Eigendecomposition-based convergence analysis of the Neumann series for laminated composites and discretization error estimation

Cédric Bellis, Hervé Moulinec, Pierre Suquet

▶ **To cite this version:**

Cédric Bellis, Hervé Moulinec, Pierre Suquet. Eigendecomposition-based convergence analysis of the Neumann series for laminated composites and discretization error estimation. International Journal for Numerical Methods in Engineering, 2020, 121 (2), pp.201-232. 10.1002/nme.6206 . hal-03522642

HAL Id: hal-03522642
https://hal.science/hal-03522642

Submitted on 19 Sep 2023

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.

Eigendecomposition-based convergence analysis of the Neumann series for laminated composites and discretization error estimation

C. Bellis, H. Moulinec, P. Suquet
Aix Marseille Univ, CNRS, Centrale Marseille, LMA, Marseille, France

Abstract

In computational homogenization for periodic composites, the Lippmann-Schwinger integral equation constitutes a convenient formulation to devise numerical methods to compute local fields and their macroscopic responses. Among them, the iterative scheme based on the Neumann series is simple and efficient. For such schemes, a priori global error estimates on local fields and effective property are not available and this is the concern of this article, which focuses on the simple, but illustrative, conductivity problem in laminated composites. The global error is split into an iteration error, associated with the Neumann series expansion, and a discretization error. The featured non-local Green's operator is expressed in terms of the averaging operator, which circumvents the use of the Fourier transform. The Neumann series is formulated in a discrete setting and the eigendecomposition of the iterated matrix is performed. The ensuing analysis shows that the local fields are computed using a particular subset of eigenvectors, the iteration error being governed by the associated eigenvalues. Quadratic error bounds on the effective property are also discussed. The discretization error is shown to be related to the accuracy of the trapezoidal quadrature scheme. These results are illustrated numerically and their extension to other configurations is discussed.

Keywords: Computational homogenization, Lippmann-Schwinger equation, Green's operator, Error estimates

1 Introduction

Since the works of Kröner (1972) and Willis (1981), volume integral formulations have shown to be convenient tools in the study of composites. In particular, starting from the set of governing equations of a given linear problem and introducing a reference homogenous comparison material together with the associated Green's operator, the solution field can be shown to satisfy the so-called Lippmann-Schwinger equation. Computing this field therefore amounts to invert this integral equation, which can be done by expressing the solution through a Neumann series expansion. In turn, such an expansion requires iterating the non-local Green's operator composed with a local material contrast function. For periodic composites, it has been recognized in Moulinec and Suquet (1994, 1998) that such computations can be advantageously performed by iterating back and forth between the physical space and the Fourier space where the Green's operator can be expressed algebraically as a frequency-dependent tensor. Building on this idea and making use of the Fast Fourier Transform for computational efficiency, an iterative fixed-point scheme has been proposed in Moulinec and Suquet (1994, 1998) to compute local fields by the Neumann series expansion, giving access in turn to macroscopic responses.

FFT-based computational homogenization methods have been flourishing since, with applications to a variety of material configurations. From an algorithmic standpoint, most developments have aimed at fast or unconditionally convergent schemes, see e.g. Eyre and Milton (1999); Michel et al. (2001); Zeman et al. (2010); Monchiet and Bonnet (2012); Brisard and Dormieux (2012); Gélébart and Mondon-Cancel (2013); Mishra et al. (2016); Moulinec et al. (2018). Recently Kabel et al. (2014), the connexion between the Lippmann-Schwinger equation and the gradient of the strain-based elastic energy functional has been highlighted. This has opened the door to the development of accelerated gradient-based algorithms, see Schneider (2017), as well as numerical schemes based on alternative geometrical variational principles Bellis and Suquet (2018). In this context and despite this abundant literature, to our knowledge, results on a priori global error estimates for this type of numerical methods are relatively scarce. This issue is therefore the subject of the present study, which is intended to be an assessment of the questions related to global error estimations on local fields and effective properties. Note

that, in the specific framework of the Fourier-Galerkin method, related results have been obtained in Vondřejc et al. (2014, 2015).

With this aim, one focuses on the original fixed-point scheme of Moulinec and Suquet (1998) that is associated with the Neumann series expansion. Given our objective to illustrate the convergence properties of this algorithm, one considers the conductivity problem in laminated periodic composites. This problem is simple enough to be solved semi-analytically but, in the mean time, it is rich enough to provide illustrative examples of the main features of the algorithm considered. Given the iterative nature of the latter the overall numerical error is split into two components that are studied separately: an iteration error that corresponds to the convergence of the Neumann series expansion, and a discretization error, which is associated with the convergence of the limit of that expansion to the exact continuous solution. Moreover, let us emphasize that, for the 1D problem considered, the non-local Green's operator is fully expressed in closed-form in the physical space using the averaging operator. As a consequence, the Fourier transform is not involved in the computation of the Green's operator in the present study, so that the issues conventionally related to the Fourier-based approximation are not at play here.

