

Combining Galerkin approximation techniques with the principle of Hashin and Shtrikman to derive a new FFT-based numerical method for the homogenization of composites

Sébastien Brisard, Luc Dormieux

▶ **To cite this version:**

Sébastien Brisard, Luc Dormieux. Combining Galerkin approximation techniques with the principle of Hashin and Shtrikman to derive a new FFT-based numerical method for the homogenization of composites. Computer Methods in Applied Mechanics and Engineering, 2012, 217-220, pp.197-212. 10.1016/j.cma.2012.01.003 . hal-00722361

HAL Id: hal-00722361

https://enpc.hal.science/hal-00722361

Submitted on 3 Sep 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Combining Galerkin approximation techniques with the principle of Hashin and Shtrikman to derive a new FFT-based numerical method for the homogenization of composites

S. Brisard[a,*], L. Dormieux[b]

[a] *Université Paris-Est, IFSTTAR, MACS, F-75732, Paris, France*
[b] *Université Paris-Est, UR Navier, Ecole des Ponts ParisTech, 6-8 av. Blaise Pascal, Cité Descartes, Champs-sur-Marne, F-77455 Marne-La-Vallée cedex 2, France*

Abstract

We report on the mathematical analysis of two different, FFT-based, numerical schemes for the homogenization of composite media within the framework of linear elasticity: the basic scheme of Moulinec and Suquet (1994, 1998), and the energy-based scheme of Brisard and Dormieux (2010). Casting these two schemes as Galerkin approximations of the same variational problem allows us to assert their well-posedness and convergence. More importantly, we extend in this work their domains of application, by relieving some stringent conditions on the reference material which were previously thought necessary. The origins of the flaws of each scheme are identified, and a *third* scheme is proposed, which seems to combine the strengths of the basic and energy-based schemes, while leaving out their weaknesses. Finally, a rule is proposed for handling heterogeneous pixels/voxels, a situation frequently met when images of real materials are used as input to these schemes.

Key words: Galerkin approximation, Heterogeneous media, Linear elasticity, Numerical homogenization, Polarization, Variational problem

1. Introduction

Determination of the effective properties is a crucial point in the analysis and design of composite materials. While closed-form homogenization schemes (such as Mori-Tanaka [1, 2], self-consistent [3], or generalized self-consistent [4, 5]) have been known to provide satisfactory estimates within the framework of linear elasticity, they may at times prove insufficient due to the limited microstructural information they account for. Indeed, the numerical input of these formulas reduces to volume fractions, any higher-order information (such as shape, relative sizes or positions) being at best accounted for in a qualitative way.

In order to faithfully account for the finest details of the microstructure, it is often necessary to resort to full-field numerical simulation of the composite. The need for such an accurate calculation is felt even more strongly when dealing with long-term effective behaviour (since creep can induce high stiffness contrasts, as argued in [6]; see also [7]), or non-linearities. Of course, the price to pay for this increased accuracy is time.

This is particularly true of the standard finite element method, which would probably first come to mind. Indeed, this approach requires each sub-domain to be meshed, a time-consuming operation for highly heterogeneous composites. In such a situation, numerical methods formulated on regular grids (e.g. custom finite element [8], finite difference [6], or FFT-based methods [9, 10]) might be preferred.

Regardless of the actual discretization scheme, inversion of a (presumably large) linear system is always required, and iterative linear solvers [11] must generally be invoked. These solvers work by iteratively performing matrix-vector products; inversion can be fast if this product is implemented efficiently.

These considerations led Moulinec and Suquet [9, 10] to a now popular numerical method based on the fast Fourier transform (FFT). The so-called basic scheme results from the introduction of a reference material leading to the Lippmann-Schwinger equation [12] (see also section 4.3), which is then discretized. In the framework of *periodic* elasticity, it is natural to solve the resulting system in the Fourier space, where costly matrix-vector products are shown to reduce to comparatively cheaper direct (element-by-element) products.

Building on the same ideas, Brisard and Dormieux [13] recently proposed another scheme, based on the energy principle of Hashin and Shtrikman [14] (see also section 2.3). It was found [13] that at a given resolution (grid fineness), inversion of the linear system underlying this energy-based scheme could be performed in much less iterations than would be necessary for the basic scheme.

Nearly simultaneously, Zeman et al. [15] also proposed a modified FFT-based scheme, where substitution of CG/BiCG iterative solvers to the fixed-point iterations initially proposed by Moulinec and Suquet [9, 10] was shown to lead to substantial acceleration.

Both the equation of Lippmann and Schwinger and the principle of Hashin and Shtrikman require the introduction of a so-called *reference material* (see below), upon which some conditions apply. For the basic scheme, these conditions are stated in [16, eq. (22)]; it is always possible to enforce them. This is not true of the energy-based scheme, for which the reference material must be either stiffer or softer (in a sense which will be

*Corresponding author.
Email addresses: sebastien.brisard@ifsttar.fr (S. Brisard), luc.dormieux@enpc.fr (L. Dormieux)

made more precise later) than all the constituants of the composite. The energy-based scheme can therefore not be applied to composites containing both pores and rigid inclusions, a situation of high practical interest. It should be noted that (for different reasons) the basic scheme suffers from the same kind of limitations, since it was shown to fail in the presence of pores [16].

Our initial purpose was to try and alleviate some of these limitations. Reviewing [16] and [13] with this goal in mind, we came to the conclusion that the requirements on the reference material stated in these papers are *sufficient*; whether they are *necessary* remained an open question at that time. In other words: is it relevant to use a reference material which *violates* the above mentioned requirements? The answer to this question demanded a rigorous mathematical analysis, in the course of which we soon realized that both basic and energy-based schemes could be regarded as Galerkin discretizations of the same variational problem [17]. This proved a very effective approach, since the classical results coming from the finite element literature could be applied.

In the present paper, the main steps of the theoretical analysis of both basic and energy-based numerical schemes are detailed. Only the essential results are stated, and some of the corresponding proofs are outlined; more details can be found in the appendices. These results have deep practical implications.

First, our variational approach results in an unambiguous separation between *discretization* of the continuous problem of Hashin and Shtrikman (see below) and *inversion* of the resulting linear system. More precisely, it will be seen that both basic- and energy-based schemes provide an estimate $\boldsymbol{\tau}^h$ of the true polarization field $\boldsymbol{\tau}$. $\boldsymbol{\tau}^h$ is a cell-wise constant field on a regular grid (where h denotes the size of each square/cubic cell); $\boldsymbol{\tau}^h$ is the unique solution of a linear system. Since this system is clearly identified, solving it via fixed-point iterations, as proposed by Moulinec and Suquet [9, 10], is no longer required. Instead, we can freely invoke more robust solvers; whether or not the iterations of the linear solver converge is therefore a well-documented problem of numerical analysis. As such, it will be completely disregarded in this paper, where we rather focus on the convergence of the approximate, cell-wise constant solution $\boldsymbol{\tau}^h$ to the true polarization field $\boldsymbol{\tau}$ *as the size h of the cells goes to* 0. When confusion might occur between convergence of the iterative solver, and convergence with h, the latter will be refered to as h–convergence.

Second, it is shown that *all* conditions on the reference material can be removed, provided that its stiffness is positive definite. More precisely, for any choice of the reference material, h–convergence (in the L^2–sense) is observed for both schemes. Their application to cases previously considered as forbidden is therefore possible, and absence of convergence of the basic scheme in the presence of pores [16] can be overcome at the cost of very limited alterations (fixed-point iterations *must* be replaced by a more robust solver).

Third, heterogeneous cells can be given equivalent elastic properties, based on a *consistent* rule. This is a critical point, which guarantees the successful coupling of these methods with experimental imaging techniques.

The paper is organized as follows. In section 2, the fundamental equations for the analysis of heterogeneous media within the framework of periodic linear elasticity are stated. Following the classical approach of Hashin and Shtrikman [14], a quadratic form $\mathcal{H}(\boldsymbol{\tau})$ on the space of polarization fields is introduced. The unique critical point of $\mathcal{H}$ gives the solution to the local problem of micromechanics. Finding this critical point is effectively a variational problem, which will be called the problem of Hashin and Shtrikman.

In section 3, the problem of Hashin and Shtrikman is studied in detail from the mathematical viewpoint. It is shown that, under mild conditions on the reference material as well as the microstructure, this problem is well-posed.

In section 4, it is shown that a slightly altered version of the basic scheme of Moulinec and Suquet [9, 10], as well as the energy-based scheme of Brisard and Dormieux [13], are well-posed, Galerkin-like approximations of the problem of Hashin and Shtrikman. h–convergence of the approximate solutions to the true solution will be proved; this is one of the essential results of this paper.

In section 5, some numerical examples are proposed to illustrate the mathematical results established theoretically. The performances of both basic and energy-based schemes are compared, and a new, intermediate scheme is proposed. This new scheme seems very promising, as it combines the assets of both its ancestors, while avoiding their weaknesses.

The paper closes with a discussion of some possible extensions of the present work.

2. Background

2.1. The local problem of micromechanics

Following Hill [18], the determination of the overall elastic properties of a heterogeneous medium $\Omega \subset \mathbb{R}^d$ amounts to finding the local stresses $\boldsymbol{\sigma}$ and strains $\boldsymbol{\varepsilon}$ at (elastic) equilibrium, subjected to appropriate boundary conditions. By definition, the effective elastic moduli $\mathbf{C}^{\text{eff}}$ then provide the relationship between $\overline{\boldsymbol{\sigma}}$ and $\overline{\boldsymbol{\varepsilon}}$

$$\overline{\boldsymbol{\sigma}} = \mathbf{C}^{\text{eff}} : \overline{\boldsymbol{\varepsilon}},$$

where $\overline{\mathcal{B}}$ denotes the volume average of the local quantity $\mathcal{B}$.

Various boundary conditions, namely kinematic or static uniform boundary conditions [19], as well as periodic boundary conditions can be adopted; for a sufficiently large domain Ω (larger than the representative volume element), the effective properties do not depend on the actual boundary conditions.

This paper is restricted to *periodic* boundary conditions. Besides a convenient formulation of the local problem in Fourier space, such conditions are known to be very favorable for the numerical homogenization of heterogeneous media. Indeed, convergence of the effective properties is observed with domains significantly smaller than kinematic or static uniform boundary would require [20].

In periodic homogenization, Ω reduces to the unit cell, $\Omega = [0, L_1] \times \ldots \times [0, L_d]$ ($d = 2$ for plane strain problems[1], $d = 3$ for three-dimensional problems); the local stiffness $\mathbf{C}(\mathbf{x})$ is $(L_1, \ldots, L_d)$-periodic. The domain is subjected to a macroscopic strain $\mathbf{E}$, and the resulting displacement field $\mathbf{u}(\mathbf{x})$ fluctuates locally about its macroscopic counterpart $\mathbf{E}\cdot\mathbf{x}$; these fluctuations are assumed to be $(L_1, \ldots, L_d)$-periodic. Within this framework, the local problem of micromechanics therefore reads

$$\operatorname{div}\left[\mathbf{C}(\mathbf{x}) : \boldsymbol{\varepsilon}(\mathbf{x})\right] = \mathbf{0}, \tag{1a}$$

$$2\varepsilon_{ij}(\mathbf{x}) = \partial_i u_j(\mathbf{x}) + \partial_j u_i(\mathbf{x}), \tag{1b}$$

$$\mathbf{u}\left(\mathbf{x} + L_i\mathbf{e}_i\right) = \mathbf{u}(\mathbf{x}) + L_i\mathbf{E}\cdot\mathbf{e}_i, \tag{1c}$$

$$\boldsymbol{\sigma}\left(\mathbf{x} + L_i\mathbf{e}_i\right)\cdot\mathbf{e}_i = \boldsymbol{\sigma}(\mathbf{x})\cdot\mathbf{e}_i, \tag{1d}$$

where $\mathbf{x} \in \mathbb{R}^d$, $i, j = 1, \ldots, d$ and $\mathbf{e}_1, \ldots, \mathbf{e}_d$ denote the basis vectors (no summation on repeated indices in the above expressions). Then

$$\mathbf{C}^{\text{eff}} : \mathbf{E} = \overline{\mathbf{C} : \boldsymbol{\varepsilon}},$$

where $\boldsymbol{\varepsilon}$ solves problem (1).

Irrespective of the boundary conditions, exact solution of the local problem of micromechanics is generally untractable due to the heterogeneity of the medium. In this paper, we present a unified mathematical formulation of two FFT-based numerical methods for the approximate solution of problem (1). The principle of Hashin and Shtrikman (see section 2.3) is a valuable tool for the analysis of both schemes. This principle requires the introduction of the Green operator for strains, see section 2.2.

2.2. The Green operator for strains [3]

Formulation of the principle of Hashin and Shtrikman requires the introduction of a so-called *reference* medium, of homogeneous stiffness $\mathbf{C}_0$, occupying the same domain Ω as the real, heterogeneous material. We further define the fourth-rank Green operator for strains $\boldsymbol{\Gamma}_0$ associated with $\mathbf{C}_0$ and the shape of Ω. This operator is formally defined [3] as the resolvent of the following auxiliary problem ($\mathbf{x} \in \mathbb{R}^d$, $i, j = 1, \ldots, d$; no summation on repeated indices)

$$\operatorname{div}\left[\mathbf{C}_0 : \boldsymbol{\varepsilon}(\mathbf{x}) + \boldsymbol{\tau}(\mathbf{x})\right] = \mathbf{0}, \tag{2a}$$

$$2\varepsilon_{ij}(\mathbf{x}) = \partial_i u_j(\mathbf{x}) + \partial_j u_i(\mathbf{x}), \tag{2b}$$

$$\mathbf{u}\left(\mathbf{x} + L_i\mathbf{e}_i\right) = \mathbf{u}(\mathbf{x}), \tag{2c}$$

$$\boldsymbol{\sigma}\left(\mathbf{x} + L_i\mathbf{e}_i\right)\cdot\mathbf{e}_i = \boldsymbol{\sigma}(\mathbf{x})\cdot\mathbf{e}_i, \tag{2d}$$

where the so-called (given) *polarization* field $\boldsymbol{\tau}(\mathbf{x})$ is a second-rank symmetric tensor, defined on Ω. It should be noted that problem (2) merely corresponds to the elastic equilibrium of a linearly elastic, homogeneous body, subjected to periodic boundary conditions. By definition, the strain field $\boldsymbol{\varepsilon}(\mathbf{x})$ which solves problem (2) reads

$$\boldsymbol{\varepsilon}(\mathbf{x}) = -\left(\boldsymbol{\Gamma}_0 * \boldsymbol{\tau}\right)(\mathbf{x}),$$

where '$*$' is to be understood as a convolution product, reading in Fourier space

$$\left(\boldsymbol{\Gamma}_0 * \boldsymbol{\tau}\right)(\mathbf{x}) = \sum_{\mathbf{b}\in\mathbb{Z}^d} \hat{\boldsymbol{\Gamma}}_0\left(\mathbf{k}_\mathbf{b}\right) : \hat{\boldsymbol{\tau}}\left(\mathbf{k}_\mathbf{b}\right) \exp\left(\imath\mathbf{k}_\mathbf{b}\cdot\mathbf{x}\right), \tag{3}$$

with

$$\mathbf{k}_\mathbf{b} = \frac{2\pi b_1}{L_1}\mathbf{e}_1 + \cdots + \frac{2\pi b_d}{L_d}\mathbf{e}_d, \quad \text{for } \mathbf{b} \in \mathbb{Z}^d. \tag{4}$$

In the remainder of this paper, the following rule will be adopted: *greek* multi-indices (such as $\boldsymbol{\beta} \in \mathbb{Z}^d$) refer to the *real* space, while *latin* multi-indices (such as $\mathbf{b} \in \mathbb{Z}^d$) refer to the Fourier space.

Assuming the reference medium to be isotropic with shear modulus μ_0 and Poisson ratio ν_0, $\hat{\boldsymbol{\Gamma}}_0(\mathbf{k})$ is known in closed-form [21]

$$\begin{aligned}\hat{\Gamma}_{0,ijhl}\left(\mathbf{k}\right) = {} & \frac{1}{4\mu_0}\left(\delta_{ih}n_j n_l + \delta_{il}n_j n_h + \delta_{jh}n_i n_l + \delta_{jl}n_i n_h\right) \\ & - \frac{1}{2\mu_0\left(1-\nu_0\right)} n_i n_j n_h n_l,\end{aligned} \tag{5}$$

with $\mathbf{n} = \mathbf{k}/k$ and $k = \|\mathbf{k}\|$ ($\mathbf{k} \neq \mathbf{0}$). Equation (5) can be recast

$$\begin{aligned}\boldsymbol{\varpi} : \hat{\boldsymbol{\Gamma}}_0\left(\mathbf{k}\right) : \boldsymbol{\tau} = {} & \frac{1}{\mu_0}\mathbf{n}\cdot\boldsymbol{\varpi}\cdot\boldsymbol{\tau}\cdot\mathbf{n} \\ & - \frac{1}{2\mu_0\left(1-\nu_0\right)}\left(\mathbf{n}\cdot\boldsymbol{\varpi}\cdot\mathbf{n}\right)\left(\mathbf{n}\cdot\boldsymbol{\tau}\cdot\mathbf{n}\right),\end{aligned}$$

for any two symmetric tensors $\boldsymbol{\tau}$ and $\boldsymbol{\varpi}$ and wave-vector $\mathbf{k}$. Then, from the Cauchy-Schwarz inequality

$$|\boldsymbol{\varpi} : \hat{\boldsymbol{\Gamma}}_0\left(\mathbf{k}\right) : \boldsymbol{\tau}| \leq \frac{3-2\nu_0}{2\mu_0\left(1-\nu_0\right)}\|\boldsymbol{\varpi}\|\|\boldsymbol{\tau}\|, \tag{6}$$

where $\|\boldsymbol{\tau}\|$ denotes the usual hermitian norm for symmetric, second-rank tensors $\|\boldsymbol{\tau}\| = \left(\boldsymbol{\tau}^* : \boldsymbol{\tau}\right)^{1/2}$ (where $\boldsymbol{\tau}^*$ stands for the component-wise complex conjugate of $\boldsymbol{\tau}$). Finally (substituting $\boldsymbol{\varpi} = \hat{\boldsymbol{\Gamma}}_0(\mathbf{k}) : \boldsymbol{\tau}$)

$$\|\hat{\boldsymbol{\Gamma}}_0\left(\mathbf{k}\right) : \boldsymbol{\tau}\| \leq \frac{3-2\nu_0}{2\mu_0\left(1-\nu_0\right)}\|\boldsymbol{\tau}\|, \tag{7}$$

for any symmetric tensor $\boldsymbol{\tau}$.

