Numer. Math. (2001) 89: 1–29
Digital Object Identifier (DOI) 10.1007/s002110000247

Numerische
Mathematik

A new approach for the optimal distribution of assemblies in a nuclear reactor

Grégoire Allaire[1,2], **Carlos Castro**[1,3]

[1] CEA Saclay, DRN/DMT/SERMA, 91191 Gif-sur-Yvette, France
[2] Laboratoire d'Analyse Numérique, Université Paris 6, 75252 Paris Cedex 5, France
[3] Departamento de Matematica Aplicada, Universidad Complutense, 28040 Madrid, Spain

Received January 13, 2000 / Published online February 5, 2001 – © Springer-Verlag 2001

Summary. The aim of this paper is to propose a new approach for optimizing the position of fuel assemblies in a nuclear reactor core. This is a control problem for the neutronic diffusion equation where the control acts on the coefficients of the equation. The goal is to minimize the power peak (i.e. the neutron flux must be as spatially uniform as possible) and maximize the reactivity (i.e. the efficiency of the reactor measured by the inverse of the first eigenvalue). Although this is truly a discrete optimization problem, our strategy is to embed it in a continuous one which is solved by the homogenization method. Then, the homogenized continuous solution is numerically projected on a discrete admissible distribution of assemblies.

Mathematics Subject Classification (1991): 65K10; 65N99

1. Introduction

This paper is concerned with an optimal design problem in nuclear reactor cores: the so-called optimal fuel re-loading problem. In most reactor cores, the nuclear fuel is made of a few hundreds of so-called assemblies, periodically distributed in a cross-section of the core (see Fig. 1). Each assembly is a very heterogeneous medium composed by a regular array of fuel pins (mainly made of uranium) and control rods immersed in water. During the fission process, the fissile isotope of uranium is consumed and other products appear. This so-called depletion process progressively decreases the efficiency of the nuclear fuel. Therefore, it must be changed periodically by

Correspondence to: G. Allaire

fresh one (such a period, also called a cycle, is about a few months). However, the fuel depletion is not spatially uniform in the core. This has two consequences: first, only part of the old assemblies (typically one fourth) are removed at the end of each cycle, second, it is not desirable to put the new assemblies exactly at the location of the removed ones. In order to maintain the maximal performance of the reactor, it is rather preferable to optimize the position of each type of assemblies. In other words, the fuel re-loading process not only consists in replacing the used assemblies by fresh ones but also in a rearrangement of all the assemblies in the core to make the most efficient use of the nuclear fuel. As such, it is a discrete optimization problem, but the large number of assemblies make it highly non-trivial since the computation of all possible combinations to find the best one is out of reach. For more details on this problem, we refer e.g. to [6,9,12].

In order to give a precise mathematical statement of this optimization problem, we now describe the state equation that models the fission process in the nuclear reactor and allows to quantify the efficiency of the assemblies distribution. The power distribution in a nuclear reactor core is usually obtained by solving an eigenvalue problem for a diffusion equation. For simplicity, in this paper we content ourselves with the one energy group diffusion equation (multiple energy groups diffusion will be considered in a next paper [2]). In a steady-state regime, this problem gives the balance between neutrons produced by fission and neutrons absorbed or diffused by the medium. Denoting by Ω the radial section of the core ($\Omega \subset \mathbb{R}^2$ is a bounded open set), our state equation is

$$\left\{ \begin{array}{ll} -\operatorname{div}\left(D(x)\nabla u(x)\right) + \Sigma(x)u(x) = \lambda\sigma(x)u(x), & x \in \Omega, \\ u(x) = 0, & x \in \partial\Omega, \end{array} \right. \tag{1}$$

where the unknowns are the neutronic flux u (i.e. the density of neutrons) and the eigenvalue $\lambda = 1/k_{eff}$ (k_{eff} is the criticality parameter which gives the ratio between produced and consumed neutrons). More precisely, λ is the first eigenvalue and u the first eigenvector of (1), which is the only one to have a physical meaning since it is positive. The diffusion coefficient $D(x)$, the absorption cross section $\Sigma(x)$, and the fission cross section $\sigma(x)$, are positive data determined by the type of assemblies. The eigenvalue λ measures the criticity of the reactor in a quasistatic limit. If $\lambda = 1$, the reactor is said to be critical and can safely be operated: a perfect balance between production and removal of neutrons is obtained. If $\lambda > 1$, too many neutrons are diffused or absorbed in the core compared to their production by fission : the nuclear chain reaction dies out, and the reactor, being sub-critical, can not operate. If $\lambda < 1$, too many neutrons are created by fission, and the reactor, being super-critical, can nevertheless be operated by introducing absorbing media in the core (with control rods, or diluted in the water). Remark that (1) gives only the spatial distribution of the neutron flux (which in turn yields the power distribution) but not its intensity since an eigenvector is defined up to a multiplicative constant.

We can now describe the objective function of the fuel re-loading optimization problem. As already said, a reactor can produce energy if its criticality eigenvalue λ is equal to or smaller than 1. However, as time goes by, the fuel depletion has a tendency to increase this eigenvalue. Therefore, at the beginning of a cycle it is highly desirable to have the smallest possible value of λ (or criticality reserve), ensuring that the reactor will be working for the longest possible time. Minimizing the eigenvalue λ may cause unusual oscillations in the profile of the first eigenvector u (the neutron flux), and produce a highly non-uniform power distribution in the core (which is proportional to σu). For efficiency and safety reasons, it is rather an undesirable feature. Indeed, at peak points of the power distribution, the surrounding flow of water could be unable to cool down the fuel pins, yielding a strong increase of the temperature that may eventually cause damage in the assembly. A major issue for safety is thus to have the most uniform power distribution in the core. This can be enforced by minimizing the maximal value of σu (the so-called peak power point). Such a criterion is non differentiable, and we approximate it by minimizing instead the $L^r(\Omega)$ norm of σu with $1 < r < +\infty$. Since u is defined up to a multiplicative constant, we take care of normalizing this $L^r(\Omega)$ norm by the $L^1(\Omega)$ norm. Finally, introducing a positive Lagrange multiplier $\ell \geq 0$, our objective function is

$$\min \left\{ \ell\lambda + \frac{(\mathcal{M}(|\sigma u|^r))^{1/r}}{\mathcal{M}(\sigma u)} \right\}, \tag{2}$$

where $\mathcal{M}$ denotes the average operator in Ω

$$\mathcal{M}(f) = \frac{1}{|\Omega|} \int_\Omega f(x)dx. \tag{3}$$

For simplicity, we outrageously simplified the constraints and requirements used in practice for fuel re-loading optimization. In particular, we optimize the assemblies distribution just for one cycle, regardless of what may happen afterwards, and we do not take into account the cost of permuting assemblies. We also do not try to minimize the production of undesirable isotopes or species in the fission process. For more informations on the actual constraints and objectives, we refer e.g. to [12].

To finish the mathematical statement of our optimization problem, it remains to define a space of admissible configurations $\mathcal{U}_{ad}$ of assemblies in the core. Then, the minimization of the objective function (2) take place on this space $\mathcal{U}_{ad}$. Assuming that at each cycle, one out of I assemblies is removed, there are I type of assemblies in the core, having different

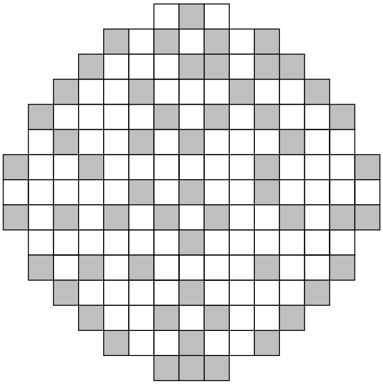

Fig. 1. A discrete configuration of two types of assemblies in a 900 Mw PWR nuclear reactor core (having 157 assemblies)

physical characteristics (D, Σ, σ) due to their different past time in the core (their so-called burnup). Typical values of I that we shall deal with in this paper are $I = 2$ or 4 (the case $I = 2$ is much simpler but not realistic, while $I = 4$ is typical and not much easier than any $I \geq 3$). For simplicity, we assume that all assemblies of the same type are identical, and that the coefficients (D, Σ, σ) are constant inside one assembly (i.e. it is homogeneous). Of course, the proportions of each type of assemblies are given. We make no special assumptions on the ordering of the physical properties of the assemblies, although physically speaking the freshest fuel produce the smallest criticity eigenvalue λ. Finally, since all assemblies have the same size, the core Ω contains a finite number of them (see Fig. 1). Thus, $\mathcal{U}_{ad}$ is a finite set of all possible permutations of these assemblies.

Since the space of admissible configurations $\mathcal{U}_{ad}$ has a finite number of elements, the minimization of the objective function (2) is a combinatorial optimization problem. There are many numerical methods proposed in the literature for solving it, based on linear programming, simulated annealing, neural networks or genetic algorithms [8,11,16,19,14,15]. However, the huge number of possible permutations, the non-convexity of the objective function make it a very hard problem to solve. We propose yet another approach in two steps. First, we transform this discrete problem in a continuous one by removing any size and shape constraints on the assemblies (see Fig. 2). In other words, we keep the prescribed amount of fuel (or material) in each of their I types, but it can now be placed in the core as freely as we want, and its repartition does not necessarily follow an assembly pattern. This idea of generalizing the fuel re-loading optimization problem as a continuous one is not new (see e.g. [6]). It has the advantage of being more tractable from a numerical standpoint. In a second step, we project a continuous optimal configuration onto the discrete space $\mathcal{U}_{ad}$, in the hope

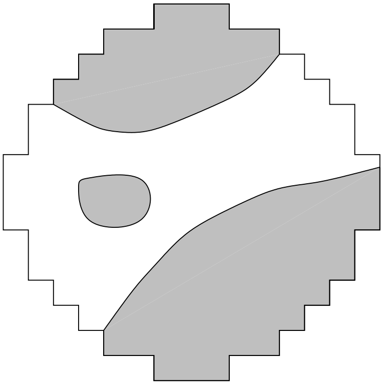

Fig. 2. A continuous configuration of two types of assemblies in a 900 Mw PWR nuclear reactor core

that it will lead to a nearly optimal admissible configuration of assemblies. Transforming an admissible configuration into a continuous one is obtained through a numerical method of penalization. This second step is therefore purely based on numerical heuristics and has no firm theoretical ground. On the contrary, we perform a detailed mathematical analysis of the first step. It turns out that the continuous optimization problem is ill-posed in the sense that it does not admit a solution in the space of all possible continuous distributions of the I materials. The reason for this is that minimizing sequences of almost optimal configurations have a tendency to exhibit very fine mixture of the I components. On a macroscopic scale these mixtures are composite materials having effective properties different from that of its phase constituents. Their effective or averaged cross sections and diffusion tensors are found by using the homogenization theory. To make this problem well-posed, one must enlarge the space of admissible designs by allowing for composite materials obtained by mixing microscopically the I different fuels. We then obtain the existence of a composite optimal configuration, as well as very efficient numerical algorithm for computing them.