The conductivity problem considered is introduced in Section 2 alongside with the definition of the Green's operator and its key properties in connection to energetic variational principles. In this context, the Lippmann-Schwinger equation is obtained through a first-order optimality condition and then inverted in the form of a conditionally convergent Neumann series. These elements are then transposed in a discrete setting for which the discretization scheme reduces to the introduction of the trapezoidal quadrature rule by consistency with the Fourier-based implementations of the fixed-point scheme that are commonly employed. Two test cases are also introduced, i.e. two specific spatial distributions of the conductivity field, as illustrative examples to be used throughout the study. In the discrete setting considered, a detailed eigenanalysis of the matrix iterated in the Neumann series expansion is provided and illustrated numerically in Section 3. With this eigendecomposition at hand, the convergence properties of the Neumann series are investigated in Section 4, first in terms of local field and then effective property. It is shown in particular that the solution field is constructed iteratively using a particular subset of eigenvectors, or modes, whose associated eigenvalues being, in some cases, smaller in absolute value than the spectral radius of the iterated matrix. This allows to provide a mode-based decomposition of the solution that is computed using the Neumann series expansion and to characterize the associated convergence rate, see Sec. 4.1. The convergence estimates of the iteration error are extended to the effective property by discussing some quadratic upper and lower bounds on the latter, making use of results which are standard in the field of numerical optimization, see Sec. 4.2. The next step in Section 5 consists in evaluating the discretization error that characterizes the convergence of the discrete solution, computed by the Neumann series, to the exact continuous one. This analysis relies on the evaluation of the accuracy of the trapezoidal quadrature rule that constitutes the core of the discretization scheme considered. Reminding some key results on this numerical integration scheme, it is shown that the discretization error is directly related to the spectral properties of the conductivity field that characterizes the composite considered, and that high level of accuracy can be reached for smooth material distributions. Owing to this analysis, aliasing effects are also shown to be entirely accountable for the discretization errors. Section 6 summarizes the main points of this study on 1D laminated composites and discusses the possible extensions of these results to other configurations. In Section 6.2, a 2D conductivity problem, for which the exact field solution and the effective property are known analytically, and a 3D elasticity problem are investigated numerically to illustrate this discussion.

2 Laminated composites

2.1 Problem setting and effective property

Consider an isotropic conductive material in $\mathbb{R}^D$, periodically laminated with layers orthogonal to the unit direction $\boldsymbol{d}$ and ℓ being the period length. The conductivity field γ satisfies

$$\gamma(\boldsymbol{x}) = \gamma(x) \text{ for all } x = \boldsymbol{x}\cdot\boldsymbol{d}, \qquad \gamma \in L^\infty_{\mathrm{per}}(0,\ell), \qquad \gamma(x) \geq \kappa > 0.$$

With the same abuse of notations, the governing equations of the conductivity problem read

$$\begin{cases} \text{(i)} & \boldsymbol{e}(x) = \overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}(x), \quad \tilde{\boldsymbol{e}}(x) = \boldsymbol{\nabla} u(x), \quad u \in H^1_{\mathrm{per}}(0,\ell), \\ \text{(ii)} & \boldsymbol{j}(x) = \gamma(x)\boldsymbol{e}(x), \\ \text{(iii)} & \operatorname{div}\boldsymbol{j}(x) = \boldsymbol{0}, \quad \boldsymbol{j}(0)\cdot\boldsymbol{d} = \boldsymbol{j}(\ell)\cdot\boldsymbol{d}, \end{cases} \tag{1}$$

where $\overline{\boldsymbol{e}} = \overline{e}\,\boldsymbol{d} + \overline{\boldsymbol{e}}^{\perp}$ with $\boldsymbol{d}\cdot\overline{\boldsymbol{e}}^{\perp} = 0$ is the imposed macroscopic intensity.