2.3. The principle of Hashin and Shtrikman [14]

The so-called "energy principle" of Hashin and Shtrikman [14] is a theorem for the characterization of the critical points of the following functional

$$\mathcal{H}(\boldsymbol{\varpi}) = \mathbf{E} : \overline{\boldsymbol{\varpi}} - \tfrac{1}{2}\overline{\boldsymbol{\varpi} : \left(\mathbf{C}-\mathbf{C}_0\right)^{-1} : \boldsymbol{\varpi}} - \tfrac{1}{2}\overline{\boldsymbol{\varpi} : \left(\boldsymbol{\Gamma}_0 * \boldsymbol{\varpi}\right)},$$

defined for any test field $\boldsymbol{\varpi}$. Let $\boldsymbol{\varepsilon}$ be the solution of the initial problem (1), and

$$\boldsymbol{\tau}(\mathbf{x}) = \left[\mathbf{C}(\mathbf{x}) - \mathbf{C}_0\right] : \boldsymbol{\varepsilon}(\mathbf{x}), \tag{8}$$

the corresponding polarization field within the heterogeneous medium at equilibrium. Theorems 1 and 2 below were first proved by Hashin and Shtrikman [14].

[1] The present work also applies to plane stress elasticity, provided that the usual substitutions for the shear modulus and Poisson ratio are performed.

Theorem 1. *For any reference material* $\mathbf{C}_0$*, the polarization field* $\boldsymbol{\tau}$ *corresponding to the solution* $\boldsymbol{\varepsilon}$ *of problem* (1) *is a critical point of* $\mathcal{H}$*. In other words, for any test field* $\boldsymbol{\varpi}$

$$\overline{\boldsymbol{\varpi} : (\mathbf{C} - \mathbf{C}_0)^{-1} : \boldsymbol{\tau}} + \overline{\boldsymbol{\varpi} : (\boldsymbol{\Gamma}_0 * \boldsymbol{\tau})} = \mathbf{E} : \overline{\boldsymbol{\varpi}}. \tag{9}$$

Furthermore

$$\mathcal{H}(\boldsymbol{\tau}) = \tfrac{1}{2}\mathbf{E} : (\mathbf{C}^{\text{eff}} - \mathbf{C}_0) : \mathbf{E}.$$

Theorem 1 states that regardless of the actual stiffness $\mathbf{C}_0$ of the reference material, $\mathcal{H}$ is *stationary* at $\boldsymbol{\tau}$ defined by (8), where $\boldsymbol{\varepsilon}$ solves (1). Theorem 2 below provides sufficient conditions on the stiffness $\mathbf{C}_0$ of the reference material for $\mathcal{H}$ to be *extremum*.

Theorem 2. *If the reference material* $\mathbf{C}_0$ *is stiffer (resp. softer) than the heterogeneous material, then* $\mathcal{H}$ *is minimum (resp. maximum) at* $\boldsymbol{\tau}$*. More precisely,*

i. *if* $\mathbf{C}(\mathbf{x}) - \mathbf{C}_0$ *is positive semidefinite at every point* $\mathbf{x} \in \Omega$*, then* $\mathcal{H}$ *is maximum at* $\boldsymbol{\tau}$*,*
ii. *if* $\mathbf{C}(\mathbf{x}) - \mathbf{C}_0$ *is negative semidefinite at every point* $\mathbf{x} \in \Omega$*, then* $\mathcal{H}$ *is minimum at* $\boldsymbol{\tau}$*.*

It is recalled that, given two fourth-rank, symmetric tensors $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \geq \mathbf{B}$ (resp. $\mathbf{A} > \mathbf{B}$) stands for "$\mathbf{A} - \mathbf{B}$ is positive semidefinite" (resp. positive definite).

2.4. Variational formulation of theorem 1

The remainder of this paper is dedicated to the analysis of two numerical schemes for the approximate solution of problem (1). In this section, we introduce the variational formulation of theorem 1, which will prove for both schemes to be a natural and convenient mathematical framework.

We start with the energy-based scheme of Brisard and Dormieux [13], which is based on theorem 2. Instead of seeking the absolute optimum of $\mathcal{H}$, this functional is optimized on the sub-space of cell-wise constant polarization fields, where it can be computed *exactly* and *efficiently* by FFT. Besides providing an approximate map of the local polarization field, this approach also leads to rigorous (*exact*) bounds on the macroscopic properties of the composite.

Despite its efficiency, this scheme has some limitations. Indeed, in practical applications, the assumptions of theorem 2 cannot always be fulfilled. For example, when the composite contains both pores *and* rigid phases, no reference material (with finite stiffness) can be found, which is stiffer (resp. softer) than all constituants of the composite. In such a case, only theorem 1 remains relevant, and the polarization field $\boldsymbol{\tau}$ is no longer an extremum of $\mathcal{H}$, but merely a *saddle-point*, which we would like to estimate. In other words, general analysis of the scheme proposed in [13] must rely on theorem 1, and not on theorem 2.

While in theorem 1, equation (9) was derived with reference to the original elasticity problem (1), the point of departure adopted in the remainder of this work is slightly different. Indeed, in what follows, (9) will be considered as the equation to be solved, *with no direct reference to the original elasticity problem*. From this perspective, the principal unknown is now the polarization field $\boldsymbol{\tau}$ (in place of the strain field $\boldsymbol{\varepsilon}$), and theorem 1 takes the standard variational form

$$\text{Find } \boldsymbol{\tau} \in \mathbb{V} \quad \text{such that } a(\boldsymbol{\tau}, \boldsymbol{\varpi}) = \ell(\boldsymbol{\varpi}) \quad \text{for all } \boldsymbol{\varpi} \in \mathbb{V}, \tag{10}$$

where the following bilinear (resp. linear) form a (resp. ℓ) has been introduced

$$a(\boldsymbol{\tau}, \boldsymbol{\varpi}) = \overline{\boldsymbol{\varpi} : (\mathbf{C} - \mathbf{C}_0)^{-1} : \boldsymbol{\tau}} + \overline{\boldsymbol{\varpi} : (\boldsymbol{\Gamma}_0 * \boldsymbol{\tau})}, \tag{11a}$$

$$\ell(\boldsymbol{\varpi}) = \mathbf{E} : \overline{\boldsymbol{\varpi}}, \tag{11b}$$

and the functional space $\mathbb{V}$ remains to be defined. Using a formulation such as (10) is attractive for several reasons. It will obviously allow us to resort to standard mathematical tools of the finite element theory [17, 22].

Most importantly, it can easily be verified that (10) is in fact the weak form of the Lippmann-Schwinger equation, on which the basic scheme of Moulinec and Suquet [9, 10] is based. From this perspective, the variational approach (10) unifies the basic and energy-based schemes (see section 4.3).

Before we address its discretization, we must first carry out the mathematical analysis of the initial problem (10) itself. This is done in section 3.

3. Mathematical analysis of problem (10)

3.1. Outline of this section

The aim of section 3 is to show that (10) is a well-posed variational problem in the sense of Hadamard [17]. Such analysis requires an appropriate mathematical setting. First of all, the space $\mathbb{V}$ of polarization fields must be defined. This is done in section 3.2, where the boundedness of a is further established under assumption 1 stated below.

In section 3.3, two alternative expressions (namely (14a) and (14b)) of a are then obtained. Although similar expressions can already be found in [23], the underlying regularity assumptions are incompatible with the functional space $\mathbb{V}^h$; new, detailed proofs are therefore provided in Appendix A.1. Expressions (14a) and (14b) in turn lead to bounds on a, see equation (15).

All ingredients are then gathered to study the well-posedness of problem (10), the proof of which is sketched in section 3.4 (all details being provided in Appendix A.2). It reduces to the verification of the hypotheses of the Banach–Nečas–Babuška theorem (see e.g. [17, theorem 2.6], or [22, theorem 5.2.1]).

3.2. On the space of polarization fields $\mathbb{V}$

The appropriate functional space $\mathbb{V}$ is hinted at by the very structure of (10). Given that the following volume average

$$\overline{\boldsymbol{\tau}_1 : \boldsymbol{\tau}_2} = \frac{1}{|\Omega|} \int_\Omega \boldsymbol{\tau}_1(\mathbf{x}) : \boldsymbol{\tau}_2(\mathbf{x}) \, \mathrm{d}\Omega \tag{12}$$

defines a scalar product on the space of symmetric, square integrable tensors, it is natural to define $\mathbb{V}$ as follows

$$\mathbb{V} = \left\{ \boldsymbol{\tau}, \quad \tau_{ij} = \tau_{ji} \in \mathrm{L}^2(\Omega), \quad i, j = 1, \dots, d \right\}.$$

It is well-known [24] that the scalar product (12) confers the structure of Hilbert space to $\mathbb{V}$; the associated norm reads

$$\|\boldsymbol{\tau}\|_{\mathbb{V}} = (\overline{\boldsymbol{\tau} : \boldsymbol{\tau}})^{1/2} = \left[\frac{1}{|\Omega|}\int_{\Omega}\|\boldsymbol{\tau}(\mathbf{x})\|^2 \,\mathrm{d}\,\Omega\right]^{1/2}.$$

It has already been assumed that the reference material $\mathbf{C}_0$ was isotropic, with shear modulus μ_0, and Poisson ratio ν_0. For convenience, the present analysis is restricted to isotropic composites. More precisely, the material at every point $\mathbf{x} \in \Omega$ is isotropic, with bulk (resp. shear) modulus $\kappa(\mathbf{x})$ (resp. $\mu(\mathbf{x})$)

$$\mathbf{C}(\mathbf{x}) = d\kappa(\mathbf{x})\mathbf{J} + 2\mu(\mathbf{x})\mathbf{K}$$

where $d = 2, 3$ is the dimension of the physical space, $\mathbf{J} = \mathbf{i}\otimes\mathbf{i}\,/d$ is the fourth rank spherical projector, and $\mathbf{K} = \mathbf{I} - \mathbf{J}$ is the fourth rank deviatoric projector. In the remainder of this paper, we will also require $\mathbf{C}$ to be square-integrable.

It is clear from (6) and Parseval's theorem that the bilinear form

$$(\boldsymbol{\tau}, \boldsymbol{\varpi}) \in \mathbb{V}^2 \mapsto \overline{\boldsymbol{\varpi} : (\boldsymbol{\Gamma}_0 * \boldsymbol{\tau})}$$

is well-defined, and bounded. In order to prove that the first term of a (see equation (11a)) is also well-defined, and bounded, some further assumptions must be made.

Indeed, because of the factor $[\mathbf{C}(\mathbf{x}) - \mathbf{C}_0]^{-1}$, occurences of the case $\mathbf{C}(\mathbf{x}) = \mathbf{C}_0$ must be eliminated. In fact, in order to ensure boundedness, a minimum constrast between the stiffness of the reference material and the local stiffness of the composite will be required.

Assumption 1. *There exists $\lambda > 0$ such that at any point $\mathbf{x} \in \Omega$*

$$\begin{aligned} &\textit{either} \quad [\mathbf{C}(\mathbf{x}) - \mathbf{C}_0 - \lambda\mathbf{I}] \quad \textit{is positive semidefinite},\\ &\quad \textit{or} \quad [\mathbf{C}(\mathbf{x}) - \mathbf{C}_0 + \lambda\mathbf{I}] \quad \textit{is negative semidefinite}. \end{aligned}$$

Under assumption 1, it is readily verified that for any polarization fields $\boldsymbol{\tau}, \boldsymbol{\varpi} \in \mathbb{V}$

$$|\overline{\boldsymbol{\varpi} : (\mathbf{C} - \mathbf{C}_0)^{-1} : \boldsymbol{\tau}}| \leq \frac{1}{\lambda}\|\boldsymbol{\varpi}\|_{\mathbb{V}}\|\boldsymbol{\tau}\|_{\mathbb{V}},$$

from which the boundedness of a follows.

Assumption 1 might seem stringent. However, the proofs presented here could be readily extended to a slightly more general case, which would cover most practical applications. In this more general case, the local stiffness $\mathbf{C}(\mathbf{x})$ is allowed to coincide with the stiffness of the reference materials $\mathbf{C}_0$ for all $x \in \Omega_0 \subset \Omega$. For all $\mathbf{x} \notin \Omega_0$, assumption 1 must be verified. Then all results presented here remain valid, provided that an additional constraint is imposed on the polarization field $\boldsymbol{\tau}$

$$\text{For all } \mathbf{x} \in \Omega_0, \quad \boldsymbol{\tau}(\mathbf{x}) = \mathbf{0},$$

which is easily implemented numerically. For the sake of simplicity, this presentation is restricted to the less general framework of assumption 1.

3.3. Alternative expressions of a

The following theorem 3 and its corollary are the key to the proof of the well-posedness of problem (10). They have previously been stated by Willis [23], and are proved here without making any regularity assumptions on the polarization field (besides square-integrability).

Theorem 3. *Let $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2 \in \mathbb{V}$ be two arbitrary polarization fields, and consider the associated strain fields $\boldsymbol{\varepsilon}_i \in \mathbb{V}$ and stress fields $\boldsymbol{\sigma}_i \in \mathbb{V}$ ($i = 1, 2$)*

$$\boldsymbol{\varepsilon}_i = -\boldsymbol{\Gamma}_0 * \boldsymbol{\tau}_i, \qquad \boldsymbol{\sigma}_i = \mathbf{C}_0 : \boldsymbol{\varepsilon}_i + \boldsymbol{\tau}_i. \tag{13}$$

Then

$$a(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2) = \overline{\boldsymbol{\tau}_1 : (\mathbf{C} - \mathbf{C}_0)^{-1} : \boldsymbol{\tau}_2} + \overline{\boldsymbol{\varepsilon}_1 : \mathbf{C}_0 : \boldsymbol{\varepsilon}_2}, \tag{14a}$$

$$= \overline{\boldsymbol{\tau}_1 : \mathbf{S}_0 : (\mathbf{S}_0 - \mathbf{S})^{-1} : \mathbf{S}_0 : \boldsymbol{\tau}_2} - \overline{\boldsymbol{\sigma}_1 : \mathbf{S}_0 : \boldsymbol{\sigma}_2}, \tag{14b}$$

where $\mathbf{S}(\mathbf{x}) = \mathbf{C}(\mathbf{x})^{-1}$, and $\mathbf{S}_0 = \mathbf{C}_0^{-1}$ are the local and reference compliances.

Proof. First, equations (A.9) and (A.10), found in the paper by Willis [23] must be extended to $\mathbb{V}$. This is carried out in Appendix A.1 (see lemmas 1 and 2). The proof proposed in [23] is then unchanged. □

The following corollary is an immediate consequence of theorem 3 and the positive-definiteness of the reference (resp. local) stiffness $\mathbf{C}_0$ (resp. $\mathbf{C}(\mathbf{x})$) and compliance $\mathbf{S}_0$ (resp. $\mathbf{S}(\mathbf{x})$).

Corollary 1. *a is self-adjoint, and for all $\boldsymbol{\tau} \in \mathbb{V}$*

$$\overline{\boldsymbol{\tau} : (\mathbf{C} - \mathbf{C}_0)^{-1} : \boldsymbol{\tau}} \leq a(\boldsymbol{\tau}, \boldsymbol{\tau}) \leq \overline{\boldsymbol{\tau} : \mathbf{S}_0 : (\mathbf{S}_0 - \mathbf{S})^{-1} : \mathbf{S}_0 : \boldsymbol{\tau}}. \tag{15}$$

3.4. Well-posedness of problem (10)

The well-posedness of problem (10), essential for the present analysis, is asserted by use of the Banach–Nečas–Babuška theorem. According to this theorem [17, theorem 2.6], two necessary and sufficient conditions must be met by the bilinear form a. That these conditions are indeed verified in the present case is stated in theorem 4 below, the proof of which is established in the present paper under the following assumption.

Assumption 2. *There exists $\kappa_{\min} > 0$ and $\mu_{\min} > 0$ such that at any point $\mathbf{x} \in \Omega$*

$$\kappa(\mathbf{x}) \geq \kappa_{\min}, \qquad \mu(\mathbf{x}) \geq \mu_{\min}.$$

It should be noted that, since $\kappa_{\min} > 0$ and $\mu_{\min} > 0$, *porous* media are excluded from the present discussion. Although numerical experiments show that in situations of practical interest, involving porous media, problem (10) seems well-posed, and the numerical schemes developed hereafter are well-behaved, the formal proof of these assertions is probably more involved than the present argument, since connectivity of the non-empty regions of Ω must presumably play a role.

Theorem 4. *Under assumption 2, a has the following properties*

i. *There exists* $\alpha > 0$ *such that*

$$\inf_{\boldsymbol{\tau}\in\mathbb{V}} \sup_{\boldsymbol{\varpi}\in\mathbb{V}} \frac{a(\boldsymbol{\tau},\boldsymbol{\varpi})}{\|\boldsymbol{\tau}\|_{\mathbb{V}}\|\boldsymbol{\varpi}\|_{\mathbb{V}}} \geq \alpha.$$

ii. *Let* $\boldsymbol{\tau} \in \mathbb{V}$. *If, for any* $\boldsymbol{\varpi} \in \mathbb{V}$, $a(\boldsymbol{\tau},\boldsymbol{\varpi}) = 0$, *then* $\boldsymbol{\tau} = \mathbf{0}$.

Outline of the proof. The first statement will be proved if we exhibit $\alpha > 0$ such that, for any $\boldsymbol{\tau} \in \mathbb{V}$, there exists $\boldsymbol{\varpi} \in \mathbb{V}$ verifying

$$a(\boldsymbol{\tau},\boldsymbol{\varpi}) \geq \alpha\|\boldsymbol{\tau}\|_{\mathbb{V}}\|\boldsymbol{\varpi}\|_{\mathbb{V}}. \tag{16}$$

The main argument of the proof comes from the fact that (14a) states the positivity of a, if $\mathbf{C} \geq \mathbf{C}_0$ everywhere in Ω; then $\boldsymbol{\varpi} = \boldsymbol{\tau}$ can be shown to satisfy (16) for an appropriate choice of α. Conversely, if $\mathbf{C} \leq \mathbf{C}_0$ everywhere in Ω, then (14b) shows that a is negative, and $\boldsymbol{\varpi} = -\boldsymbol{\tau}$ again verifies (16)

In the most general case, the complete proof of this theorem must account for the fact that the set of stiffness tensors is not *totally* ordered. In other words, points $\mathbf{x} \in \Omega$ may be found, where neither $\mathbf{C}(\mathbf{x}) > \mathbf{C}_0$, nor $\mathbf{C}(\mathbf{x}) < \mathbf{C}_0$.