This approach is called the homogenization method for optimal design. It has been successfully implemented in structural optimization (see e.g. [1,3,4]). Our work must be seen as a generalization of this method to the fuel re-loading optimization problem. In structural design the homogenization method is regarded as a method for topology optimization, which is not incompatible, but rather complementary, with other classical methods. Likewise in the present setting, our approach should be taken as a topology optimizer, i.e., whatever the starting configuration, it is able to find a quasi-optimal distribution of assemblies, possibly very remote from the starting one. The homogenization method is not a concurrent of other methods, but rather a pre-processor, since its final output could still be refined by these methods.

Finally, we conclude this introduction by a brief description of the content of this paper. Although our goal is to address the case of $I > 2$ types of assemblies, for simplicity our exposition starts with the easier case $I = 2$. In Sect. 2, a mathematical setting is introduced for the original continuous problem. Section 3 is devoted to its relaxation, and Sect. 4 deals with optimality conditions. Eventually Sect. 5 generalizes the previous results for more than two type of assemblies. Numerical results are presented in Sect. 6.

2. Setting of the problem

We first recall that the state equation of our optimization problem is the spectral equation for the one energy group diffusion model. Denoting by Ω a bounded open set in $\mathbb{R}^2$, it reads

$$\begin{cases} -\operatorname{div}(D(x)\nabla u(x)) + \Sigma(x)u(x) = \lambda\sigma(x)u(x), & x \in \Omega, \\ u(x) = 0, & x \in \partial\Omega, \end{cases} \tag{4}$$

where (λ, u) is the first eigenvalue and eigenvector. The coefficients in (4) are assumed to be measurable and bounded (but not necessarily smooth), i.e. $D, \Sigma, \sigma \in L^\infty(\Omega)$. Furthermore, we assume that, almost everywhere in Ω, they satisfy

$$\Sigma(x) \geq 0,\ \sigma(x) \geq \sigma_0 > 0,\ D(x) \geq d_0 > 0.$$

Under these assumptions, it is well known that (4) admits a unique solution in the sense given by the following result (see e.g. [20]).

Theorem 2.1 *There exists a countable infinite number of eigenvalues for (4), that we label by increasing order as $(\lambda_i)_{i\geq 1}$, and corresponding eigenvectors $(u_i)_{i\geq 1} \in H_0^1(\Omega)$. Furthermore, the first eigenvalue λ_1 (i.e. the smallest one) is positive and simple, and its eigenvector is the only one that can be chosen to be positive in Ω.*

Remark 2.2 The only solution of (4) which has a physical meaning is the first eigenvalue and eigenvector (λ_1, u_1) since u_1 is the only eigenvector to be positive (a necessary feature for a density function). From now on, we drop the subscript 1, and denote by $(\lambda, u) = (\lambda_1, u_1)$ the solution of (4). Of course, u is unique only up to a multiplicative constant.

In this section we assume that there are only two types of assemblies, characterized by constant positive coefficients $(d^j, \Sigma^j, \sigma^j)$ with $j = 1, 2$, given in prescribed proportions $\gamma_j \geq 0$ with $\gamma_1 + \gamma_2 = |\Omega|$. We consider only the continuous optimization problem (as defined in the introduction), i.e. we do not require that each type of material $j = 1, 2$ fit into assemblies, but it may rather fill the domain Ω in any possible shape. In other words, denoting by Ω_j the part of Ω occupied by material j, there is no restrictions on (Ω_1, Ω_2) except the obvious ones

$$\Omega_1 \cap \Omega_2 = 0,\ \Omega_1 \cup \Omega_2 = \Omega,\ |\Omega_j| = \gamma_j, j = 1, 2. \tag{5}$$

Introducing the characteristic functions (χ_1, χ_2) of these subsets (Ω_1, Ω_2) (i.e. $\chi_j(x) = 1$ if $x \in \Omega_j$ and $\chi_j(x) = 0$ if $x \not\in \Omega_j$), the coefficients of (4) are given by

$$\begin{cases} D(x) = d^1\chi_1(x) + d^2\chi_2(x), \\ \Sigma(x) = \Sigma^1\chi_1(x) + \Sigma^2\chi_2(x), \\ \sigma(x) = \sigma^1\chi_1(x) + \sigma^2\chi_2(x). \end{cases} \tag{6}$$

Since $\chi_1 + \chi_2 = 1$, a single characteristic function $\chi_1 = \chi$ defines completely the distribution of the two fuel types. Therefore, the space of admissible configurations can now be defined in a very simple way by

$$\mathcal{U}_{ad} = \left\{ \chi \in L^\infty(\Omega; \{0, 1\}) \text{ such that } \int_\Omega \chi(x)dx = \gamma_1 \right\}. \tag{7}$$

Our fuel re-load optimization problem is to find a minimizer of

$$\min_{\chi \in \mathcal{U}_{ad}} J(\chi) = \left(\ell\lambda + \frac{(\mathcal{M}(|\sigma u|^r))^{1/r}}{\mathcal{M}(\sigma u)} \right), \tag{8}$$

where (λ, u) is the solution of (4), $\mathcal{M}$ is the average operator in Ω defined by (3), $1 < r < +\infty$, and the coefficients of (4) are given by (6).

To solve this optimization problem we can try the direct method of the calculus of variations. It amounts to proceed in the following order

1. We prove that minimizing sequences are relatively compact for a suitable topology.
2. We prove a lower semicontinuous result for the objective function, which yields that it attains its minimum.
3. We differentiate the cost function to obtain optimality conditions.

As remarked by [18], two main problems arise when we try to carry out this process for (8). First, the set $\mathcal{U}_{ad}$ is not closed in the topologies for which minimizing sequences are compact. This means that, in general, minimizing sequences can converge to limits outside from $\mathcal{U}_{ad}$, i.e. they are not characteristic functions. In this case there is no minimizer of (8) in $\mathcal{U}_{ad}$ (explicit counter-examples may be found in [17]). Second, we can not differentiate the cost function (8) because the set $\mathcal{U}_{ad}$ is not stable by standard variations, i.e. a convex combination of two characteristic functions is never a characteristic function.

To overcome these two obstacles, one can use the so-called relaxation procedure (see e.g. [5,7]). It amounts to extend the original space of admissible solutions into a space of generalized, or relaxed, admissible solutions, denoted by $\mathcal{U}_{ad}^*$, as well as the objective function J that becomes a relaxed objective function J^*. This extension is built to guarantee the existence of an optimal relaxed solution, but it should not be too ”large” in order to keep track of the behavior of minimizing sequences for the original problem. In other words, relaxing a problem does not change its physical significance. More precisely, a relaxed formulation must satisfy the following conditions

1. $\mathcal{U}_{ad} \subset \mathcal{U}_{ad}^*$ and the relaxed cost function coincides with the original one over $\mathcal{U}_{ad}$,
2. there exists at least one minimizer of the relaxed problem, and the minimal values of the original and relaxed objective functions are equal,
3. any minimizer of the relaxed problem is attained by a minimizing sequence of the original problem,
4. any minimizing sequence of the original problem converges to a minimizer of the relaxed problem.

In the next section we introduce such a relaxation for our problem using homogenization.

3. The relaxed problem

In this section we introduce a relaxed problem associated to (8). We follow the homogenization method introduced in [18].

The set of characteristic functions is bounded in $L^\infty(\Omega)$ and therefore relatively compact for the weak * convergence. Thus, from any sequence $(\chi_n)_{n\geq 1}$ in $\mathcal{U}_{ad}$, we can extract a subsequence, still denoted χ_n, such that it converges weakly * in $L^\infty(\Omega)$ to a limit $\theta(x)$. Since the convergence is weak and not strong, θ is usually not anymore a characteristic function but a density, i.e. $\theta(x)$ may take its values in the full range $[0,1]$. Defining the corresponding coefficients

$$\begin{aligned} D_n(x) &= \chi_n(x)d^1 + (1-\chi_n(x))d^2, \\ \Sigma_n(x) &= \chi_n(x)\Sigma^1 + (1-\chi_n(x))\Sigma^2, \\ \sigma_n(x) &= \chi_n(x)\sigma^1 + (1-\chi_n(x))\sigma^2, \end{aligned}$$

the state equation is rewritten

$$\left\{ \begin{array}{ll} -\operatorname{div}(D_n \nabla u_n) + \Sigma_n u_n = \lambda_n \sigma_n u_n, & \text{in } \Omega, \\ u_n = 0, & \text{on } \partial\Omega, \end{array} \right. \tag{9}$$

where (λ_n, u_n) are the first eigenvalue and normalized eigenvector. In order to pass to the limit in (9), we use the theory of H-convergence (also called

G-convergence, see e.g. [10,18]), which states that, up to a subsequence, the limit of (9) is the following homogenized problem

$$\begin{cases} -\operatorname{div}(D^*\nabla u(x)) + \overline{\Sigma} u(x) = \lambda\overline{\sigma} u(x), & x \in \Omega, \\ u(x) = 0, & x \in \partial\Omega, \end{cases} \tag{10}$$

where (λ, u) are the first eigenvalue and normalized eigenvector, and the homogenized coefficients are defined by

$$\overline{\Sigma}(x) = \theta(x)\Sigma^1 + (1-\theta(x))\Sigma^2, \;\; \overline{\sigma}(x) = \theta(x)\sigma^1 + (1-\theta(x))\sigma^2,$$

and D^* is the H-limit (i.e. the limit in the sense of homogenization) of the sequence $D_n = \chi_n d^1 + (1-\chi_n)d^2$. This H-convergence has to be understood in the following sense

$$\begin{aligned} &\lim_{n\to\infty} \lambda_n = \lambda, \\ &u_n \rightharpoonup u \text{ in } H_0^1(\Omega) \text{ weakly.} \end{aligned} \tag{11}$$