By definition, for any scalar function $w \in H^1_{\rm per}(0,\ell)$ one has $\boldsymbol{\nabla} w(x) = \frac{\mathrm{d}w}{\mathrm{d}x}(x)\boldsymbol{d}$, so that one introduces the functional space $\boldsymbol{\mathscr{E}}_0$ of mean-free compatible gradients as:

$$\boldsymbol{\mathscr{E}}_0 = \left\{ \boldsymbol{f} \ : \ \exists\, f \in L^2_{\rm per}(0,\ell) \text{ with } \langle f\rangle = 0 \text{ such that } \boldsymbol{f}(x) = f(x)\boldsymbol{d} \text{ in } [0,\ell[\right\}, \tag{2}$$

where the averaging operator $\langle\cdot\rangle$ is defined by

$$\langle f\rangle = \frac{1}{\ell}\int_0^\ell f(x)\,\mathrm{d}x.$$

Given $\alpha > 0$, one defines a weighted scalar product $(\cdot,\cdot)_\alpha$ as

$$(\boldsymbol{f}_1, \boldsymbol{f}_2)_\alpha = \langle \boldsymbol{f}_1 \cdot \alpha\, \boldsymbol{f}_2\rangle \qquad \forall \boldsymbol{f}_1,\ \boldsymbol{f}_2 \in L^2_{\rm per}(0,\ell)^D, \tag{3}$$

the case $\alpha = 1$ being the standard $L^2_{\rm per}$-scalar product. Equipped with the scalar product (3), the space $\boldsymbol{\mathscr{E}}_0$ is a Hilbert space. Its polar space $\mathscr{S}$, which is defined as the space of functions $\boldsymbol{h}$ that satisfy the $\langle \boldsymbol{h}\cdot\boldsymbol{f}\rangle = 0$ for all $\boldsymbol{f} \in \boldsymbol{\mathscr{E}}_0$, is given by

$$\mathscr{S} = \Big\{ \boldsymbol{h} \ : \ \boldsymbol{h}\cdot\boldsymbol{d} \text{ is a constant} \Big\}. \tag{4}$$

The elements of $\mathscr{S}$ are divergence-free fields as for any vector-valued function $\boldsymbol{h}(x)$ one has $\operatorname{div}\boldsymbol{h}(x) = \frac{\mathrm{d}\boldsymbol{h}}{\mathrm{d}x}(x)\cdot\boldsymbol{d}$. Note that the spaces $\boldsymbol{\mathscr{E}}_0$ and $\mathscr{S}$ differ by the physical dimension of their elements, which are intensity and current fields respectively, and the term $\langle \boldsymbol{h}\cdot\boldsymbol{f}\rangle$ constitutes a duality product. In this context, the local equations (1) can be rewritten in the condensed form:

$$\text{(i) } \boldsymbol{e}(x) = \overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}(x), \quad \tilde{\boldsymbol{e}} \in \boldsymbol{\mathscr{E}}_0, \qquad \text{(ii) } \boldsymbol{j}(x) = \gamma(x)\boldsymbol{e}(x), \qquad \text{(iii) } \boldsymbol{j} \in \mathscr{S}. \tag{5}$$

Reference can be made to, e.g., (Milton, 2002, Chap. 9) for the study of the problem (5). It is a 1D problem where $(\boldsymbol{I} - \boldsymbol{d}\otimes\boldsymbol{d})\cdot\boldsymbol{e}(x) = \overline{\boldsymbol{e}}^{\perp}$, with $\boldsymbol{I}$ being the identity tensor. Moreover, in the direction of lamination the solution is denoted as $e(x) = \boldsymbol{e}(x)\cdot\boldsymbol{d}$. Since (5.iii) entails that $\gamma(x)e(x)$ is a constant and given that $\langle e\rangle = \overline{e}$ from (5.i), one arrives at:

$$e(x) = \left\langle \frac{1}{\gamma} \right\rangle^{-1} \frac{\overline{e}}{\gamma(x)}. \tag{6}$$

The problem (5) is also equivalent to the minimum energy principle:

$$\tilde{\boldsymbol{e}} = \arg\min_{\boldsymbol{e}'\in\boldsymbol{\mathscr{E}}_0} W(\boldsymbol{e}') \qquad \text{with} \qquad W(\boldsymbol{e}') = \frac{1}{2}\left\langle \big(\overline{\boldsymbol{e}} + \boldsymbol{e}'(x)\big)\cdot\gamma(x)\big(\overline{\boldsymbol{e}} + \boldsymbol{e}'(x)\big)\right\rangle, \tag{7}$$

which allows to define the effective conductivity tensor $\boldsymbol{\gamma}_{\rm eff}$ from the solution $\boldsymbol{e} = \overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}$ using the following identity:

$$\overline{\boldsymbol{e}}\cdot\boldsymbol{\gamma}_{\rm eff}\cdot\overline{\boldsymbol{e}} = \langle \boldsymbol{e}(x)\cdot\gamma(x)\boldsymbol{e}(x)\rangle. \tag{8}$$