Instead of a direct comparison of the stiffness *tensors* $\mathbf{C}(\mathbf{x})$ and $\mathbf{C}_0$, the *moduli* must then be compared, which requires the decomposition of $\boldsymbol{\tau}$ into hydrostatic ($\boldsymbol{\tau}^{\mathrm{hyd}}$) and deviatoric ($\boldsymbol{\tau}^{\mathrm{dev}}$) parts; at any point $\mathbf{x} \in \Omega$, $\boldsymbol{\varpi}(\mathbf{x})$ is then defined as $\boldsymbol{\varpi}(\mathbf{x}) = \pm\boldsymbol{\tau}^{\mathrm{hyd}}(\mathbf{x}) \pm \boldsymbol{\tau}^{\mathrm{dev}}(\mathbf{x})$.

Detailed construction of $\boldsymbol{\varpi}$, as well as verification of (16) can be found in Appendix A.2, where the second assertion is also proved by contradiction. □

A simple application of the Banach–Nečas–Babuška theorem then leads to the following conclusion.

Corollary 2. *Problem* (10) *is well-posed.*

The results presented in section 3 have very important fundamental consequences. In a way, they generalize the results established by Hashin and Shtrikman [14] and Willis [23], since existence and uniqueness of the solution of (10) is true whether $\mathbf{C}(\mathbf{x}) - \mathbf{C}_0$ is positive, negative, or undetermined.

In general, problem (10) is not solvable analytically, and numerical schemes must be devised to find an approximate solution. The FFT-based schemes proposed by Moulinec and Suquet [9, 10, 16] and Brisard and Dormieux [13] are two examples of such numerical schemes. In the next section, it will be proved that both methods can be viewed as Galerkin finite-elements approximations of the initial problem (10). With this new approach of existing schemes, new convergence results will be established, and some requirements on the reference material, which were previously thought to be necessary, will be alleviated.

4. Numerical approximation of problem (10)

4.1. General setting

Galerkin-like discretizations of problem (10) are obtained by selecting a finite-dimensional space of *trial* and *test* functions $\mathbb{V}^h$. This space depends on a numerical parameter h, which will ultimately tend to 0. Then (10) is replaced by the following discretized problem

$$\begin{aligned} &\text{Find } \boldsymbol{\tau}^h \in \mathbb{V}^h \quad \text{such that } a^h(\boldsymbol{\tau}^h,\boldsymbol{\varpi}^h) = \ell(\boldsymbol{\varpi}^h) \\ &\quad \text{for all } \boldsymbol{\varpi}^h \in \mathbb{V}^h, \end{aligned} \tag{17}$$

where the bilinear form a^h approximate (in a sense which will be made more precise) a. Problem (17) evidently reduces to a linear system, since $\mathbb{V}^h$ is finite-dimensional.

In this section we show that the numerical scheme of Moulinec and Suquet [10], as well as the numerical scheme of Brisard and Dormieux [13] can be viewed as two different Galerkin-like approximations of problem (10), based on the *same* space of trial functions $\mathbb{V}^h$, but *different* approximations of a.

We start with the definition of the discretization parameter h and the space of trial and test functions $\mathbb{V}^h$. Then the two versions of the discretized problem (17) are studied separately, in order to assert their well-posedness and h–convergence.

It should be noted that the present analysis is greatly simplified by two facts. First, the spaces of trial and test functions coincide, and are *included* in the initial space $\mathbb{V}$, from which they therefore inherit the norm. Second, unlike the bilinear form a, the linear form ℓ is computed *exactly* in both schemes on $\mathbb{V}^h$.

We now come to the definition of the space of trial and test functions, $\mathbb{V}^h$, which is to remain unchanged throughout this section. The origin of the success of numerical schemes based on the discretization of the Lippmann-Schwinger equation [9, 10, 13, 15, 16, 25, 26] lies in the fact that the matrix of the underlying linear system is the sum of two matrices with noteworthy properties. The first matrix is block-diagonal, and its product with a vector is performed efficiently in the real space. The second matrix corresponds to a translation-invariant linear operator, and its product with a vector is computed most efficiently in the Fourier space by use of the FFT.

Invoking the FFT requires the use of a d-dimensional regular mesh on Ω, each cell of this mesh being a d-dimensional cube of measure h^d (pixel in plane elasticity, voxel in three-dimensional elasticity). Let N_i be the number of cells in the i-th direction ($i = 1,\dots,d$), and $N = N_1 \cdots N_d$ the total number of cells. Taking advantage of the periodic boundary conditions, it is convenient (and equivalent) to seek an estimate of the exact polarization field $\boldsymbol{\tau}$ on Ω, translated by the vector $-\frac{h}{2}(\mathbf{e}_1 + \cdots + \mathbf{e}_d)$, rather than Ω itself. Under these conditions, cell $\Omega^h_{\boldsymbol{\beta}}$ of the current mesh is centered at point $\mathbf{x}^h_{\boldsymbol{\beta}}$, given by

$$\mathbf{x}^h_{\boldsymbol{\beta}} = \beta_1 h\mathbf{e}_1 + \cdots + \beta_d h\mathbf{e}_d, \tag{18}$$

where $\boldsymbol{\beta}$ denotes any multi-index in the following set

$$\mathcal{I}^h = \{0,\dots,N_1-1\} \times \cdots \times \{0,\dots,N_d-1\}.$$

Finally, the characteristic function of cell $\Omega^h_{\boldsymbol{\beta}}$ will be denoted $\mathbf{x} \mapsto \chi^h_{\boldsymbol{\beta}}(\mathbf{x})$. Having defined the mesh on which the approximate solution is to be computed, we chose to use cell-wise constant functions as trial functions. In other words, $\mathbb{V}^h$ is defined as the space of polarization fields $\boldsymbol{\tau}^h(\mathbf{x})$ of the form

$$\boldsymbol{\tau}^h(\mathbf{x}) = \sum_{\boldsymbol{\beta}\in\mathcal{I}^h} \chi^h_{\boldsymbol{\beta}}(\mathbf{x})\boldsymbol{\tau}^h_{\boldsymbol{\beta}}, \tag{19}$$

where the $\boldsymbol{\tau}^h_\beta$ are constant, symmetric tensors. $\mathbb{V}^h$ thus defined is obviously a sub-space of $\mathbb{V}$, and can be equipped with the same norm. The following important theorem easily follows from the density of continuous functions in $\mathrm{L}^2(\Omega)$.

Theorem 5. *The approximation setting has the* approximability property *in the sense of Ern and Guermond [17], definition 2.14. In other words*

$$\textit{For all } \boldsymbol{\tau} \in \mathbb{V}, \quad \lim_{h\to 0} \inf_{\boldsymbol{\tau}^h \in \mathbb{V}^h} \|\boldsymbol{\tau} - \boldsymbol{\tau}^h\|_{\mathbb{V}} = 0.$$

It is emphasized that any cell-wise constant polarization field $\boldsymbol{\tau}^h$ is uniquely associated with the set of its indexed values $\boldsymbol{\tau}^h_\beta$ on cell Ω^h_β. In the remainder of this paper, $\boldsymbol{\tau}^h_\beta$ will abusively refer either to the discrete set of values, or to the corresponding cell-wise constant polarization field.

Sub-sections 4.2 and 4.3 are devoted to the separate analysis of two different Galerkin-type discretizations of problem (10); the two methods differ by the underlying approximation of the bilinear form a. From the historical viewpoint, the basic scheme of Moulinec and Suquet [9, 10] should logically be addressed first. However, being a non-consistent Galerkin approximation of (10), its theoretical analysis is more involved than the energy-based scheme of Brisard and Dormieux [13], which is consistent. We will therefore first discuss the energy-based scheme, then the basic scheme.

4.2. Consistent Galerkin discretization [13]

The method proposed in [13] was based on theorem 2: an approximate solution of (10) was found by optimization of the energy $\mathcal{H}$ of Hashin and Shtrikman [14] on the space of cell-wise constant polarization fields. Using the notation of the present paper, this amounts to solving the following problem

$$\text{Optimize} \quad \mathcal{H}(\boldsymbol{\tau}^h) = \frac{1}{2} a(\boldsymbol{\tau}^h, \boldsymbol{\tau}^h) - \ell(\boldsymbol{\tau}^h), \quad \text{for } \boldsymbol{\tau}^h \in \mathbb{V}^h. \tag{20}$$

It was shown that this discretized problem could be efficiently solved by a combination of the conjugate gradient method for inversion of the underlying (symmetric, definite) linear system, as well as the FFT for the evaluation of the necessary matrix-vector products. Furthermore, operators a and ℓ were computed *exactly* on $\mathbb{V}^h$ [13].

However, one restriction of this approach lies in the assumptions made on the reference material (see theorem 2). Indeed, in order to ensure existence and uniqueness of the optimum of $\mathcal{H}$, the reference material must be either stiffer, or softer than all phases in the composite.

The initial goal of the present paper was to address the following question: is it mathematically sound to select a reference material which *violates* the above condition? As shown below, the answer to this question turns out to be "yes". However, the discretized problem at hand is now a *saddle-point* problem, and well-posedness, as well as h–convergence towards the solution of the initial problem (10) must be carefully proved. So, in place of (20), the following discretized problem is now considered

$$\begin{aligned} &\text{Find } \boldsymbol{\tau}^h \in \mathbb{V}^h \quad \text{such that } a(\boldsymbol{\tau}^h, \boldsymbol{\varpi}^h) = \ell(\boldsymbol{\varpi}^h) \\ &\text{for all } \boldsymbol{\varpi}^h \in \mathbb{V}^h. \end{aligned} \tag{21}$$

Comparison with (17) shows that the discretized problem (21) is indeed of the Galerkin type. It is *conformal*, since $\mathbb{V}^h \subset \mathbb{V}$, and *consistent*, since operators a and ℓ are computed *exactly* on $\mathbb{V}^h$. In other words the solution $\boldsymbol{\tau} \in \mathbb{V}$ of the exact problem (10) satisfies the approximate problem (21)

$$\text{For all } \boldsymbol{\varpi}^h \in \mathbb{V}^h, \quad a(\boldsymbol{\tau}, \boldsymbol{\varpi}^h) = \ell(\boldsymbol{\varpi}^h).$$

In order to prove the well-posedness of problem (21), use will again be made of the Banach–Nečas–Babuška theorem. Before we proceed to verify that this theorem applies in the present case, a few words must be said on the *exact* evaluation of a and ℓ on $\mathbb{V}^h$. For any trial (resp. test) field $\boldsymbol{\tau}^h \in \mathbb{V}^h$ (resp. $\boldsymbol{\varpi}^h \in \mathbb{V}^h$), $\boldsymbol{\tau}^h_\beta$ (resp. $\boldsymbol{\varpi}^h_\beta$) is defined as in (19). It is then readily verified that

$$\ell(\boldsymbol{\tau}^h) = \frac{1}{N} \sum_{\beta \in \mathcal{I}^h} \mathbf{E} : \boldsymbol{\tau}^h_\beta,$$

and similarly

$$\overline{\boldsymbol{\varpi}^h : (\mathbf{C} - \mathbf{C}_0)^{-1} : \boldsymbol{\tau}^h} = \frac{1}{N} \sum_{\beta \in \mathcal{I}^h} \boldsymbol{\varpi}^h_\beta : \left(\mathbf{C}^h_\beta - \mathbf{C}_0\right)^{-1} : \boldsymbol{\tau}^h_\beta, \tag{22}$$

where $\mathbf{C}^h_\beta$ is the *consistent* equivalent stiffness of the cell Ω^h_β, defined by

$$\left(\mathbf{C}^h_\beta - \mathbf{C}_0\right)^{-1} = \frac{1}{|\Omega^h_\beta|} \int_{\Omega^h_\beta} [\mathbf{C}(\mathbf{x}) - \mathbf{C}_0]^{-1} \, \mathrm{d}\,\Omega. \tag{23}$$

From the practical point of view, the previous expression is of great interest, since it allows the exact evaluation on $\mathbb{V}^h \times \mathbb{V}^h$ of the first term of a (11a), *even in the frequent case of an heterogeneous cell*. From the mathematical point of view, (23) means that for trial and test functions in $\mathbb{V}^h$, the initial composite $\mathbf{x} \mapsto \mathbf{C}(\mathbf{x})$ is strictly equivalent to a fictitious composite $\mathbf{x} \mapsto \mathbf{C}^h(\mathbf{x})$ with cell-wise constant stiffness $\mathbf{C}^h_\beta$

$$\mathbf{C}^h(\mathbf{x}) = \sum_{\beta \in \mathcal{I}^h} \chi^h_\beta(\mathbf{x}) \mathbf{C}^h_\beta.$$

This equivalence will be invoked below to prove theorem 6 by analogy with theorem 4. Full evaluation of $a(\boldsymbol{\tau}^h, \boldsymbol{\varpi}^h)$ also requires the evaluation of the non-local term. In [13], the following *exact* expression was derived

$$\overline{\boldsymbol{\varpi}^h : (\boldsymbol{\Gamma}_0 * \boldsymbol{\tau}^h)} = \frac{1}{N^2} \sum_{\mathbf{b} \in \mathbb{Z}^d} [F(h\mathbf{k}_\mathbf{b})]^2 \, \hat{\boldsymbol{\varpi}}^{h*}_\mathbf{b} : \hat{\boldsymbol{\Gamma}}_0(\mathbf{k}_\mathbf{b}) : \hat{\boldsymbol{\tau}}^h_\mathbf{b}, \tag{24}$$

where the DFT $\hat{\boldsymbol{\varpi}}^h_\mathbf{b}$ (resp. $\hat{\boldsymbol{\tau}}^h_\mathbf{b}$) of the *finite* sequence $\boldsymbol{\varpi}^h_\beta$ ($\boldsymbol{\tau}^h_\beta$) has been introduced

$$\hat{\boldsymbol{\tau}}^h_\mathbf{b} = \sum_{\beta \in \mathcal{I}^h} \exp\left[-2\imath\pi \left(\frac{\beta_1 b_1}{N_1} + \cdots \frac{\beta_d b_d}{N_d}\right)\right] \boldsymbol{\tau}^h_\beta, \tag{25}$$

as well as the following product of sine cardinal functions

$$F(\mathbf{K}) = \operatorname{sinc} \frac{K_1}{2} \cdots \operatorname{sinc} \frac{K_d}{2}. \tag{26}$$

Owing to the periodicity of the DFT, it was further shown in [13] that the *infinite series* (24) on $\mathbb{Z}^d$ could be reduced to the *finite sum* on $\mathcal{I}^h$

$$\overline{\boldsymbol{\varpi}^h : (\boldsymbol{\Gamma}_0 * \boldsymbol{\tau}^h)} = \frac{1}{N^2} \sum_{\mathbf{b} \in \mathcal{I}^h} \hat{\boldsymbol{\varpi}}^{h*}_{\mathbf{b}} : \hat{\boldsymbol{\Gamma}}^{h,\mathrm{c}}_{0,\mathbf{b}} : \hat{\boldsymbol{\tau}}^h_{\mathbf{b}}, \tag{27}$$

introducing the Fourier components of the *consistent* discrete Green operator

$$\hat{\boldsymbol{\Gamma}}^{h,\mathrm{c}}_{0,\mathbf{b}} = \sum_{\mathbf{n} \in \mathbb{Z}^d} [F(h\mathbf{k}_{\mathbf{b}+\mathbf{nN}})]^2 \, \hat{\boldsymbol{\Gamma}}_0(\mathbf{k}_{\mathbf{b}+\mathbf{nN}}), \tag{28}$$

where $\mathbf{b}+\mathbf{nN}$ denotes the multi-index $(b_1+n_1N_1, \ldots, b_d+n_dN_d)$, and $\mathbf{k}_{\mathbf{b}}$ is defined by (4). In a more compact form, we have, for $\boldsymbol{\tau}^h, \boldsymbol{\varpi}^h \in \mathbb{V}^h$

$$\overline{\boldsymbol{\varpi}^h : (\boldsymbol{\Gamma}_0 * \boldsymbol{\tau}^h)} = \overline{\boldsymbol{\varpi}^h : \left(\boldsymbol{\Gamma}^{h,\mathrm{c}}_0 * \boldsymbol{\tau}^h\right)}.$$

It should be noted that this operator was referred to as the *periodized Green operator* in [13]. The new terminology emphasizes the difference with the *non-consistent* discrete Green operator, defined in section 4.3 below.

The consistent discrete Green operator is pre-computed and stored for further use. It can be evaluated very efficiently in plane strain elasticity[2]. It is noted however that in the three-dimensional case, the infinite series involved in (28) converge very slowly, making the numerical evaluation of the consistent discrete Green operator rather involved.

Finally, gathering (22), (23), (27) and (28), it is found that $a(\boldsymbol{\tau}^h, \boldsymbol{\varpi}^h)$ can be computed *exactly* and *efficiently* (DFTs being computed by means of the FFT) on $\mathbb{V}^h \times \mathbb{V}^h$.

We now turn to the well-posedness of the discretized problem (21). In view of invoking the Banach–Nečas–Babuška theorem, we must verify the conditions under which it is stated. Since $\mathbb{V}^h$ is of finite dimension, it is sufficient to verify one of the two conditions stated in theorem 4. This is done in theorem 6 below.

Theorem 6. *Under assumption 2, there exists $\alpha > 0$ such that*

$$\inf_{\boldsymbol{\tau}^h \in \mathbb{V}^h} \sup_{\boldsymbol{\varpi}^h \in \mathbb{V}^h} \frac{a\left(\boldsymbol{\tau}^h, \boldsymbol{\varpi}^h\right)}{\|\boldsymbol{\tau}^h\|_{\mathbb{V}} \|\boldsymbol{\varpi}^h\|_{\mathbb{V}}} \geq \alpha.$$

Proof. The equivalence between $\mathbf{C}$ and $\mathbf{C}^h$ is used to prove this theorem. Selecting a fixed trial field $\boldsymbol{\tau}^h \in \mathbb{V}^h$, we construct $\boldsymbol{\varpi}^h$ using the same procedure as for the construction of $\boldsymbol{\varpi}$ in theorem 4, except that decisions are now made depending on the sign of $\mathbf{C}^h - \mathbf{C}_0$, instead of $\mathbf{C} - \mathbf{C}_0$. $\boldsymbol{\varpi}^h$ thus constructed evidently belongs to $\mathbb{V}^h$.