It turns out that, although the homogenized cross sections $\overline{\Sigma}, \overline{\sigma}$ are uniquely defined by the limit density θ, the homogenized diffusion coefficient D^* is not explicitly characterized by θ. Indeed, depending on the geometry of the mixture represented by the sequence χ_n, D^* may be any symmetric positive definite matrix in a set G_θ. This set of all possible homogenized diffusion tensors associated to the density θ has been characterized in [13,18]. We assume, with no loss of generality, that $0 < d^1 \leq d^2$. At any point $x \in \Omega$, $D^*(x)$ is any symmetric matrix with eigenvalues $(\mu_1(x), \mu_2(x))$ satisfying (see Fig. 3)

$$\frac{1}{\mu_1 - d^1} + \frac{1}{\mu_2 - d^1} \leq \frac{1}{\mu_\theta^- - d^1} + \frac{1}{\mu_\theta^+ - d^1}, \tag{12}$$

$$\frac{1}{d^2 - \mu_1} + \frac{1}{d^2 - \mu_2} \leq \frac{1}{d^2 - \mu_\theta^-} + \frac{1}{d^2 - \mu_\theta^+}. \tag{13}$$

where μ_θ^+ and μ_θ^- are the arithmetic and harmonic means of the phase diffusion coefficients

$$\begin{aligned} \mu_\theta^+ &= \theta d^1 + (1-\theta)d^2, \\ \mu_\theta^- &= \frac{1}{\theta/d^1 + (1-\theta)/d^2}. \end{aligned} \tag{14}$$

One can easily check that (12) implies that $\mu_\theta^- \leq \mu_i \leq \mu_\theta^+$ for $i = 1, 2$.

Since the homogenized state equation (10) depends on two design parameters, namely θ and D^*, the set of generalized admissible configurations $\mathcal{U}_{ad}^*$ must be the set of such couples, namely

$$\mathcal{U}_{ad}^* = \left\{ (\theta, D^*) \in L^\infty(\Omega) \text{ such that } 0 \leq \theta \leq 1, D^* \in G_\theta, \int_\Omega \theta = \gamma_1 \right\}, \tag{15}$$

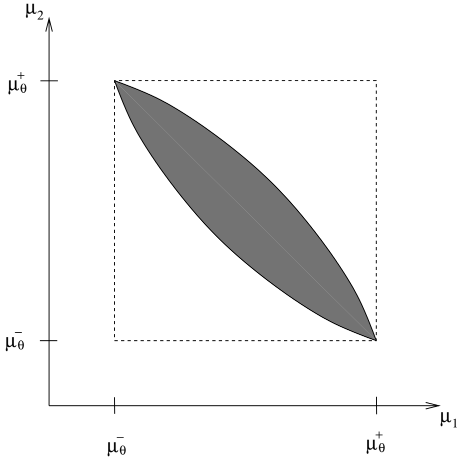

Fig. 3. Set G_θ of all homogenized diffusion tensors

where the constraints on (θ, D^*) are pointwise in Ω. Remark that we have $\mathcal{U}_{ad} \subset \mathcal{U}_{ad}^*$ if we associate to each characteristic function $\chi \in \mathcal{U}_{ad}$ a diffusion tensor $D = d^1\chi + d^2(1-\chi)$. By Rellich theorem, the sequence u_n, which converges weakly to u in $H_0^1(\Omega)$, converges strongly to u in $L^r(\Omega)$ for any $1 \le r < +\infty$ in two space dimensions (and for any $1 \le r < 6$ in three dimensions). Since χ_n converges weakly * to θ in $L^\infty(\Omega)$, we deduce that $(\sigma_n)^r$ converges weakly * to $\theta(\sigma^1)^r + (1-\theta)(\sigma^2)^r$ in $L^\infty(\Omega)$, and therefore

$$\lim_{n\to\infty} \int_\Omega \sigma_n(x) u_n(x) dx = \int_\Omega \overline{\sigma}(x) u(x) dx,$$

while

$$\lim_{n\to\infty} \int_\Omega |\sigma(x) u_n(x)|^r \, dx = \int_\Omega |s(x) u(x)|^r \, dx,$$

with s being usually different from $\overline{\sigma}$

$$s(x) = \left(\theta(x)(\sigma^1)^r + (1-\theta(x))(\sigma^2)^r\right)^{1/r}.$$

Thus, combined with (11) we obtain

$$\lim_{n\to\infty} J(\chi_n) = J^*(\theta, D^*),$$

and J^* is a relaxed objective function defined by

$$J^*(\theta, D^*) = \left(\ell\lambda + \frac{(\mathcal{M}(|su|^r))^{1/r}}{\mathcal{M}(\overline{\sigma}u)}\right), \tag{16}$$

where (λ, u) is the first eigenvalue and eigenvector of (10) (remark that Theorem 2.1 also applies to (10)). Our relaxed problem is finally to minimize J^* over $\mathcal{U}_{ad}^*$, i.e.

$$\min_{(\theta,D^*)\in\mathcal{U}^*_{ad}} J^*(\theta,D^*). \tag{17}$$

We can now state the main result of relaxation.

Theorem 3.1 *Assume that* $1 \le r < +\infty$ *in two space dimensions, and* $1 < r < 6$ *in three dimensions. The relaxation of the original optimization problem (8) is (16) in the sense that*

1. *there exists at least one minimizer in* $\mathcal{U}^*_{ad}$ *of* J^**,*
2. *any minimizer* (θ, D^*) *of the relaxed problem is attained by a minimizing sequence* χ_n *of the original problem in the sense that*

$$\begin{cases} \chi_n \rightharpoonup \theta \text{ weakly * in } L^\infty(\Omega), \\ D_n = \chi_n d^1 + (1-\chi_n)d^2 \ H\text{-converges to } D^*, \end{cases}$$

and

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

3. *any minimizing sequence* χ_n *of the original problem converges to a minimizer* (θ, D^*) *of the relaxed problem.*

Proof. This proof is an adaptation of that in [18]. It is a simple consequence of the convergence result (11). Indeed, if χ_n is a minimizing sequence for J, (11) implies that χ_n converges to θ_∞ and $J(\chi_n)$ converges, up to a subsequence, to $J^*(\theta_\infty, D^*_\infty)$ which is therefore equal to $\min_{\chi\in\mathcal{U}_{ad}} J(\chi)$. Since any (θ, D^*) is attained by a sequence χ_n (not necessarily minimizing), we deduce that $(\theta_\infty, D^*_\infty)$ is a minimizer of J^*. This finishes the proof of Theorem 3.1.

4. Optimality conditions

One advantage of the relaxed formulation is that it allows to derive optimality conditions that are of both theoretical and numerical interest. The results of this section are a variation of those in [18]. The relaxed cost function J^* is defined by

$$J^*(\theta,D^*) = \ell\lambda + \frac{(\mathcal{M}(|su|^r))^{1/r}}{\mathcal{M}(\overline{\sigma}u)}. \tag{18}$$

with $\mathcal{M}(f) = |\Omega|^{-1}\int_\Omega f(x)dx$. By theorem 2.1 the first eigenvalue λ and the first normalized eigenvector u are simple : therefore, they are Gâteaux-differentiable, as well as J^*, with respect to (θ, D^*) in the admissible set

$$\mathcal{U}^*_{ad} = \left\{ (\theta,D^*) \in L^\infty(\Omega) \text{ such that } 0 \le \theta \le 1,\ \int_\Omega \theta = \gamma_1,\ D^* \in G_\theta \right\}. \tag{19}$$

If $(\delta\theta, \delta D^*)$ is an admissible increment in $\mathcal{U}^*_{ad}$, the derivative of J^* is

$$\begin{aligned} \delta J^* = \ell\delta\lambda + \frac{(\mathcal{M}(|su|^r))^{(1-r)/r}}{\mathcal{M}(\overline{\sigma}u)}\mathcal{M}\left((su)^{r-1}\left(s\delta u + u\delta s\right)\right) \\ -\frac{(\mathcal{M}(|su|^r))^{1/r}}{(\mathcal{M}(\overline{\sigma}u))^2}\mathcal{M}\left(\overline{\sigma}\delta u + u\delta\overline{\sigma}\right), \end{aligned} \tag{20}$$

where $rs^{r-1}\delta s = ((\sigma^1)^r - (\sigma^2)^r)\delta\theta$, $\delta\overline{\sigma} = (\sigma^1 - \sigma^2)\delta\theta$, $\delta\lambda$ is the increment in the first eigenvalue, and δu is the increment in the first eigenvector solution of (10). To compute $\delta\lambda$, we first remark that multiplying equation (10) by a test function $v \in H^1_0(\Omega)$ yields

$$\lambda = \frac{\int_\Omega \left(D^*\nabla u \cdot \nabla v + \overline{\Sigma}uv\right)dx}{\int_\Omega \overline{\sigma}uvdx}. \tag{21}$$

Thus, differentiating (21) gives after some easy algebra

$$\delta\lambda = \frac{\int_\Omega \delta D^*\nabla u \cdot \nabla u dx + \int_\Omega |u|^2\delta\overline{\Sigma}dx}{\int_\Omega \overline{\sigma}|u|^2dx} - \lambda\frac{\int_\Omega |u|^2\delta\overline{\sigma}dx}{\int_\Omega \overline{\sigma}|u|^2dx}. \tag{22}$$

On the other hand, differentiating (10) shows that δu is the unique solution in $H^1_0(\Omega)$ of

$$\begin{cases} -\operatorname{div}\left(D^*\nabla(\delta u)\right) + \overline{\Sigma}\delta u - \lambda\overline{\sigma}\delta u = \operatorname{div}\left(\delta D^*\nabla u\right) - \delta\overline{\Sigma}u \\ \qquad\qquad\qquad\qquad\qquad\qquad + (\lambda\delta\overline{\sigma} + \overline{\sigma}\delta\lambda)u \quad \text{in } \Omega, \\ \delta u = 0 \quad \text{on } \partial\Omega. \end{cases} \tag{23}$$

Remark that the right hand side of (23) is orthogonal to the first eigenvector u which implies that it admits a solution, unique up to the addition of a multiple of u.

