma\cdot n=g \text{ on } \Gamma_N,\\ \sigma\cdot n=0 \text{ on } \partial D\setminus(\Gamma_N\cup\Gamma_D)}} \min_{(\theta,A^*)\in\mathcal{U}_{\text{ad}}^*} \left\{ \int_D (A^*)^{-1}\sigma\cdot\sigma \, dx + \ell \int_D \theta(x)\, dx \right\}.$$

The minimization with respect to the design parameters (θ, A^*) is local. Hence the above minimization becomes

$$\min_{\substack{\operatorname{div}\sigma=0 \text{ in } D,\\ \sigma\cdot n=g \text{ on } \Gamma_N,\\ \sigma\cdot n=0 \text{ on } \partial D\setminus(\Gamma_N\cup\Gamma_D)}} \int_D \min_{\substack{0\le\theta\le 1,\\ A^*\in G_\theta}} ((A^*)^{-1}\sigma\cdot\sigma + \ell\theta)\, dx. \tag{5.19}$$

For a given stress tensor σ, the minimization of complementary energy

$$\min_{A^*\in G_\theta} (A^*)^{-1}\sigma\cdot\sigma$$

is a classical problem in homogenization of finding optimal bounds on the effective properties of composite materials. It turns out that we can restrict ourselves to sequential laminates which form an explicit subset L_θ of G_θ by Proposition 4.22. Recall that in the conductivity setting, it was enough to consider only the case of rank-1 laminates. On the other hand, Proposition 4.22 tells us that, in the elasticity setting, rank-1 laminates are not enough and G_θ has to be replaced by the set of rank-N sequential laminates instead.

As in the case of the conductivity setting in Sect. 5.2.1, one can prove that the problem (5.17) is a relaxation of the original shape optimization in the following sense (for the details of the proof, see [Al2002, Theorem 4.1.12]).

Theorem 5.10. *The homogenized formulation (5.17) is the relaxation of the original problem (5.15) in the sense where*

1. *there exists, at least, one optimal composite (θ, A^*) minimizing (5.17),*
2. *any minimizing sequence of classical shapes Ω for (5.15) converges, in the sense of homogenization, to a minimizer (θ, A^*) of (5.17),*
3. *the minimal values of the original and homogenized objective functions coincide.*

5.3.2 An explicit optimal bound $HS(\sigma)$

As we stated in Sect. 5.3.1, we can restrict the set G_θ to the sequential laminates L_θ. In this subsection, we will show the explicit computation of $HS(\sigma)$ for a special case. We will consider the case $B = 0$ for the simplification of algebra. Note that the case $B = 0$ is natural since the weak material is actually degenerate.

In two dimension case, we can obtain an explicit formula for the bound (see [Al2002, Theorem 2.3.35]):

Theorem 5.11. *Assume that $N = 2$, $B = 0$ and $\theta \neq 0$. Then for any stress tensor σ, the optimal bound $HS(\sigma)$ is rewritten as*

$$HS(\sigma) = A^{-1}\sigma \cdot \sigma + \frac{1-\theta}{\theta} g^*(\sigma), \tag{5.20}$$

where

$$g^*(\sigma) = \frac{\kappa + \mu}{4\mu\kappa}(|\sigma_1| + |\sigma_2|)^2 \tag{5.21}$$

and σ_1, σ_2 are the eigenvalues of σ. Furthermore, an optimal rank-2 sequential laminate is given by

$$m_1 = \frac{|\sigma_2|}{|\sigma_1| + |\sigma_2|}, \quad m_2 = \frac{|\sigma_1|}{|\sigma_1| + |\sigma_2|},$$

where m_i is the parameters appeared in Lemma 4.13.

In the three dimension case, we can also obtain the explicit formula for the bound (see [Al2002, Theorem 2.3.36]), however, it is more complicated than the two dimension case. Hence we restrict ourselves to the simple case of zero Poisson ratio, i.e., $\kappa = 2\mu/3$.

Theorem 5.12. *Assume that $N = 3$, $B = 0$ and $\theta \neq 0$. We also assume that the constants κ and μ satisfies $\kappa = 2\mu/3$. For any σ, we label the eigenvalues of σ as $|\sigma_1| \leq |\sigma_2| \leq |\sigma_3|$. Then, we obtain (5.20) with*

$$g^*(\sigma) = \begin{cases} \dfrac{1}{4\mu}(|\sigma_1| + |\sigma_2| + |\sigma_3|)^2 & \text{if } |\sigma_3| \leq |\sigma_1| + |\sigma_2|, \\ \dfrac{1}{2\mu}((|\sigma_1| + |\sigma_2|)^2 + |\sigma_3|^2) & \text{if } |\sigma_3| > |\sigma_1| + |\sigma_2|. \end{cases}$$

Furthermore, in the first regime, an optimal rank-3 sequential laminate is given by

$$m_1 = \frac{|\sigma_3| + |\sigma_2| - |\sigma_1|}{|\sigma_1| + |\sigma_2| + |\sigma_3|}, \quad m_2 = \frac{|\sigma_1| - |\sigma_2| + |\sigma_3|}{|\sigma_1| + |\sigma_2| + |\sigma_3|}, \quad m_3 = \frac{|\sigma_1| + |\sigma_2| - |\sigma_3|}{|\sigma_1| + |\sigma_2| + |\sigma_3|},$$

and in the second regime, an optimal rank-2 sequential laminate is

$$m_1 = \frac{|\sigma_2|}{|\sigma_1| + |\sigma_2|}, \quad m_2 = \frac{|\sigma_1|}{|\sigma_1| + |\sigma_2|}, \quad m_3 = 0,$$

where m_i is the parameters appeared in Lemma 4.13.

5.3.3 Optimality conditions

We consider the optimality condition for the minimization problem (5.19). In this subsection, we assume the same condition as in Sect. 5.3.2, i.e., $B = 0$ and $\theta \neq 0$. If (θ, A^*, σ) is a minimizer, then by Proposition 4.22, A^* is a rank-N sequential laminate aligned with σ. Moreover, Proposition 4.19 leads the explicit proportions

$$(A^*)^{-1} = A^{-1} + \frac{1-\theta}{\theta}\left(\sum_{i=1}^{N} m_i f_A^c(e_i)\right)^{-1}.$$

If we consider the case $N = 2$, then, by (5.21), we can rewrite the minimization appearing in the integrand (5.19) as

$$A^{-1}\sigma \cdot \sigma + \min_{0<\theta\leq 1}\left(\frac{(\kappa+\mu)(1-\theta)}{4\mu\kappa\theta}(|\sigma_1| + |\sigma_2|)^2 + \ell\theta\right).$$

Hence we obtain the explicit optimality formula for θ as follows:

$$\theta_{\text{opt}} = \min\left(1, \sqrt{\frac{\kappa+\mu}{4\mu\kappa\ell}}(|\sigma_1| + |\sigma_2|)\right). \tag{5.22}$$

The explicit formula for θ in the case $N = 3$ follows by a similar argument.

5.3.4 Numerical algorithm

In this subsection, we will introduce a numerical algorithm for the minimization problem (5.17). We use the following double "alternating" minimization in σ and in (θ, A^*):

Algorithm 5 Double alternating minimization for (5.17)

1. Initialization of the shape (θ_0, A_0^*) by a finite element method.
2. Iterations until convergence, for $n \geq 1$:
   - Given a shape $(\theta_{n-1}, A_{n-1}^*)$, we compute the stress σ_n by solving the linear elasticity problem (5.16),
   - Given the stress field σ_n, we update the new design parameters (θ_n, A_n^*) by the explicit optimality formula (5.22) in terms of σ_n.

Since the problem is self-adjoint, we can exchange the problem which we can consider the explicit optimality formula and the fact allows us to consider the above numerical algorithm. The algorithm uses a local microstructure A^* and a global density θ. Such algorithm is called micro-macro method.

Remark 5.13. *The objective function always decreases. Indeed, since (θ_n, A_n^*) minimizes the compliance under the stress σ_n, we have*

$$\int_D (A_{n-1}^*)^{-1}\sigma_n \cdot \sigma_n \, dx + \ell \int_D \theta_{n-1} \, dx \geq \int_D (A_n^*)^{-1}\sigma_n \cdot \sigma_n \, dx + \ell \int_D \theta_n \, dx.$$

On the other hand, σ_{n+1} minimizes the elastic complementary energy corresponding to the Hooke's law A_n^*, we see that

$$\int_D (A_n^*)^{-1}\sigma_n \cdot \sigma_n \, dx \geq \int_D (A_n^*)^{-1}\sigma_{n+1} \cdot \sigma_{n+1} \, dx.$$

Combining the above inequalities, we obtain the claim

$$J(\theta_{n-1}, A_{n-1}^*) \geq J(\theta_n, A_n^*).$$

We show the numerical results for the case of the cantilever (Fig. 27). The optimal shape of the short cantilever, i.e., the domain size 10×20, is displayed on the left of Fig. 28. Moreover, the left figure of Fig. 29 shows the convergence history of the objective function J. Here, the horizontal axis means the iteration number. The right figure of Fig. 29 shows the transition of the quantity

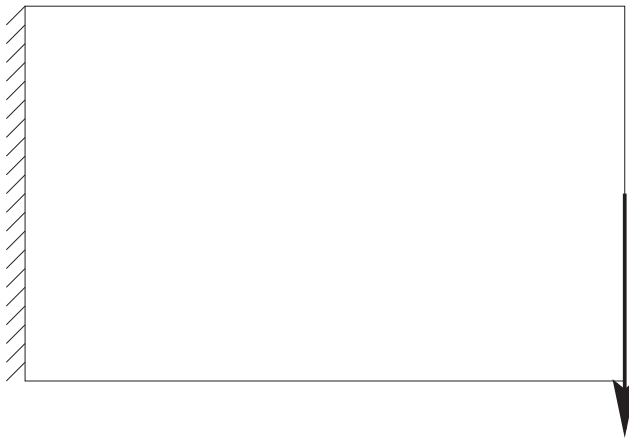


Fig. 27. Boundary conditions for the cantilever problem.

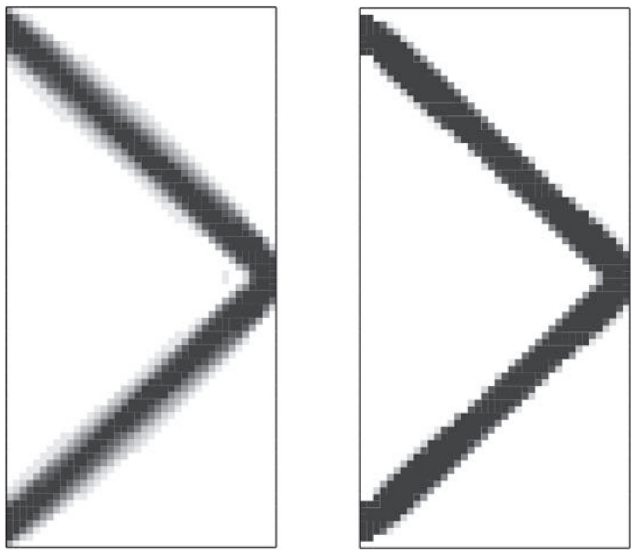


Fig. 28. Optimal shape of the short cantilever (left: composite, right: penalized).

$$\max\left(\max_i |\theta_i^{k+1}-\theta_i^k|, 1-\frac{\displaystyle\int_\Omega (A_{k+1}^*)^{-1}\sigma^k\cdot\sigma^k\,dx + l\int_\Omega \theta^{k+1}\,dx}{\displaystyle\int_\Omega (A_k^*)^{-1}\sigma^{k-1}\cdot\sigma^{k-1}\,dx + l\int_\Omega \theta^k\,dx}\right),$$

where the index i refers to the cell number. We also show the optimal shape of the medium cantilever, i.e., the domain size is 20×10 in the left of Fig. 30.

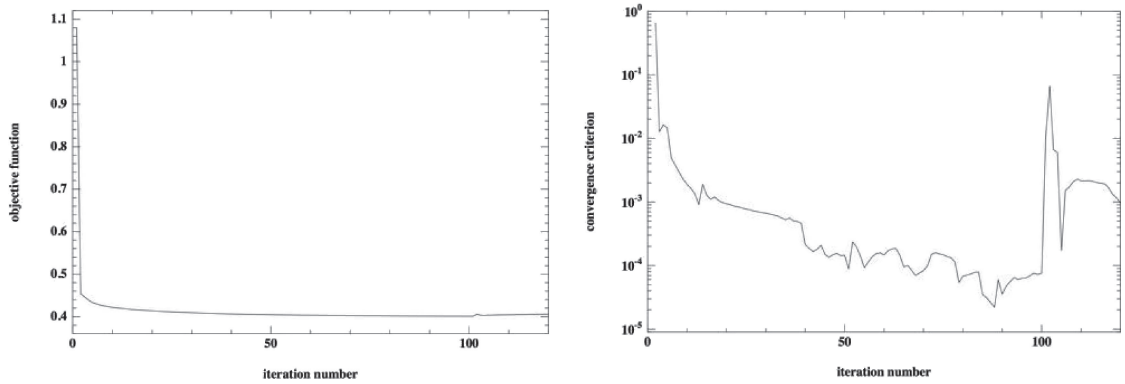


Fig. 29. Convergence history of the short cantilever (left: objective function, right: convergence criterion).