Since in the consistent approach, a is computed exactly, (14a) and (14b) apply to $\boldsymbol{\tau}^h$ and $\boldsymbol{\varpi}^h$. The proof then proceeds as in theorem 4. □

Application of the Banach–Nečas–Babuška theorem in finite dimension leads to the well-posedness of problem (21) for *any* choice of the reference material. This means that the linear system resulting from the discretized problem (21) always has a *unique* solution, even if the reference material fails to be stiffer (or softer) than all the constituants of the composite.

Besides, the assumptions for Céa's lemma are satisfied (see [17], lemma 2.28), and we therefore conclude that the piecewise-constant solution $\boldsymbol{\tau}^h$ to the discretized problem (21) converges (in the L^2–sense) to the solution $\boldsymbol{\tau}$ to the initial problem (10).

4.3. Non-consistent Galerkin discretization [9, 10]

As noted in the previous section, the consistent Galerkin discretization of (10) requires the calculation of the consistent discrete Green operator (28), which is difficult –in the three-dimensional case– because of the slow convergence of the underlying series.

Moulinec and Suquet [9, 10] proposed a discretization of the Lippmann-Schwinger equation, in which the Green operator is simply approximated by a *truncated* Fourier series. Again, this approach leads to an efficient implementation, because of the use of the FFT for matrix-vector products.

The purpose of this section is to cast the basic scheme of Moulinec and Suquet [9, 10] in an appropriate mathematical framework, so as to prove its well-posedness and h–convergence to the solution of the initial problem (10). We first show that the basic scheme can be viewed as a Galerkin discretization of problem (10); the bilinear form a is not computed exactly and the approximation is in fact *non-consistent*. We then show that the approximate bilinear form a^h is *asymptotically consistent*, which leads to convergence results of the approximate solution.

The basic scheme of Moulinec and Suquet [9, 10] finds its roots in the strong form of equation (9)

$$(\mathbf{C} - \mathbf{C}_0)^{-1} : \boldsymbol{\tau} + \boldsymbol{\Gamma}_0 * \boldsymbol{\tau} = \mathbf{E}$$

which, upon substitution of the strain field $\boldsymbol{\varepsilon}$ to the polarization field $\boldsymbol{\tau}$, can be seen as a *fixed-point problem*

$$\boldsymbol{\varepsilon} = \mathbf{E} - \boldsymbol{\Gamma}_0 * [(\mathbf{C} - \mathbf{C}_0) : \boldsymbol{\varepsilon}],$$

this equation being known as the *Lippmann-Schwinger* equation [12]. The classical iterative algorithm reads

$$\boldsymbol{\varepsilon}^{n+1} = \mathbf{E} - \boldsymbol{\Gamma}_0 * [(\mathbf{C} - \mathbf{C}_0) : \boldsymbol{\varepsilon}^n],$$

it is known to be only *conditionally* convergent [16].

Each iterate $\boldsymbol{\varepsilon}^n$ is discretized on a regular grid, leading to a discrete set of values $\boldsymbol{\varepsilon}^{h,n}_{\boldsymbol{\beta}}$, $\boldsymbol{\beta} \in \mathcal{I}^h$. Moulinec and Suquet [9, 10] suggest that $\boldsymbol{\varepsilon}^{h,n}_{\boldsymbol{\beta}}$ should be understood as a point-wise estimate of $\boldsymbol{\varepsilon}^n$ at point $\mathbf{x}_{\boldsymbol{\beta}}$ defined by (18). However, in the present mathematical framework, it is more natural to consider that $\boldsymbol{\varepsilon}^{h,n}_{\boldsymbol{\beta}}$ is the step value on $\Omega^h_{\boldsymbol{\beta}}$ of a cell-wise constant function $\boldsymbol{\varepsilon}^{h,n}$ which approximates $\boldsymbol{\varepsilon}^h$ in the L^2-sense, rather than a point-wise estimate of $\boldsymbol{\varepsilon}^{h,n}(\mathbf{x}_{\boldsymbol{\beta}})$

$$\boldsymbol{\varepsilon}^n(\mathbf{x}) \simeq \boldsymbol{\varepsilon}^{h,n}(\mathbf{x}), \quad \text{with } \boldsymbol{\varepsilon}^{h,n}(\mathbf{x}) = \sum_{\boldsymbol{\beta} \in \mathcal{I}^h} \chi^h_{\boldsymbol{\beta}}(\mathbf{x}) \boldsymbol{\varepsilon}^{h,n}_{\boldsymbol{\beta}}.$$

In order to compute the direct product $(\mathbf{C} - \mathbf{C}_0) : \boldsymbol{\varepsilon}^n$, the local stiffness $\mathbf{C}$ should also be discretized. However, to the best

[2] The required formulas will be reported in a paper to come.

of our knowledge, no consistent rule has yet been proposed to carry out this critical operation. The analysis below shows that the *consistent* discretization $\mathbf{C}^h$ of $\mathbf{C}$ defined by (23) should be used. In other words, $(\mathbf{C}-\mathbf{C}_0):\boldsymbol{\varepsilon}^n$ is approximated as follows

$$(\mathbf{C}-\mathbf{C}_0):\boldsymbol{\varepsilon}^n \simeq \left(\mathbf{C}^h-\mathbf{C}_0\right):\boldsymbol{\varepsilon}^{h,n},$$

the right-hand side being of course cell-wise constant (product of two cell-wise constant tensors).

Finally, the convolution product occuring in equation (3) is approximated by *truncated* Fourier series, where only the modes of lowest frequency are retained. Introducing the *non-consistent* discrete Green operator $\boldsymbol{\Gamma}_0^{h,\mathrm{nc}}$, the resulting approximation reads

$$\boldsymbol{\Gamma}_0 * \left[(\mathbf{C}-\mathbf{C}_0):\boldsymbol{\varepsilon}^n\right] \simeq \boldsymbol{\Gamma}_0^{h,\mathrm{nc}} * \left[\left(\mathbf{C}^h-\mathbf{C}_0\right):\boldsymbol{\varepsilon}^{h,n}\right].$$

The non-consistent discrete Green operator is defined by its Fourier components

$$\hat{\boldsymbol{\Gamma}}^{h,\mathrm{nc}}_{0,\mathbf{b}+\mathbf{nN}} = \hat{\boldsymbol{\Gamma}}_0(\mathbf{k_b}), \quad \text{for } \mathbf{b}\in\mathcal{J}^h \quad \text{and } \mathbf{n}\in\mathbb{Z}^d, \tag{29}$$

where multi-index $\mathbf{b}\in\mathcal{J}^h$ selects the modes of lowest-frequency

$$\mathcal{J}^h = \left\{-\frac{N_1}{2}+1,\dots,\frac{N_1}{2}\right\}\times\cdots\times\left\{-\frac{N_d}{2}+1,\dots,\frac{N_d}{2}\right\}, \tag{30}$$

while multi-index $\mathbf{n}$ enforces periodicity of the discrete Green operator (in view of invoking the FFT). For the sake of simplicity, it is assumed that $N_1,\dots,N_d$ are even[3].

Finally, for any cell-wise constant polarization field $\boldsymbol{\tau}^h$, $\boldsymbol{\Gamma}_0^{h,\mathrm{nc}}*\boldsymbol{\tau}^h$ is defined as the cell-wise constant tensor $\boldsymbol{\eta}^h$

$$\boldsymbol{\eta}^h(\mathbf{x}) = \sum_{\boldsymbol{\beta}\in\mathcal{I}^h}\chi^h_{\boldsymbol{\beta}}(\mathbf{x})\boldsymbol{\eta}^h_{\boldsymbol{\beta}}, \quad \text{with } \boldsymbol{\eta}^h_{\boldsymbol{\beta}} = \left[\mathrm{DFT}^{-1}\left(\hat{\boldsymbol{\Gamma}}^{h,\mathrm{nc}}_{0,\mathbf{b}}:\hat{\boldsymbol{\tau}}^h_{\mathbf{b}}\right)\right]_{\boldsymbol{\beta}}. \tag{31}$$

The mapping $\boldsymbol{\tau}^h \mapsto \boldsymbol{\Gamma}_0^{h,\mathrm{nc}}*\boldsymbol{\tau}^h$ evidently defines a linear operator from $\mathbb{V}^h$ onto $\mathbb{V}^h$. The iterative scheme of Moulinec and Suquet [9, 10] finally reads

$$\boldsymbol{\varepsilon}^{h,n+1} = \mathbf{E} - \boldsymbol{\Gamma}_0^{h,\mathrm{nc}} * \left[\left(\mathbf{C}^h-\mathbf{C}_0\right):\boldsymbol{\varepsilon}^{h,n}\right]. \tag{32}$$

It should be noted that, unlike its consistent counterpart (28), the non-consistent discrete Green operator is known in closed-form; this is a great asset of the basic scheme of Moulinec and Suquet [9, 10], as no precomputation is necessary.

In order to retrieve a Galerkin formulation of this scheme, we first observe that if (32) converges with n, its limit $\boldsymbol{\varepsilon}^h$ is the solution of the linear system of equations

$$\boldsymbol{\varepsilon}^h + \boldsymbol{\Gamma}_0^{h,\mathrm{nc}} * \left[\left(\mathbf{C}^h-\mathbf{C}_0\right):\boldsymbol{\varepsilon}^h\right] = \mathbf{E}. \tag{33}$$

Michel et al. have shown that if the reference material $\mathbf{C}_0$ fails to satisfy the following conditions [16, eq. (22)]

$$2\kappa_0 > \sup_{\mathbf{x}\in\Omega}\kappa(\mathbf{x}), \quad \text{and } 2\mu_0 > \sup_{\mathbf{x}\in\Omega}\mu(\mathbf{x}), \tag{34}$$

the iterations (32) do not converge. However, it will be shown below that the system (33) *always* has a unique solution, regardless of the reference material $\mathbf{C}_0$. Clearly, the weaknesses of the basic scheme are to be attributed to the method used to solve the discretized problem (namely, the fixed-point algorithm), not to the discretization itself. This point was already recognized by Zeman et al. [15], who replaced the fixed-point iterations with the conjugate and bi-conjugate gradient methods. According to these authors, convergence of these iterative solvers is insensitive to the actual stiffness of the reference material (in other words, Zeman et al. [15] showed experimentally that violation of (34) seems to be allowed).

Keeping in mind that a better suited linear iterative solver may be substituted to (32), we will now focus on the approximation (33) of problem (10), setting aside the iterative aspect of the basic scheme. Using the discretized polarization $\boldsymbol{\tau}^h$ instead of the discretized strain $\boldsymbol{\varepsilon}^h$ as main unknown, (33) reads

$$\left(\mathbf{C}^h-\mathbf{C}_0\right)^{-1}:\boldsymbol{\tau}^h + \boldsymbol{\Gamma}_0^{h,\mathrm{nc}}*\boldsymbol{\tau}^h = \mathbf{E}.$$

Equating those cell-wise constant tensors on each cell $\Omega^h_{\boldsymbol{\beta}}$ gives

$$\left(\mathbf{C}^h_{\boldsymbol{\beta}}-\mathbf{C}_0\right)^{-1}:\boldsymbol{\tau}^h_{\boldsymbol{\beta}} + \boldsymbol{\eta}^h_{\boldsymbol{\beta}} = \mathbf{E}, \quad \text{for all } \boldsymbol{\beta}\in\mathcal{I}^h,$$

where $\boldsymbol{\eta}^h = \boldsymbol{\Gamma}_0^{h,\mathrm{nc}}*\boldsymbol{\tau}^h$. Both sides of the previous equation are contracted with an arbitrary test function $\boldsymbol{\varpi}^h\in\mathbb{V}^h$, and all cell-values are summed

$$\begin{aligned}\frac{1}{N}\sum_{\boldsymbol{\beta}\in\mathcal{I}^h}\boldsymbol{\varpi}^h_{\boldsymbol{\beta}}:\left(\mathbf{C}^h_{\boldsymbol{\beta}}-\mathbf{C}_0\right)^{-1}:\boldsymbol{\tau}^h_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{\boldsymbol{\beta}\in\mathcal{I}^h}\boldsymbol{\varpi}^h_{\boldsymbol{\beta}}:\boldsymbol{\eta}^h_{\boldsymbol{\beta}}\\ = \frac{1}{N}\sum_{\boldsymbol{\beta}\in\mathcal{I}^h}\mathbf{E}:\boldsymbol{\varpi}^h_{\boldsymbol{\beta}}.\end{aligned}$$

Each individual sum can be recognized as the volume average (on Ω) of the underlying cell-wise constant tensor, leading to the compact expression

$$\overline{\boldsymbol{\varpi}^h:\left(\mathbf{C}^h-\mathbf{C}_0\right)^{-1}:\boldsymbol{\tau}^h} + \overline{\boldsymbol{\varpi}^h:\boldsymbol{\eta}^h} = \mathbf{E}:\overline{\boldsymbol{\tau}^h},$$

which can be recognized as a variational (discrete) problem

$$\begin{aligned}&\text{Find } \boldsymbol{\tau}^h\in\mathbb{V}^h \quad \text{such that } a^h(\boldsymbol{\tau}^h,\boldsymbol{\varpi}^h)=\ell(\boldsymbol{\varpi}^h)\\ &\quad\text{for all } \boldsymbol{\varpi}^h\in\mathbb{V}^h,\end{aligned} \tag{35}$$

where the discretized bilinear form a^h reads

$$a^h(\boldsymbol{\tau}^h,\boldsymbol{\varpi}^h) = \overline{\boldsymbol{\varpi}^h:\left(\mathbf{C}^h-\mathbf{C}_0\right)^{-1}:\boldsymbol{\tau}^h} + \overline{\boldsymbol{\varpi}^h:\left(\boldsymbol{\Gamma}_0^{h,\mathrm{nc}}*\boldsymbol{\tau}^h\right)}.$$

Setting aside the question of the (sub-optimal) iterative solver, the above discussion shows that the basic scheme of Moulinec and Suquet [9, 10] is a Galerkin-like discretization of the variational problem (10). We will now proceed to the mathematical analysis (well posedness and convergence as $h\to 0$) of this scheme.

Well-posedness of problem (35) is easily assessed; detailed proofs can be found in Appendix C.1. Quite remarkably, a^h verifies on $\mathbb{V}^h$ the same property (15) as its continuous counterpart

[3] If one of the N_i is odd, then the corresponding $\{-N_i/2+1,\dots,N_i/2\}$ in (30) must be replaced with $\{-(N_i-1)/2,\dots,(N_i-1)/2\}$; the subsequent mathematical analysis is unchanged.

on $\mathbb{V}$. More precisely, for any trial field $\boldsymbol{\tau}^h \in \mathbb{V}^h$,

$$\begin{aligned}\overline{\boldsymbol{\tau}^h : \left(\mathbf{C}^h - \mathbf{C}_0\right)^{-1} : \boldsymbol{\tau}^h} \leq a^h\left(\boldsymbol{\tau}^h, \boldsymbol{\tau}^h\right) \\ \leq \overline{\boldsymbol{\tau}^h : \mathbf{S}_0 : \left(\mathbf{S}_0 - \mathbf{S}^h\right)^{-1} : \mathbf{S}_0 : \boldsymbol{\tau}^h},\end{aligned} \tag{36}$$

where $\mathbf{S}^h$ denotes the element-wise constant compliance $\mathbf{S}^h = (\mathbf{C}^h)^{-1}$. The discrete counterpart of theorem 4 can therefore be deduced: indeed, in the proof of this theorem, the local stiffness $\mathbf{C}(\mathbf{x})$ need only be replaced by the cell-wise constant consistent stiffness $\mathbf{C}^h_\beta$. In other words, problem (35) is well-posed, regardless of the choice of the reference material $\mathbf{C}_0$.

Turning now to the h–convergence of the (unique) solution of (35) to the solution of (10), it must first be proved that the discrete bilinear form a^h approximates (in a way that will be made more precise later) the initial bilinear form a. Since a^h is defined on $\mathbb{V}^h \times \mathbb{V}^h$ (not on $\mathbb{V} \times \mathbb{V}$), a linear mapping Π^h between $\mathbb{V}$ and $\mathbb{V}^h$ must first be defined. In order to apply standard theorems of the finite elements theory [17], this mapping must have the following property

$$\text{For any } \boldsymbol{\tau} \in \mathbb{V}, \quad \|\Pi^h \boldsymbol{\tau} - \boldsymbol{\tau}\|_{\mathbb{V}} \leq c \inf_{\boldsymbol{\tau}^h \in \mathbb{V}^h} \|\boldsymbol{\tau} - \boldsymbol{\tau}^h\|_{\mathbb{V}},$$

where c is independent of the polarization field $\boldsymbol{\tau} \in \mathbb{V}$. It is shown in Appendix B.2 that orthogonal projection onto $\mathbb{V}^h$ is in the present case a convenient choice for Π^h. Then $\Pi^h \boldsymbol{\tau}$ is the cell-average of $\boldsymbol{\tau}$, and $c = 1$.

Theorem 7. *Problem* (35) *is* asymptotically consistent *in the sense of Ern and Guermond [17, definition 2.15]. More precisely, the non-consistent Galerkin approximation of problem* (10) *has the following property*

$$\lim_{h \to 0} \sup_{\boldsymbol{\varpi}^h \in \mathbb{V}^h} \frac{|\ell(\boldsymbol{\varpi}^h) - a^h(\Pi^h \boldsymbol{\tau}, \boldsymbol{\varpi}^h)|}{\|\boldsymbol{\varpi}^h\|_{\mathbb{V}}} = 0, \tag{37}$$

where $\boldsymbol{\tau} \in \mathbb{V}$ denotes the unique solution to problem (10)*, and the linear mapping Π^h has been defined above.*

The proof of this theorem can be found in Appendix C.2. It should be noted that, since $\boldsymbol{\tau}$ is the solution of (10), (37) reduces to

$$\lim_{h \to 0} \sup_{\boldsymbol{\varpi}^h \in \mathbb{V}^h} \frac{|a(\boldsymbol{\tau}, \boldsymbol{\varpi}^h) - a^h(\Pi^h \boldsymbol{\tau}, \boldsymbol{\varpi}^h)|}{\|\boldsymbol{\varpi}^h\|_{\mathbb{V}}} = 0. \tag{38}$$

Application of the lemma of Strang [17, lemma 2.27] then shows that the solution $\boldsymbol{\tau}^h$ to the non-consistent, discretized problem (35) converges (in the L^2-sense) to the solution of the initial problem (10) when $h \to 0$, and closes the analysis of the basic scheme.