As usual, to eliminate δu an adjoint state q is introduced. It is defined as the unique solution in $H^1_0(\Omega)$ of

$$\begin{cases} -\operatorname{div}\left(D^*\nabla q\right) + \overline{\Sigma}q - \lambda\overline{\sigma}q \\ \qquad = \frac{(\mathcal{M}(|su|^r))^{(1-r)/r}}{\mathcal{M}(\overline{\sigma}u)}\frac{s^r u^{r-1}}{|\Omega|} - \frac{(\mathcal{M}(|su|^r))^{1/r}}{(\mathcal{M}(\overline{\sigma}u))^2}\frac{\overline{\sigma}}{|\Omega|} \quad \text{in } \Omega, \\ q = 0 \quad \text{on } \partial\Omega. \end{cases} \tag{24}$$

Remark that the right hand side of (24) is orthogonal to u which implies that it admits a solution, unique up to a multiple of u. Then, multiplying equation (24) by δu and equation (23) by q leads to

$$\begin{aligned} \frac{(\mathcal{M}(|su|^r))^{(1-r)/r}}{\mathcal{M}(\overline{\sigma}u)}\frac{1}{|\Omega|}\int_\Omega s^r u^{r-1}\delta u dx - \frac{(\mathcal{M}(|su|^r))^{1/r}}{(\mathcal{M}(\overline{\sigma}u))^2}\frac{1}{|\Omega|}\int_\Omega \overline{\sigma}\delta u dx \\ = -\int_\Omega \delta D^*\nabla u \cdot \nabla q dx - \int_\Omega \left(\delta\overline{\Sigma} - \lambda\delta\overline{\sigma} - \overline{\sigma}\delta\lambda\right)uq dx. \end{aligned} \tag{25}$$

Thus

$$\delta J^*(\theta, D^*) = \ell\delta\lambda + \frac{(\mathcal{M}(|su|^r))^{(1-r)/r}}{\mathcal{M}(\overline{\sigma}u)}\frac{1}{|\Omega|}\int_\Omega u^r s^{r-1}\delta s dx$$
$$-\frac{(\mathcal{M}(|su|^r))^{1/r}}{(\mathcal{M}(\overline{\sigma}u))^2}\frac{1}{|\Omega|}\int_\Omega u\delta\overline{\sigma}dx$$
$$-\int_\Omega \delta D^*\nabla u\cdot\nabla q dx - \int_\Omega \left(\delta\overline{\Sigma} - \lambda\delta\overline{\sigma} - \overline{\sigma}\delta\lambda\right) uqdx.$$

Introducing a combination function $p \in H_0^1(\Omega)$ defined by

$$p = \frac{\ell + \int_\Omega \overline{\sigma}uqdx}{\int_\Omega \overline{\sigma}|u|^2 dx}u - q, \tag{26}$$

the derivative of J^* becomes

$$\delta J^*(\theta, D^*) = \int_\Omega \delta D^*\nabla u\cdot\nabla p dx + \int_\Omega \left(\delta\overline{\Sigma} - \lambda\delta\overline{\sigma}\right) upd x$$
$$+\frac{(\mathcal{M}(|su|^r))^{(1-r)/r}}{\mathcal{M}(\overline{\sigma}u)}\frac{1}{|\Omega|}\int_\Omega u^r s^{r-1}\delta s dx \tag{27}$$
$$-\frac{(\mathcal{M}(|su|^r))^{1/r}}{(\mathcal{M}(\overline{\sigma}u))^2}\frac{1}{|\Omega|}\int_\Omega u\delta\overline{\sigma}dx,$$

where $\delta\overline{\sigma} = (\sigma^1 - \sigma^2)\delta\theta$, $\delta\overline{\Sigma} = (\Sigma^1 - \Sigma^2)\delta\theta$, and $rs^{r-1}\delta s = ((\sigma^1)^r - (\sigma^2)^r)\delta\theta$.

Lemma 4.1 *A necessary condition for (θ, D^*) to be a minimizer of J^* in $\mathcal{U}_{ad}^*$ is*

$$\delta J^*(\theta, D^*) \geq 0 \tag{28}$$

for any admissible increment $(\delta\theta, \delta D^*)$.

According to the structure of$\mathcal{U}_{ad}^*$, the minimization process can be carried out in two steps: firstly in D^* and secondly in θ. In other words,

$$\min_{(\theta,D^*)\in\mathcal{U}_{ad}^*} J^*(\theta, D^*) = \min_{0\leq\theta\leq 1}\min_{D^*\in G_\theta} J^*(\theta, D^*). \tag{29}$$

Minimizing first in D^*, i.e. taking $\delta\theta = 0$ in Lemma 4.1 yields

Proposition 4.2 *When $\delta\theta = 0$, the optimality conditions becomes*

$$\delta J^*(\theta, D^*) = \int_\Omega \delta D^* \nabla u \cdot \nabla p dx \geq 0. \tag{30}$$

It implies that, outside the set where $|\nabla u||\nabla p| = 0$, an optimal diffusion tensor D^* satisfies

$$\left\{ \begin{array}{l} D^*\nabla u = \frac{1}{2}(\mu_\theta^+ + \mu_\theta^-)\nabla u - \frac{1}{2}(\mu_\theta^+ - \mu_\theta^-)\frac{|\nabla u|}{|\nabla p|}\nabla p, \\ D^*\nabla p = -\frac{1}{2}(\mu_\theta^+ - \mu_\theta^-)\frac{|\nabla p|}{|\nabla u|}\nabla u + \frac{1}{2}(\mu_\theta^+ + \mu_\theta^-)\nabla p. \end{array} \right. \tag{31}$$

Besides, defining an angle φ by $\nabla u \cdot \nabla p = |\nabla u||\nabla p| \cos\varphi$, we obtain

$$D^*\nabla u \cdot \nabla p = |\nabla u||\nabla p| \left[\mu_\theta^- \cos^2\frac{\varphi}{2} - \mu_\theta^+ \sin^2\frac{\varphi}{2} \right]. \tag{32}$$

Remark 4.3 As a byproduct of Proposition 4.2, it turns out that, at the points where $\nabla u \neq 0$ and $\nabla p \neq 0$, an optimal diffusion tensor D^* can always be found in the class of so-called simple laminates or layered materials (see the proof below). A careful investigation of the remaining points $|\nabla u||\nabla p| = 0$ shows that, everywhere, an optimal D^* can be found in the class of layered materials (this remarks is due to [18,21]). A layered material is obtained by stacking slices of the two components 1 and 2 and computing its effective or homogenized properties (see Fig. 4). It turns out that there is an explicit formula for its homogenized tensor D^* which depends on the volume fraction θ of phase 1 and on the unit direction e which is normal to the slices or layers. In 2-D, in the basis $(e, \mathrm{e}^\perp)$ it reads

$$D^* \; = \; \text{diag}\left(\mu_\theta^-, \mu_\theta^+\right), \tag{33}$$

i.e. the diffusion is the harmonic average in the normal direction of the layers while it is the arithmetic average in the direction parallel to the layers.

Taking into account the optimal diffusion tensor D^* furnished by Proposition 4.2, we know vary the volume fraction θ to obtain

Proposition 4.4 *Defining a function $Q(x)$ by*

$$\begin{aligned} Q(x) = {} & |\nabla u||\nabla p| \left[\frac{d\mu_\theta^-}{d\theta} \cos^2\frac{\varphi}{2} - \frac{d\mu_\theta^+}{d\theta} \sin^2\frac{\varphi}{2} \right] \\ & + \left[(\Sigma^1 - \Sigma^2) - \lambda(\sigma^1 - \sigma^2) \right] up \\ & + \frac{(\mathcal{M}(|su|^r))^{(1-r)/r}}{\mathcal{M}(\overline{\sigma}u)} \frac{\left[(\sigma^1)^r - (\sigma^2)^r \right] u^r}{r|\Omega|} \\ & - \frac{(\mathcal{M}(|su|^r))^{1/r}}{(\mathcal{M}(\overline{\sigma}u))^2} \frac{(\sigma^1 - \sigma^2)u}{|\Omega|}, \end{aligned} \tag{34}$$

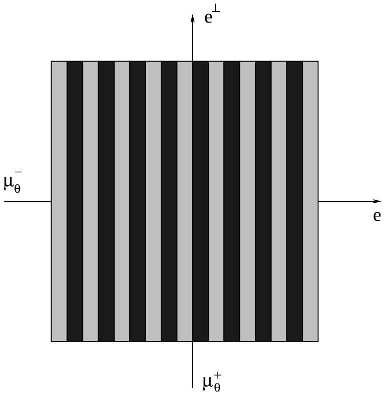

Fig. 4. A two-phase layered material

there exists a constant C_0 such that a minimizer (θ, D^*) for J^* satisfies

$$(35)\qquad \begin{cases} \theta(x) = 0 \textit{ if } Q(x) < C_0, \\ 0 \leq \theta(x) \leq 1 \textit{ if } Q(x) = C_0, \\ \theta(x) = 1 \textit{ if } Q(x) > C_0, \end{cases}$$

and reciprocally

$$(36)\qquad \begin{cases} Q(x) \leq C_0 \textit{ if } \theta(x) = 0, \\ Q(x) = C_0 \textit{ if } 0 < \theta(x) < 1, \\ Q(x) \geq C_0 \textit{ if } \theta(x) = 1. \end{cases}$$

Proof of Proposition 4.2. Taking $\delta\theta = 0$ implies that $\delta\overline{\sigma} = \delta\overline{\Sigma} = \delta s = 0$, which in turn yields (30). It also implies that the variation δD^* stays inside the set G_θ. It turns out that G_θ is a convex set of symmetric matrices since it is defined as a convex set of eigenvalues (see [18]). Therefore, δD^* can be parallel to any straight line in G_θ passing through D^*. In other words, for any $C^* \in G_\theta$, we can choose

$$\delta D^* = C^* - D^*.$$

Thus, (30) implies

$$(37)\quad \int_\Omega C^*(x)\nabla u \cdot \nabla p dx \geq \int_\Omega D^*(x)\nabla u \cdot \nabla p dx, \quad \forall C^* \in G_\theta.$$

Equation (37) implies that the optimal $D^*(x)$ is at each point $x \in \Omega$ a minimizer of $C^*(x)\nabla u(x) \cdot \nabla p(x)$ over all matrices $C^*(x) \in G_{\theta(x)}$. This minimization problem has been solved in [18]. If ∇u or ∇p is equal to 0, any matrix is a minimizer. If $\nabla u \neq 0$ and $\nabla p \neq 0$, then, upon defining two unit vectors $e = \nabla u/|\nabla u|$ and $e' = \nabla p/|\nabla p|$, any minimizer D^* satisfies

$$(38)\qquad D^*e = \frac{1}{2}(\mu_\theta^+ + \mu_\theta^-)e - \frac{1}{2}(\mu_\theta^+ - \mu_\theta^-)\mathrm{e}'$$

and

$$D^*\mathrm{e}' = \frac{1}{2}(\mu_\theta^+ + \mu_\theta^-)\mathrm{e}' - \frac{1}{2}(\mu_\theta^+ - \mu_\theta^-)e. \tag{39}$$

Furthermore, a minimizer is necessarily a so-called rank-one laminate in the direction of $(e+e')$ if $e+e' \neq 0$, or in a direction orthogonal to e if $e+e'=0$. This leads to the desired result.