According to the solution (6), the homogenized medium is *anisotropic* with

$$\boldsymbol{\gamma}_{\rm eff} = \gamma_{\rm eff}\,\boldsymbol{d}\otimes\boldsymbol{d} + \gamma^{\perp}_{\rm eff}(\boldsymbol{I} - \boldsymbol{d}\otimes\boldsymbol{d}) \qquad \text{where} \qquad \gamma_{\rm eff} = \left\langle\frac{1}{\gamma}\right\rangle^{-1} \qquad \text{and} \qquad \gamma^{\perp}_{\rm eff} = \langle\gamma\rangle. \tag{9}$$

2.2 Optimization-based approach

Green's operator Given a reference homogeneous conductivity $\gamma_0 > 0$ then the corresponding Green's operator $\boldsymbol{\Gamma}_0$ is defined as:

$$\boldsymbol{\Gamma}_0 : \boldsymbol{s} \mapsto \boldsymbol{\Gamma}_0\boldsymbol{s} = \boldsymbol{e}' \text{ such that } \boldsymbol{e}' \in \boldsymbol{\mathscr{E}}_0 \text{ and } (\gamma_0\boldsymbol{e}' - \boldsymbol{s}) \in \mathscr{S}. \tag{10}$$

By solving the problem corresponding to $\boldsymbol{e}'$ as in Section 2.1, one can show that the Green's operator is a non-local operator that is given in closed-form in the physical space by

$$\boldsymbol{\Gamma}_0\boldsymbol{s}(x) = \frac{\boldsymbol{d}\otimes\boldsymbol{d}}{\gamma_0}\cdot\big(\boldsymbol{s}(x) - \langle\boldsymbol{s}\rangle\big). \tag{11}$$

The Green's operator is self-adjoint in $L^2_{\rm per}(0,\ell)^D$ for the scalar product (3) for any $\alpha > 0$ since

$$(\boldsymbol{\Gamma}_0\boldsymbol{s}_1, \boldsymbol{s}_2)_\alpha = \frac{\alpha}{\gamma_0}\left\langle \big(s_1 - \langle s_1\rangle\big)s_2\right\rangle = \frac{\alpha}{\gamma_0}\big(\langle s_1 s_2\rangle - \langle s_1\rangle\langle s_2\rangle\big) = (\boldsymbol{s}_1, \boldsymbol{\Gamma}_0\boldsymbol{s}_2)_\alpha,$$

where one has noted $s_i = \boldsymbol{s}_i\cdot\boldsymbol{d}$.

Optimality condition It has been shown, see e.g. Bellis and Suquet (2018), that the gradient of the energy functional W in (7) for the energetic scalar product $(\cdot,\cdot)_{\gamma_0}$ is given for all $\boldsymbol{e}' \in \mathscr{E}_0$ by:

$$\boldsymbol{\nabla} W(\boldsymbol{e}') = \boldsymbol{\Gamma}_0 \boldsymbol{h} \qquad \text{with} \qquad \boldsymbol{h}(x) = \gamma(x)\big(\overline{\boldsymbol{e}} + \boldsymbol{e}'(x)\big). \tag{12}$$

Therefore, the solution $\tilde{\boldsymbol{e}} \in \mathscr{E}_0$ to the variational problem (7) satisfies the first-order optimality condition:

$$\boldsymbol{\nabla} W(\tilde{\boldsymbol{e}}) = \boldsymbol{\Gamma}_0 \boldsymbol{j} = \boldsymbol{0} \qquad \text{with} \qquad \boldsymbol{j}(x) = \gamma(x)\boldsymbol{e}(x) = \gamma(x)\big(\overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}(x)\big). \tag{13}$$

Note that identities such as (12) and (13) hold beyond the scope of the case of laminates investigated in the present study, see Kabel et al. (2014); Bellis and Suquet (2018).

It can be checked that the total field $\boldsymbol{e}$ in (13) satisfies $\langle \boldsymbol{e} \rangle = \overline{\boldsymbol{e}}$ and $(\boldsymbol{I} - \boldsymbol{d} \otimes \boldsymbol{d}) \cdot \boldsymbol{e}(x) = \overline{\boldsymbol{e}}^{\perp}$. Moreover, from the optimality condition $\boldsymbol{\Gamma}_0[\gamma\, \boldsymbol{e}] = \boldsymbol{0}$ one gets

$$\frac{\boldsymbol{d} \otimes \boldsymbol{d}}{\gamma_0} \cdot \big(\gamma(x)\boldsymbol{e}(x) - \langle \gamma\, \boldsymbol{e} \rangle\big) = \boldsymbol{0},$$

so that $\gamma(x)e(x) = \langle \gamma\, e \rangle$. This entails that $\gamma(x)e(x)$ is a constant, which can be shown to be equal to $\langle 1/\gamma \rangle^{-1}\, \overline{e}$. Therefore the solution to (13) is the one given by (6).