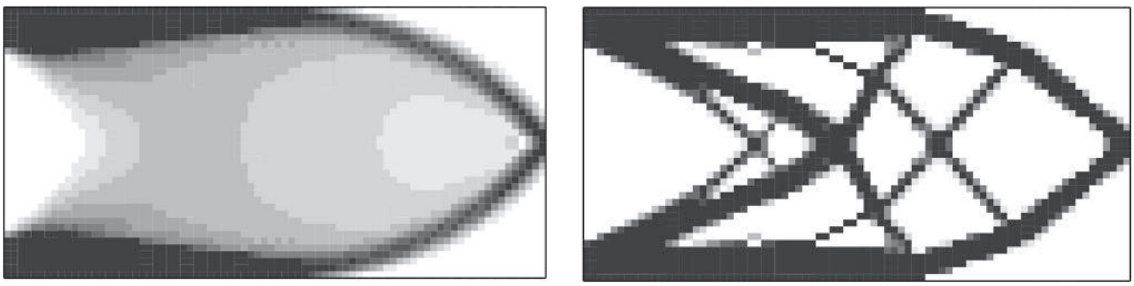


Fig. 30. Optimal shape of the medium cantilever (left: composite, right: penalized).

Penalization

The algorithm, we considered in this subsection, compute composite shapes instead of classical shapes. We thus use a penalization technique to force the density to take values close to 0 or 1 as in Sect. 5.2.3. The algorithm is the following:

Algorithm 6 Penalization process for (5.17)

After convergence to a composite shape, we perform a few more iterations with a penalized density

$$\theta_{\text{pen}} = \frac{1-\cos(\pi\theta_{\text{opt}})}{2}.$$

Note that if $0<\theta_{\text{opt}}<1/2$, then $\theta_{\text{pen}}<\theta_{\text{opt}}$, while, if $1/2<\theta_{\text{opt}}<1$, then $\theta_{\text{pen}}>\theta_{\text{opt}}$. By using this algorithm, we can obtain the penalized resulting shape of the short cantilever and the medium cantilever (see the right figure of Figs. 28 and 30 respectively). We also show the numerical result for the bridge problem (Figs. 31 and 32).

5.4 Convexification, "fictitious materials" and SIMP

In the homogenization method, composite materials are introduced, however, discarded at the end by penalization. In this section, we will consider whether we can simplify the approach by introducing merely a density θ. We will use the following "convexification" approach. A classical shape is parametrized by characteristic functions $\chi(x)$. If we convexify this admissible set, we obtain $\theta(x)\in[0,1]$. Replacing the admissible set by the convexified set, the Hooke's law, which was $\chi(x)A$, becomes $\theta(x)A$. We will call these $\theta(x)A$ "fictitious materials," because one can not realize them by a true homogenization process in general. Combined with a penalization scheme, this method is called SIMP method.

We consider the elasticity setting. For $\theta\in L^\infty(D;[0,1])$, the convexified formulation for the problem reads as follows:

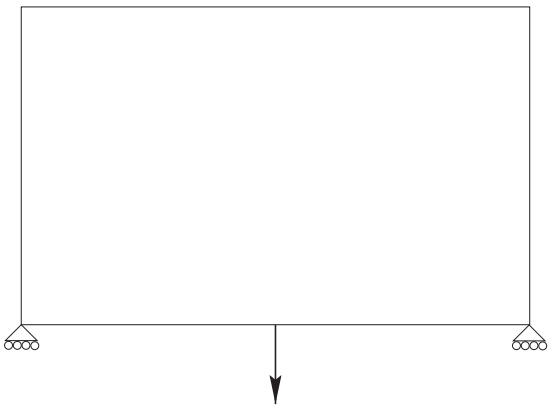


Fig. 31. Boundary conditions for the bridge problem.

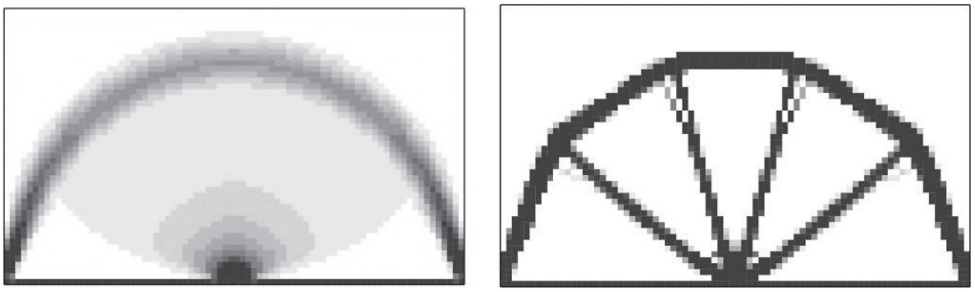


Fig. 32. Optimal shape of the bridge (left: composite, right: penalized).

$$\begin{cases} \operatorname{div}\sigma = 0 & \text{in } D, \\ \sigma = \theta A e(u) & \text{in } D, \\ u = 0 & \text{on } \Gamma_D, \\ \sigma \cdot n = g & \text{on } \Gamma_N, \\ \sigma \cdot n = 0 & \text{on } \partial D \setminus (\Gamma_D \cup \Gamma_N), \end{cases}$$

where $e(u) := (\nabla u + (\nabla u)^t)/2$. Moreover, the compliance minimization becomes

$$\min_{\theta \in L^\infty(D;[0,1])} \left\{ c(\theta) + \ell \int_D \theta(x)\, dx \right\} \tag{5.23}$$

with

$$c(\theta) = \int_{\Gamma_N} g \cdot u\, ds = \int_D (\theta(x)A)^{-1}\sigma \cdot \sigma\, dx = \min_{\substack{-\operatorname{div}\tau = 0 \text{ in } D, \\ \tau \cdot n = g \text{ on } \Gamma_N, \\ \tau \cdot n = 0 \text{ on } \partial D \setminus (\Gamma_N \cup \Gamma_D)}} \int_D (\theta(x)A)^{-1}\tau \cdot \tau\, dx.$$

There is only one single design parameter, the material density θ. In other words, any information concerning the microstructure A^* has disappeared.

5.4.1 Existence of solutions

In this section, we will show the existence of the minimizer of (5.23).

Theorem 5.14. *The convexified formulation*

$$\min_{\theta \in L^\infty(D;[0,1])} \min_{\substack{-\operatorname{div}\tau = 0 \text{ in } D, \\ \tau \cdot n = g \text{ on } \Gamma_N, \\ \tau \cdot n = 0 \text{ on } \partial D \setminus (\Gamma_N \cup \Gamma_D)}} \left(\int_D (\theta(x)A)^{-1}\tau \cdot \tau\, dx + \ell \int_D \theta(x)\, dx \right)$$

admits at least one solution.

Proof. Let $\mathcal{M}_N^s$ be the set of symmetric squared matrices of order N. The function

$$\phi(a, \sigma) = a^{-1}A^{-1}\sigma \cdot \sigma, \quad (a, \sigma) \in \mathbb{R}_{\geq 0} \times \mathcal{M}_N^s$$

is convex because

$$\begin{aligned}\phi(a,\sigma) &= \phi(a_0,\sigma_0) + D\phi(a_0,\sigma_0)\cdot(a-a_0,\sigma-\sigma_0) + \phi(a,\sigma - aa_0^{-1}\sigma_0)\\ &\geq \phi(a_0,\sigma_0) + D\phi(a_0,\sigma_0)\cdot(a-a_0,\sigma-\sigma_0),\end{aligned} \tag{5.24}$$

where the derivative $D\phi$ is given by

$$D\phi(a_0,\sigma_0)\cdot(b,\tau) = -\frac{b}{a_0^2}A^{-1}\sigma_0\cdot\sigma_0 + 2a_0^{-1}A^{-1}\sigma_0\cdot\tau.$$

Then, by Theorem 2.7 we can see that there exists a minimizer of (5.23). □

5.4.2 Optimality condition

If we exchange the minimizations in τ and in θ, we can compute the optimal θ which is

$$\theta(x) = \begin{cases} 1 & \text{if } A^{-1}\tau\cdot\tau \geq \ell,\\ \sqrt{\ell^{-1}A^{-1}\tau\cdot\tau} & \text{if } A^{-1}\tau\cdot\tau < \ell.\end{cases} \tag{5.25}$$

By using this explicit optimality formula, we can use again an "alternating" double minimization algorithm which will be shown in the following subsection.

5.4.3 Numerical algorithm

By the use of the explicit optimality formula (5.25) for θ, we can apply the following double minimization algorithm.

Algorithm 7 Double minimization algorithm and penalization for (5.23)

1. Initialization of the shape θ_0,
2. Iterations $k \geq 1$ until convergence
   - given a shape θ_{k-1}, we compute the stress τ_k by solving an elasticity problem by a finite element method,
   - given a stress field τ_k, we update the new material density θ_k with the explicit optimality formula in terms of τ_k.

As a penalization, we use the penalized density

$$\theta_{\text{pen}} = \frac{1-\cos(\pi\theta_{\text{opt}})}{2} \quad \text{or} \quad \theta_{\text{pen}} = \theta^p,$$

where $p > 1$ if we consider the SIMP method.

In practice, it is extremely simple, however, the numerical results are not as good. This can be explained as follows: since the SIMP method uses very little information on the given composites (in particular, it neglects its microstructure all together), we have a lack of a relaxation theorem. In other words, applying the SIMP method changes the problem, and, as a consequence, we have no guarantee that the optimal solution obtained is really an approximation of the optimal solution of the original problem. Moreover, we have to be careful, as it could be very delicate to monitor the penalization. We show the numerical result for the case of the bridge problem (Fig. 31). In Fig. 33, we show the convergence history of the objective function. Here, the horizontal axis means the iteration number. In Fig. 34, we show the numerical result of the optimal shape of the bridge by the algorithm which is mentioned in this subsection.

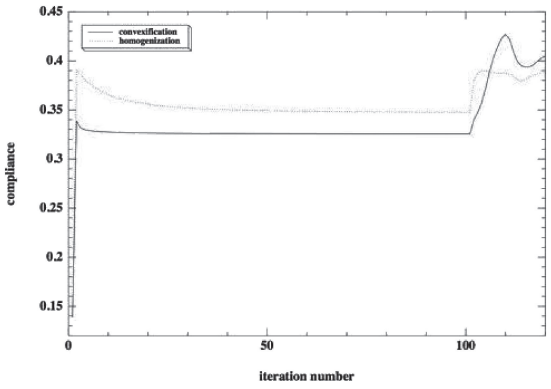


Fig. 33. Convergence history of the objective function for the bridge problem.

5.4.4 Concluding remarks on the SIMP method

SIMP (or convexification, or "fictitious materials") is very simple and very popular. Actually, many commercial codes are using the method. However, since SIMP uses very little information on composites, its simplicity comes at a cost. In particular, contrary to the homogenization method, SIMP is a convexification method and not a relaxation

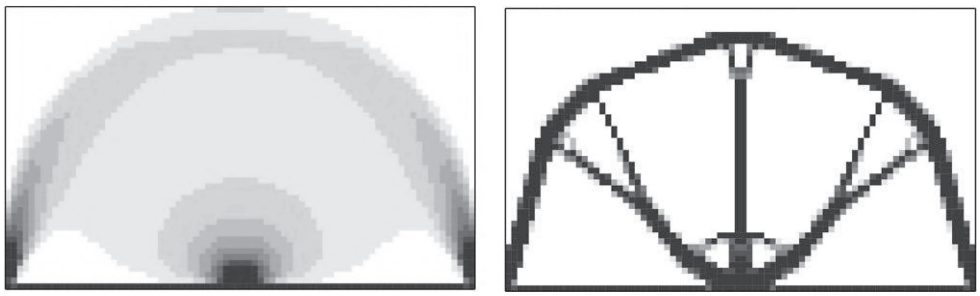


Fig. 34. Optimal shape of the bridge (left: convexified, right: penalized).

method, i.e., it changes the problem. Hence there is a gap between the true minimal value of the objective function and that of SIMP.

5.5 Generalizations of the homogenization method

As possible generalizations of the homogenization method, we give the following three examples:

- Multiple loads,
- Vibration eigenfrequency,
- General criterion of the least-square type.

The first two cases are self-adjoint and we have a complete understanding and justification of the relaxation process. The third case is, however, not self-adjoint and only a partial relaxation is known. For the detail of these cases, see [Al2002].

5.6 Exercises

Problem 5.6.1. *Implement the optimal radiator test case (with penalization).*

Problem 5.6.2. *Implement the SIMP method (start with exponent $p = 1$ and increase to $p = 3$) for compliance minimization: cantilever, bridge, MBB beam, L-beam.*

Problem 5.6.3. *Implement SIMP method with an adjoint for the objective function*

$$J(\theta) = \int_D \theta |u|^2 \, dx.$$

Problem 5.6.4. *Minimize compliance for the cantilever problem with a parametrized cell (rectangular hole in a square cell).*

6. Resurrection of the Homogenization Method: Lattice Materials in Additive Manufacturing

6.1 Introduction

As we mentioned in the previous sections, the homogenization method uses true composite materials, possibly anisotropic, and this makes it complicated to implement. Therefore, in practice, the method was replaced by its much simplified version, the so-called SIMP method, which uses only fictitious isotropic materials. (For the detail of the SIMP method, see [BS2003].) Since intermediate densities (between full material and void) are penalized in the end, there is indeed no need to have a detailed knowledge and optimization of microstructures.