Proof of Proposition 4.4. We now take an optimal D^* defined by Proposition 4.2. This implies that δD^* is a function of $\delta\theta$. In particular, we have

$$\delta D^* \nabla u \cdot \nabla p = |\nabla u||\nabla p| \left[\frac{d\mu_\theta^-}{d\theta} \cos^2 \frac{\varphi}{2} - \frac{d\mu_\theta^+}{d\theta} \sin^2 \frac{\varphi}{2} \right] \delta\theta, \tag{40}$$

where φ is defined by (32). Introducing the function $Q(x)$, we obtain

$$\delta J^*(\theta, D^*) = \int_\Omega Q(x) \delta\theta(x) dx \geq 0,$$

which, upon taking into account the volume constraint $\int_\Omega \delta\theta(x)dx = 0$, yields the desired result (the constant C_0 is the corresponding Lagrange multiplier).

Remark 4.5 The above optimality conditions are at the root of a numerical algorithm which is described in Sect. 6.

5. Generalization to more types of assemblies

In this section, we keep the same model of one energy group diffusion equation (4) and of objective function (8), but we change the definition of admissible configurations $\mathcal{U}_{ad}$ by allowing for more than two types of assemblies. Let I denotes the number of different types of assemblies. Each type i, with $1 \leq i \leq I$, is characterized by a diffusion coefficient d^i and cross sections σ^i, Σ^i, and occupies a given volume γ_i in the domain Ω with

$$\sum_{i=1}^{I} \gamma_i = |\Omega|, \quad \gamma_i \geq 0.$$

We denote by $\chi_i(x)$ the characteristic function of that part of Ω occupied by assembly i. Clearly, it satisfies

$$\sum_{i=1}^{I} \chi_i(x) = 1, \quad \chi_i(x)\chi_j(x) = 0 \text{ if } i \neq j, \quad \int_\Omega \chi_i(x) dx = \gamma_i. \tag{41}$$

The coefficients of (4) are now given by

$$\begin{cases} D(x) = \sum_{i=1}^{I} d^i \chi_i(x), \\ \Sigma(x) = \sum_{i=1}^{I} \Sigma^i \chi_i(x), \\ \sigma(x) = \sum_{i=1}^{I} \sigma^i \chi_i(x). \end{cases} \tag{42}$$

The space of admissible configurations is therefore defined as

$$\mathcal{U}_{ad} = \left\{ (\chi_i)_{1\le i\le I} \in L^\infty(\Omega; \{0,1\}) \text{ satisfying (41)} \right\}. \tag{43}$$

As in the case of two type of assemblies, the minimization of (8) is not well-posed in $\mathcal{U}_{ad}$, i.e. there exist no minimizers. As before we introduce a relaxation of this problem by considering generalized designs in a space $\mathcal{U}^*_{ad}$.

The relaxation is built by using the homogenization theory (see e.g. [10, 18]) which works equally well for a mixture of any number I of components. The same arguments of H-convergence leads to the homogenized state equation

$$\begin{cases} -\operatorname{div}(D^* \nabla u(x)) + \overline{\Sigma} u(x) = \lambda \overline{\sigma} u(x), & x \in \Omega, \\ u(x) = 0, & x \in \partial\Omega, \end{cases} \tag{44}$$

where (λ, u) are the first eigenvalue and eigenvector, and the homogenized coefficients are defined by

$$\overline{\Sigma}(x) = \sum_{i=1}^{I} \theta_i(x) \Sigma^i, \quad \overline{\sigma}(x) = \sum_{i=1}^{I} \theta_i(x) \sigma^i,$$

where each θ_i is a density function which is the weak * limit in $L^\infty(\Omega; [0,1])$ of a sequence of characteristic functions $(\chi^i_n)_{n\ge 1}$ of the subdomain occupied by assembly i. These proportion functions satisfy

$$\sum_{i=1}^{I} \theta_i(x) = 1, \quad \int_\Omega \theta_i(x) dx = \gamma_i, \quad 0 \le \theta_i(x) \le 1. \tag{45}$$

The homogenized diffusion tensor $D^*(x)$ is the H-limit (i.e. the limit in the sense of homogenization) of the sequence $D_n = \sum_{i=1}^{I} \chi^i_n d^i$. The relaxed objective function is defined by

$$J^*(\theta, D^*) = \left(\ell\lambda + \frac{(\mathcal{M}(|su|^r))^{1/r}}{\mathcal{M}(\overline{\sigma} u)} \right), \tag{46}$$

where (λ, u) is the solution of (44). The set of generalized admissible configurations is defined by

$$\mathcal{U}^*_{ad} = \left\{ (\theta_1, \cdots, \theta_I, D^*) \in L^\infty(\Omega) \text{ satisfying (45), such that } D^* \in G_\theta \right\}, \tag{47}$$

where the constraint $D^* \in G_\theta$ holds almost everywhere in Ω, and G_θ is the set of all possible homogenized diffusion tensors associated to the family of proportions $\theta = (\theta_1, \cdots, \theta_I)$. Our relaxed problem is therefore to minimize J^* over $\mathcal{U}^*_{ad}$, i.e.

$$\min_{(\theta,D^*)\in\mathcal{U}^*_{ad}} J^*(\theta, D^*). \tag{48}$$

One can prove a relaxation result completely similar to Theorem 3.1 which states that (48) is the true relaxation of the original optimization problem. For the sake of brevity, we shall not dwell on this. Unfortunately, this relaxed formulation is not very useful since we do not know an algebraic characterization of the set G_θ when $I \geq 3$, on the contrary of the previous case $I = 2$. Nevertheless, as was first remarked by Raitum [21] (see also [22]), we do not need the full set G_θ for our optimization problem. It turns out that optimal arrangements of the components can always be found in the smaller subset of simple laminates or layered materials (this was already the case when $I = 2$, see Remark 4.3).

Let us explain this ”miracle” that allows to treat the case $I \geq 3$ as the previous one $I = 2$ (a similar argument has already been given in [22]). We define a set C_θ of all symmetric matrices with eigenvalues bounded between the harmonic and arithmetic means, i.e. $M \in C_\theta$ if and only if its eigenvalues μ_1, μ_2 satisfy

$$\mu^-_\theta \leq \mu_i \leq \mu^+_\theta, \text{ for } i = 1, 2, \tag{49}$$

where

$$\frac{1}{\mu^-_\theta} = \sum_{i=1}^{I} \frac{\theta_i}{d^i} \text{ and } \mu^+_\theta = \sum_{i=1}^{I} \theta_i d^i. \tag{50}$$

A well-known result of homogenization (see e.g. [10,18]) states that G_θ is included in C_θ, i.e. any homogenized tensor $D^* \in G_\theta$ satisfies $\mu^-_\theta Id \leq D^* \leq \mu^+_\theta Id$ in the sense of quadratic forms. This inclusion is also known to be strict, i.e. $G_\theta \neq C_\theta$. We then introduce a larger set of admissible designs

$$\begin{aligned} \mathcal{U}^c_{ad} = \{(\theta_1, \cdots, \theta_I, D^*) \in L^\infty(\Omega) & \text{ satisfying (45)}, \\ & \text{such that } D^* \in C_\theta\} . \end{aligned} \tag{51}$$

The set $\mathcal{U}^c_{ad}$ has no physical meaning: its tensors D^* are usually not homogenized tensors corresponding to a fine mixture of the phase components. It is just a mathematical artefact. Since $\mathcal{U}^*_{ad} \subset \mathcal{U}^c_{ad}$, we have

$$\min_{(\theta,D^*)\in\mathcal{U}^c_{ad}} J^*(\theta, D^*) \leq \min_{(\theta,D^*)\in\mathcal{U}^*_{ad}} J^*(\theta, D^*). \tag{52}$$

It turns out that J^* attains its minimum also in the set $\mathcal{U}^c_{ad}$. Indeed if (θ_n, D^*_n) is a minimizing sequence, up to a subsequence, there exists a limit $(\theta_\infty, D^*_\infty)$

such that θ_n converges to θ_∞ in $L^\infty(\Omega;[0,1])$ weak *, and D_n^* H-converges to D_∞^*. Furthermore, by the properties of H-convergence (see (11)), we have

$$\lim_{n\to+\infty} J^*(\theta_n, D_n^*) = J^*(\theta_\infty, D_\infty^*).$$

The only thing to check is that $(\theta_\infty, D_\infty^*)$ does indeed belong to $\mathcal{U}_{ad}^c$. Another classical result of homogenization tells us that

$$D_\infty^* \leq D_+^*, \quad (D_\infty^*)^{-1} \leq \left(D_+^*\right)^{-1},$$

where D_+^* (respectively $(D_-^*)^{-1}$) is the weak * limit of D_n^* (respectively $(D_n^*)^{-1}$) in $L^\infty(\Omega)$. Since $D_n^* \in C_\theta$ it satisfies

$$D_n^* \leq \mu_{\theta_n}^+ Id, \quad (D_n^*)^{-1} \leq \left(\mu_{\theta_n}^-\right)^{-1} Id,$$

where both right hand sides of the inequalities are affine function of θ_n, which implies that in the limit

$$D_\infty^* \leq \mu_{\theta_\infty}^+ Id, \quad (D_\infty^*)^{-1} \leq \left(\mu_{\theta_\infty}^-\right)^{-1} Id,$$

i.e. $D_\infty^* \in C_{\theta_\infty}$ as desired. In other words, J^* attains its minimum at $(\theta_\infty, D_\infty^*)$ in $\mathcal{U}_{ad}^c$. As proved in the next theorem, the optimality conditions in $\mathcal{U}_{ad}^c$ furnish a minimizer in $\mathcal{U}_{ad}^*$. Therefore, it is sufficient to find minimizers in $\mathcal{U}_{ad}^c$ for obtaining a particular minimizer in $\mathcal{U}_{ad}^*$, which yields a tractable relaxation for numerical computations.