4.4. Discussion: consistent vs. non-consistent approaches

We have shown in sub-sections 4.2 and 4.3 that two apparently different numerical schemes for the simulation of heterogeneous materials can be reconciled. Indeed, both of these schemes can be viewed as Galerkin approximations of the same variational problem. The *basic scheme* of Moulinec and Suquet [9, 10] is a *non-consistent* Galerkin approximation of problem (10), while the energy-based scheme of Brisard and Dormieux [13] is a *consistent* approximation of the same problem. Both approximations share many attractive properties.

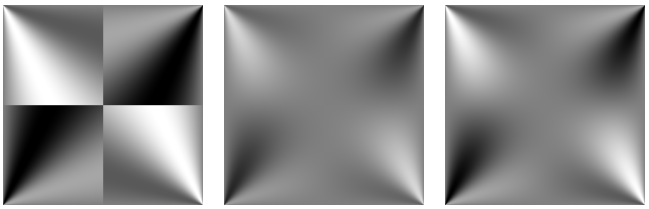


Figure 1: map in Fourier space of the *xyxy* component of the non-consistent (left), consistent (middle), and filtered, non-consistent (right), discrete Green operators. The calculation corresponds to a reference material with unit shear modulus $\mu_0 = 1$, and Poisson ratio $\nu_0 = 0.3$, in 2d elasticity (plane strain). The consistent operator is smooth, while the non-consistent operator exhibits strong discontinuities at the center of the image (corresponding to the highest frequencies). The filtered, non-consistent operator (described in section 5.2) is a good compromise, combining smoothness and ease of computation.

First and foremost, the Green operator $\boldsymbol{\Gamma}_0$ is discretized in both cases in such a way as to allow the use of the FFT, resulting in very efficient schemes (in terms of CPU time). While the calculation of the non-consistent discrete Green operator $\boldsymbol{\Gamma}_0^{h,\mathrm{nc}}$ given by (29) is straightforward, the evaluation of its consistent counterpart $\boldsymbol{\Gamma}_0^{h,\mathrm{c}}$ (28) is more involved. In our experience, the benefit of the latter over the former resides in the fact that it leads to generally better behaved numerical solutions at high contrast, where spurious oscillations might develop with the non-consistent discrete Green operator [27]. The reason for this is obvious on figure 1, where the *xyxy* component of each discrete operator is represented in Fourier space. Indeed, while the consistent operator is smooth in Fourier space, the non-consistent operator exhibits a strong discontinuity at the highest frequencies.

Another benefit of the use of the consistent operator resides in the fact that it leads to *rigorous* bounds on the macroscopic elastic properties of the composite, provided that the reference material is stiffer or softer than all phases [13].

While of great practical interest, the derivation of bounds is not the main concern of this paper, which rather aims at a faithfull calculation of the local polarization $\boldsymbol{\tau}(\mathbf{x})$. From this perspective, one of the two main results of this paper is the fact that the discrete problems (21) and (35) are *always* well-posed, regardless of the reference material $\mathbf{C}_0$. This result has important consequences for both non-consistent and consistent schemes.

For the non-consistent approach of Moulinec and Suquet [9, 10], this means that conditions (34) on the reference material are *not* necessary to ensure existence and uniqueness of solution to the discrete problem (35); since the Neumann iterations no longer converge [16], this solution should be computed by means of an appropriate linear iterative solver.

For the consistent approach of Brisard and Dormieux [13], this means that the discrete problem (21) has a unique solution, even if the conditions of theorem 2 are not fulfilled. However, this solution no longer provides a bound on the macroscopic properties of the composites.

The second of the main results of this paper is the *convergence* to the exact solution $\boldsymbol{\tau}$ of (10) of the discrete solution

$\boldsymbol{\tau}^h$ of (21) or (35) as the size h of the cells tends to zero (h–convergence). In other words, for any choice of h, $\boldsymbol{\tau}^h$ can be viewed as a cellwise-constant estimate of $\boldsymbol{\tau}$. It is emphasized again that this result has been proved under assumptions 1 and 2, the latter excluding porous media. However, numerical experiments shown below tend to indicate that the theoretical results presented here might be extended to this case, which should be investigated further.

A by-product of the previous study is rule (23), which was previously derived in the context of the consistent approach [13], but is shown in the present work to be more general. This rule states how consistent equivalent properties of a heterogeneous cell can be computed. It is of great practical importance when the present numerical schemes are coupled with real-life experiments (e.g. micro-tomography). Indeed, resolution of the imaging instruments being finite, observed pixels or voxels are *always* heterogeneous, so that they cannot be attributed the elastic stiffness of one of the pure phases. From this point of view, it is interesting to note that (23) depends on the *composition* of the heterogeneous cell, but not on the *spatial organization* of the different phases within the cell. While the latter is by definition inacessible, the former can be retrieved in carefully conducted experiments by an adequate inverse analysis using minimum prior knowledge [28].

It would appear from the above discussion that both consistent and non-consistent approaches are equivalent. This is not strictly true from the practical point of view, for two already stated reasons: on the one hand, pre-computation of the consistent discrete Green operator is difficult, while on the other hand, the non-consistent discrete Green operator is discontinuous. In the next section, the theoretical results stated previously are illustrated on a simple two-dimensional (plane strain) application. A third discrete operator is then introduced, on purely heuristic grounds. Being smooth and easily computable, this operator is shown to combine assets of both consistent and non-consistent discrete Green operators.

5. Numerical examples

All numerical examples in this section are based on the same bidimensional (plane strain) geometry, shown on figure 2. It should be noted that, although this example has already been considered in [13], the results presented here are new. A square inclusion of size $a \times a$ is embedded in a square unit-cell, of size $L \times L$; in the present work, the size of the inclusion is fixed, as well as the elastic properties of the matrix

$$a = \frac{L}{2}, \qquad \mu_\mathrm{m} = 1, \qquad \nu_\mathrm{m} = 0.3,$$

while the elastic properties of the inclusion are variable, and denoted μ_i and ν_i. This simple composite material is submitted to a unit macroscopic shearing strain

$$\mathbf{E} = E\,(\mathbf{e}_1 \otimes \mathbf{e}_2 + \mathbf{e}_2 \otimes \mathbf{e}_1) \qquad (E = 1).$$

In view of studying the convergence of the approximate solution $\boldsymbol{\tau}^h$ as $h \to 0$, several values of h will be considered, corresponding to power-of-two square grids, $N_1 = N_2 = 4, \ldots, 1024$.

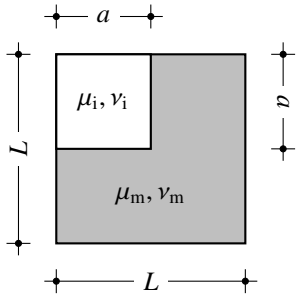


Figure 2: geometry of the examples considered in section 5. The size of the inclusion is $a = L/2$.

A reference solution is also necessary, in order to compute the discretization error. As such reference solution does not exist in closed-form for the problem at hand, we computed a numerical approximation on a very fine grid (2048×2048), using the consistent approach. In other words, we approximate the true solution $\boldsymbol{\tau}$ with $\boldsymbol{\tau}^{h_0,\mathrm{c}}$, $h_0 = L/2048$. For a fixed value of $h > h_0$, we then compute the approximate relative error

$$\epsilon[\boldsymbol{\tau}^h] = \frac{\|\boldsymbol{\tau}^h - \boldsymbol{\tau}^{h_0}\|_\mathbb{V}}{\|\boldsymbol{\tau}^{h_0}\|_\mathbb{V}}. \tag{39}$$

It shoul be noted that the (approximate) polarization field $\boldsymbol{\tau}^h$ is not intrinsic, since it depends on the reference material $\mathbf{C}_0$. As the latter will be varied in the following study, it will prove more convenient to compute the relative error on the stress field

$$\epsilon[\boldsymbol{\sigma}^h] = \frac{\|\boldsymbol{\sigma}^h - \boldsymbol{\sigma}^{h_0}\|_\mathbb{V}}{\|\boldsymbol{\sigma}^{h_0}\|_\mathbb{V}}, \tag{40}$$

where $\boldsymbol{\sigma}^h$ is the approximate stress field associated with the approximate polarization field $\boldsymbol{\tau}^h$

$$\boldsymbol{\sigma}^h = \mathbf{C}_0 : \boldsymbol{\varepsilon}^h + \boldsymbol{\tau}^h = \mathbf{C}_0 : \left(\mathbf{E} - \boldsymbol{\Gamma}_0^h * \boldsymbol{\tau}^h\right) + \boldsymbol{\tau}^h,$$

where $\boldsymbol{\Gamma}_0^h$ denotes either the consistent or the non-consistent discrete Green operator.

Before we proceed to the quantitative analysis of the relative error on the stress field, a few words must be said on the implementation of the numerical schemes. We have already mentioned that problems (21) and (35) reduce to linear systems, the matrix of which cannot be expressed in closed-form, whereas matrix-vector products are easily computed as follows

Input $\boldsymbol{\tau}_{\boldsymbol{\beta}}^h$,

$$\hat{\boldsymbol{\tau}}_{\mathbf{b}}^h \leftarrow \mathrm{FFT}[\boldsymbol{\tau}_{\boldsymbol{\beta}}^h]_{\mathbf{b}}, \tag{41a}$$

$$\text{For all } \mathbf{b} \in \mathcal{I}^h, \quad \hat{\boldsymbol{\eta}}_{\mathbf{b}}^h \leftarrow \hat{\boldsymbol{\Gamma}}_{0,\mathbf{b}}^h : \boldsymbol{\tau}_{\mathbf{b}}^h, \tag{41b}$$

$$\boldsymbol{\eta}_{\boldsymbol{\beta}}^h \leftarrow \mathrm{FFT}^{-1}[\hat{\boldsymbol{\eta}}_{\mathbf{b}}^h]_{\boldsymbol{\beta}}, \tag{41c}$$

$$\text{For all } \boldsymbol{\beta} \in \mathcal{I}^h, \quad \boldsymbol{\eta}_{\boldsymbol{\beta}}^h \leftarrow \boldsymbol{\eta}_{\boldsymbol{\beta}}^h + \left(\mathbf{C}_{\boldsymbol{\beta}}^h - \mathbf{C}_0\right)^{-1} : \boldsymbol{\eta}_{\boldsymbol{\beta}}^h, \tag{41d}$$

Return $\boldsymbol{\eta}_{\boldsymbol{\beta}}^h$.

The discrete Fourier transform of the polarization field $\boldsymbol{\tau}_{\boldsymbol{\beta}}^h$ is first computed (41a). Then, the discrete (consistent or non-consistent) Green operator $\hat{\boldsymbol{\Gamma}}_{0,\mathbf{b}}^h$ is applied to each Fourier component (41b), and the inverse discrete Fourier transform is taken

(41c). Finally, the local part of the bilinear operator a^h is added (41d). In the above equations, the tensor field $\boldsymbol{\eta}^h_{\boldsymbol{\beta}}$ is the result of the product of the matrix to be inverted, and the input vector $\boldsymbol{\tau}^h_{\boldsymbol{\beta}}$. Then the linear system reads

$$\text{For all } \boldsymbol{\beta} \in \mathcal{I}^h, \quad \boldsymbol{\eta}^h_{\boldsymbol{\beta}} = \mathbf{E}. \tag{42}$$

Standard iterative linear solvers [11] are invoked to solve (42). These solvers need to be passed the implementation of the matrix-vector product calculation (41), as well as a stopping-criterion. In the present application, the iterations are stopped when the residual $\boldsymbol{\rho}^h_{\boldsymbol{\beta}} = \mathbf{E} - \boldsymbol{\eta}^h_{\boldsymbol{\beta}}$ is small enough

$$\|\boldsymbol{\rho}^h\|_{\mathbb{V}} \leq \delta \|\mathbf{E}\|, \tag{43}$$

where δ is a user-specified relative tolerance.

At this point, it is worth emphasizing again that (as argued in [13]) the entries of the matrix of the linear system to be solved need *never* be computed and stored: indeed, only the implementation of a routine for the computation of matrix-vector products (following the procedure (41)) is required.

In sections 5.1 and 5.2, two applications are considered, with two different values of μ_i.

5.1. The case of finite contrast

By finite contrast, we mean that the shear modulus of the inclusion is here neither null (pore) nor infinite (rigid inclusion). The elastic properties of the inclusion selected in the present application are

$$\mu_\mathrm{i} = 0.01, \quad \nu_\mathrm{i} = 0.2,$$

hence assumption 2 is satisfied. The theoretical analysis of section 4 then shows that any reference material is permitted. All calculations presented here were carried out with $\delta = 10^{-10}$, resulting in a very stringent stopping criterion.

For the first series of calculations, we selected $\mathbf{C}_0 = \mathbf{C}_\mathrm{m}$. As previously mentioned, this requires a slight modification of (21) and (35), which must then be solved under the additional constraint that the discretized polarization field be null in the matrix. Such a constraint is easily accounted for within the framework of linear iterative solvers. Inequalities (15) and (36) then show that both consistent and non-consistent approches lead to negative-definite systems, to which the (unpreconditioned) conjugate gradient method can be applied. Figure 3 clearly shows the h–convergence of both consistent (C01) and non-consistent (NC01) approaches. It is experimentally observed that the consistent method is slightly more accurate than the non-consistent method, both methods being approximately of order one in h.

Figure 4 shows how the relative error on the *stresses* tends to zero as h tends to zero. This graph is important, as it allows the comparison of simulations carried out with different reference materials (in which case comparing the polarization fields becomes meaningless).

The previous choice $\mathbf{C}_0 = \mathbf{C}_\mathrm{m}$ was consistent with previously published requirements for both the non-consistent [16] and

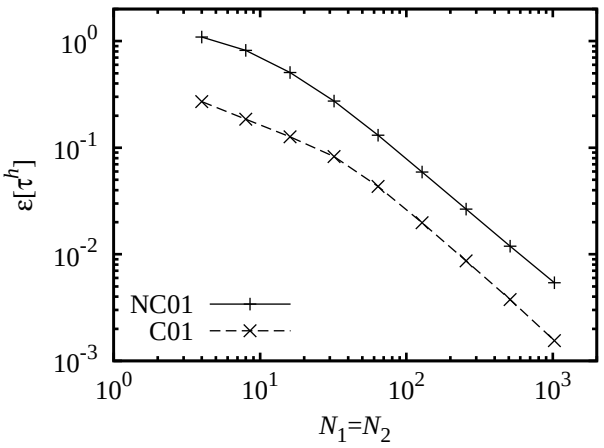


Figure 3: relative error (39) on the polarization for the problem sketched on figure 2, with $\mu_\mathrm{i} = 0.01$, $\nu_\mathrm{i} = 0.3$. NC01: non-consistent scheme, $\mu_0 = 1.0$, $\nu_0 = 0.3$; C01: consistent scheme, $\mu_0 = 1.0$, $\nu_0 = 0.3$.

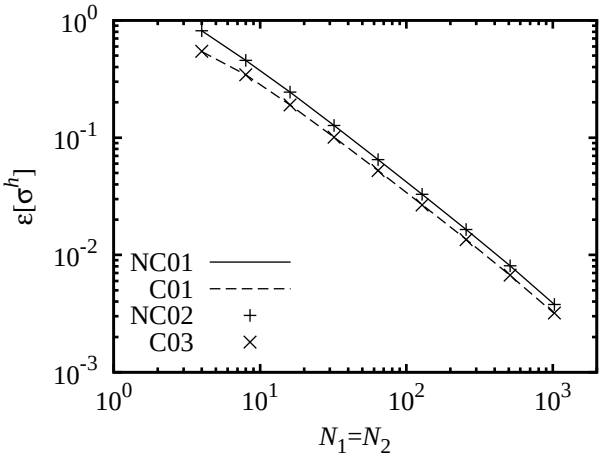


Figure 4: relative error (40) on the stresses for the problem sketched on figure 2, with $\mu_\mathrm{i} = 0.01$, $\nu_\mathrm{i} = 0.3$. NC01: non-consistent scheme, $\mu_0 = 1.0$, $\nu_0 = 0.3$; C01: consistent scheme, $\mu_0 = 1.0$, $\nu_0 = 0.3$; NC02: non-consistent scheme, $\mu_0 = 0.001$, $\nu_0 = 0.3$; C03: consistent scheme, $\mu_0 = 0.5$, $\nu_0 = 0.3$.

	Solver	δ	256^2	512^2	1024^2
NC01	CG	10^{-10}	143	142	139
C01	CG	10^{-10}	140	138	136
NC02	SYMMLQ	10^{-10}	302	296	285
C03	SYMMLQ	10^{-10}	343	343	337
NC04	CG	5×10^{-4}	545	779	574
AL04	[16]	5×10^{-4}	2591	686	267
C04	CG	5×10^{-4}	68	44	29
FNC04	CG	5×10^{-4}	73	50	33

Table 1: Performance of the iterative solvers on each case, for 256×256, 512×512 and 1024×1024 meshes. Reported here is the number of iterations to reach the desired accuracy.

consistent approaches [13]. The following experiments show that h–convergence still occurs *when these requirements are not met*.

For the non-consistent scheme first, we selected a reference material which fails to have elastic constants greater than half the elastic constants of any phase in the composite [16, equation (23)]: $\mu_0 = 0.001$ and $\nu_0 = 0.3$; the corresponding curve on figure 4 is labelled NC02.

We then selected for the consistent scheme a reference material which is neither stiffer nor softer than the phases in the composite: $\mu_0 = 0.5$ and $\nu_0 = 0.3$; the corresponding curve on figure 4 is labelled C03.

As expected from the theoretical analysis of section 4, both curves indicate h–convergence of the corresponding schemes as $h \to 0$. It should be noticed that curves NC01 and NC02 on the one hand, C01 and C03 on the other hand are barely distinguishable. This means that both non-consistent and consistent schemes are not very sensitive to the actual choice of the reference material (the consistent scheme being in any case slightly more accurate than the non-consistent scheme).