Nevertheless, the recent progress of additive manufacturing techniques has revived the interest for the use of graded or microstructured materials since they are now manufacturable (see Fig. 6, page 6). Since the homogenization method is the right technique to deal with microstructured materials, where anisotropy plays a key role (a feature which is absent from SIMP), we could well see a resurrection of the homogenization method for such applications. There is however one final hurdle to overcome, once an optimal composite structure has been obtained, that is the projection of the optimal microstructure at a chosen finite lengthscale to get a global and detailed picture of the optimal microstructure. This is the most delicate part of this homogenization approach and the one where this section is most contributing.

Often (but not always) lattice materials are periodic structures, with macroscopically varying parameters. Hence we will restrict to periodic homogenization and macroscopically modulated periodic structures, i.e., the material parameters are of the type

$$A\left(x, \frac{x}{\varepsilon}\right),$$

where $y \mapsto A(x, y)$ is Y-periodic and $x \mapsto A(x, y)$ describes the macroscopic variations. (We discuss how to choose the period and the holes in Sect. 6.2.) The orientation of the microstructures is rarely taken into account and optimized, although it is well-known that their orientation is a crucial and determining parameter in topology optimization (see [Al2002] and [Pe1989]). Actually, even if optimizing the microstructure orientation is not difficult, reconstructing the oriented periodic structure is a challenging issue. We propose a method to settle the difficulty by projecting the optimal microstructure on a fine mesh of the overall structure in a smoothly varying way. This idea was first introduced by [PT2008].

In this section, we consider the post-treatment of 2-d compliance minimization, which is based on [AGP2018]. We will improve the pioneer work [PT2008] in several aspects, which we will note below. See also [GS2018] for another homogenization method in the spirit of [PT2008]. Note that this methods can be extended for the case of 3-d compliance minimization as well [GAP2018]. One of the difficulties of this extension is to treat the rotation of the materials. Indeed, compared to the 2-d case, where rotations are parametrized by a single angle (see Sect. 6.4), the 3-d case is more involved and requires new ingredients.

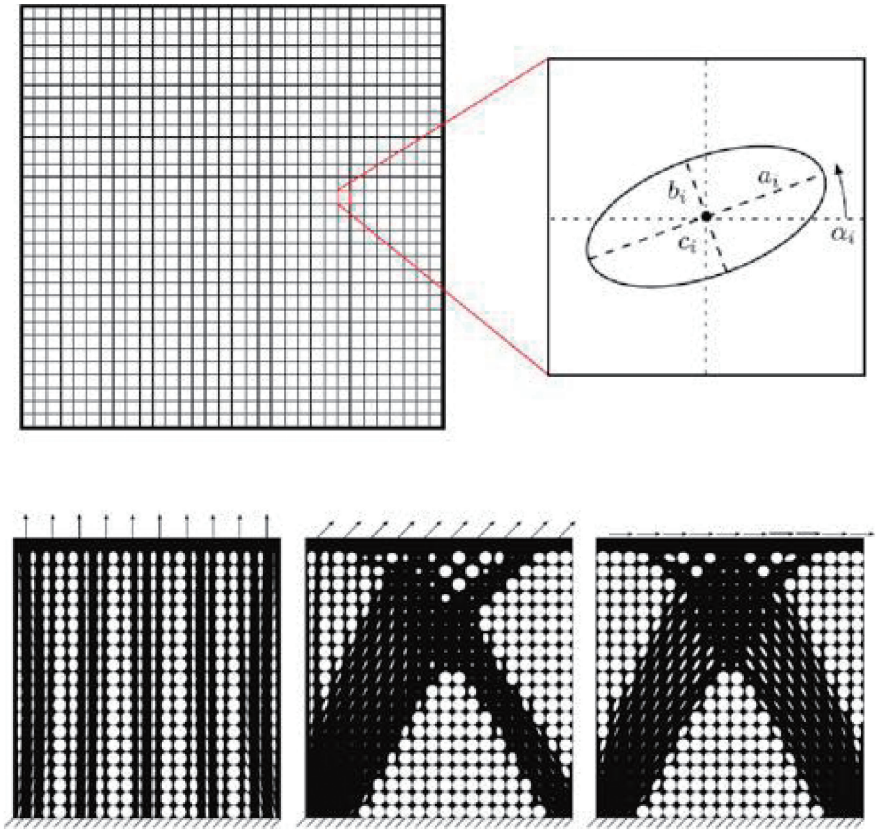


Fig. 35. Example of a macroscopically varying microstructure (extracted from [GLRS2013]).

6.2 Setting of the problem

As in the previous sections, we will consider compliance minimization problems in $N = 2$. Let $D \subset \mathbb{R}^N$ be a smooth bounded domain and $\Omega \subset D$ be the reference configuration of a homogeneous isotropic linear elastic body whose Hooke's law A is defined by (4.32), with Lamé coefficients λ and μ. We assume that Ω is clamped on $\Gamma_D \subset \partial\Omega$ and subject to surface loads g on $\Gamma_N \subset \partial\Omega$. Also, for simplicity, these parts Γ_D and Γ_N of the boundary $\partial\Omega$ are assumed to be fixed and subsets of ∂D. The displacement vector field u and the stress tensor σ are given by the following system

$$\begin{cases} \operatorname{div}\sigma = 0 & \text{in } \Omega, \\ \sigma = Ae(u) & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_D, \\ \sigma \cdot n = g & \text{on } \Gamma_N, \\ \sigma \cdot n = 0 & \text{on } \Gamma = \partial\Omega \setminus (\Gamma_D \cup \Gamma_N), \end{cases}$$

where $e(u) := (\nabla u + (\nabla u)^t)/2$ is the strain tensor. Now, let us consider the following compliance minimization problem:

$$\min_{\substack{|\Omega|\le V,\\ \Gamma_D\cup\Gamma_N\subset\partial\Omega}} J(\Omega), \tag{6.1}$$

where $V > 0$ is the maximum admissible volume and the objective function J is the compliance

$$J(\Omega) = \int_{\Gamma_N} g \cdot u \, ds.$$

As we have already seen in Sect. 5, the compliance minimization problem (6.1) does not admit a classical solution. This is why we consider the homogenized problem. We introduce composite structures characterized by the local volume density $\theta(x)$ of the material and a homogenized elasticity tensor $A^*(x)$, corresponding to its microstructure. Then, the homogenized or macroscopic displacement u^* is the solution of the system

$$\begin{cases} \operatorname{div}\sigma = 0 & \text{in } D, \\ \sigma = A^* e(u^*) & \text{in } D, \\ u^* = 0 & \text{on } \Gamma_D, \\ \sigma \cdot n = g & \text{on } \Gamma_N, \\ \sigma \cdot n = 0 & \text{on } \partial D \setminus (\Gamma_D \cup \Gamma_N). \end{cases} \tag{6.2}$$

By the above setting, the relaxed or homogenized optimization problem is obtained as follows:

$$\min_{\substack{\int_D \theta(x)\,dx \le V,\\ A^*(x)\in P_\theta(x)}} \left\{ J^*(\theta, A^*) = \int_{\Gamma_N} g \cdot u^* \, ds \right\}, \tag{6.3}$$

where u^* is the solution of (6.2) and $P_\theta(x)$ is a given subset of effective or homogenized Hooke's laws for some well-chosen microstructures of density $\theta(x)$.

Our aim in Sect. 6 is to propose a specific subset P_θ of periodic composites and to construct a minimizing sequence for (6.3). A typical example is that of a rectangular hole in a square cell (Fig. 21, page 39), where the cell parameters are the lengths $m_1, m_2 > 0$ and the rotation angle α (acting either on the hole or the whole cell). Also, we let the homogenized tensors be denoted by $A^*(m_1, m_2, \alpha)$. We note that the same ideas in the following are applicable to other geometries as well.

6.3 A three steps approach

To achieve our purpose, we take a three steps approach for the optimization. The first step is to pre-compute the homogenized properties $A^*(m_1, m_2, \alpha)$ for all values of the parameters. The second step is to apply a simple parametric optimization process to the homogenized problem. Compared to Sect. 3, we replace the thickness field h by new parameter fields m_1, m_2 and α which vary in space. The third step is to choose a length scale ε and reconstruct a periodic domain $A(x, \frac{x}{\varepsilon})$ approximating the optimal A^*. We remark that the third step is the trickiest part of the whole process.

The most delicate point is the combined problem of the orientation of the microstructure and the reconstruction of a macroscopically varying periodic lattice. In order to avoid these difficulties, there are two possible approaches. The first one is a "naive" approach, that is, we assume that the periodic grid is never deformed and the holes are simply rotated. The main advantage of the "naive" approach is that the reconstruction of the periodic perforated structure is very easy (see e.g., [GLRS2013]). We remark that this approach is naive because the "skeleton" of the reconstructed structure is fixed and thus it does not follow the supported stresses or forces.

The second one is a deeper approach initiated by Pantz and Trabelsi [PT2008]. The main advantage of this new approach is that the reconstructed structure adapts its geometry to the supported stresses or forces (in some sense, it looks like Michell trusses [ROZ1989] in the case of dimension 2).

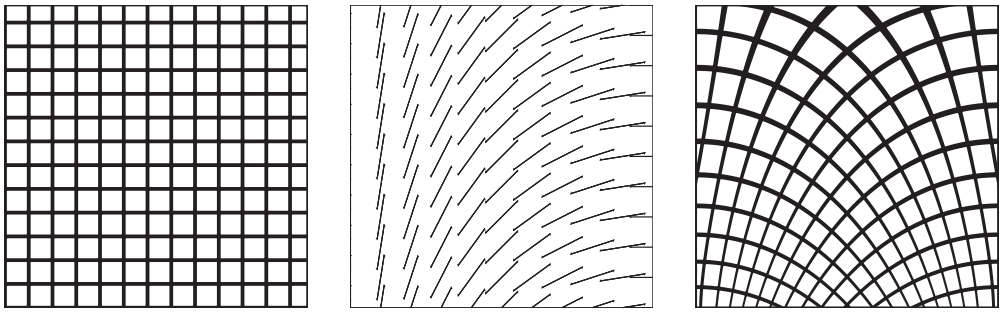


Fig. 36. A regular grid (left) is associated to an orientation field (middle), giving the local orientation of each cell: it yields a distorted grid (right).

The main difficulty is then to reconstruct such a macroscopically deformed periodic perforated structure. There may be issues on the regularity of the orientation field for stresses or forces. This is the approach that we follow in the sequel.

6.4 1st step: pre-computing the homogenized properties

For the square cell with a rectangular hole, with a fixed orientation, we compute the homogenized properties $A^*(m_1, m_2)$ for a discrete sampling of $0 \leq m_1, m_2 \leq 1$. We also compute the derivatives of $A^*(m_1, m_2)$ with respect to (m_1, m_2) by using a shape derivative and an adjoint approach in the same manner in Sect. 3.3 (see [AGP2018, Sects. 3.3 and 3.4] for the details and [HP2018, SK1992] for an introduction on shape derivatives). We note that we can also compute the derivatives of $A^*(m_1, m_2)$ with respect to (m_1, m_2) numerically by means of a finite difference method.

In the following, for simplicity, we will only consider the case of dimension $N = 2$. Assume that the cell is rotated by an angle α. We define $Y_\alpha(m)$ as the periodic cells with hole $m = (m_1, m_2)$, together with the orientation α of the cell. Then the dependency of the homogenized properties $A^*(m_1, m_2, \alpha)$ with respect to the angle α is given by

$$A^*(m_1, m_2, \alpha) = R(\alpha)^T A^*(m_1, m_2, 0) R(\alpha), \tag{6.4}$$

where $R(\alpha)$ is the fourth-order tensor defined by

$$\forall \xi \in \mathcal{M}_2^s, \quad R(\alpha)\xi = Q(\alpha)^T \xi Q(\alpha),$$

and $Q(\alpha)$ is the rotation matrix of angle α. Indeed, the variational formulation (4.33) and the characterization of A^* (4.34) imply that the quadratic form $A^*(m_1, m_2, \alpha)\xi : \xi$ satisfies

$$A^*(m_1, m_2, \alpha)\xi \cdot \xi = \min_{w \in H^1_\#(Y_\alpha(m))^N} \int_{H^1_\#(Y_\alpha(m))^N} A(\xi + e(w)) \cdot (\xi + e(w))\, dy$$

for any $\xi \in \mathcal{M}_2^s$. On the other hand, we see that, for $\xi \in \mathcal{M}_2^s$,

$$\begin{aligned}
[R(\alpha)^T A^*(m_1, m_2, 0) R(\alpha)]\xi \cdot \xi &= A^*(m_1, m_2, 0)(R(\alpha)\xi) \cdot (R(\alpha)\xi) \\
&= \min_{w \in H^1_\#(Y_0(m))^N} \int_{H^1_\#(Y_0(m))^N} A(R(\alpha)\xi + e(w)) \cdot (R(\alpha)\xi + e(w))\, dy \\
&= \min_{w \in H^1_\#(Y_0(m))^N} \int_{H^1_\#(Y_0(m))^N} A(\xi + Q(\alpha)e(w)Q(\alpha)^T) \cdot (\xi + Q(\alpha)e(w)Q(\alpha)^T)\, dy \\
&= \min_{w \in H^1_\#(Y_\alpha(m))^N} \int_{H^1_\#(Y_\alpha(m))^N} A(\xi + e(w)) \cdot (\xi + e(w))\, dy.
\end{aligned}$$

Thus, the numerical computation of the homogenized properties $A^*(m_1, m_2, \alpha)$ can be restricted to the case $\alpha = 0$. Note that (6.4) implies that a rotation of the cell by an angle π does not change its Hooke's law as $R(\pi) = -\mathrm{Id}$. Hence the optimal orientation can only be defined modulo π.