From (11), it can be seen that $\boldsymbol{\Gamma}_0 \gamma_0 \overline{\boldsymbol{e}} = \boldsymbol{0}$ and $\boldsymbol{\Gamma}_0 \gamma_0 \boldsymbol{e}' = \boldsymbol{e}'$ for all $\boldsymbol{e}' \in \mathscr{E}_0$. As a consequence, the optimality condition (13) is equivalent to the so-called Lippmann-Schwinger equation given by

$$\boldsymbol{e}(x) = \overline{\boldsymbol{e}} - \boldsymbol{\Gamma}_0[\delta\gamma\, \boldsymbol{e}](x), \tag{14}$$

where $\delta\gamma(x) = \gamma(x) - \gamma_0$.

Stationary iterative scheme: Neumann series The solution to the variational problem (7) can be computed using a gradient-based algorithm, the simplest one being the gradient-descent scheme with fixed step given by:

$$\begin{cases} \tilde{\boldsymbol{e}}_0(x) = \boldsymbol{0} \\ \tilde{\boldsymbol{e}}_{k+1}(x) = \tilde{\boldsymbol{e}}_k(x) - t_k \boldsymbol{\nabla} W(\tilde{\boldsymbol{e}}_k) \end{cases} \qquad \text{with} \qquad t_k = 1, \tag{15}$$

which has been highlighted in Kabel et al. (2014) as being equivalent to the iterative fixed-point scheme introduced in Moulinec and Suquet (1998) to solve the Lippmann-Schwinger equation (14). In the sequel, one focuses on the computation of the component $e(x)$ of the total intensity field solution along $\boldsymbol{d}$, so that the above scheme reduces to:

$$\begin{cases} e_0(x) = \overline{e}, \\ e_{k+1}(x) = \overline{e} - \Gamma_0[\delta\gamma\, e_k](x), \end{cases} \tag{16}$$

where the scalar Green's operator Γ_0 is defined as

$$\Gamma_0 s = \frac{1}{\gamma_0}(s - \langle s \rangle), \qquad \text{i.e.,} \qquad \Gamma_0 s = \boldsymbol{\Gamma}_0[s\, \boldsymbol{d}] \cdot \boldsymbol{d}. \tag{17}$$

The scheme (16) guarantees that $\langle e_k \rangle = \overline{e}$ for all $k \geq 0$. Moreover, the output at the iterate k can be expressed as a Neumann series as:

$$e_k(x) = \sum_{j=0}^{k} (-\mathcal{G})^j\, \overline{e} \qquad \text{with} \qquad \mathcal{G} = \Gamma_0\, \delta\gamma = \Gamma_0 \gamma_0 \frac{\delta\gamma}{\gamma_0}. \tag{18}$$

Note that in (18), the notation $\mathcal{G}^j$ denotes the j-th composition of the operator $\mathcal{G}$ with itself. Introducing the subspace of mean-free fields as

$$L^2_{\text{per},0}(0,\ell) = \big\{ f \in L^2_{\text{per}}(0,\ell) \ : \ \langle f \rangle = 0 \big\} \tag{19}$$

and, with an abuse of notation, upon using (3) as a scalar product on the space L^2_{per} of *scalar* functions, then for all $f_1, f_2 \in L^2_{\text{per},0}(0,\ell)$ one has

$$(\mathcal{G} f_1, f_2)_\alpha = \frac{\alpha}{\gamma_0} \big\langle \big(\delta\gamma f_1 - \langle \delta\gamma f_1 \rangle\big) f_2 \big\rangle = \frac{\alpha}{\gamma_0} \big(\langle \delta\gamma f_1 f_2 \rangle - \langle \delta\gamma f_1 \rangle \langle f_2 \rangle \big) = \frac{\alpha}{\gamma_0} \langle \delta\gamma f_1 f_2 \rangle = (f_1, \mathcal{G} f_2)_\alpha,$$

which proves the self-adjointness of $\mathcal{G}$ in $L^2_{\text{per},0}$. Supplementing this result with the known geometrical interpretation of the Green's operator, see Milton (2002); Bellis and Suquet (2018), one arrives at the following property.