While the h–convergence is not really affected by the choice of the reference material, the situation is more contrasted for the actual *inversion* of the linear system (h being fixed). Indeed, the linear systems arising from the cases NC01 and C01 are negative definite, and can be solved by means of the conjugate gradient method. This is no longer true of the cases NC02 and C03, for which convergence of the conjugate gradient method is not guaranteed; the results presented here were obtained with the solver SYMMLQ [29]. Table 1 shows that inversion of the linear system required more iterations for cases NC02 and C03 than for cases NC01 and C01. It should be noted at this point that each iteration of either CG or SYMMLQ requires one matrix-vector product, the cost of which is dominated by the two FFTs. Therefore, comparison of the different cases gathered in table 1 is fair.

5.2. *The case of infinite contrast and the filtered, non-consistent Green operator*

In this section, we address the case of infinite contrast. More precisely, we consider here that the inclusion is a pore, $\mu_\mathrm{i} = 0$. Assumption 2 is no longer valid, and the theoretical results from sections 3 and 4 do not apply. The numerical experiments presented here should therefore be considered as exploratory, prior to more rigorous mathematical backing.

In its original form, the basic scheme of Moulinec and Suquet [9, 10] is known not to be convergent at fixed h. While in the case of finite contrast, this difficulty was overcome in section 5.1 by an appropriate change of the linear iterative solver, this no longer holds in the case of infinite contrast. In fact, the linear system arising from the non-consistent approach seems to be ill-conditioned. We propose an alternative non-consistent approach (the *filtered* non-consistent approach) which is apparently more robust.

As for the consistent approach, it has already been demonstrated [13] that the conjugate gradient iterations converge with porous composites. In the present work, the focus is put on the h–convergence, and we will quantify how the relative error (40) on the stress tensor tends to zero as $h \to 0$.

Generally speaking, convergence of the iterative solver is much slower than in the previous case; we therefore allowed for a higher value of the residual, selecting $\delta = 5 \times 10^{-4}$; still, the number of iterations is rather high (see table 1). Four different schemes were tested, the results being shown on figure 5. Obviously, all four calculations converge when $h \to 0$, and all are approximately of the same order in h.

The non-consistent approach (NC04) is closely related to the basic scheme of Moulinec and Suquet [9, 10]; however, at fixed h, the former is convergent (albeit slowly), while the latter is not.

As a comparison, we also implemented the augmented lagrangian scheme (AL04) first proposed by Michel et al. [16] to overcome the incompatibility of the basic scheme with infinite contrast. For the AL04 calculation, we also used (43) as a stopping criterion, so that direct comparisons in table 1 are meaningful. Interestingly, for both NC04 and AL04 calculations, the number of iterations decreases significantly as the mesh gets finer.

Contrary to the non-consistent approach, the consistent approach (C04) is much better behaved, and can be seen to converge in less than 100 iterations for any refinement h of the mesh. Figure 5 furthermore shows that this scheme is slightly more accurate than both the non-consistent and the augmented Lagrangian approaches.

This indicates again that from the purely numerical perspective, the consistent scheme is superior to its non-consistent counterpart. However, the major drawback of the former lies in the complexity of the calculation of the consistent discrete Green operator (28). This led us to try and derive an alternative, non-consistent discrete Green operator, which would be fairly easy to compute, while leading to well-behaved (easily invertible) linear systems.

The starting point of the heuristic process which led us to the so-called *filtered, non-consistent discrete Green operator*, is the qualitative comparison of the stress fields obtained in calculations NC04 and C04. Figure 6 shows the xy component of $\boldsymbol{\sigma}^h$, for a 32×32 grid. While the result of the consistent calculation (C04, middle) is smooth, the result of the non-consistent calculation (NC04, left) exhibits a "checkerboard" pattern.

This observation suggests that the shortcomings of the non-

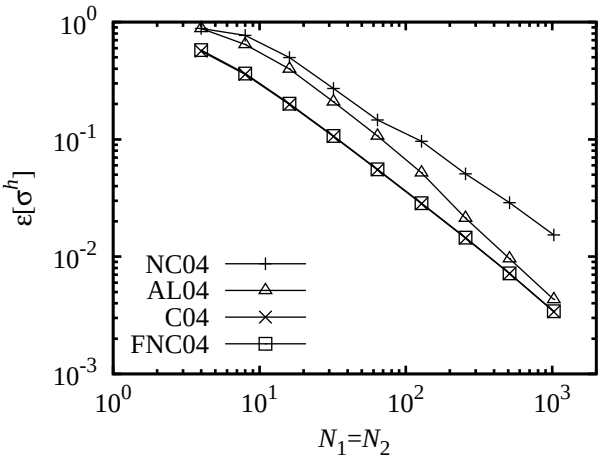


Figure 5: relative error (40) on the stresses for the problem sketched on figure 2, with $\mu_\mathrm{i} = 0$, $\mu_0 = 1.0$ and $\nu_0 = 0.3$. NC04: non-consistent scheme; AL04: augmented Lagrangian scheme [16]; C04: consistent scheme; FNC04: filtered, non-consistent scheme.

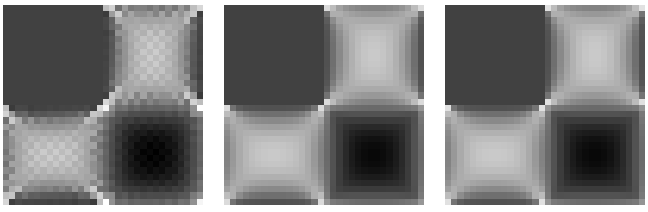


Figure 6: the *xy* component of the stress tensor $\boldsymbol{\sigma}^h$, resulting from a calculation on a 32×32 grid. In this calculation, $\mu_\mathrm{i} = 0$ (pore), $\mu_0 = \mu_\mathrm{m}$, $\nu_0 = \nu_\mathrm{m}$. Three different schemes were used: the non-consistent scheme (NC04, left), the consistent scheme (C04, middle) and the filtered, non-consistent scheme (FNC04, right).

consistent scheme originate in an inaccurate treatment of the highest frequencies, which is in fact confirmed by the sharp discontinuity in Fourier space of the non-consistent discrete Green operator (see figure 1). Comparison of the corners of the left and middle images on figure 1 indicate that the lowest frequencies (say, up to $N_i/4$, $i = 1, \ldots, d$) of the consistent and non-consistent operators are close enough.

To sum up, discretizing $\boldsymbol{\tau}$ on a $N_1 \times \cdots \times N_d$ grid theoretically gives access to frequencies up to $N_i/2$, $i = 1, \ldots, d$. However, if we use the non-consistent discrete Green operator $\boldsymbol{\Gamma}_0^{h,\mathrm{nc}}$ instead of the consistent discrete Green operator, the highest frequencies get polluted. It is then tempting to use the non-consistent Green operator discretized on a *finer* grid, and filter out the (unreliable) high frequencies. This is done in three steps

i. $\boldsymbol{\tau}^h \in \mathbb{V}^h$ is cell-wise constant on cells of size h. It is therefore also cell-wise constant on cells of size $h/2$. In other words, $\boldsymbol{\tau}^h \in \mathbb{V}^{h/2}$,

ii. $\boldsymbol{\Gamma}_0^{h/2,\mathrm{nc}}$ can then be applied to $\boldsymbol{\tau}^h \in \mathbb{V}^{h/2}$

$$\boldsymbol{\eta}^{h/2} = \boldsymbol{\Gamma}_0^{h/2,\mathrm{nc}} * \boldsymbol{\tau}^h,$$

iii. finally, an element $\boldsymbol{\eta}^h \in \mathbb{V}^h$ is constructed by averaging $\boldsymbol{\eta}^{h/2}$ on all 2^d sub-cells of size $h/2$ of one cell of size h.

We define the *filtered, non-consistent, discrete Green operator* $\boldsymbol{\Gamma}_0^{h,\mathrm{fnc}}$ as the operator mapping $\boldsymbol{\tau}^h$ onto $\boldsymbol{\eta}^h$ thus derived

$$\boldsymbol{\eta}^h = \boldsymbol{\Gamma}_0^{h,\mathrm{fnc}} * \boldsymbol{\tau}^h,$$

and straightforward manipulations lead to the following simple expression of the Fourier components of this new discrete Green operator

$$\hat{\boldsymbol{\Gamma}}_{0,\mathbf{b}}^{h,\mathrm{fnc}} = \sum_{\mathbf{n}\in\{-1,0\}^d} [G(h\mathbf{k}_{\mathbf{b}+\mathbf{n}\mathbf{N}})]^2\, \hat{\boldsymbol{\Gamma}}_0(\mathbf{k}_{\mathbf{b}+\mathbf{n}\mathbf{N}}), \quad \text{for } \mathbf{b} \in \mathcal{I}^h, \tag{44}$$

with

$$G(\mathbf{K}) = \cos\frac{K_1}{4} \cdots \cos\frac{K_d}{4}. \tag{45}$$

The formal similarity of (44) and (45) with (28) and (26) is striking. The benefit of the new operator lies of course in the fact that the sum in (44) is finite (it contains only 2^d terms). The filtered, non-consistent discrete Green operator can therefore be evaluated almost as cheaply as the non-filtered, non-consistent discrete Green operator. Besides, this new operator can also be proved to be asymptotically consistent, which means that *under assumptions 1 and 2*, this operator leads to estimates $\boldsymbol{\tau}^h$ that tend to the solution $\boldsymbol{\tau}$ of (10) as $h \to 0$. Figure 1 (right) shows in Fourier space a map of the *xyxy* component (plane strain elasticity) of $\boldsymbol{\Gamma}_0^{h,\mathrm{fnc}}$. Obviously, the discontinuity has been removed; on the whole, $\boldsymbol{\Gamma}_0^{h,\mathrm{fnc}}$ appears to be a much better approximation of the consistent operator $\boldsymbol{\Gamma}_0^{h,\mathrm{c}}$ than $\boldsymbol{\Gamma}_0^{h,\mathrm{nc}}$.

The filtered, non-consistent discrete Green operator was used to compute a Galerkin approximation of the solution to the problem at hand (the calculation is labelled FNC04). Again, the theoretical results of section 4 do not apply to this case.

However, numerical experiments show that the new scheme behaves very satisfactorily, even in the case of infinite contrast. Indeed, both consistent and filtered, non-consistent schemes have similar properties in terms of number of iterations for h fixed (see table 1), and relative error $\epsilon[\boldsymbol{\sigma}^h]$ as $h \to 0$ (see figure 5, where the curves corresponding to C04 and FNC04 are practically undistinguishable). This application suggests that the new operator $\boldsymbol{\Gamma}_0^{h,\mathrm{fnc}}$ realizes the desired compromise between accuracy and ease of computation.

6. Conclusion

In this paper, we have presented a mathematical analysis of two FFT-based schemes for the numerical homogenization of composites within the framework of linear elasticity: the basic scheme of Moulinec and Suquet [9, 10] and the energy-based scheme of Brisard and Dormieux [13]. This work was motivated by practical considerations, and led to important practical conclusions.

We have shown that a slightly modified version of the basic scheme, as well as the energy-based scheme can be regarded as Galerkin discretizations of the same continuous problem (namely, the Lippmann-Schwinger equation); the former is non-consistent, whereas the latter is consistent. We then proved that both approaches lead to L^2 estimates of the true polarization field within the composite.

We focused on the *discretization* of the continuous problem, while the *inversion* of the discretized problem is obtained with the help of standard iterative solvers (conjugate gradient, SYMMLQ, ...) and was not considered here. This distinction revealed that the shortcomings of the basic scheme lie mainly in the inversion step. Theoretical analysis indeed shows that replacing the fixed-point iterations by a more appropriate iterative solver effectively removes the difficulties encountered by Michel et al. [16]; in other words, the basic scheme provides a satisfactory solution for *any choice of the reference material*. However, simple examples indicate that at high contrast, the numerical solution exhibits undesirable oscillations ("checkerboard" pattern). This is to be attributed to the discretization of the Green operator, which poorly reproduces the high-frequencies.

As for the energy-based scheme, the theoretical analysis again led to the result that *any* reference material was admissible. Of course, if the conditions of the principle of Hashin and Shtrikman [14] are not fulfilled, then the estimate of the macroscopic properties is no longer a bound on the real effective properties. The main drawback of the energy-based scheme is the necessary precomputation of the consistent discrete Green operator.

It is worth noting at this point that the present work focused on h–convergence. From this perspective, any reference material is satisfactory, in the sense that it is always true that $\boldsymbol{\tau}^h \to \boldsymbol{\tau}$ when $h \to 0$. However, at fixed resolution h, the quality of the estimates of the local mechanical fields (stresses, strains) can sometimes improve if the reference material is carefully selected. Determination of the optimal reference material at fixed resolution is one of the perspectives of this paper.

Our work allowed to reconcile both basic and energy-based schemes from the theoretical point of view as well as the practical point of view. Indeed their implementations are almost identical, the only difference being the discrete Green operator itself. This led us to try and derive a *third* discrete Green operator, which would combine the strengths of the non-consistent discrete Green operator (ease of computation) with those of the consistent discrete Green operator (absence of spurious oscillations). We thus proposed the filtered, non-consistent Green operator, which realizes a very satisfactory compromise.

In this work, we also proposed a consistent rule for the determination of the equivalent properties of heterogeneous cells. This is of paramount importance in the context of homogenization of real materials, whose microstructure has been obtained by finite-resolution imaging techniques. We have shown that the rule previously introduced in [13] for the energy-based scheme can be extended to the basic scheme as well.

As a final remark, we note that all the mathematical results presented here are established under two assumptions, which are not verified with *porous* media. Numerical experiments presented here and elsewhere indicate that these results remain valid even when pores are present. It is our goal to try and extend the present work to this case. We believe that this further mathematical analysis will improve our practical understanding of the two numerical schemes.

Appendix A. On the mathematical analysis of the continuous problem

Appendix A.1. Two lemmas supporting the proof of theorem 3

The proof of theorem 3 is directly inspired by [23] (appendix). In its original form however, it is established with reference to the initial boundary-value problem of linear elasticity, and makes use of some celebrated differential geometry identities (namely, Stokes' theorem).

In contrast, in the present work, problem (10) is considered independently from the initial elasticity problem (2); besides, (10) is stated in $\mathbb{V}$, where derivatives are not necessarily meaningful, and application of Stokes' theorem would be questionable. It was therefore deemed necessary to rewrite this proof, in order to make sure that theorem 3 remains valid in $\mathbb{V}$.

We start by extending to $\mathbb{V}$ two results (lemmas 1 and 2) which are well-known in the framework of continuum mechanics [30].

Lemma 1. *For any polarization field* $\boldsymbol{\tau} \in \mathbb{V}$

$$\boldsymbol{\Gamma}_0 * [\mathbf{C}_0 : (\boldsymbol{\Gamma}_0 * \boldsymbol{\tau})] = \boldsymbol{\Gamma}_0 * \boldsymbol{\tau}. \tag{A.1}$$

Proof. Starting from (5), simple algebra shows that for any $\mathbf{b} \in \mathbb{Z}^d$,

$$\hat{\boldsymbol{\Gamma}}_0(\mathbf{k_b}) : \mathbf{C}_0 : \hat{\boldsymbol{\Gamma}}_0(\mathbf{k_b}) = \hat{\boldsymbol{\Gamma}}_0(\mathbf{k_b}). \tag{A.2}$$

Summation of the corresponding Fourier series shows that equality (A.1) holds in the L^2-sense. □

Lemma 2 (A particular case of Hill's lemma.)**.** *With the same notation as in theorem 3, the following identities hold*

$$\overline{\boldsymbol{\sigma}_1 : \boldsymbol{\varepsilon}_2} = \overline{\boldsymbol{\sigma}_2 : \boldsymbol{\varepsilon}_1} = 0. \tag{A.3}$$

Proof. It is first noted that taking the Fourier transform of both relations in (13) leads to

$$\hat{\boldsymbol{\varepsilon}}_i(\mathbf{k}) = -\hat{\boldsymbol{\Gamma}}_0(\mathbf{k}) : \hat{\boldsymbol{\tau}}_i(\mathbf{k}), \qquad \hat{\boldsymbol{\sigma}}_i(\mathbf{k}) = \mathbf{C}_0 : \hat{\boldsymbol{\varepsilon}}_i(\mathbf{k}) + \hat{\boldsymbol{\tau}}_i(\mathbf{k}),$$

which, combined with (A.2), brings

$$\hat{\boldsymbol{\sigma}}_1^*(\mathbf{k}) : \hat{\boldsymbol{\varepsilon}}_2(\mathbf{k}) = \hat{\boldsymbol{\tau}}_1(\mathbf{k}) : \left[\hat{\boldsymbol{\Gamma}}_0(\mathbf{k}) : \mathbf{C}_0 : \hat{\boldsymbol{\Gamma}}_0(\mathbf{k}) - \hat{\boldsymbol{\Gamma}}_0(\mathbf{k})\right] : \hat{\boldsymbol{\tau}}_2(\mathbf{k}) = 0,$$

by application of lemma 1. Invoking Parseval's theorem, the scalar product $\overline{\boldsymbol{\sigma}_1 : \boldsymbol{\varepsilon}_2}$ is evaluated in Fourier space, which proves (A.3), since each term of the Parseval series is zero. □

With these two lemmas at hand, the proof of theorem 3 is straightforward, and can be found in e.g. [23].

Appendix A.2. Proof of theorem 4

Proof. It will prove convenient to introduce the following indicator functions, defined for all $\mathbf{x} \in \Omega$

$$\mathbf{1}_{\kappa>\kappa_0}(\mathbf{x}) = \begin{cases} 1 & \text{if } \kappa(\mathbf{x}) > \kappa_0 \\ 0 & \text{if } \kappa(\mathbf{x}) \leq \kappa_0 \end{cases}, \qquad \mathbf{1}_{\kappa<\kappa_0}(\mathbf{x}) = \begin{cases} 1 & \text{if } \kappa(\mathbf{x}) < \kappa_0 \\ 0 & \text{if } \kappa(\mathbf{x}) \geq \kappa_0 \end{cases},$$

as well as the corresponding functions $\mathbf{1}_{\mu>\mu_0}$ and $\mathbf{1}_{\mu<\mu_0}$.