We show the numerical results for the entires of the homogenized tensor $A^*(m_1, m_2, 0)$ and their derivatives as functions of m (see Fig. 37). When $m = 0$, then the homogenized tensor $A^*(m_1, m_2, 0)$ is equal to A. On the other hand, if m is close to $(1, 1)$, then the homogenized tensor is converging to the null tensor. Moreover, one can see easily check, that the entries of $A^*(m_1, m_2, 0)$ decrease, when m_1 is fixed and m_2 is increasing (and vice versa). In other words, the cell is globally weaker when its hole is widening in one direction or the other. However, the sensitivity of the component $A^*(m_1, m_2, 0)_{1111}$ to the parameter m_2 is greater than the one to the parameter m_1 (see Fig. 37). Fig. 37 shows the numerical results in the case $A_{1111} = A_{2222} = 24.07$, $A_{1122} = 12.96$ and $A_{1212} = 11.11$. That is explained by the fact that, along the y_1 axis, the strength of the cell is mainly ensured by the material in the areas above and below the hole, whose sizes depend on m_2. As one could expect, the homogenized elasticity tensor is quite smooth with respect to the parameter m, so it is amenable to a gradient based optimization method.

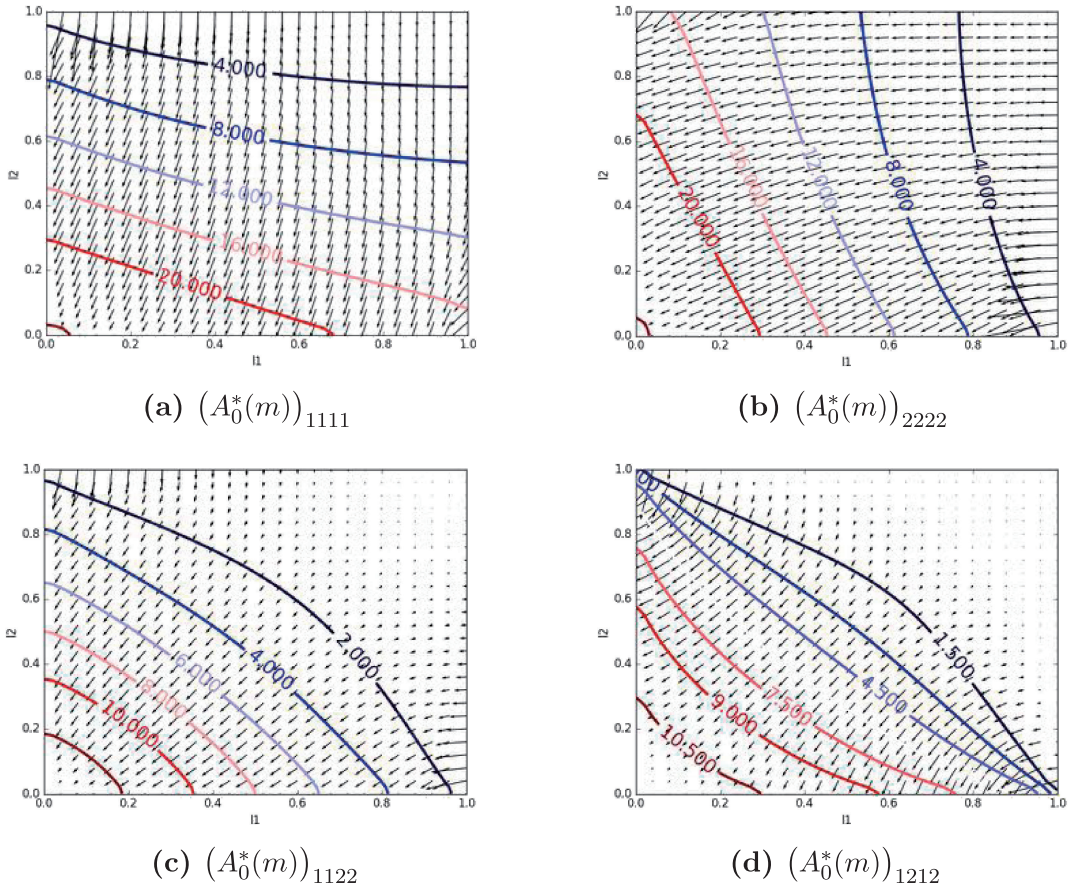
6.5 2nd step: parametric optimization of the homogenized problem

Let us recall that the homogenized equation in a box D is

$$\begin{cases}
\operatorname{div} \sigma = 0 & \text{in } D, \\
\sigma = A^* e(u) & \text{in } D, \\
u = 0 & \text{on } \Gamma_D, \\
\sigma \cdot n = g & \text{on } \Gamma_N, \\
\sigma \cdot n = 0 & \text{on } \Gamma = \partial D \setminus (\Gamma_D \cup \Gamma_N)
\end{cases}$$

and the compliance minimization problem is

$$\min_{m_1, m_2, \alpha} \left\{ J(A^*) = \int_{\Gamma_N} g \cdot u\, ds \right\}.$$



(a) $\left(A_0^*(m)\right)_{1111}$ (b) $\left(A_0^*(m)\right)_{2222}$

(c) $\left(A_0^*(m)\right)_{1122}$ (d) $\left(A_0^*(m)\right)_{1212}$

Fig. 37. Isolines of the entries of the homogenized tensor $A^*(m_1, m_2, 0)$ and their gradient (small arrows) according to the parameters m_1 (x-axis) and m_2 (y-axis).

Here, the minimum is taken over all functions m_1, $m_2 \in L^\infty(D; [0,1])$ and $\alpha \in L^\infty(D)$. One can rewrite the compliance as the minimum of complementary energy (see Sect. 2.4) as follows

$$J(A^*) = \min_{\tau \in H_0} \int_D (A^*)^{-1}\, \tau \cdot \tau \, dx,$$

where

$$H_0 = \left\{ \tau \in L^2(D; \mathcal{M}_2^s) : \begin{array}{ll} \operatorname{div} \tau = 0 & \text{in } D, \\ \tau \cdot n = g & \text{on } \Gamma_N, \\ \tau \cdot n = 0 & \text{on } \Gamma \end{array} \right\}.$$

This is interesting for algorithmic purposes because we can apply the optimality criteria or alternate minimization algorithm of Sect. 3. We note that the orientation optimization with respect to α is very simple, due to a result of Pedersen [Pe1989]. Pedersen proved that the optimal orientation of an orthotropic cell for a given displacement field is the one where the cell is aligned with the principal eigen-direction of the strain tensor. We can easily show a similar result for stress field in the same way.

We compute the volume fraction of material in a single unit cell as

$$\theta(x) = 1 - m_1(x) m_2(x),$$

also, the total volume of the lattice structure is thus

$$\text{Vol} = \int_D (1 - m_1(x) m_2(x))\, dx.$$

To implement a volume constraint, we rely on a Lagrangian algorithm

$$\mathcal{L}(m, \alpha, \sigma, \ell) = \int_D (A^*)^{-1}(m, \alpha)\, \sigma \cdot \sigma \, dx + \ell \left(\int_D (1 - m_1(x) m_2(x))\, dx - V_0 \right),$$

where ℓ is the Lagrange multiplier associated to the volume constraint $\text{Vol} = V_0$.

Now let us consider a numerical algorithm. To minimize with respect to the microstructure m, we use the following

algorithm of alternate minimization (or optimality criteria). Moreover, to minimize with respect to the orientation α, we could use the same method as for the minimization with respect to the microstructure m, but Pedersen's result [Pe1989] is a better (more efficient) algorithm than the gradient descent method to compute the optimal orientation because it is a global minimization method, proving an optimal orientation at each iteration. However, this method can usually not be generalized to other objective functions.

Algorithm 8 Algorithm of alternate minimization (or optimality criteria)

1. Initialization of the cell parameters m_1^0, m_2^0, α^0. For example, we take $m_1 = m_2$, constant satisfying the volume constraint, and $\alpha = 0$.
2. Iterations until convergence, for $n \geq 0$:
   (a) Computation of the stress tensor σ^n, unique solution of the (dual) elasticity equations with $A^*_{\alpha^n}(m^n)$.
   (b) Update of the parameters:
   - perform one iteration of the projected gradient algorithm for hole parameters
   $$m_i^{n+1} = \mathcal{P}\left(m_i^n - \mu_m \frac{\partial \mathcal{L}}{\partial m_i}(m^n, \alpha^n, \sigma^n, \ell^n) \right), \tag{6.5}$$
   where $\mu_m > 0$ is the step size and $\mathcal{P}$ is the projection operator to satisfy the constraints.
   - by Pedersen's result [Pe1989], for a given stress tensor σ^n, the optimal orientation angle α^n is the one where the cell is aligned with the principal eigen-directions of the strain tensor.

Let us focus on the technical details of the algorithm. The partial derivative of the Lagrangian $\mathcal{L}$ with respect to the parameter m_i $(i = 1, 2)$, is given by

$$\frac{\partial \mathcal{L}}{\partial m_i} = -\frac{\partial A^*}{\partial m_i}(m, \alpha)(A^*)^{-1}(m, \alpha)\,\sigma \cdot (A^*)^{-1}(m, \alpha)\sigma - \ell m_{3-i}.$$

We remark that the derivative of A^* with respect to m_i can be obtained by employing the use of shape derivatives. For the details of the computation, see [AGP2018].

Finally, $\mathcal{P}$ is the projection operator onto the interval $[0, 1]$. In the process of this projection, we have to update the Lagrange multiplier ℓ, which is constant in D, by a dichotomy process designed to respect the volume constraint.

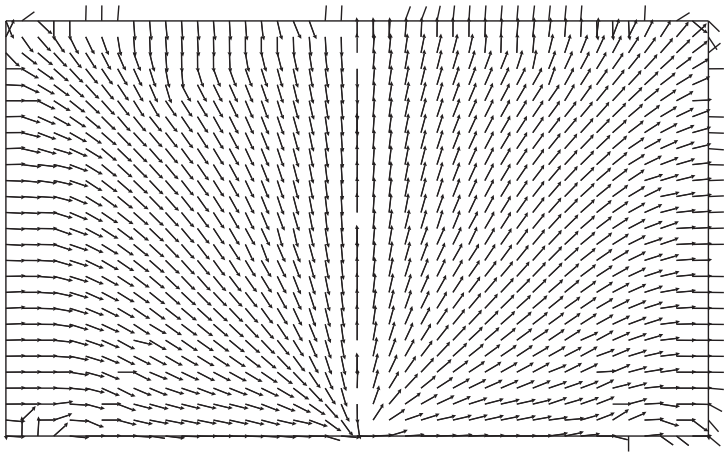


Fig. 38. Regularity issues for the computed optimal orientation α (bridge case).

Note that except when σ is proportional to the identity, the optimal orientation angle α is unique up to the addition of a multiple of π. As shown in Fig. 38, this non-uniqueness creates a regularity issue for α. For example,

- α or $\alpha + \pi$ correspond to the same orientation,
- where the material density is close to 0 or 1, the orientation does not play any role (cf. the corners in Fig. 38),
- there are real singularities of the orientation, like a fan (cf. the bottom middle in Fig. 38),
- if the values of m_1 and m_2 are exchanged, then the optimal orientation switches from α to $\alpha + \pi/2$, but it does not seem to appear in our results.

These issues create some numerical difficulties, that we will explain in the next section.

6.6 3rd step: reconstruction of an optimal periodic structure

In the previous sections we computed an optimal homogenized design (with an underlying modulated periodic structure). In this section, let us enter a post-processing step. We choose a length scale ε for this projection step and reconstruct a periodic shape with length scale ε, approximating the optimal one.

6.6.1 Projection in the simple case without varying orientation, $\alpha \equiv 0$

First, we consider the case where there is no varying orientation, that is, $\alpha \equiv 0$. The unit cells (rectangular hole in a square) are defined by

$$Y(m) = \left\{ y \in [0,1]^2 : \begin{array}{c} \cos(2\pi y_1) \geq \cos(\pi(1-m_1)), \\ \text{or} \\ \cos(2\pi y_2) \geq \cos(\pi(1-m_2)) \end{array} \right\}.$$

The domain D is paved with cells $\varepsilon Y(m)$. Since the hole size $m(x)$ is varying in D, the periodicity cell is macroscopically modulated and we define a projected lattice shape $\Omega_\varepsilon(m)$ as

$$\Omega_\varepsilon(m) := \left\{ x \in D : \begin{array}{c} \cos\left(\dfrac{2\pi x_1}{\varepsilon}\right) \geq \cos(\pi(1-m_1(x))), \\ \text{or} \\ \cos\left(\dfrac{2\pi x_2}{\varepsilon}\right) \geq \cos(\pi(1-m_2(x))) \end{array} \right\},$$

where $m_1(x), m_2(x)$ are functions defined on D with values in $[0,1]$.

Remark 6.1. *The values of m_1 m_2 are not necessarily constant in each cell of the structure. Hence, the holes in the cellular structure $\Omega_\varepsilon(m)$ are not exactly rectangles. But, when ε goes to 0, the sequence of cellular structures converges to the composite with local Hooke's law equal to $A_0^*(m)$.*

The cellular structures can be defined using level-sets. We introduce two functions $f_{\varepsilon,i}^m$, one for each direction

$$f_{\varepsilon,i}^m(x) := -\cos\left(\frac{2\pi x_i}{\varepsilon}\right) + \cos(\pi(1-m_i(x))),$$

and the level-set function

$$F_\varepsilon^m := \min(f_{\varepsilon,1}^m, f_{\varepsilon,2}^m).$$

The final structure $\Omega_\varepsilon(m)$ is then defined by

$$\Omega_\varepsilon(m) = \{x \in D : F_\varepsilon^m(x) \leq 0\}.$$

The construction of a minimizing sequence of shapes is immediate: we just have to update the size ε in the previous level-set function.