Property 1. *The operator $\mathcal{G}$ is self-adjoint from $L^2_{per,0}(0,\ell)$ into itself for* any *scalar product $(\cdot,\cdot)_\alpha$. Moreover, it can be decomposed as $\mathcal{G} = \mathcal{P}_0 g$ where $\mathcal{P}_0 = \Gamma_0\gamma_0$ is the orthogonal projection operator from $L^2_{per,0}(0,\ell)$ onto itself and g is the normalized material contrast defined as $g(x) = \delta\gamma(x)/\gamma_0$.*

Finally, the series (18) is conditionally convergent, depending on the choice of the reference value γ_0 and, when it does, it converges to the solution $e = \boldsymbol{e}\cdot\boldsymbol{d}$ of (14). For any given norm $\|\cdot\|_\alpha$ on L^2_{per}, evaluating the subordinate operator norm $\|\mathcal{G}\|_\alpha$ provides a valuable information as it yields an upper bound on the convergence rate of the continuous Neumann series (18) according to the global estimate:

$$\|e_k - e\|_\alpha \leq \|\mathcal{G}^k\|_\alpha \, \|e_0 - e\|_\alpha.$$

Remark 1. *According to Property 1, the case of laminated composites is peculiar in that the overall properties of the operators considered are independent of the weight α featured in the scalar product (3). However, by consistency with earlier studies Kabel et al. (2014); Moulinec et al. (2018); Bellis and Suquet (2018), the* energetic *scalar product $(\cdot,\cdot)_{\gamma_0}$ is preferred and will be used thereafter.*

2.3 Discretization

Discretization grid In this study, convergence issues will be investigated at the discrete level. To do so, consider a regular N-points discretization of the period $[0,\ell[$ with associated grid size $h = \ell/N$ and points $x_n = nh$ with $n = 0,\dots,N-1$. Note that the point $x = \ell$ is not included in the discretization owing to the periodicity of the functions considered. The notation γ^h is also used to denote the vector in $\mathbb{R}^N$ whose components are the conductivity values at the grid points, i.e. $\gamma^h_n = \gamma(x_n)$. Moreover, one introduces the vector $\mathbf{g}^h$, whose components are given by the values $g(x_n)$ of the normalized material contrast, i.e.

$$\mathrm{g}^h_n = \frac{\delta\gamma^h_n}{\gamma_0} \qquad \text{for all } n = 0,\dots,N-1 \tag{20}$$

where $\delta\gamma^h_n = \gamma^h_n - \gamma_0$.

Material phases As some components of the conductivity vector γ^h can be equal, one can define a number $\mathcal{P}$ of *phases* ϕ_p, consisting each in the union of the points x_n that share a given value $\gamma^h_{n_p}$. To do so one can rearrange the values γ^h_n for $n = 0,\dots,N-1$ by indexing them by phase numbers, i.e. $\gamma^h_{n_p}$ for $p = 1,\dots,\mathcal{P}$ with $1 \leq \mathcal{P} \leq N$ and such that $\gamma^h_{n_p} \neq \gamma^h_{n_q}$ if $p \neq q$. This allows to define the phases as the discrete set

$$\phi_p = \left\{ x_n \ : \ \gamma^h_n = \gamma^h_{n_p} \right\}, \tag{21}$$

with the corresponding phase fraction being equal to N_p/N with $N_p = \mathrm{card}(\phi_p)$ and $\sum_{p=1}^{\mathcal{P}} N_p = N$. This definition extends to the normalized material contrast vector $\mathbf{g}^h$. Lastly, for further purposes, in the case where $\mathbf{g}^h$ takes zero values, the associated *reference* phase will be denoted as

$$\phi_* = \left\{ x_n \ : \ \mathrm{g}^h_n = 0 \right\}, \tag{22}$$

with $N_* = \mathrm{card}(\phi_*)$ and, for convenience, it will then be designated by the index value $p = \mathcal{P}$, i.e. $\phi_* \equiv \phi_{\mathcal{P}}$, $n_* \equiv n_{\mathcal{P}}$ and $N_* \equiv N_{\mathcal{P}}$.

Discrete operators and scheme Considering periodic functions, one defines a discrete averaging operator $\langle\cdot\rangle_h$ based on a numerical integration scheme by the trapezoidal rule, i.e.

$$\langle f\rangle \approx \langle \mathbf{f}\rangle_h = \frac{1}{N}\sum_{n=0}^{N-1} \mathrm{f}_n \qquad \text{for } \mathbf{f}\in\mathbb{R}^N \text{ such that } \mathrm{f}_n = f(x_n). \tag{23}$$

Since $\langle f\rangle = \hat{f}(0)$ with $\hat{f}$ being the Fourier transform of f, the approximation (23) is chosen as it coincides with the computation of averages using the discrete Fourier transform, the latter being a widely used tool in computational homogenization for the numerical implementation of iterative schemes such as (16), see Moulinec

and Suquet (1998); Moulinec et al. (2018). The question of the convergence of the approximation (23) will be returned to in Section 5. One also defines the *energetic* scalar product $(\cdot,\cdot)_{\gamma_0}$ as

$$\big(\mathbf{f},\mathbf{g}\big)_{\gamma_0} = \frac{\gamma_0}{N}\mathbf{f}^{\mathtt{t}}\cdot\mathbf{g} \qquad \forall \mathbf{f},\mathbf{g}\in\mathbb{R}^N, \tag{24}$$

where the exponent $\mathtt{t}$ denotes the transposition operator, so that the corresponding energetic norm reads

$$\|\mathbf{f}\|^2_{\gamma_0} = \frac{\gamma_0}{N}\sum_{n=0}^{N-1}(\mathrm{f}_n)^2. \tag{25}$$