The first statement in theorem 4 will be proved if we exhibit $\alpha > 0$ such that, for any $\boldsymbol{\tau} \in \mathbb{V}$, there exists $\boldsymbol{\varpi} \in \mathbb{V}$ verifying

$$a(\boldsymbol{\tau}, \boldsymbol{\varpi}) \geq \alpha \|\boldsymbol{\tau}\|_{\mathbb{V}} \|\boldsymbol{\varpi}\|_{\mathbb{V}}.$$

Let $\boldsymbol{\tau} \in \mathbb{V}$ be an arbitrary polarization field; a specific polarization field $\boldsymbol{\varpi} \in \mathbb{V}$ is then built upon $\boldsymbol{\tau}$. In order to do so, $\boldsymbol{\tau}$ is first decomposed into hydrostatic ($\boldsymbol{\tau}^{\text{hyd}}$) and deviatoric ($\boldsymbol{\tau}^{\text{dev}}$) parts

$$\boldsymbol{\tau}^{\text{hyd}} = \frac{1}{d} \operatorname{tr} \boldsymbol{\tau}\, \mathbf{i}, \qquad \boldsymbol{\tau}^{\text{dev}} = \boldsymbol{\tau} - \boldsymbol{\tau}^{\text{hyd}}.$$

and the following polarization fields $\boldsymbol{\tau}^+$ and $\boldsymbol{\tau}^-$ are introduced

$$\boldsymbol{\tau}^+ = \mathbf{1}_{\kappa>\kappa_0}\boldsymbol{\tau}^{\text{hyd}} + \mathbf{1}_{\mu>\mu_0}\boldsymbol{\tau}^{\text{dev}}, \qquad \boldsymbol{\tau}^- = \mathbf{1}_{\kappa<\kappa_0}\boldsymbol{\tau}^{\text{hyd}} + \mathbf{1}_{\mu<\mu_0}\boldsymbol{\tau}^{\text{dev}},$$

so that $\boldsymbol{\tau} = \boldsymbol{\tau}^+ + \boldsymbol{\tau}^-$. Introducing the polarization field $\boldsymbol{\varpi} = \boldsymbol{\tau}^+ - \boldsymbol{\tau}^-$, it is readily verified that $\boldsymbol{\varpi} \in \mathbb{V}$. Owing to the symmetry of the bilinear form a, and making use of corollary 1

$$a(\boldsymbol{\tau}, \boldsymbol{\varpi}) = a(\boldsymbol{\tau}^+ + \boldsymbol{\tau}^-, \boldsymbol{\tau}^+ - \boldsymbol{\tau}^-) = a(\boldsymbol{\tau}^+, \boldsymbol{\tau}^+) - a(\boldsymbol{\tau}^-, \boldsymbol{\tau}^-)$$
$$\geq \overline{\boldsymbol{\tau}^+ : (\mathbf{C} - \mathbf{C}_0)^{-1} : \boldsymbol{\tau}^+} + \overline{\boldsymbol{\tau}^- : \mathbf{S}_0 : (\mathbf{S} - \mathbf{S}_0)^{-1} : \mathbf{S}_0 : \boldsymbol{\tau}^-}.$$

Taking advantage of the isotropy of both local and reference materials, the above volume averages can be expanded

$$\overline{\boldsymbol{\tau}^+ : (\mathbf{C} - \mathbf{C}_0)^{-1} : \boldsymbol{\tau}^+} = \frac{1}{|\Omega|} \int_{\kappa(\mathbf{x})>\kappa_0} \frac{\|\boldsymbol{\tau}^{\text{hyd}}(\mathbf{x})\|^2}{d\,[\kappa(\mathbf{x}) - \kappa_0]}\, \mathrm{d}\,\Omega + \frac{1}{|\Omega|} \int_{\mu(\mathbf{x})>\mu_0} \frac{\|\boldsymbol{\tau}^{\text{dev}}(\mathbf{x})\|^2}{2\,[\mu(\mathbf{x}) - \mu_0]}\, \mathrm{d}\,\Omega,$$

and

$$\overline{\boldsymbol{\tau}^- : \mathbf{S}_0 : (\mathbf{S} - \mathbf{S}_0)^{-1} : \mathbf{S}_0 : \boldsymbol{\tau}^-} = \frac{1}{|\Omega|} \int_{\kappa(\mathbf{x})<\kappa_0} \frac{\kappa(\mathbf{x})\, \|\boldsymbol{\tau}^{\text{hyd}}(\mathbf{x})\|^2}{d\kappa_0\,[\kappa(\mathbf{x}) - \kappa_0]}\, \mathrm{d}\,\Omega + \frac{1}{|\Omega|} \int_{\mu(\mathbf{x})<\mu_0} \frac{\mu(\mathbf{x})\, \|\boldsymbol{\tau}^{\text{dev}}(\mathbf{x})\|^2}{2\mu_0\,[\mu(\mathbf{x}) - \mu_0]}\, \mathrm{d}\,\Omega,$$

from which the following bound results

$$a(\boldsymbol{\tau}, \boldsymbol{\varpi}) \geq \frac{\alpha}{|\Omega|} \int_\Omega \left[\|\boldsymbol{\tau}^{\text{hyd}}(\mathbf{x})\|^2 + \|\boldsymbol{\tau}^{\text{dev}}(\mathbf{x})\|^2\right] \mathrm{d}\,\Omega = \alpha \|\boldsymbol{\tau}\|_{\mathbb{V}}^2, \quad \text{(A.4)}$$

with

$$\alpha = \min\left\{ \inf_{\kappa>\kappa_0} \frac{1}{d\,[\kappa(\mathbf{x}) - \kappa_0]}, \inf_{\kappa<\kappa_0} \frac{\kappa(\mathbf{x})}{d\kappa_0\,[\kappa_0 - \kappa(\mathbf{x})]}, \inf_{\mu>\mu_0} \frac{1}{2\,[\mu(\mathbf{x}) - \mu_0]}, \inf_{\mu<\mu_0} \frac{\mu(\mathbf{x})}{2\mu_0\,[\mu_0 - \mu(\mathbf{x})]} \right\},$$

and the proof of the first statement is complete, since $\|\boldsymbol{\varpi}\|_{\mathbb{V}} = \|\boldsymbol{\tau}\|_{\mathbb{V}}$, and assumption 1 ensures that $\alpha > 0$.

Proof of the second statement is not needed, as the first statement is necessary and sufficient when the bilinear form a is symmetric. □

Appendix B. On the set of trial and test functions, $\mathbb{V}^h$

In this appendix, we prove some useful properties of cell-wise constant functions. In particular, we establish for $\boldsymbol{\varpi}^h \in \mathbb{V}^h$ a link between the Fourier coefficients $\hat{\boldsymbol{\varpi}}^h(\mathbf{k_b})$ and the discrete Fourier transform $\hat{\boldsymbol{\varpi}}^h_{\mathbf{b}}$ of the indexed values $\boldsymbol{\varpi}^h_{\boldsymbol{\beta}}$.

Appendix B.1. Fourier coefficients of $\boldsymbol{\varpi}^h \in \mathbb{V}^h$

The natural setting of problem (10) is the space of square-integrable functions. It is therefore natural to seek the expression of the Fourier coefficients of any test function $\boldsymbol{\varpi}^h \in \mathbb{V}^h$. Straightforward calculations show that, for any multi-index $\mathbf{b} \in \mathbb{Z}^d$

$$\hat{\boldsymbol{\varpi}}^h(\mathbf{k_b}) = \frac{1}{|\Omega|} \int_\Omega \boldsymbol{\varpi}^h(\mathbf{x}) \exp(-\imath \mathbf{k_b} \cdot \mathbf{x})\, \mathrm{d}\,\Omega = \frac{1}{N} F(h\mathbf{k_b}) \sum_{\boldsymbol{\beta} \in \mathcal{I}^h} \exp(-\imath \mathbf{k_b} \cdot \mathbf{x}^h_{\boldsymbol{\beta}}) \boldsymbol{\varpi}^h_{\boldsymbol{\beta}} = \frac{1}{N} F(h\mathbf{k_b}) \hat{\boldsymbol{\varpi}}^h_{\mathbf{b}}, \quad \text{(B.1)}$$

where we have introduced the discrete Fourier transform $\hat{\boldsymbol{\varpi}}^h_{\mathbf{b}}$ of the sequence $\boldsymbol{\varpi}^h_{\boldsymbol{\beta}}$, defined as in (25), as well as function F, defined by (26).

We also note that the norm of the cell-wise constant test function $\boldsymbol{\varpi}^h$ can be indifferently computed in the real space, or in the Fourier space, thanks to the Plancherel theorem

$$\|\boldsymbol{\varpi}^h\|_{\mathbb{V}}^2 = \frac{1}{N} \sum_{\boldsymbol{\beta} \in \mathcal{I}^h} \|\boldsymbol{\varpi}^h_{\boldsymbol{\beta}}\|^2 = \frac{1}{N^2} \sum_{\mathbf{b} \in \mathcal{J}^h} \|\hat{\boldsymbol{\varpi}}^h_{\mathbf{b}}\|^2. \quad \text{(B.2)}$$

Finally, a straightforward application of Parseval's theorem leads to the following useful identity, valid for any $\mathbf{b} \in \mathcal{J}^h$

$$\sum_{\mathbf{n} \in \mathbb{Z}^d} [F(h\mathbf{k}_{\mathbf{b}+\mathbf{n}\mathbf{N}})]^2 = 1. \quad \text{(B.3)}$$

Appendix B.2. Approximation in $\mathbb{V}^h$

For any polarization field $\boldsymbol{\tau} \in \mathbb{V}$, we now address the problem of approximating $\boldsymbol{\tau}$ by a cell-wise constant polarization field $\boldsymbol{\tau}^h \in \mathbb{V}^h$. Theorem 5 states that the approximation error $\|\boldsymbol{\tau}-\boldsymbol{\tau}^h\|_{\mathbb{V}}$ can be made as small as desired, provided that h is small enough; this simply results from the density of continuous functions in $\mathrm{L}^2(\Omega)$, and the fact that continuous functions on Ω are uniformly continuous (being closed and bounded, Ω is compact).

For the analysis of the well-posedness of problem (35), it will prove convenient to provide explicit expressions of the *best* estimate on $\mathbb{V}^h$ of any polarization field $\boldsymbol{\tau} \in \mathbb{V}$. This best estimate $\boldsymbol{\tau}^h \in \mathbb{V}^h$ minimizes $\|\boldsymbol{\tau} - \boldsymbol{\varpi}^h\|_{\mathbb{V}}$ for $\boldsymbol{\varpi}^h \in \mathbb{V}^h$. In other words, it is the *orthogonal projection* of $\boldsymbol{\tau}$ onto $\mathbb{V}^h$, which will be denoted $\Pi^h\boldsymbol{\tau}$.

For any polarization field $\boldsymbol{\tau} \in \mathbb{V}$, the following cell-wise constant polarization field $\boldsymbol{\tau}^h \in \mathbb{V}^h$ is defined

$$\boldsymbol{\tau}^h_{\boldsymbol{\beta}} = \frac{1}{|\Omega^h_{\boldsymbol{\beta}}|} \int_{\Omega^h_{\boldsymbol{\beta}}} \boldsymbol{\tau}(\mathbf{x}) \,\mathrm{d}\,\Omega. \tag{B.4}$$

Simple algebra shows that $\overline{(\boldsymbol{\tau} - \boldsymbol{\tau}^h) : \boldsymbol{\tau}^h} = 0$, that is to say $\boldsymbol{\tau}^h$ is the orthogonal projection of $\boldsymbol{\tau}$ onto $\mathbb{V}^h$. In other words $\Pi^h\boldsymbol{\tau} = \boldsymbol{\tau}^h$, and the best estimate of $\boldsymbol{\tau}$ on $\mathbb{V}^h$ is given by the cell-averages (B.4).

To close this section, the Fourier coefficients of $\boldsymbol{\tau}^h$ are expressed as a function of the Fourier coefficients of $\boldsymbol{\tau}$

Theorem 8. *For any trial field $\boldsymbol{\tau} \in \mathbb{V}$*

$$\hat{\boldsymbol{\tau}}^h_{\mathbf{b}} = N \sum_{\mathbf{n}\in\mathbb{Z}^d} F(h\mathbf{k}_{\mathbf{b}+\mathbf{n}\mathbf{N}})\hat{\boldsymbol{\tau}}(\mathbf{k}_{\mathbf{b}+\mathbf{n}\mathbf{N}}), \tag{B.5}$$

where $\boldsymbol{\tau}^h$ denotes the cell-average of $\boldsymbol{\tau}$, defined by (B.4).

Proof. To prove this identity, we first invoke (B.3), as well as periodicity of the discrete Fourier Transform $\hat{\boldsymbol{\tau}}^h_{\mathbf{b}}$. In what follows, $\mathbf{b}$ is a fixed multi-index

$$\begin{aligned}
&\hat{\boldsymbol{\tau}}^h_{\mathbf{b}} - N \sum_{\mathbf{n}\in\mathbb{Z}^d} F(h\mathbf{k}_{\mathbf{b}+\mathbf{n}\mathbf{N}})\hat{\boldsymbol{\tau}}(\mathbf{k}_{\mathbf{b}+\mathbf{n}\mathbf{N}}) \\
&\quad= \sum_{\mathbf{n}\in\mathbb{Z}^d} \left[F(h\mathbf{k}_{\mathbf{b}+\mathbf{n}\mathbf{N}})^2\hat{\boldsymbol{\tau}}^h_{\mathbf{b}+\mathbf{n}\mathbf{N}} - NF(h\mathbf{k}_{\mathbf{b}+\mathbf{n}\mathbf{N}})\hat{\boldsymbol{\tau}}(\mathbf{k}_{\mathbf{b}+\mathbf{n}\mathbf{N}})\right], \\
&\quad= N \sum_{\mathbf{n}\in\mathbb{Z}^d} F(h\mathbf{k}_{\mathbf{b}+\mathbf{n}\mathbf{N}}) \left[\frac{1}{N} F(h\mathbf{k}_{\mathbf{b}+\mathbf{n}\mathbf{N}})\hat{\boldsymbol{\tau}}^h_{\mathbf{b}+\mathbf{n}\mathbf{N}} - \hat{\boldsymbol{\tau}}(\mathbf{k}_{\mathbf{b}+\mathbf{n}\mathbf{N}})\right], \\
&\quad= N \sum_{\mathbf{n}\in\mathbb{Z}^d} F(h\mathbf{k}_{\mathbf{b}+\mathbf{n}\mathbf{N}}) \left[\hat{\boldsymbol{\tau}}^h(\mathbf{k}_{\mathbf{b}+\mathbf{n}\mathbf{N}}) - \hat{\boldsymbol{\tau}}(\mathbf{k}_{\mathbf{b}+\mathbf{n}\mathbf{N}})\right],
\end{aligned} \tag{B.6}$$

where (B.1) has been used. Multi-index $\mathbf{b}$ still being fixed, we then introduce the auxiliary function

$$\Phi_{\mathbf{b}}(\mathbf{x}) = \sum_{\boldsymbol{\beta}\in\mathcal{I}^h} \chi^h_{\boldsymbol{\beta}}(\mathbf{x}) \exp(\imath\mathbf{k}_{\mathbf{b}} \cdot \mathbf{x}^h_{\boldsymbol{\beta}}) \qquad (\mathbf{x} \in \Omega).$$

It should be noted that at any point $\mathbf{x} \in \Omega$, at most one term in the above sum is non-zero; besides, straightforward algebra leads to the Fourier coefficients of $\Phi_{\mathbf{b}}$

$$\hat{\Phi}_{\mathbf{b}}(\mathbf{k}_{\mathbf{a}}) = \begin{cases} F(h\mathbf{k}_{\mathbf{a}}) & \text{if } \mathbf{a} = \mathbf{b} + \mathbf{n}\mathbf{N}, \\ 0 & \text{otherwise,} \end{cases}$$

and equation (B.6) can be recast as

$$\begin{aligned}
&\hat{\boldsymbol{\tau}}^h_{\mathbf{b}} - N \sum_{\mathbf{n}\in\mathbb{Z}^d} F(h\mathbf{k}_{\mathbf{b}+\mathbf{n}\mathbf{N}})\hat{\boldsymbol{\tau}}(\mathbf{k}_{\mathbf{b}+\mathbf{n}\mathbf{N}}) \\
&\quad= N \sum_{\mathbf{a}\in\mathbb{Z}^d} \hat{\Phi}^*_{\mathbf{b}}(\mathbf{k}_{\mathbf{a}}) \left[\hat{\boldsymbol{\tau}}^h(\mathbf{k}_{\mathbf{a}}) - \hat{\boldsymbol{\tau}}(\mathbf{k}_{\mathbf{a}})\right], \\
&\quad= \frac{N}{|\Omega|} \int_{\mathbf{x}\in\Omega} \Phi_{\mathbf{b}}(\mathbf{x}) \left[\boldsymbol{\tau}^h(\mathbf{x}) - \boldsymbol{\tau}(\mathbf{x})\right] \mathrm{d}\,\Omega,
\end{aligned}$$

where the last equality results from Parseval's theorem. The above integral is the scalar product of the cell-wise constant auxiliary function $\Phi_{\mathbf{b}}$ with $\boldsymbol{\tau}^h - \boldsymbol{\tau}$, which is orthogonal to the subspace of cell-wise constant functions. This integral is therefore null, which proves (B.5). □

Appendix C. On the mathematical analysis of the non-consistent approximation

Appendix C.1. Well-posedness

The proof of the well-posedness of the discrete problem (35) is very similar to the proof of the well-posedness of the continuous problem (10), because lemmas 1 and 2 can be stated for a^h as well as a.