6.6.2 Projection in the general case with orientation, $\alpha \neq 0$

This section is based on the paper of Pantz and Trabelsi [PT2008]. The main idea of Pantz and Trabelsi is to find a map $\varphi = (\varphi_1, \varphi_2)$ from D into $\mathbb{R}^2$ which distorts the regular square grid in order to orientate each square at the optimal angle α (or $\alpha + \pi$). Once φ is found, we can proceed as before.

The final shape, now denoted $\Omega_\varepsilon(\varphi, m)$, is still defined by a level-set function:

$$\Omega_\varepsilon(\varphi, m) = \{x \in D : F_\varepsilon^{\varphi,m}(x) \leq 0\}$$

with $F_\varepsilon^{\varphi,m} = \min(f_{\varepsilon,1}^{\varphi,m}, f_{\varepsilon,2}^{\varphi,m})$ and

$$f_{\varepsilon,i}^{\varphi,m}(x) = -\cos\left(\frac{2\pi\varphi_i(x)}{\varepsilon}\right) + \cos(\pi(1-m_i(x))).$$

Geometrically (in 2-d), the Jacobian $\nabla\varphi$ should be proportional to the rotation matrix defined by

$$Q(\alpha) = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}. \tag{6.6}$$

One possibility (Pantz and Trabelsi) is to find φ which (roughly) minimizes

$$\int_D |\nabla\varphi - Q(\alpha)|^2 \, dx.$$

However, it is not straightforward because $Q(\alpha)$ is not smooth (the orientation is not coherent a priori). Pantz and Trabelsi proposed a complicated trick to avoid this coherent orientation issue. We also mention that Groen and Sigmund [GS2018] suggested another trick from image processing to obtain a coherent orientation.

6.6.3 The new approach of Allaire–Geoffroy–Pantz

In this section, we propose a new approach, based on the paper of Allaire–Geoffroy–Pantz [AGP2018].

Let us recall that geometrically (in 2-d), at every point $x \in D$, the Jacobian $\nabla\varphi$ should be proportional to the rotation

matrix $Q(\alpha)$ defined by (6.6). Moreover, the proportions of the cell have to be preserved in order to converge to a true square and not simply to a rectangle. For this purpose, we impose $|\nabla\varphi_1| = |\nabla\varphi_2| = e^r$, where $r \in H^1(D)$ is a scalar dilation field. Then the Jacobian $\nabla\varphi$ should be

$$\nabla\varphi = e^r Q(\alpha) \quad \text{in } D. \tag{6.7}$$

This equation can be satisfied only if α satisfies the following conformality condition.

Lemma 6.2. *Let α be a regular orientation field and D be a simply connected domain. There exists a mapping function φ and a dilation field r satisfying $\nabla\varphi = e^r Q(\alpha)$ if and only if*

$$\Delta\alpha = 0 \quad \textit{in } D. \tag{6.8}$$

We recall that, for a vector field $u = (u_1, u_2)$, its curl is defined as $\operatorname{curl} u = \nabla \wedge u = \dfrac{\partial u_2}{\partial x_1} - \dfrac{\partial u_1}{\partial x_2}$, where $\wedge$ is the 2-d cross product of vectors. Of course, $\operatorname{curl} \nabla\varphi = (\operatorname{curl} \nabla\varphi_1, \operatorname{curl} \nabla\varphi_2) = 0$.

Proof of Lemma 6.2. Since the domain $D \subset \mathbb{R}^2$ is simply connected, By Poincaré's lemma, the map φ exists if and only if the right hand side of (6.7) is curl-free. That is,

$$\operatorname{curl}(e^r Q(\alpha)) = 0.$$

Let a_1, a_2 be the columns of $Q(\alpha)$. Then

$$\operatorname{curl}(e^r Q(\alpha)) = 0 \Longleftrightarrow \nabla r \wedge a_i = -\nabla \wedge a_i, \quad i = 1, 2. \tag{6.9}$$

Since, for fixed α, (a_1, a_2) is an orthonormal basis of $\mathbb{R}^2$, by (6.9) we see that the vector ∇r can be decomposed as

$$\nabla r = (-\nabla \wedge a_2)a_1 + (\nabla \wedge a_1)a_2.$$

We compute

$$\nabla \wedge a_1 = \frac{\partial\alpha}{\partial x_1}\cos(\alpha) + \frac{\partial\alpha}{\partial x_2}\sin(\alpha) \quad \text{and} \quad \nabla \wedge a_2 = -\frac{\partial\alpha}{\partial x_1}\sin(\alpha) + \frac{\partial\alpha}{\partial x_2}\cos(\alpha).$$

It leads to

$$\nabla r = \left(-\frac{\partial\alpha}{\partial x_2}, \frac{\partial\alpha}{\partial x_1}\right)^T.$$

Thus, again by Poincaré's lemma, the dilation factor r exists if and only if the right hand side in the above is curl-free, which leads to the harmonic condition on α. □

The following proposition is a very useful property of conformal orientations.

Proposition 6.3. *If there exists a map $\varphi = (\varphi_1, \varphi_2)$ from $D \subset \mathbb{R}^2$ to $\mathbb{R}^2$ such that*

$$\nabla\varphi = e^r Q(\alpha) \quad \textit{in } D,$$

then all angles are preserved by the map φ. In particular, small square cells are deformed into almost square cells locally.

Sketch of the proof. Let $x = (x_1, x_2)$ be the origin of a small square S of side length $\varepsilon > 0$ and edges εe_1, εe_2 (in other words, the vertices of S are $x, x + \varepsilon e_1, x + \varepsilon e_2$ and $x + \varepsilon e_1 + \varepsilon e_2$). The map φ then transforms S into an "almost" square $\varphi(S)$ of vertices $\varphi(x), \varphi(x + \varepsilon e_1), \varphi(x + \varepsilon e_2)$ and $\varphi(x + \varepsilon e_1 + \varepsilon e_2)$. By a Taylor expansion, we see that $\varphi(S)$ is, up to terms of order ε^2, equal to a parallelogram of origin $\phi(x)$ and edges $\varepsilon\dfrac{\partial\varphi}{\partial x_1}$ and $\varepsilon\dfrac{\partial\varphi}{\partial x_2}$. Since $\nabla\varphi = e^r Q(\alpha)$, the two edges are orthogonal, so $\varphi(S)$ is "almost" a square with side length $\varepsilon e^r(x)$ (possibly rotated with respect to the initial square S). □

Nevertheless, in the applications we face a problem. Since α is a stress eigen-direction, it has no reason to be a harmonic function in general. Even worse, α might not even be smooth at some places (for example, at corners or at the junction point of different boundary conditions, but at other places as well). A more profound reason lies in the following observation: both α and $\alpha + \pi$ give rise to the same orientation. By Pedersen's result, we can show that the rotated Hooke's law $A^*(m_1, m_2, \alpha) = R(\alpha)^T A^*(m_1, m_2, 0)R(\alpha)$ depends only on the double angle 2α.

From now on, we are going to be working with the double angle $\beta = 2\alpha$, thus, removing the indeterminate additive constant π. In what follows, we shall regularize the double angle $\beta = 2\alpha$ and make it harmonic.

6.6.4 Regularization of the double angle $\beta = 2\alpha$

As we mentioned before, the orientation α given by the optimization does not necessarily satisfy the conformality condition $\Delta\alpha = 0$. Thus, at each iteration of the algorithm, instead of minimizing locally (by using Pedersen's result) the following quantity

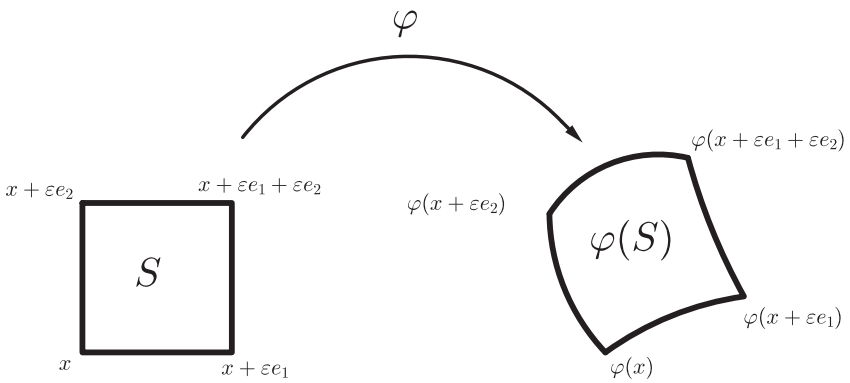


Fig. 39. A small square S and the distorted square $\varphi(S)$.

$$A^*(m_1, m_2, \beta)^{-1}\sigma \cdot \sigma,$$

let us consider to minimize globally

$$\int_D (A^*(m_1, m_2, \beta)^{-1}\sigma \cdot \sigma + \eta^2 |\nabla\beta|^2)\, dx$$

for a small parameter $\eta > 0$, under the harmonic constraint

$$\int_D \nabla\beta \cdot \nabla q\, dx = 0 \quad \forall q \in H_0^1(D).$$

This is a non-linear (and non-quadratic) constrained optimization problem. It turns out that working with the angle β is not so easy since the Hooke's law $A^*(m_1, m_2, \beta)$ is highly nonlinear in terms of β. It is however quadratic with respect to the vector b, defined by

$$b = (\cos\beta, \sin\beta) \quad \text{and} \quad A^*(m_1, m_2, \beta) = S(b)^T A(m_1, m_2, 0) S(b),$$

where the matrix $S(b)$ is defined by $S(b) = R(\alpha)$. It can be shown that $S(b)$ is affine with respect to b by a careful computation based on Pedersen's result. Therefore, from now on we shall work with b as the main unknown. We remark that $\nabla\beta = b \wedge \nabla b$. Indeed,

$$b \wedge \nabla b = (\cos\beta, \sin\beta) \wedge (-\sin\beta\, \nabla\beta, \cos\beta\, \nabla\beta) = \nabla\beta.$$

Therefore, the objective function becomes

$$\int_D (A^*(m_1, m_2, b)^{-1}\sigma \cdot \sigma + \eta^2 |b \wedge \nabla b|^2)\, dx, \tag{6.10}$$

which in turn has to be minimized under the harmonic constraint

$$\int_D (b \wedge \nabla b) \cdot \nabla q\, dx = 0 \quad \forall q \in H_0^1(D). \tag{6.11}$$

We iteratively solve the non-linear problem (6.10) under the minimization constraint (6.11) by a Newton-type approximation with an increment δb as follows.

Find a step $\delta b^n \in H^1(D; \mathbb{R}^2)$ and a Lagrange multiplier $p^{n+1} \in H_0^1(D)$ such that, for any $\delta c \in H^1(D; \mathbb{R}^2)$ and $q \in H_0^1(D)$,

$$\begin{aligned} &\int_D A^*(m)^{-1} S(b^n + \delta b^n)\sigma \cdot S'(\delta c)\sigma\, dx + \eta^2 \int_D (b^n \wedge \nabla(b^n + \delta b^n)) \cdot (b^n \wedge \nabla \delta c)\, dx \\ &\quad + \int_D (b^n \wedge \nabla \delta c) \cdot \nabla p^{n+1}\, dx = 0 \end{aligned} \tag{6.12}$$

and

$$\int_D (b^n \wedge \nabla(b^n + \delta b^n)) \cdot \nabla q\, dx = 0, \tag{6.13}$$

where $S'(\delta c)$ is the directional derivative of $S(b)$ in the direction δc. Notice that, since S is an affine function, then $S'(\delta c) = S(c + \delta c) - S(c)$. At each iteration, we update the vector field b as follows

$$b^{n+1} = \frac{b^n + \delta b^n}{|b^n + \delta b^n|}. \tag{6.14}$$

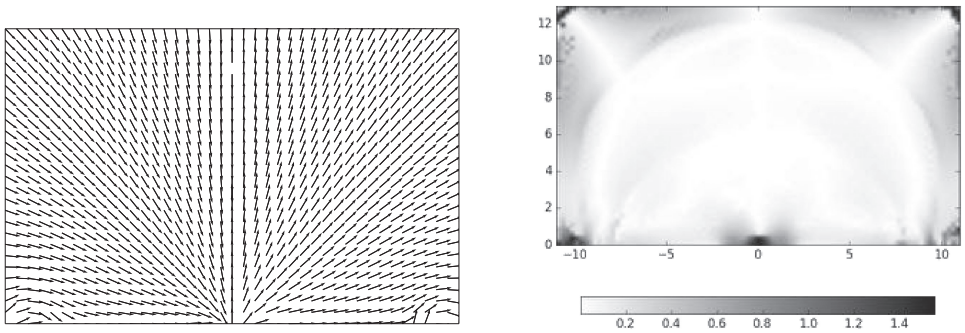


Fig. 40. Regularized orientation for the bridge case (left) and angle difference between optimized and regularized orientation.

We now apply this regularization process together with the alternate minimization algorithm.