The discrete counterpart of (19) is the subset:

$$\mathbb{R}^N_0 = \big\{\mathbf{f}\in\mathbb{R}^N \;:\; \langle\mathbf{f}\rangle_h = 0\big\}, \tag{26}$$

which satisfies $\dim(\mathbb{R}^N_0) = N-1$. Moreover, one introduces the following orthogonal decomposition for all $\mathbf{f}\in\mathbb{R}^N$:

$$\mathbf{f} = \mathbf{A}\cdot\mathbf{f} + \mathbf{P}_0\cdot\mathbf{f} \tag{27}$$

with $\mathbf{A}\cdot\mathbf{f}$ being a constant *averaged* vector of $\mathbb{R}^N$ which components are equal to the mean value $\langle\mathbf{f}\rangle_h$, see (65). Moreover, $\mathbf{P}_0$ is the orthogonal projection matrix onto $\mathbb{R}^N_0$ defined by (66).

In this setting and from the definitions of Appendix A, the discrete form of the energetic variational principle (7) reads:

$$\tilde{\mathbf{e}}^h = \arg\min_{\boldsymbol{e}'\in\mathbb{R}^N_0} \mathrm{W}^h(\mathbf{e}') \qquad \text{with} \qquad \mathrm{W}^h(\mathbf{e}') = \frac{1}{2N}(\overline{\mathbf{e}}+\mathbf{e}')^{\mathtt{t}}\cdot\mathbf{Diag}[\gamma^h]\cdot(\overline{\mathbf{e}}+\mathbf{e}'), \tag{28}$$

and where $\overline{\mathbf{e}}$ denotes the constant vector in $\mathbb{R}^N$ whose components are equal to $\overline{e}$. For the energetic scalar product (24), the gradient of the energy functional W^h is defined through the following identity

$$\big(\boldsymbol{\nabla}\mathrm{W}^h(\mathbf{e}'),\tilde{\mathbf{f}}\big)_{\gamma_0} = \frac{1}{N}(\overline{\mathbf{e}}+\mathbf{e}')^{\mathtt{t}}\cdot\mathbf{Diag}[\gamma^h]\cdot\tilde{\mathbf{f}} \qquad \forall\tilde{\mathbf{f}}\in\mathbb{R}^N_0.$$

As a consequence, the vector $\mathbf{b}^h = \big(\gamma_0\boldsymbol{\nabla}\mathrm{W}^h(\mathbf{e}') - \mathbf{Diag}[\gamma^h]\cdot(\overline{\mathbf{e}}+\mathbf{e}')\big)$ belongs to the subspace orthogonal to $\mathbb{R}^N_0$ so that its projection onto the latter is zero, i.e. $\mathbf{P}_0\cdot\mathbf{b}^h = \mathbf{0}$. Since $\boldsymbol{\nabla}\mathrm{W}^h(\mathbf{e}')\in\mathbb{R}^N_0$ for all $\mathbf{e}'\in\mathbb{R}^N_0$ one obtains the discrete counterpart of (12) as

$$\boldsymbol{\nabla}\mathrm{W}^h(\mathbf{e}') = \frac{1}{\gamma_0}\mathbf{P}_0\cdot\mathbf{Diag}[\gamma^h]\cdot(\overline{\mathbf{e}}+\mathbf{e}'). \tag{29}$$

This shows in particular, that the matrix $\mathbf{P}_0$ is the discretized version of the orthogonal projection operator $\mathcal{P}_0 = \Gamma_0\gamma_0$ introduced in Property 1. Moreover, from the definition (66) it holds

$$\mathbf{P}_0\cdot\overline{\mathbf{e}} = \mathbf{0} \qquad \text{and} \qquad \mathbf{P}_0\cdot\mathbf{e}' = \mathbf{e}', \quad \forall\mathbf{e}'\in\mathbb{R}^N_0.$$