Theorem 5.1 *There exists at least one couple $(\theta_\infty, D_\infty^*) \in \mathcal{U}_{ad}^*$, such that D_∞^* is the effective tensor of a layered material, which is a minimizer of J^* both in $\mathcal{U}_{ad}^*$ and in $\mathcal{U}_{ad}^c$.*

Remark 5.2 A layered material is obtained by stacking slices of the I components. Its effective or homogenized properties D^* can be computed explicitly in terms of the proportions $\theta = (\theta_1, \cdots, \theta_I)$ and the unit direction e which is normal to the slices or layers. In 2-D, in the basis $(e, \mathrm{e}^\perp)$ it reads

$$D^* = \text{diag}\left(\mu_\theta^-, \mu_\theta^+\right), \tag{53}$$

where μ_θ^- and μ_θ^+ are defined by (50).

Proof of Theorem 5.1. It is completely similar to that of Proposition 4.2. The key point is that the minimizer of (37) are the same in G_θ or in C_θ. As before, the derivative of J^* is given by (27) with the only difference that we now have $\delta\overline{\sigma} = \sum_{i=1}^I \sigma^i \delta\theta_i$, $\delta\overline{\Sigma} = \sum_{i=1}^I \Sigma^i \delta\theta_i$, and $rs^{r-1}\delta s = \sum_{i=1}^I (\sigma^i)^r \delta\theta_i$. Let (θ, D^*) be an optimal family of proportions and diffusion tensor in $\mathcal{U}_{ad}^c$. Let us denote by u the state, solution of (44), and by p the adjoint state, solution of (26). Since C_θ is a convex set, arguing as in Proposition 4.2, the optimal tensor D^* satisfies, outside the set where $|\nabla u||\nabla p| = 0$,

$$\left\{ \begin{array}{l} D^*\nabla u = \frac{1}{2}(\mu_\theta^+ + \mu_\theta^-)\nabla u - \frac{1}{2}(\mu_\theta^+ - \mu_\theta^-)\frac{|\nabla u|}{|\nabla p|}\nabla p, \\ D^*\nabla p = -\frac{1}{2}(\mu_\theta^+ - \mu_\theta^-)\frac{|\nabla p|}{|\nabla u|}\nabla u + \frac{1}{2}(\mu_\theta^+ + \mu_\theta^-)\nabla p, \end{array} \right. \tag{54}$$

and it is necessarily a layered material whose orientation is given in terms of the angle between ∇u and ∇p. Tartar [22] generalized the clever trick of [21], mentioned in Remark 4.3, to the case of $I > 2$ phases. It enables to prove that, even where $|\nabla u||\nabla p| = 0$, one can find an optimal D^* which is a layered material. This shows that at least one optimal tensor D^* in $\mathcal{U}_{ad}^c$ corresponds to a layered material which implies that it belongs to $\mathcal{U}_{ad}^*$ and is therefore optimal in $\mathcal{U}_{ad}^*$ too. Let us show that, if (θ, D^*) is a minimizer, then there exists another minimizer $(\tilde{\theta}, \tilde{D}^*)$ with the same state u and adjoint state p such that $\tilde{D}^*$ is a simple laminate. If $\nabla u = 0$, we can clearly take $\tilde{\theta} = \theta$ and replace D^* by any simple laminate $\tilde{D}^*$: it does not change u although p may change. When $\nabla p = 0$, we are going to change both θ and D^* while keeping the same state u and adjoint state p (a symmetric argument would also work in the case $\nabla u = 0$). Indeed, by Lemma 5.3 below, there exist I families of proportions $(\theta^k)_{1\le k\le I}$ with components $\theta^k = (\theta_i^k)_{1\le i\le I}$ such that $\sum_{i=1}^I \theta_i^k = 1$, $\theta_i^k \ge 0$, the rank of this family is $I-1$, and there exist simple laminates $D^k \in G_{\theta^k}$ satisfying $D^*\nabla u = D^k\nabla u$ for all $1 \le k \le I$. Denoting by ω the measurable subset of Ω where $\nabla p = 0$, we build $\tilde{\theta}$ in ω by partionning ω in subsets ω^k where it takes only the value θ^k (depending on ∇u and $D^*\nabla u$), which minimizes the "restriction" of the cost function to ω, while satisfying the volume constraints on each phase (this is possible since the rank of $(\theta^k)_{1\le k\le I}$ is $I-1$). Similarly, $\tilde{D}^*$ is defined as the simple laminate D^k in each ω^k. We have thus obtained another minimizer $(\tilde{\theta}, \tilde{D}^*)$ which is a simple laminate.

Lemma 5.3 *Let B be a symmetric matrix, with eigenvalues $(\mu_i)_{1\le i\le N}$ satisfying $\mu_\theta^- \le \mu_i \le \mu_\theta^+$. Let e and $\overline{e}$ be two vectors in $\mathbb{R}^N$ such that $\overline{e} = Be$. Then, we have*

$$\|\overline{e} - \frac{1}{2}(\mu_\theta^+ + \mu_\theta^-)e\| \le \frac{1}{2}(\mu_\theta^+ - \mu_\theta^-)\|e\|. \tag{55}$$

Furthermore, there exist I families of proportions $(\theta^k)_{1\le k\le I}$, depending only on e and $\overline{e}$, with components $\theta^k = (\theta_i^k)_{1\le i\le I}$ such that $\sum_{i=1}^I \theta_i^k = 1$ and $\theta_i^k \ge 0$, and there exist simple laminates $D^k \in G_{\theta^k}$ such that, for any $1 \le k \le I$,

$$\overline{e} = Be = D^k e.$$

Proof. By definition, the eigenvalues of $(B - \frac{1}{2}(\mu_\theta^+ + \mu_\theta^-)Id)$ belong to the interval $[-\frac{1}{2}(\mu_\theta^+ - \mu_\theta^-), \frac{1}{2}(\mu_\theta^+ - \mu_\theta^-)]$ which implies (55). For $t \in \mathbb{R}^I$, let us introduce a function $f(t)$ defined by

$$\begin{aligned} f(t) &= \frac{1}{4}(\mu_t^+ - \mu_t^-)^2 \|e\|^2 - \|\overline{e} - \frac{1}{2}(\mu_t^+ + \mu_t^-)e\|^2 \\ &= - \left(\mu_t^+ e - \overline{e}\right) \cdot \left(\mu_t^- e - \overline{e}\right). \end{aligned}$$

An easy computation shows that, for $t^k = (\delta_{ik})_{1 \le i \le I}$ (corresponding to pure phase k), $f(t^k) = -\|\overline{e} - d^k e\|^2$, while $f(\theta) \ge 0$ by virtue of (55). Since $f(t)$ is continuous, on each segment $[\theta, t^k]$ there exists a point θ^k such that $f(\theta^k) = 0$, $\sum_{i=1}^I \theta_i^k = 1$ and $\theta_i^k \ge 0$. The collection $(\theta^k)_{1 \le k \le I}$ is of rank $I - 1$ if $f(t^k) < 0$ and $f(\theta) > 0$ (if this is not the case, B can be replaced by a pure phase or a simple laminate). For such θ^k we have from $f(\theta^k) = 0$

$$f^k \cdot g^k = 0 \text{ with } f^k = \frac{\mu_{\theta^k}^+ e - \overline{e}}{\mu_{\theta^k}^+ - \mu_{\theta^k}^-} \text{ and } g^k = \frac{\mu_{\theta^k}^- e - \overline{e}}{\mu_{\theta^k}^+ - \mu_{\theta^k}^-}.$$

Then, defining a rank-one laminate $D^k \in G_{\theta^k}$ in the direction f^k if $f^k \neq 0$, or in any direction orthogonal to g^k if $f^k = 0$, we have

$$D^k f^k = \mu_{\theta^k}^- f^k \text{ and } D^k g^k = \mu_{\theta^k}^+ g^k.$$

Since $e = f^k - g^k$ and $\overline{e} = \mu_{\theta^k}^- f^k - \mu_{\theta^k}^+ g^k$, we thus deduce that $D^k e = \overline{e}$, as desired.

We now turn to the optimality conditions for the volume fractions θ_i. Let us define an angle φ between ∇u and ∇p by

$$\nabla u \cdot \nabla p = |\nabla u| |\nabla p| \cos \varphi, \tag{56}$$

where u and p are solutions of (44) and (26) respectively. By taking the optimal layered material D^* furnished by Theorem 5.1, the optimal proportions $\theta = (\theta_1, \cdots, \theta_I)$ satisfy the following optimality conditions

Proposition 5.4 *For* $1 \le i \le I$*, let* $Q_i(x)$ *be functions defined by*

$$\begin{aligned} Q_i(x) = -|\nabla u||\nabla p| \left[\frac{\partial \mu_\theta^-}{\partial \theta_i} \cos^2 \frac{\varphi}{2} - \frac{\partial \mu_\theta^+}{\partial \theta_i} \sin^2 \frac{\varphi}{2} \right] + \left(\Sigma^i - \lambda \sigma^i\right) up \\ + \frac{(\mathcal{M}(|su|^r))^{(1-r)/r}}{\mathcal{M}(\overline{\sigma} u)} \frac{(\sigma^i)^r u^r}{r|\Omega|} - \frac{(\mathcal{M}(|su|^r))^{1/r}}{(\mathcal{M}(\overline{\sigma} u))^2} \frac{\sigma^i u}{|\Omega|}. \end{aligned}$$

There exist a function $C_0(x)$ and constants C_i such that a minimizer (θ, D^*) for J^* satisfies

$$\begin{cases} \theta_i(x) = 0 \textit{ if } Q_i(x) - C_0(x) < C_i, \\ 0 \le \theta_i(x) \le 1 \textit{ if } Q_i(x) - C_0(x) = C_i, \\ \theta_i(x) = 1 \textit{ if } Q_i(x) - C_0(x) > C_i, \end{cases}$$

and reciprocally

$$\begin{cases} Q_i(x) - C_0(x) \le C_i \textit{ if } \theta_i(x) = 0, \\ Q_i(x) - C_0(x) = C_i \textit{ if } 0 < \theta_i(x) < 1, \\ Q_i(x) - C_0(x) \ge C_i \textit{ if } \theta_i(x) = 1. \end{cases}$$

The proof of Proposition 5.4 is completely similar to that of Proposition 4.4. In particular, taking the optimal D^* implies that

$$\delta J^* = \int_\Omega \sum_{i=1}^{I} Q_i(x)\delta\theta_i(x)dx \ge 0,$$

which, upon taking into account the I volume constraints $\int_\Omega \delta\theta_i(x)dx = 0$ and the pointwise constraint $\sum_{i=1}^I \theta_i = 1$, yields the desired result.