The above mentioned algorithm is structured as follows:

Algorithm 9 Regularization algorithm

1. Initialization of the design parameters m_1^0, m_2^0, b with the results of the optimization without the harmonic constraint.
2. Iterations until convergence, for $n \geq 0$:
   (a) Computation of the stress tensor σ^n through a problem of linear elasticity.
   (b) Updating of the hole parameters m^n by using the projected gradient algorithm (6.5) with the orientation b^n.
   (c) Computation of the increment δb^n by solving (6.12) and (6.13).
   (d) Updating of the orientation with (6.14).

Remark 6.4. *In numerical practice, starting from the optimal (but not necessarily smooth) orientation, a few tens of iterations of this regularization process are enough.*

Remark 6.5. *One advantage of this b-formulation is that it is insensitive to the 2π-modulo of β.*

As we can see by Fig. 40, the regularization occurs mainly in areas where the density is close to 0 or 1, i.e., where the homogenized material is almost isotropic and the orientation has no significant impact.

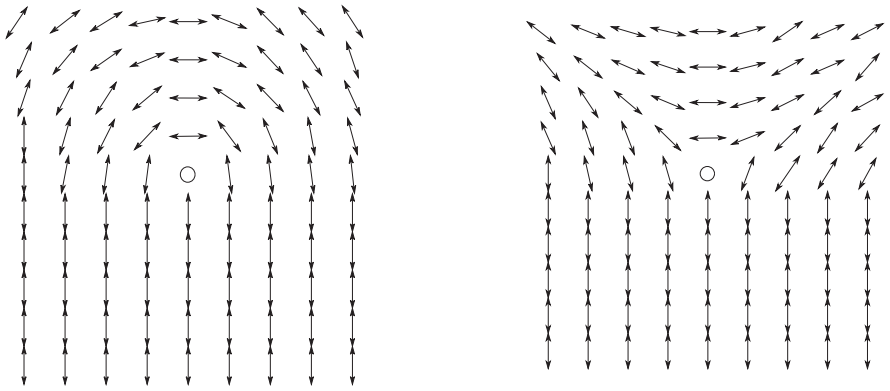


Fig. 41. Positive singularity (left) and negative singularity (right).

Remark 6.6. *Unfortunately this process does not always work in general. In other words, it does not always yield a smooth harmonic angle β. This is because of "true" singularities, i.e., singularities of the orientation that remain and thus do not allow the angle to be harmonic. The vector field is not coherently orientable in these cases, as the vector rotates by an angle of $\pm\pi$ along circles which are enclosing the singularities (see Fig. 41). Such cases might be handled by means of other regularization processes (e.g., minimizing a Ginzburg–Landau energy [GE2018]).*

6.6.5 Computation of the map φ

Once a harmonic angle $\alpha = \beta/2$ has been found, one needs to compute r and φ such that

$$\nabla\varphi = e^r Q(\alpha) \quad \text{in } D.$$

Since the dilation field r satisfies

$$\nabla r = (-\nabla \wedge a_2)a_1 + (\nabla \wedge a_1)a_2 \quad \text{with } (a_1, a_2) = Q(\alpha),$$

one computes r as the minimizer in $H^1(D)$ of

$$\int_D |\nabla r + (\nabla \wedge a_2)a_1 - (\nabla \wedge a_1)a_2|^2 \, dx.$$

Once r has been computed, a naive idea would be to compute φ as a minimizer in $H^1(D; \mathbb{R}^2)$ of

$$\int_D |\nabla \varphi - e^r Q(\alpha)|^2 \, dx.$$

However, we know that, even if β is smooth, α may have jumps of the type $\pm\pi$ and thus $Q(\alpha)$ may have jumps of its sign (recall that $Q(\alpha + \pi) = -Q(\alpha)$).

To compute φ there are two possibilities.

1. Find a coherent orientation of α (i.e., choose between α and $\alpha + \pi$ at every point): this is possible only if there are no singularities (this is the approach of Groen and Sigmund [GS2018]).
2. Leave the angle α as it is and extend φ to be defined in an abstract manifold. This is the approach of Allaire–Geoffroy–Pantz [AGP2018] and it works also in the presence of singularities.

6.6.6 An abstract manifold setting

Let us introduce the cover space of D.

Definition 6.7. *Denote by T a rotation matrix field which is a candidate for being $Q(\alpha)$. Then we define*

$$\mathcal{D} = \{(x, T) \in D \times \mathrm{SO}(2) \text{ such that } T^2 = Q(\beta)\},$$

where SO(2) *is the set of rotations in* $\mathbb{R}^2$.

We note that at every point $x \in D$ the rotation satisfies $T(x)^2 = Q(\beta)(x)$. If the angle α is globally orientable, then $T(x) = Q(\alpha)(x)$ or $T(x) = -Q(\alpha)(x)$, and that $\mathcal{D}$ is simply the union of two copies of D, consisting of the two possible signs of $Q(\alpha)$. We assume the simple case where α could be globally oriented (no singularity) but extend it to the singular cases.

We change our working space from D to $\mathcal{D}$. The map $\varphi(x, T)$ is now defined on the manifold $\mathcal{D}$ by

$$\nabla \varphi = e^r T, \tag{6.15}$$

and the gradient operator in (6.15) defined by

$$\nabla \varphi(x, T) = \nabla \varphi_U(x),$$

where U is an orientable open subset of D,

$$\varphi_U(x) = \varphi \circ g_U(x),$$

and g_U is one of the charts

$$\begin{aligned} g_U^{\pm} \colon U &\longrightarrow \mathcal{D} \\ x &\longmapsto (x, \pm T_U(x)) \end{aligned}$$

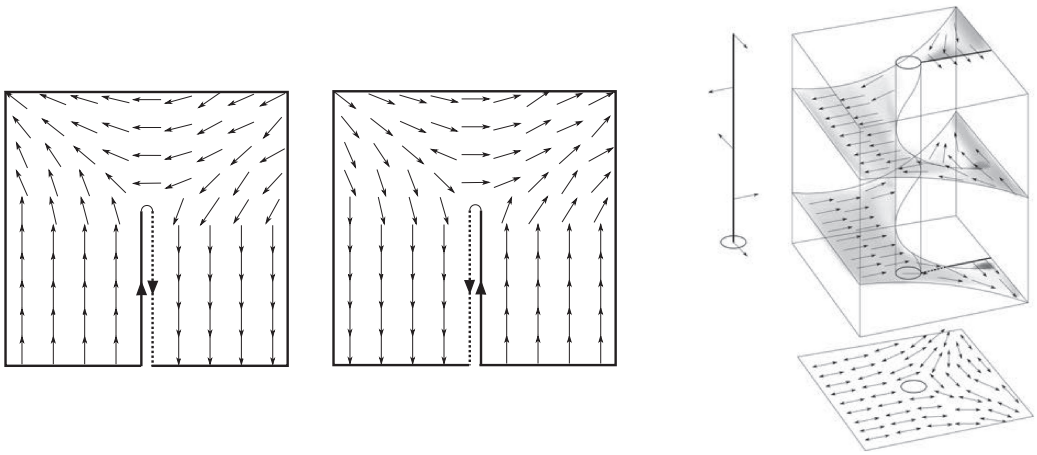


Fig. 42. Two possible orientation (left) and the manifold obtained by gluing them together.

with $T_U^2 = Q(\beta)$ and $T_U \in C(U, \mathrm{SO}(2))$. Moreover, without loss of generality, we can assume the antisymmetric property

$$\varphi(x, -T) = -\varphi(x, T).$$

Indeed, if φ satisfies $\nabla\varphi = e^r T$, then the map $(\varphi(x,T) - \varphi(x,-T))/2$ still satisfies (6.15) and is antisymmetric. Thus, if the orientation α satisfies the conformality condition (6.8), the map φ can be defined as a minimizer of

$$\min_{\varphi \in \mathcal{V}} \int_{\mathcal{D}} |\nabla\varphi - e^r T|^2 \, dx,$$

over all maps φ in

$$\mathcal{V} := \{\varphi \in H^1(D, \mathbb{R}^2) \ : \ \varphi(x, -T) = -\varphi(x, T) \text{ for all } (x, T) \in \mathcal{D}\}.$$

In practice we face the problem of making the actual computations on the abstract manifold $\mathcal{D}$. In order to solve this problem we use a new idea, namely, non-conformal finite elements on D.

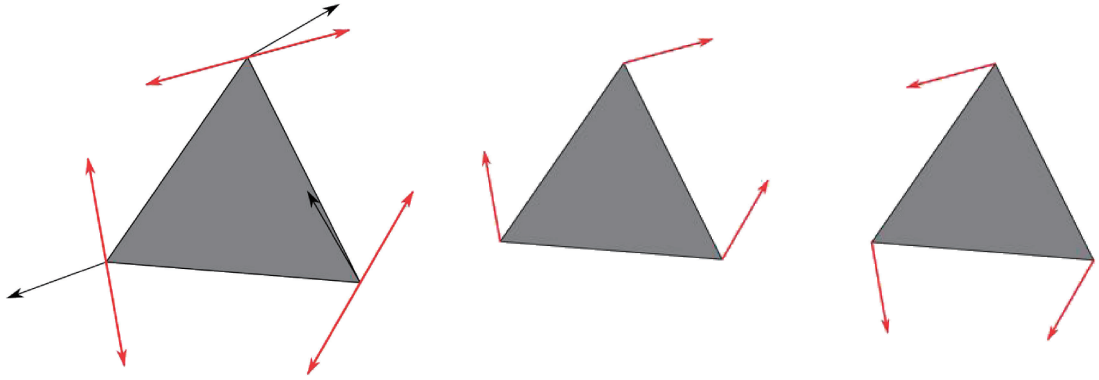


Fig. 43. Left: orientation of β (black arrows) and of α (red arrows). Right: two possible coherent orientations of α.

On each triangle K of the mesh we compute one continuous orientation T_K such that $T_K^2 = Q(\beta)$. Then, we glue together (with P_1 discontinuous finite elements) these orientations. We compute

$$\begin{aligned}\int_{\mathcal{D}} |\nabla\varphi - e^r T|^2 \, dx &= \sum_K \int_{g_K^+(K) \cup g_K^-(K)} |\nabla\varphi - e^r T|^2 \, dx \\ &= \sum_K \int_K |\nabla(\varphi \circ g_K^+) - e^r T_K(x)|^2 \, dx \\ &\quad + \sum_K \int_K |\nabla(\varphi \circ g_K^-) - e^r T_K(x)|^2 \, dx\end{aligned}$$

with $g_K^\pm = \mathrm{Id} \times (\pm T_K)$. By the antisymmetry of φ, we obtain

$$\int_{\mathcal{D}} |\nabla\varphi - e^r T|^2 \, dx = 2 \sum_K \int_K |\nabla(\varphi \circ g_K^+) - e^{r(x)} T_K(x)|^2 \, dx.$$

Then, we minimize with respect to φ in the space of P_1 discontinuous finite elements.

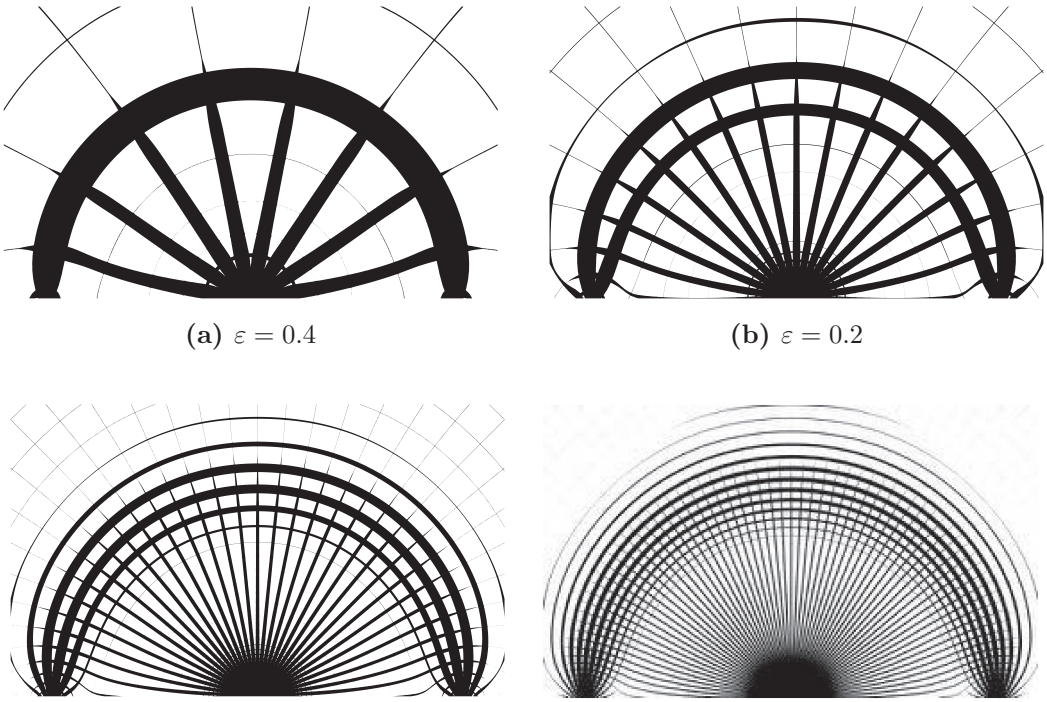
6.7 A final post-processing/cleaning of the lattice reconstruction

The shapes we obtained in Sect. 6.6 are not straightforwardly manufacturable. Indeed, there are disconnected components of the lattice structure and/or too thin members that should be removed. A final post-processing is made to cure these defects. Note that there is room for improvement in the process.