Lemma 3. *For any trial field $\boldsymbol{\tau}^h \in \mathbb{V}^h$*

$$\boldsymbol{\Gamma}^{h,\mathrm{nc}}_0 * \left[\mathbf{C}_0 : \left(\boldsymbol{\Gamma}^{h,\mathrm{nc}}_0 * \boldsymbol{\tau}^h\right)\right] = \boldsymbol{\Gamma}^{h,\mathrm{nc}}_0 * \boldsymbol{\tau}^h.$$

Outline of the proof. This is a simple application, in Fourier space, of (31), as well as (A.2). □

Lemma 4. *Let $\boldsymbol{\tau}^h_1, \boldsymbol{\tau}^h_2 \in \mathbb{V}^h$ be two arbitrary trial fields, and consider the element-wise constant fields $\boldsymbol{\varepsilon}^h_i \in \mathbb{V}^h$ and $\boldsymbol{\sigma}^h_i \in \mathbb{V}^h$*

$$\boldsymbol{\varepsilon}^h_i = -\boldsymbol{\Gamma}^{h,\mathrm{nc}}_0 * \boldsymbol{\tau}_i, \qquad \boldsymbol{\sigma}^h_i = \mathbf{C}_0 : \boldsymbol{\varepsilon}^h_i + \boldsymbol{\tau}_i.$$

Then

$$\overline{\boldsymbol{\sigma}^h_1 : \boldsymbol{\varepsilon}^h_2} = \overline{\boldsymbol{\sigma}^h_2 : \boldsymbol{\varepsilon}^h_1} = 0.$$

Outline of the proof. The proof is similar to that of lemma 2, using discrete Fourier transforms and (31) instead of continuous Fourier transforms, and (3). Finally, lemma 3 and Plancherel's identity are invoked instead of lemma 1 and Parseval's identity. □

Theorem 9. *With the same notation as in lemma 4, the following alternative expressions of the approximate bilinear form a^h hold for any two trial fields $\boldsymbol{\tau}^h_1, \boldsymbol{\tau}^h_2 \in \mathbb{V}^h$*

$$a^h(\boldsymbol{\tau}^h_1, \boldsymbol{\tau}^h_2) = \overline{\boldsymbol{\tau}^h_1 : \left(\mathbf{C}^h - \mathbf{C}_0\right)^{-1} : \boldsymbol{\tau}^h_2} + \overline{\boldsymbol{\varepsilon}^h_1 : \mathbf{C}_0 : \boldsymbol{\varepsilon}^h_2},$$
$$a^h(\boldsymbol{\tau}^h_1, \boldsymbol{\tau}^h_2) = \overline{\boldsymbol{\tau}^h_1 : \mathbf{S}_0 : \left(\mathbf{S}_0 - \mathbf{S}^h\right)^{-1} : \mathbf{S}_0 : \boldsymbol{\tau}^h_2} - \overline{\boldsymbol{\sigma}^h_1 : \mathbf{S}_0 : \boldsymbol{\sigma}^h_2}.$$

Outline of the proof. Invoking lemma 3 (resp. lemma 4), in place of lemma 1 (resp. lemma 2), the proof is identical to theorem 3. □

Inequality (36) follows immediately from the above theorem, and well-posedness of problem (35) can be established in a way almost identical to problem (10).

Appendix C.2. Asymptotic consistency

Proof of theorem 7. Let $\boldsymbol{\tau} \in \mathbb{V}$ be the unique solution of problem (10), and $\boldsymbol{\tau}^h = \Pi^h \boldsymbol{\tau}$ its orthogonal projection onto $\mathbb{V}^h$. To prove the asymptotic consistency of (35), we need to compare, for any $\boldsymbol{\varpi}^h \in \mathbb{V}^h$, the value of $a(\boldsymbol{\tau}, \boldsymbol{\varpi}^h)$ with $a^h(\Pi^h\boldsymbol{\tau}, \boldsymbol{\varpi}^h) = a^h(\boldsymbol{\tau}^h, \boldsymbol{\varpi}^h)$ (see equation (38)).

The difference $a(\boldsymbol{\tau}, \boldsymbol{\varpi}^h) - a^h(\boldsymbol{\tau}^h, \boldsymbol{\varpi}^h)$ comprises two families of terms: the *local* terms, which involve the local stiffness of the composite, and the *non-local* terms, which involve the Green operators $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\Gamma}_0^{h,\mathrm{nc}}$. We first address the local terms, namely

$$\begin{aligned}&\overline{\boldsymbol{\varpi}^h : (\mathbf{C} - \mathbf{C}_0)^{-1} : \boldsymbol{\tau}} - \overline{\boldsymbol{\varpi}^h : (\mathbf{C}^h - \mathbf{C}_0)^{-1} : \boldsymbol{\tau}^h} \\ &\quad = \overline{\boldsymbol{\varpi}^h : (\mathbf{C} - \mathbf{C}_0)^{-1} : (\boldsymbol{\tau} - \boldsymbol{\tau}^h)},\end{aligned}$$

where the equivalence on $\mathbb{V}^h$ between $\mathbf{C}^h$ and $\mathbf{C}$ has been used. From assumption 1,

$$|\overline{\boldsymbol{\varpi}^h : (\mathbf{C} - \mathbf{C}_0)^{-1} : (\boldsymbol{\tau} - \boldsymbol{\tau}^h)}| \leq \frac{1}{\lambda} \|\boldsymbol{\varpi}^h\|_{\mathbb{V}} \|\boldsymbol{\tau} - \boldsymbol{\tau}^h\|_{\mathbb{V}}. \tag{C.2}$$

The non-local term of $a(\boldsymbol{\tau}, \boldsymbol{\varpi}^h)$ is transformed with the help of (B.1), taking advantage of the fact that the discrete Fourier transform is periodic

$$\begin{aligned}\overline{\boldsymbol{\varpi}^h : (\boldsymbol{\Gamma}_0 * \boldsymbol{\tau})} &= \sum_{\mathbf{b} \in \mathbb{Z}^d} \hat{\boldsymbol{\varpi}}^{h*}(\mathbf{k_b}) : \hat{\boldsymbol{\Gamma}}_0(\mathbf{k_b}) : \hat{\boldsymbol{\tau}}(\mathbf{k_b}) \\ &= \frac{1}{N} \sum_{\mathbf{b} \in \mathcal{J}^h} \hat{\boldsymbol{\varpi}}_{\mathbf{b}}^{h*} : \sum_{\mathbf{n} \in \mathbb{Z}^d} F(h\mathbf{k_{b+nN}}) \hat{\boldsymbol{\Gamma}}_0(\mathbf{k_{b+nN}}) : \hat{\boldsymbol{\tau}}(\mathbf{k_{b+nN}}),\end{aligned} \tag{C.3}$$

while definition (31), combined with Plancherel's theorem and (B.5) lead to

$$\begin{aligned}\overline{\boldsymbol{\varpi}^h : (\boldsymbol{\Gamma}_0^{h,\mathrm{nc}} * \boldsymbol{\tau}^h)} &= \frac{1}{N^2} \sum_{\mathbf{b} \in \mathcal{J}^h} \hat{\boldsymbol{\varpi}}_{\mathbf{b}}^{h*} : \hat{\boldsymbol{\Gamma}}_0(\mathbf{k_b}) : \hat{\boldsymbol{\tau}}_{\mathbf{b}}^h \\ &= \frac{1}{N} \sum_{\mathbf{b} \in \mathcal{J}^h} \hat{\boldsymbol{\varpi}}_{\mathbf{b}}^{h*} : \sum_{\mathbf{n} \in \mathbb{Z}^d} F(h\mathbf{k_{b+nN}}) \hat{\boldsymbol{\Gamma}}_0(\mathbf{k_b}) : \hat{\boldsymbol{\tau}}(\mathbf{k_{b+nN}}),\end{aligned} \tag{C.4}$$

and, gathering (C.3) and (C.4)

$$\overline{\boldsymbol{\varpi}^h : (\boldsymbol{\Gamma}_0 * \boldsymbol{\tau})} - \overline{\boldsymbol{\varpi}^h : (\boldsymbol{\Gamma}_0^{h,\mathrm{nc}} * \boldsymbol{\tau}^h)} = \frac{1}{N} \sum_{\mathbf{b} \in \mathcal{J}^h} \hat{\boldsymbol{\varpi}}_{\mathbf{b}}^{h*} : \hat{\boldsymbol{\eta}}_{\mathbf{b}}^h, \tag{C.5}$$

where

$$\hat{\boldsymbol{\eta}}_{\mathbf{b}}^h = \sum_{\mathbf{n} \in \mathbb{Z}^d} F(h\mathbf{k_{b+nN}}) \left[\hat{\boldsymbol{\Gamma}}_0(\mathbf{k_{b+nN}}) - \hat{\boldsymbol{\Gamma}}_0(\mathbf{k_b}) \right] : \hat{\boldsymbol{\tau}}(\mathbf{k_{b+nN}}). \tag{C.6}$$

Applying the inequality of Cauchy-Schwarz to (C.5), and substituting (B.2)

$$\begin{aligned}&|\overline{\boldsymbol{\varpi}^h : (\boldsymbol{\Gamma}_0 * \boldsymbol{\tau})} - \overline{\boldsymbol{\varpi}^h : (\boldsymbol{\Gamma}_0^{h,\mathrm{nc}} * \boldsymbol{\tau}^h)}| \\ &\quad \leq \frac{1}{N} \left(\sum_{\mathbf{b} \in \mathcal{J}^h} \|\hat{\boldsymbol{\varpi}}_{\mathbf{b}}^h\|^2 \right)^{1/2} \left(\sum_{\mathbf{b} \in \mathcal{J}^h} \|\hat{\boldsymbol{\eta}}_{\mathbf{b}}^h\|^2 \right)^{1/2} \\ &\quad \leq \|\boldsymbol{\varpi}^h\|_{\mathbb{V}} \left(\sum_{\mathbf{b} \in \mathcal{J}^h} \|\hat{\boldsymbol{\eta}}_{\mathbf{b}}^h\|^2 \right)^{1/2}.\end{aligned} \tag{C.7}$$

Then, from (C.6), (B.3), (7) and the inequality of Cauchy-Schwarz

$$\begin{aligned}\|\hat{\boldsymbol{\eta}}_{\mathbf{b}}^h\| &\leq \frac{3 - 2\nu_0}{\mu_0 (1 - \nu_0)} \sum_{\substack{\mathbf{n} \in \mathbb{Z}^d \\ \mathbf{n} \neq (0,\ldots,0)}} |F(h\mathbf{k_{b+nN}})| \, \|\hat{\boldsymbol{\tau}}(\mathbf{k_{b+nN}})\| \\ &\leq \frac{3 - 2\nu_0}{\mu_0 (1 - \nu_0)} \left[\sum_{\substack{\mathbf{n} \in \mathbb{Z}^d \\ \mathbf{n} \neq (0,\ldots,0)}} |F(h\mathbf{k_{b+nN}})|^2 \right]^{1/2} \left[\sum_{\substack{\mathbf{n} \in \mathbb{Z}^d \\ \mathbf{n} \neq (0,\ldots,0)}} \|\hat{\boldsymbol{\tau}}(\mathbf{k_{b+nN}})\|^2 \right]^{1/2} \\ &\leq \frac{3 - 2\nu_0}{\mu_0 (1 - \nu_0)} \left[\sum_{\substack{\mathbf{n} \in \mathbb{Z}^d \\ \mathbf{n} \neq (0,\ldots,0)}} \|\hat{\boldsymbol{\tau}}(\mathbf{k_{b+nN}})\|^2 \right]^{1/2},\end{aligned} \tag{C.8}$$

Regrouping (C.7) and (C.8) finally leads to the following upper-bound

$$\begin{aligned}&|\overline{\boldsymbol{\varpi}^h : (\boldsymbol{\Gamma}_0 * \boldsymbol{\tau})} - \overline{\boldsymbol{\varpi}^h : (\boldsymbol{\Gamma}_0^{h,\mathrm{nc}} * \boldsymbol{\tau}^h)}| \\ &\quad \leq \frac{3 - 2\nu_0}{\mu_0 (1 - \nu_0)} \|\boldsymbol{\varpi}^h\|_{\mathbb{V}} \left[\sum_{\mathbf{b} \in \mathcal{J}^h} \sum_{\substack{\mathbf{n} \in \mathbb{Z}^d \\ \mathbf{n} \neq (0,\ldots,0)}} \|\hat{\boldsymbol{\tau}}(\mathbf{k_{b+nN}})\|^2 \right]^{1/2} \\ &\quad \leq \frac{3 - 2\nu_0}{\mu_0 (1 - \nu_0)} \|\boldsymbol{\varpi}^h\|_{\mathbb{V}} \left[\|\boldsymbol{\tau}\|_{\mathbb{V}}^2 - \sum_{\mathbf{b} \in \mathcal{J}^h} \|\hat{\boldsymbol{\tau}}(\mathbf{k_b})\|^2 \right]^{1/2},\end{aligned} \tag{C.9}$$

which obviously tends to 0 as $h \to 0$ (see e.g. Parseval's identity). Asymptotic consistency of a^h, that is

$$\lim_{h \to 0} \sup_{\boldsymbol{\varpi}^h \in \mathbb{V}^h} \frac{|a(\boldsymbol{\tau}, \boldsymbol{\varpi}^h) - a^h(\boldsymbol{\tau}^h, \boldsymbol{\varpi}^h)|}{\|\boldsymbol{\varpi}^h\|_{\mathbb{V}}} = 0,$$

then results from (C.2) and (C.9). □

References

[1] T. Mori, K. Tanaka, Average stress in matrix and average elastic energy of materials with misfitting inclusions, Acta Metallurgica 21 (1973) 571–574.

[2] Y. Benveniste, A new approach to the application of Mori-Tanaka's theory in composite materials, Mechanics of Materials 6 (1987) 147–157.

[3] E. Kröner, Bounds for effective elastic moduli of disordered materials, Journal of the Mechanics and Physics of Solids 25 (1977) 137–155.

[4] R. M. Christensen, K. H. Lo, Solutions for effective shear properties in three phase sphere and cylinder models, Journal of the Mechanics and Physics of Solids 27 (1979) 315–330.

[5] R. M. Christensen, Two theoretical elasticity micromechanics models, Journal of Elasticity 50 (1998) 15–25.

[6] J. Sanahuja, C. Toulemonde, Numerical homogenization of concrete microstructures without explicit meshes, Cement and Concrete Research 41 (2011) 1320–1329.

[7] V. Šmilauer, Z. P. Bažant, Identification of viscoelastic C–S–H behavior in mature cement paste by fft-based homogenization method, Cement and Concrete Research 40 (2010) 197–207.

[8] K. Terada, T. Miura, N. Kikuchi, Digital image-based modeling applied to the homogenization analysis of composite materials, Computational Mechanics 20 (1997) 331–346.

[9] H. Moulinec, P. Suquet, A fast numerical method for computing the linear and nonlinear properties of composites, Comptes-rendus de l'Académie des sciences série II 318 (1994) 1417–1423.

[10] H. Moulinec, P. Suquet, A numerical method for computing the overall response of nonlinear composites with complex microstructure, Computer Methods in Applied Mechanics and Engineering 157 (1998) 69–94.

[11] R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, H. Van der Vorst, Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods, SIAM, 2 edition, 1994.
[12] R. Zeller, P. H. Dederichs, Elastic constants of polycrystals, Physica Status Solidi (B) 55 (1973) 831–842.
[13] S. Brisard, L. Dormieux, FFT-based methods for the mechanics of composites: A general variational framework, Computational Materials Science 49 (2010) 663–671.
[14] Z. Hashin, S. Shtrikman, On some variational principles in anisotropic and nonhomogeneous elasticity, Journal of the Mechanics and Physics of Solids 10 (1962) 335–342.
[15] J. Zeman, J. Vondřejc, J. Novák, I. Marek, Accelerating a FFT-based solver for numerical homogenization of periodic media by conjugate gradients, Journal of Computational Physics 229 (2010) 8065–8071.
[16] J. C. Michel, H. Moulinec, P. Suquet, A computational scheme for linear and non-linear composites with arbitrary phase contrast, International Journal for Numerical Methods in Engineering 52 (2001) 139–160.
[17] A. Ern, J.-L. Guermond, Theory and Practice of Finite Elements, volume 159 of *Applied Mathematical Sciences*, Springer, 2004.
[18] R. Hill, Elastic properties of reinforced solids: Some theoretical principles, Journal of the Mechanics and Physics of Solids 11 (1963) 357–372.
[19] C. Huet, Application of variational concepts to size effects in elastic heterogeneous bodies, Journal of the Mechanics and Physics of Solids 38 (1990) 813–841.
[20] T. Kanit, S. Forest, I. Galliet, V. Mounoury, D. Jeulin, Determination of the size of the representative volume element for random composites: statistical and numerical approach, International Journal of Solids and Structures 40 (2003) 3647–3679.
[21] P. Suquet, A simplified method for the prediction of homogenized elastic properties of composites with a periodic structure, Comptes-rendus de l'Académie des sciences série II 311 (1990) 769–774.
[22] I. Babuška, A. K. Aziz, Variational principles, in: A. K. Aziz (Ed.), The mathematical foundations of the finite element method with applications to partial differential equations, Academic Press, 1972, pp. 111–184.
[23] J. R. Willis, Bounds and self-consistent estimates for the overall properties of anisotropic composites, Journal of the Mechanics and Physics of Solids 25 (1977) 185–202.
[24] W. Rudin, Real and Complex Analysis, McGraw-Hill, 1966.
[25] D. J. Eyre, G. W. Milton, A fast numerical scheme for computing the response of composites using grid refinement, European Physical Journal-Applied Physics 6 (1999) 41–47. Progress in Electromagnetics Research Symposium (PIERS 98), NANTES, FRANCE, JUL 13-17, 1998.
[26] V. Vinogradov, G. W. Milton, An accelerated FFT algorithm for thermoelastic and non-linear composites, International Journal for Numerical Methods in Engineering 76 (2008) 1678–1695.
[27] F. Willot, Y.-P. Pellegrini, Fast fourier transform computations and build-up of plastic deformation in 2d, elastic-perfectly plastic, pixelwise disordered porous media, in: D. Jeulin, S. Forest (Eds.), Continuum models and discrete systems CMDS 11, Presses Mines ParisTech, 2008, pp. 443–450.
[28] S. Scheiner, R. Sinibaldi, B. Pichler, V. Komlev, C. Renghini, C. Vitale-Brovarone, F. Rustichelli, C. Hellmich, Micromechanics of bone tissue-engineering scaffolds, based on resolution error-cleared computer tomography, Biomaterials 30 (2009) 2411–2419.
[29] C. C. Paige, M. A. Saunders, Solution of sparse indefinite systems of linear equations, SIAM Journal of Numerical Analysis 12 (1975) 617–629.
[30] J. R. Willis, Lectures on mechanics of random media, in: D. Jeulin, M. Ostoja-Starzewski (Eds.), Mechanics of random and multiscale microstructures, number 430 in CISM courses and lectures, Springer, 2001, pp. 221–267.