6. Numerical algorithms

This section is devoted to a gradient-type numerical algorithm for solving the proposed relaxed formulation of the re-loading optimization problem (in two space dimensions). It relies on our knowledge of the optimality conditions and, in particular, of the optimality of simply layered microstructures. By virtue of Remark 4.3, the optimal homogenized diffusion tensor D^* can be chosen to be that of a layered material which is parametrized by two variables : the volume fractions $\theta = (\theta_1, \cdots, \theta_I)$ and a rotation angle α. We therefore work with the following class of diffusion tensors

$$D^*(\theta, \alpha) = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} \mu_\theta^+ & 0 \\ 0 & \mu_\theta^- \end{pmatrix} \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix},$$

where μ_θ^- and μ_θ^+ are defined by (50). In other words, our objective function J^* is now a function of the $(I+1)$ scalar design variables θ and α, subject to the constraints

$$\sum_{i=1}^{I} \theta_i(x) = 1\,, \quad 0 \le \theta_i(x) \le 1\,, \quad \int_\Omega \theta_i(x)dx = \gamma_i\,. \tag{57}$$

The computation of the gradient of J^* with respect to (θ,α) is very similar to the derivation of the optimality conditions in Sect. 4. For an admissible increment $(\delta\theta_i,\delta\alpha)$, we find

$$\delta J^*(\theta,\alpha)=\int_\Omega \frac{\partial D^*}{\partial\alpha}\nabla u\cdot\nabla p\,\delta\alpha dx+\sum_{i=1}^{I}\int_\Omega \overline{Q}_i(x)\delta\theta_i dx,$$

where $\overline{Q}_i(x)$, being very similar to $Q_i(x)$, is defined by

$$\begin{aligned}\overline{Q}_i(x)=\frac{\partial\mu_\theta^-}{\partial\theta_i}&\left[\sin^2\alpha\frac{\partial u}{\partial x}\frac{\partial p}{\partial x}+\cos^2\alpha\frac{\partial u}{\partial y}\frac{\partial p}{\partial y}\right.\\&\left.+\cos\alpha\sin\alpha\left(\frac{\partial u}{\partial x}\frac{\partial p}{\partial y}+\frac{\partial u}{\partial y}\frac{\partial p}{\partial x}\right)\right]+\frac{\partial\mu_\theta^+}{\partial\theta_i}\left[\cos^2\alpha\frac{\partial u}{\partial x}\frac{\partial p}{\partial x}\right.\\&\left.+\sin^2\alpha\frac{\partial u}{\partial y}\frac{\partial p}{\partial y}-\cos\alpha\sin\alpha\left(\frac{\partial u}{\partial x}\frac{\partial p}{\partial y}+\frac{\partial u}{\partial y}\frac{\partial p}{\partial x}\right)\right]\\&+\left(\Sigma^i-\lambda\sigma^i\right)up+\frac{(\mathcal{M}(|su|^r))^{(1-r)/r}}{\mathcal{M}(\overline{\sigma}u)}\frac{(\sigma^i)^r u^r}{r|\Omega|}\\&-\frac{(\mathcal{M}(|su|^r))^{1/r}}{(\mathcal{M}(\overline{\sigma}u))^2}\frac{\sigma^i u}{|\Omega|},\end{aligned}$$

with

$$\frac{\partial\mu_\theta^+}{\partial\theta_i}=d^i\ ,\ \text{and}\ \frac{\partial\mu_\theta^-}{\partial\theta_i}=\frac{-(\mu_\theta^-)^2}{d^i}.$$

Once we have computed the gradient we need to add a projection step in order to satisfy the admissibility constraints (57). The gradient method is then structured as follows.

1. We **initialize** the design parameters $\theta^1=(\theta_1^1,\cdots,\theta_I^1)$ and α^1 (for example, we take a constant angle α_1 and volume fractions θ_i^1, which satisfy the volume constraints).
2. Until convergence, for $n\geq 1$ we **iteratively** compute the state u^n and the adjoint state q^n, solutions of (10) and (24) respectively with the previous design parameters (θ^n,α^n), and then update these parameters by

   $$\theta_i^{n+1}(x)=\max\left(0,\min\left(1,\theta_i^n(x)-t_n(\overline{Q}_i^n(x)-C_0^{n+1}(x)-C_i^{n+1})\right)\right)$$

   where C_i^{n+1} are Lagrange multipliers (constant throughout the domain) for the global volume constraints, and $C_0^{n+1}(x)$ is the Lagrange multiplier (varying at each point x) for the local volume constraint $\sum_{i=1}^{I}\theta_i^{n+1}(x)=1$, and

   $$\alpha^{n+1}=\alpha^n-t_n\frac{\partial D^*}{\partial\alpha}(\theta^n,\alpha^n)\nabla u^n\cdot\nabla p^n$$

   where $t_n>0$ is a small step such that $J^*(\theta^{n+1},\alpha^{n+1})<J^*(\theta^n,\alpha^n)$.

Table 1. Physical constants of the 4 types of assembly

Label of assembly	Diffusion D	Absorption Σ	Fission σ	Proportion
1	1.340	0.0245	0.0311	40/157
2	1.356	0.0250	0.0287	40/157
3	1.375	0.0254	0.0270	40/157
4	1.390	0.0258	0.0256	37/157

The Lagrange multipliers are iteratively adjusted in a inner loop at each step n of the above algorithm (this is the most delicate part of the algorithm, the case of $I \geq 3$ phases being much more time-consuming than just two phases). Such a gradient method always converges to a (local) minimum, and its speed of convergence is partly governed by the efficiency of the line search for finding a good step t_n. However, in practice we made no special efforts in optimizing the choice of t_n. Neverthelees, to improve the speed of the algorithm, we have replaced the gradient method for the angle α by an application of the optimality criteria (this is a very popular principle in structural design ; see e.g. [4]). In view of Proposition 4.2 the lamination direction α^{n+1} is determined by the angle between ∇u^n and ∇p^n rather than by the above formula.

We test our method on a core with 157 squared assemblies (with side length 21.5 cm) of 4 different types with properties given by Table 1 (these data are representative of a 900 Mw pressurized water reactor). The computation are performed on one fourth of the geometry using the Matlab software. There are 362 $P1$ finite elements in the mesh and the volume fractions are constant on each assembly. We choose $\ell = 0$ and $r = 10$ in the objective function (other choices work as well). We first compute the optimal solution for the relaxed formulation after 200 iterations. Figures 5 and 6 display the optimal volume fractions, and Fig. 7 the resulting power distribution σu. The convergence is smooth as shown by figure 8 and independent of the initialization (we believe we reached a global minimum). The power peak $\max(\sigma u)$ is globally decreasing (there is no reconstruction of the fine structure of the flux).

The above relaxed or homogenized optimal solution gives a lower bound on the minimal performance of any discrete distribution of assemblies. More than that, by penalizing the intermediate values of the volume fractions, we can recover a quasi-optimal distribution of assemblies. We introduce a penalized objective function, defined by

$$J^{pen}(\theta, \alpha) = \ell\lambda + \frac{\left(\mathcal{M}(|su|^r)\right)^{1/r}}{\mathcal{M}(\overline{\sigma}u)} + \frac{\eta}{|\Omega|}\int_\Omega \sum_{i=1}^{I} \theta_i(1-\theta_i)\,dx\,,$$

where (λ, u) is the solution of (44). For $\eta = 0$ we recover the relaxed objective function J^*, while for $\eta > 0$ we force the volume fractions to take

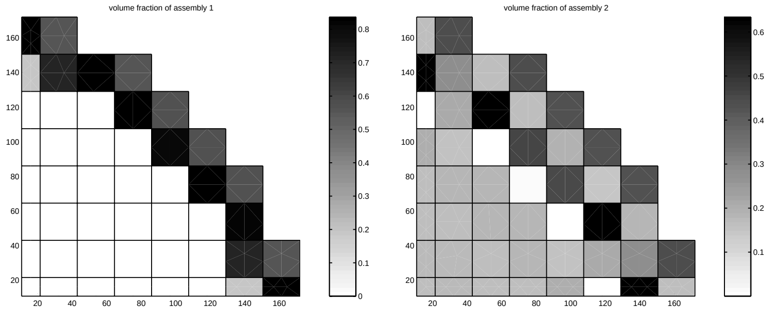

Fig. 5. Volume fractions of assembly 1 (left) and 2 (right)

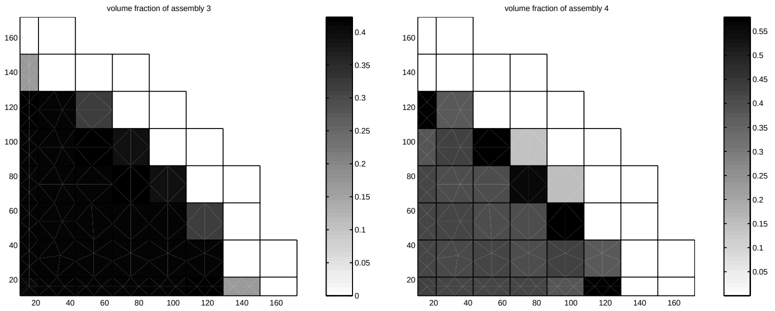

Fig. 6. Volume fractions of assembly 3 (left) and 4 (right)