Let $h_{\min}$ be the minimal manufacturable lengthscale or feature size, meaning the smallest possible width of bars and diameter of holes which can be effectively built. Recall that ε is our choice of a global size of cells. After deformation, the cell size is $h_c(x) = \varepsilon e^{-r(x)}$. Hence the local widths of the bars and holes are respectively given by $(1 - m_i(x))h_c(x)$ and $m_i(x)h_c(x)$.

In the following, we distinguish two regimes, depending of the local size of the cell $h_c(x)$. First, if the cell size is too small, a hole and a bar of minimal width cannot coexist and then we have to choose a completely full or void cell. Hence, if $h_c < 2h_{\min}$, a thresholding is applied separately to each field m_i: it is assigned the value 0 if $m_i < 0.5$ and 1 otherwise.

Second, when $h_c \geq 2h_{\min}$, our post-processing criterion is satisfied if



(a) $\varepsilon = 0.4$ (b) $\varepsilon = 0.2$

(c) $\varepsilon = 0.1$ (d) $\varepsilon = 0.05$

Fig. 44. $\Omega_\varepsilon(\varphi, m)$ for several ε in the case of the bridge.

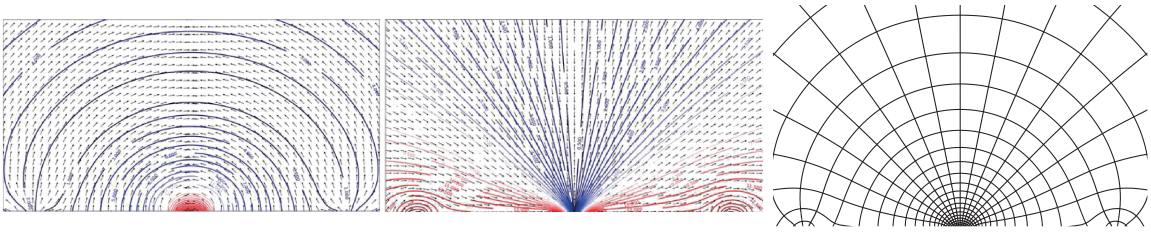


Fig. 45. The map $|\varphi_i|$ (isolines) and the vectors a_i (arrows) for $i = 1$ (left) and $i = 2$ (middle). On the right we have the projection of a regular grid by the map φ.

$$\frac{h_{\min}}{h_c} \leq m_i \leq 1 - \frac{h_{\min}}{h_c}, \quad i = 1, 2.$$

Otherwise, we simply threshold the values of m_1 and m_2, according to Fig. 46, in order to reach void or full materials. The thresholded m is then denoted by $\tilde{m}$.

Let $O_\varepsilon(\varphi, \tilde{m})$ be the shape obtained from $\Omega_\varepsilon(\varphi, \tilde{m})$ by filling its closed holes. Numerically, the complement of $\Omega_\varepsilon(\varphi, \tilde{m})$ is computed step by step, by evaluating the sign of $F_\varepsilon^{\varphi,\tilde{m}}$. If it is positive, the current vertex belongs to the complement $\Omega_\varepsilon^c(\varphi, \tilde{m})$ and then its neighbors, which are not already visited, are added to the list of vertices which should be tested. Otherwise, the current vertex does not belong to $\Omega_\varepsilon^c(\varphi, \tilde{m})$.

We will regularize the subset $O_\varepsilon(\varphi, \tilde{m})$ in order to remove the disconnected bars or the bars that have one free endpoint. Numerically, we explore all the vertices of the complement as follows. For any given vertex, we check each other vertex not further away than a distance $h_{\min}$: if this vertex belongs to the complement too, all vertices between them are added to the complement. In this way, we suppress all disconnected bars and all bars of $O_\varepsilon(\varphi, \tilde{m})$ that have one free endpoint, which are not too wide. This new subset is denoted by $\tilde{O}_\varepsilon(\varphi, \tilde{m})$.

Finally, the post-processed structure is given by the intersection $\tilde{\Omega}_\varepsilon(\varphi, \tilde{m}) := \Omega_\varepsilon(\varphi, \tilde{m}) \cap \tilde{O}_\varepsilon(\varphi, \tilde{m})$. Several post-processed structures $\tilde{\Omega}_\varepsilon(\varphi, \tilde{m})$ for the bridge case are displayed in Fig. 47.

6.8 Other numerical examples

In this section, we show some numerical examples of the application of the whole method of this section. Here, we mention that the method is not always applicable because some singularities cannot be eliminated, as we will see later.

We will show the numerical results for the optimization of a cantilever (Fig. 48), an MBB beam (Fig. 49) and an L-beam (Fig. 50). For each case, we have represented:

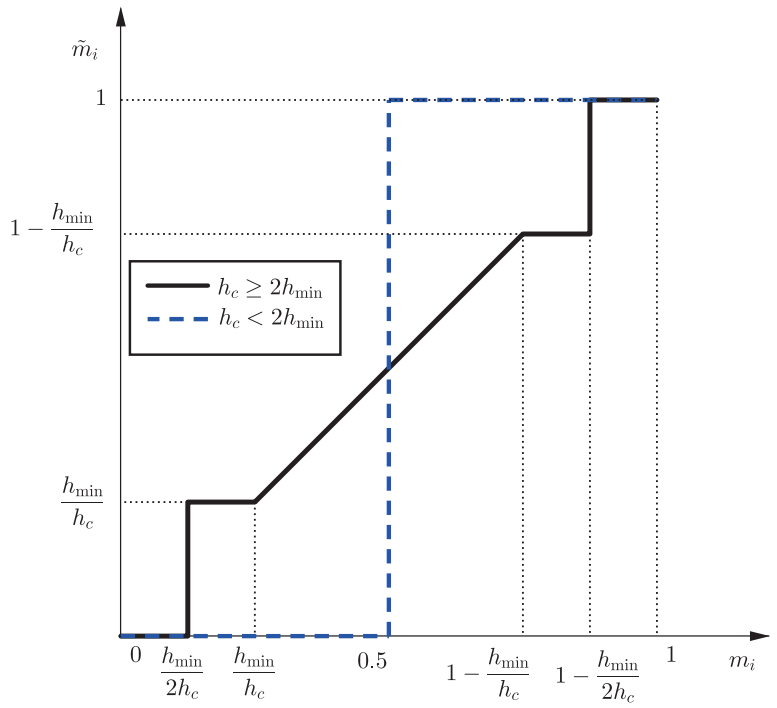
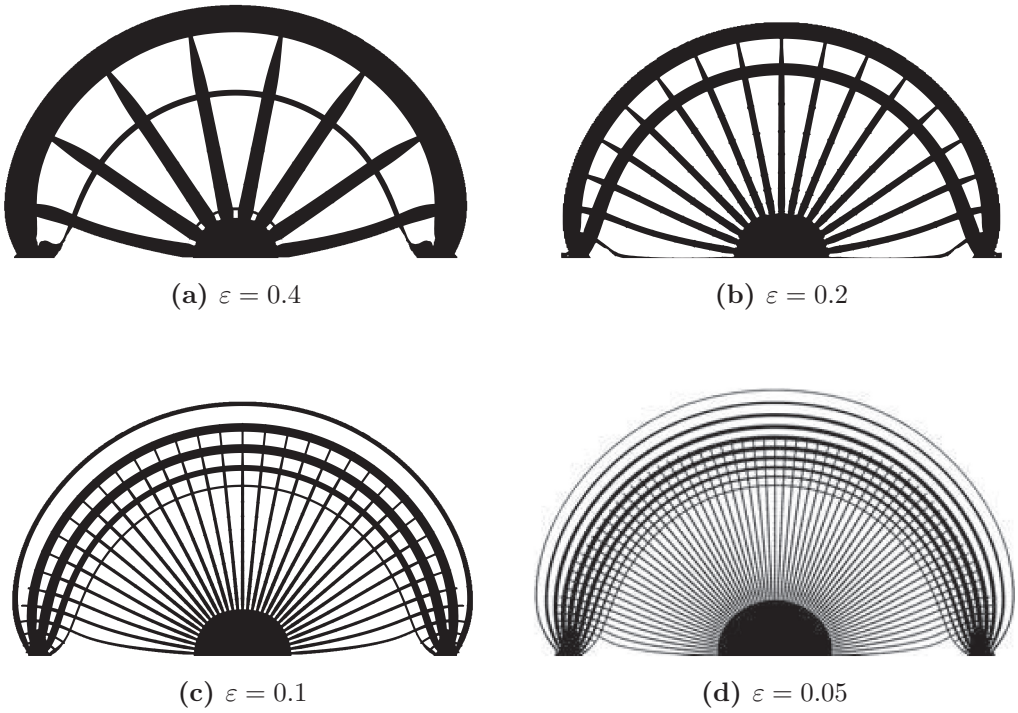


Fig. 46. Thresholding criteria.



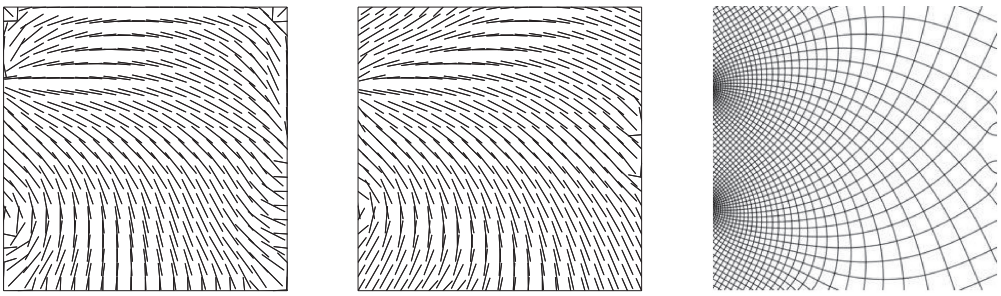
(a) $\varepsilon = 0.4$ **(b)** $\varepsilon = 0.2$

(c) $\varepsilon = 0.1$ **(d)** $\varepsilon = 0.05$

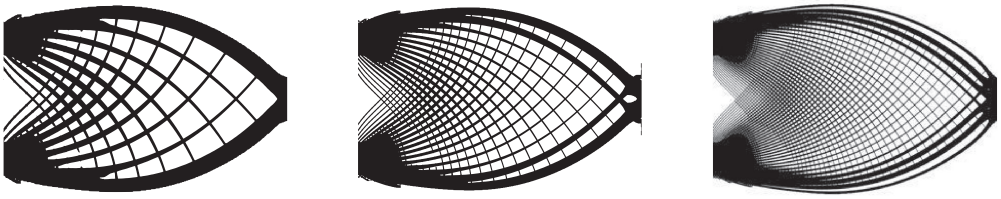
Fig. 47. Post-processed $\bar{\Omega}_\varepsilon(\varphi, \bar{m})$ for several ε in the case of the bridge.

(a) the optimal orientation of the periodicity cells before regularization,
(b) the optimal orientation of the periodicity cells after regularization,
(c) the underlying lattice on which the optimal composite is built, i.e., the projection by φ,
(d)–(f) the sequence of shapes after post-processing for the case of $\varepsilon = 0.2$, 0.1 and 0.05 respectively.

As shown in the case of L-beam (Fig. 50), the singularities which appear in Fig. 50-(a) are removed during the regularization step (Fig. 50-(b)). This is a necessary condition in order to apply our method. Indeed, as we mentioned, there is a case where the singularity cannot be removed, the so-called electrical mast (see Fig. 51). Fig. 51-(a) shows that two negative singularities, located inside the domain, cannot neither be removed nor pushed toward the boundary during the regularization step.

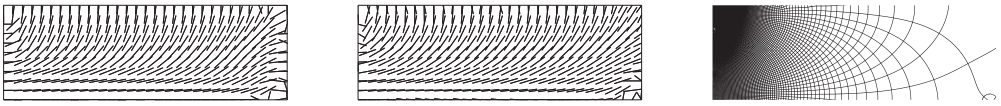


(a) Opt. orientation **(b)** Regul. orientation **(c)** Projection by φ

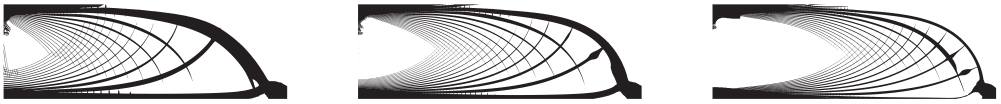


(d) $\bar{\Omega}_\varepsilon(\varphi, \bar{m})$, $\varepsilon = 0.2$ **(e)** $\bar{\Omega}_\varepsilon(\varphi, \bar{m})$, $\varepsilon = 0.1$ **(f)** $\bar{\Omega}_\varepsilon(\varphi, \bar{m})$, $\varepsilon = 0.05$

Fig. 48. Cantilever case.



(a) Opt. orientation **(b)** Regul. orientation **(c)** Projection by φ



(d) $\bar{\Omega}_\varepsilon(\varphi, \bar{m})$, $\varepsilon = 0.2$ **(e)** $\bar{\Omega}_\varepsilon(\varphi, \bar{m})$, $\varepsilon = 0.1$ **(f)** $\bar{\Omega}_\varepsilon(\varphi, \bar{m})$, $\varepsilon = 0.05$

Fig. 49. MBB beam case.

Hence, the computed grid shown in Fig. 51-(b) is clearly not correctly aligned with the optimal orientation of the cells in the vicinity of the singularities. To overcome this problem, at least two different strategies can be considered. One consists in modifying the regularization process in a way that forces more effectively the singularities to be eliminated (see Figs. 51-(c) and 51-(d)). Another approach is to adapt the projection step so that it is able to take singularities into account, and will be discussed in the next section.

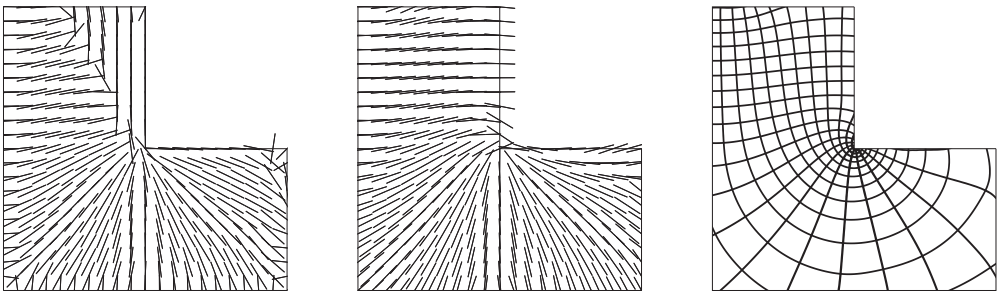
6.9 Further issues

As we mentioned in Sect. 6.6, the problem of removing the singularity might persist. To overcome this problem, some preliminary remedies are proposed in the PhD thesis of Perle Geoffroy [GE2018] by either trying to eliminate them by a Ginzburg–Landau approach or compute a map φ with the previous approach and an enriched discontinuous finite element space. In [GE2018], one may find others extensions to cases, such as different objective functions, multiple loads and 3-d problems.

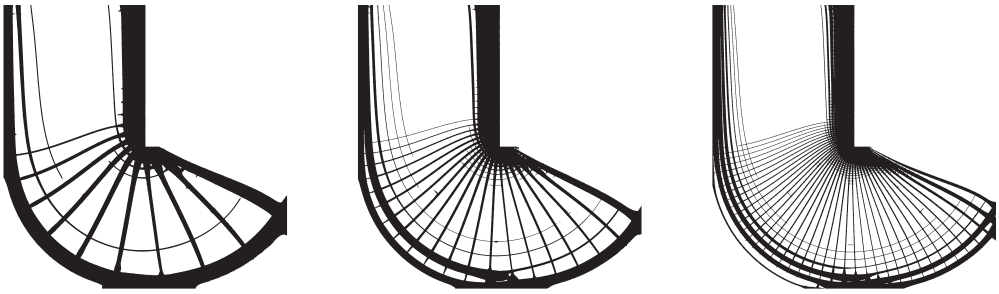
6.10 Exercises

Problem 6.10.1. *Consider a compliance minimization problem (for one test case like cantilever, bridge, MBB beam, etc.) for the homogenized formulation. Choose an homogenized tensor corresponding to a non-isotropic microstructure (for example a square cell with a rectangular hole with fixed size). Then minimize the compliance with respect to the sole orientation of the microstructure (using Pedersen result).*

Problem 6.10.2. *For the same compliance minimization problem as in the previous exercise, fix now the orientation and let the parameters of the microstructure (for example, the lengths m_1 and m_2 of the rectangular hole, see Fig. 21)*

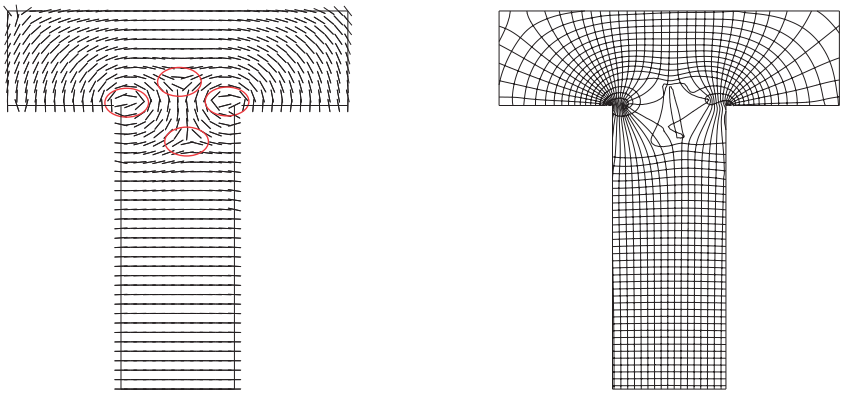


(a) Opt. orientation **(b)** Regul. orientation **(c)** Projection by φ

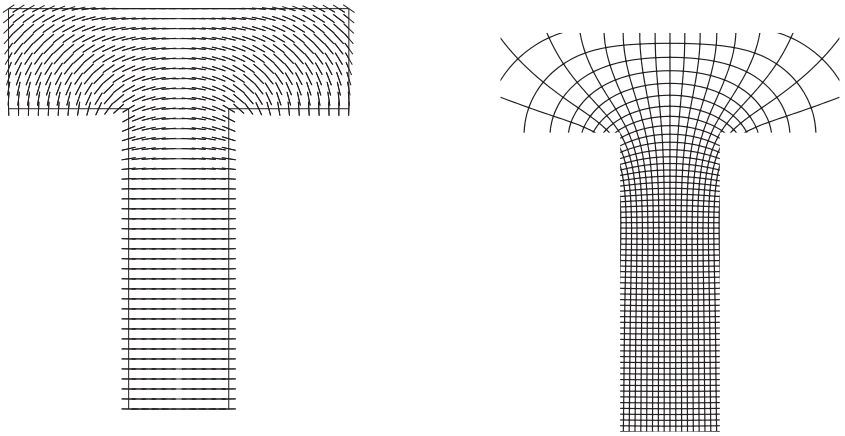


(d) $\bar{\Omega}_\varepsilon(\varphi, \bar{m})$, $\varepsilon = 0.2$ **(e)** $\bar{\Omega}_\varepsilon(\varphi, \bar{m})$, $\varepsilon = 0.1$ **(f)** $\bar{\Omega}_\varepsilon(\varphi, \bar{m})$, $\varepsilon = 0.05$

Fig. 50. L-beam case.



(a) Singularities appear **(b)** Projection by φ



(c) Improved reg. orientation **(d)** Reconstructed grid

Fig. 51. Improved regularized orientation in the case of the electrical mast.

become the optimization variables. Then minimize the compliance with respect to these parameters (with fixed orientation). Compare the optimal designs and the attained minimal compliances with the previous exercise. In particular, notice that the absence of orientation optimization yields a self-penalizing effect, namely the obtained designs feature almost no intermediate densities (somehow similarly to the SIMP method).

Problem 6.10.3. *Combine the two previous optimization (with respect to the orientation and the size parameters) and minimize the compliance for the test case of the previous exercises.*

REFERENCES

[Al1992] Allaire, G., "Homogenization and two-scale convergence," *SIAM J. Math. Anal.*, **23**: 1482–1518 (1992).

[Al2002] Allaire, G., Shape Optimization by the Homogenization Method, Vol. 146 of Applied Mathematical Sciences, Springer-Verlag, New York (2002).

[Al2007-1] Allaire, G., Conception Optimale de Structures, Vol. 58 of Mathématiques et Applications, Springer, Heidelberg (2007).

[Al2007-2] Allaire, G., Numerical Analysis and Optimization. An Introduction to Mathematical Modelling and Numerical Simulation. Translated from the French by Alan Craig, Numer. Math. Sci. Comput., Oxford University Press, Oxford, UK (2007).

[AGP2018] Allaire, G., Geoffroy-Donders, P., and Pantz, O., Topology Optimization of Modulated and Oriented Periodic Microstructures by the Homogenization Method, Computers & Mathematics with Applications, special issue SimAM (2019).

[AP2006] Allaire, G., and Pantz, O., "Structural optimization with FreeFem++," *Struct. Multidiscip. Optim.*, **32**: 173–181 (2006).

[BK1988] Bendsøe, M., and Kikuchi, N., "Generating optimal topologies in structual design using a homogenization method," *Comput. Methods Appl. Mech. Engrg.*, **71(2)**: 197–224 (1988).

[BS2003] Bendsøe, M., and Sigmund, O., Topology Optimization, Springer-Verlag, Berlin (2003).

[BLP1978] Bensoussan, A., Lions, J.-L., and Papanicolaou, G., Asymptotic Analysis for Periodic Structures, Studies in Mathematics and Its Applications, North-Holland, Amsterdam (1978).

[BGLS2006] Bonnans, J. F., Gilbert, C., Lemaréchal, C., and Sagastizábal, C., Numerical Optimization: Theoretical and Practical Aspects (Universitext), Second edition, Springer-Verlag, Berlin (2006).

[Ch2000] Cherkaev, A., Variational Methods for Structural Optimization, Springer Verlag, New York (2000).

[CD1999] Cioranescu, D., and Donato, P., An Introduction to Homogenization, Oxford Lecture Series in Mathematics and Applications, **17**, Oxford (1999).

[ET1999] Ekeland, I., and Témam, R., Convex Analysis and Variational Problems, Classics in Applied Mathematics, **28**, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (1999).

[GLRS2013] Geihe, B., Lenz, M., Rumpf, M., and Schultz, R., "Risk averse elastic shape optimization with parametrized fine scale geometry," *Math. Program., Ser. A*, **141**: 383–403 (2013).

[GE2018] Geoffroy-Donders, P., *Homogenization method for topology optimization of structures built with lattice materials*, PhD thesis, Ecole Polytechnique, Université Paris-Saclay (2018).

[GAP2018] Geoffroy-Donders, P., Allaire, G., and Pantz, O., *3-d topology optimization of modulated and oriented periodic microstructures by the homogenization method*, submitted. HAL preprint: hal-01939201 (November 2018).

[GRS2015] Gibson, I., Rosen, D., and Stucker, B., Additive Manufacturing Technologies, Springer, New York (2015).

[GS2018] Groen, J. P., and Sigmund, O., "Homogenization based topology optimization for high resolution manufacturable microstructures," *International Journal for Numerical Methods in Engineering*, **113**: 1148–1163 (2018).

[HS1963] Hashin, Z., and Shtrikman, S., "A variational approach to the theory of the elastic behavior of multiphase materials," *J. Mech. Phys. Solids*, **11**: 127–140 (1963).

[HM2003] Haslinger, J., and Mäkinen, R., Introduction to Shape Optimization: Theory, Approximation, and Computation, SIAM, Philadelphie (2003).

[He2012] Hecht, F., "New development in freefem++," *J. Numer. Math.*, **20**: 251–265 (2012).

[HP2018] Henrot, A., and Pierre, M., Shape Variation and Optimization: A Geometrical Analysis. English Version of the French Publication with Additions and Updates, EMS Tracts in Mathematics, 28. European Mathematical Society (EMS), Zürich (2018).

[H1996] Hornung, U., Editor, Homogenization and Porous Media, Springer Verlag (1996).

[JKO1995] Jikov, V., Kozlov, S., and Oleinik, O., Homogenization of Differential Operators and Integral Functionals, Springer, Berlin (1995).

[KPTZ2000] Kawohl, B., Pironneau, O., Tartar, L., and Zolésio, J.-P., Optimal shape design. Lectures given at the Joint C.I.M./C.I.M.E. Summer School held in Tróia, June 1–6, 1998. Edited by Cellina, A., and Ornelas, A. Lecture Notes in Mathematics, 1740. Fondazione CIME/CIME Foundation Subseries. Springer-Verlag, Berlin; Centro Internazionale Matematico Estivo (C.I.M.E.), Florence (2000).

[K2016] Katsikadelis, J. T., The Boundary Element Method for Engineers and Scientists: Theory and Applications, Elsevier (2016).

[MI2001] Milton, G., The Theory of Composites, Cambridge University Press (2001).

[Mu1977] Murat, F., "Contre-exemples pour divers problèmes où le contrôle intervient dans les coefficients," *Annali Mat. Pura Appli.*, **112**: 49–68 (1977).

[MT1997] Murat, F., and Tartar, L., *H-Convergence*, in Topics in the Mathematical Modeling of Compositic Materials, Progress

in Nonlinear Differential Equations and their Applieations., ed. Cherkaev, A., and Kohn, R., Birkhäuser, Boston, **31** (1997), 21–43.

[Ng1989] Nguetseng, G., "A general convergence result for a functional related to the theory of homogenization," *SIAM J. Math. Anal.*, **20**: 608–623 (1989).

[NW1999] Nocedal, J., and Wright, S., Numerical Optimization, Springer Science (1999).

[PT2008] Pantz, O., and Trabelsi, K., "A post-treatment of the homogenization method for shape optimization," *SIAM J. Control Optim.*, **47**: 1380–1398 (2008).

[Pe1989] Pedersen, P., "On optimal orientation of orthotropic materials," *Structural Optimization*, **1(2)**: 101–106 (1989).

[ROZ1989] Rozvany, G., Structural Design via Optimality Criteria, Kluwer Academic Publishers, Dordrecht (1989).

[SK1992] Sokolowski, J., and Zolésio, J.-P., Introduction to Shape Optimization. Shape Sensitivity Analysis, Springer Series in Computational Mathematics, 16, Springer, Berlin (1992).

[TA2000] Tartar, L., An introduction to the Homogenization Method in Optimal Design, in Optimal Shape Design (Tróia, 1998), Cellina, A., and Ornelas, A. eds., Lecture Notes in Mathematics 1740, pp. 47–156, Springer, Berlin (2000).