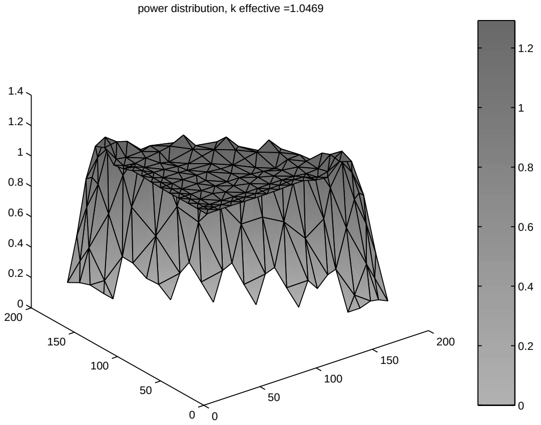

Fig. 7. Power distribution σu

only the values 0 or 1. Starting from the previous relaxed optimal design, we minimize the penalized objective function and increase progressively the value of η. Since by virtue of Theorem 3.1 any relaxed design is the limit of a sequence of closer and closer classical designs, the penalization process amounts to build such an approximating sequence for which the

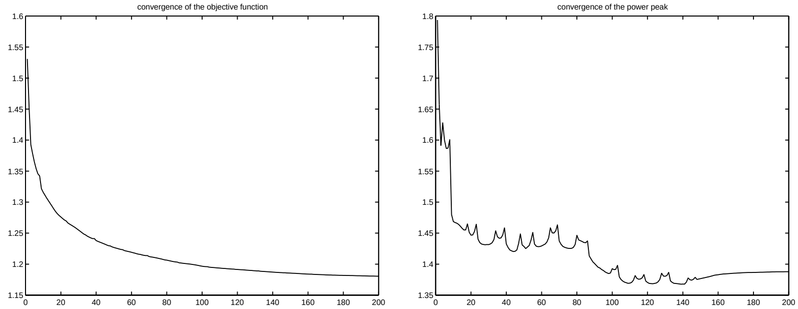

Fig. 8. Convergence history: objective function (left) and power peak (right)

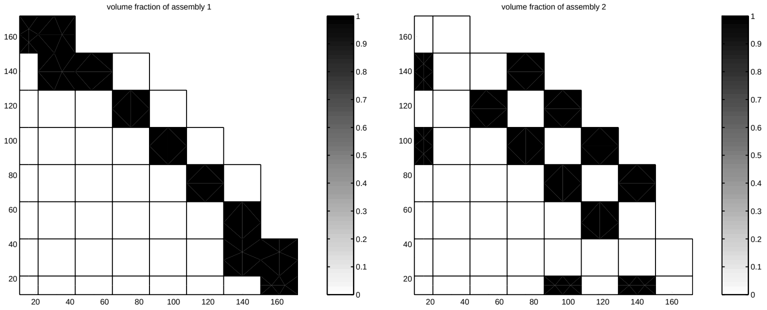

Fig. 9. Distributions of assembly 1 (left) and 2 (right)

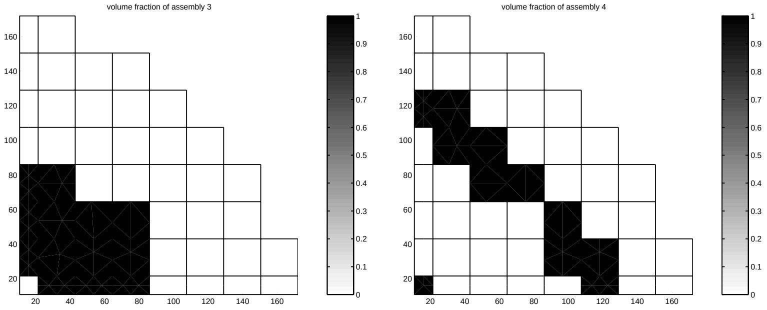

Fig. 10. Distributions of assembly 3 (left) and 4 (right)

objective function should not change too much. This procedure is now well-established in structural optimization (see [1,4]). Here, we run 50 iterations with $\eta = 1$ and 20 more with $\eta = 2$. Of course, the results are very sensitive to the choice of η which should not be too large. Figures 9 and 10 display the discrete distribution of assemblies, and Fig. 11 the resulting power distribution σu. Remark that the obtained pattern is not symmetric with respect to the first diagonal. It may indicate that an even better design could be found if we do not enforce the core symmetry by fourth.

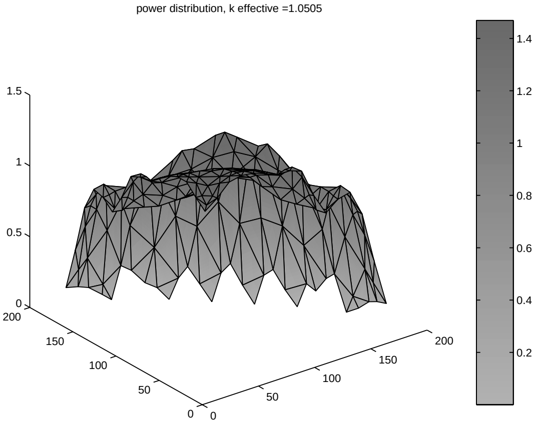

Fig. 11. Power distribution after penalization

Table 2. Comparison between the homogenized and penalized designs

	Objective function	Power peak
Homogenized design	1.180	1.387
Penalized design	1.249	1.551

In Table 2 we compare the values of the objective function for the relaxed optimal design and for the penalized one (the penalization term $J^{pen} - J^*$ is almost zero at the end of the penalization process).

In our opinion the interest of the homogenization method is twofold. First, the homogenized optimal design gives an absolute lower bound to any proposed discrete distribution of assemblies. Therefore, it is a good element of comparison with any other optimization method. Second, the homogenization algorithm is insensitive to the initial guess and the resulting penalized discrete distribution of assemblies is free of any implicit or explicit constraint on its pattern (in structural optimization this is called topology optimization, see e.g. [1,3,4]). We do not view this method as an alternative to other optimization algorithms but rather as a pre-processing step. Indeed, it gives rise to new patterns that may be different from initial guesses or intuitions, but that can be improved by local optimization using more realistic constraints or objective function.

7. Conclusion and perspectives

This paper describes a new approach for optimizing the fuel assemblies positions in a nuclear reactor core. This approach is based on the homogenization method which has already been successfully implemented for structural optimization. The work reported here is still in progress. Basically we are working in two directions. First, we generalize the present work to the more realistic model of two-groups diffusion (this is a system of two coupled diffusion equations). The principle of this generalization is the same but many new mathematical difficulties arise. In particular, we shall introduce a partial relaxation instead of the true relaxed formulation which is unfortunately untractable. Second, we have to take into account more realistic constraints in the optimization process and do more numerical comparisons with other approaches in the literature. This will be reported in a next paper [2].

References

1. Allaire G., Bonnetier E., Francfort G., Jouve F.: Shape optimization by the homogenization method. Numerische Mathematik **76**, 27–68 (1997)
2. Allaire G., Castro C. (in preparation)
3. Allaire G., Kohn R.V.: Optimal design for minimum weight and compliance in plane stress using extremal microstructures. Eur. J. Mech. A/Solids **12**(6), 839–878 (1993)
4. Bendsoe M.: Methods for optimization of structural topology, shape and material. Berlin Heidelberg New York: Springer 1995
5. Dacorogna B.: Weak continuity and weak lower semicontinuity of nonlinear functionals. Lecture Notes in Math. 922, Berlin Heidelberg New York: Springer 1982
6. Dumas M.: Optimisation du repositionnement des assemblages combustibles d'un réacteur nucléaire, in Numerical methods for engineering, 2^{nd} international congress GAMNI, E. Absi et al. eds., pp. 865-874, Paris: Dunod 1980
7. Ekeland I., Temam R.: Convex analysis and variational problems. Amsterdam: North Holland 1976
8. Gaudier F.: Modélisation par réseaux de neurones. Application à la gestion du combustible dans un réacteur, PhD thesis, ENS Cachan (1999)
9. Ho L.-W., Rohach A.: Perturbation theory in nuclear fuel management optimization. Nucl. Sc. Eng. **82**, 151–161 (1982)
10. Jikov V., Kozlov S., Oleinik O.: Homogenization of differential operators. Berlin: Springer 1995
11. Kropaczek D.J., Turinsky P.J.: In-core nuclear fuel management optimization for pressurized water reactors utilizing simulated annealing. Nucl. Technol. **95**, 9 (1991)
12. Levine S., In-core fuel management of four reactor types, in Handbook of Nuclear Reactor Calculations, vol. II, Y. Ronen ed., CRC Press, pp. 87–201 (1986).
13. Lurie K., Cherkaev A.: Exact estimates of conductivity of composites formed by two isotropically conducting media, taken in prescribed proportion. Proc. R. Soc. Edinburgh **99**A, 71–87 (1984)
14. Lysenko M.G., Wong H.I., Maldonado G.I.: Neural network and perturbation Theory hybrid Models For Eigenvalue Prediction. Nucl. Sc. Eng. **132**, (1999)
15. Lysenko M.G., Wong H.I., Maldonado G.I.: Predicting Neutron Diffusion Eigenvalues with a Query-Based Adaptive Neural Architecture. IEEE Trans. Neural Network **10** (1999)
16. Maldonado G.I., Turinsky P.J.: Application of nonlinear nodal diffusion generalized perturbation theory to nuclear fuel reload optimization. Nucl. Technol. **110**, 198–219 (1995)
17. Murat F.: Contre-exemples pour divers problèmes où le contrôle intervient dans les coefficients. Ann. Mat. Pura Appl. **112**, 49–68 (1977)

18. Murat F., Tartar L.: Calcul des variations et Hogénéisation, in Les Méthodes de l'Homogénéisation: Théorie et Applications en Physique, Eyrolles, 319-369 (1985). English translation in Topics in the mathematical modelling of composite materials. A. Cherkaev, R. Kohn, Editors, Progress in Nonlinear Differential Equations and their Applications, 31, Birkhäuser, Boston (1997)
19. Parks G.T.: Multiobjective pressurized water reactor reload core design by nondominated genetic algorithm search. Nucl. Sci. Eng. **124**, 178–187 (1996)
20. Planchard J.: Méthodes mathématiques en neutronique, Paris: Eyrolles 1995
21. Raitum U.: The extension of extremal problems connected with a linear elliptic equation. Sov. Math. **19**, 1342–1345 (1978)
22. Tartar L.: Remarks on the homogenization method in optimal design methods. In: Homogenization and Applications to Material Sciences, pp. 393–412, Gakuto International Series, Mathematical Sciences and Applications 9 (1997)