Therefore, the first-order optimality condition $\boldsymbol{\nabla}\mathrm{W}^h(\tilde{\mathbf{e}}^h) = \mathbf{0}$ associated with the variational principle (28) is equivalent to the equation:

$$\mathbf{e}^h = \overline{\mathbf{e}} - \mathbf{G}^h\cdot\mathbf{e}^h \qquad \text{with} \qquad \mathbf{G}^h = \mathbf{P}_0\cdot\mathbf{Diag}[\mathbf{g}^h] \tag{30}$$

and where $\mathbf{e}^h = \overline{\mathbf{e}} + \tilde{\mathbf{e}}^h$. The matrix $\mathbf{G}^h\in\mathbb{R}^{N\times N}$ is the discrete counterpart of the continuous operator $\mathcal{G}$, as is the equation (30) with regard to the Lippmann-Schwinger equation (14). The approximation of the solution to (28) based on an iterative gradient-descent scheme with fixed step $t_k = 1$ as in (15) is denoted as $\tilde{\mathbf{e}}^h_k$ for all $k\geq 0$. As in the continuous case, the corresponding total field $\mathbf{e}^h_k = \overline{\mathbf{e}} + \tilde{\mathbf{e}}^h_k$ is equivalent to a fixed-point approximation of the solution to (30) and it satisfies

$$\mathbf{e}^h_k = \sum_{j=0}^{k}(-\mathbf{G}^h)^j\cdot\overline{\mathbf{e}} \tag{31}$$

with $(\mathbf{G}^h)^j$ being the matrix raised to the j-th power. Lastly, one defines an approximated effective conductivity $(\gamma_{\mathrm{eff}})^h_k$ according to the energetic definition (8), which depends both on the discretization parameter h and the iteration number k in (31) as

$$(\gamma_{\mathrm{eff}})^h_k = \frac{2}{\overline{e}^2}\mathrm{W}^h(\tilde{\mathbf{e}}^h_k) \qquad \text{with} \qquad \tilde{\mathbf{e}}^h_k = \mathbf{e}^h_k - \overline{\mathbf{e}}. \tag{32}$$

2.4 Objectives

With $\mathbf{e}$ being the vector of the numerical values of the continuous solution e given by (6) at the grid points, one aims at characterizing the behavior of the overall error $\|\mathbf{e}_k^h - \mathbf{e}\|_{\gamma_0}$ when $k \to \infty$ and $h \to 0$, and likewise for the effective property $(\gamma_{\text{eff}})_k^h$ in relation to γ_{eff} in (9). To do so the global error is decomposed as follows:

$$\|\mathbf{e}_k^h - \mathbf{e}\|_{\gamma_0} \leq \|\mathbf{e}_k^h - \mathbf{e}^h\|_{\gamma_0} + \|\mathbf{e}^h - \mathbf{e}\|_{\gamma_0} \tag{33}$$

where the first right-hand side term is the convergence error of the Neumann series while the second term is the discretization error. In this context, our objectives are as follows:

1. Investigate the convergence properties of the series (31) i.e., when $\mathbf{e}_k^h \underset{k\to\infty}{\longrightarrow} \mathbf{e}^h$, provide an estimate for the discrete stationary iterative scheme error $\|\mathbf{e}_k^h - \mathbf{e}^h\|_{\gamma_0}$, and likewise for the effective property $(\gamma_{\text{eff}})_k^h$ in regard to its limit $(\gamma_{\text{eff}})^h$.
2. Assess the convergence of the discrete solution, i.e. evaluate the discretization error $\|\mathbf{e}^h - \mathbf{e}\|_{\gamma_0}$ when $h \to 0$, and similarly for $(\gamma_{\text{eff}})^h$ compared to γ_{eff}.

The investigation of these convergence properties will be illustrated numerically throughout the article on two test cases of material distributions $\gamma(x)$, see Section 3.2.

3 Eigendecomposition of the iterated matrix

3.1 Analysis

Upon introducing the vector $\mathbf{f}_0^h = \mathbf{G}^h \cdot \overline{\mathbf{e}}$, the series (31) can be rewritten as:

$$\mathbf{e}_0^h = \overline{\mathbf{e}} \qquad \text{and} \qquad \mathbf{e}_k^h = \overline{\mathbf{e}} - \sum_{j=0}^{k-1} (-\mathbf{G}^h)^j \cdot \mathbf{f}_0^h \qquad \text{for } k \geq 1. \tag{34}$$

Moreover, from the definition of $\mathbf{G}^h$ in (30) and using that the matrix $\mathbf{P}_0$ is the orthogonal projector onto the space $\mathbb{R}_0^N$ of mean-free vectors, see (66), one obtains the following property which is the discrete counterpart of Property 1.

Property 2. *The range of the matrix $\mathbf{G}^h$ satisfies $\mathcal{R}(\mathbf{G}^h) \subset \mathbb{R}_0^N$. Its restriction to the subspace $\mathbb{R}_0^N$ is the matrix $\mathbf{G}_0^h = \mathbf{G}^h \cdot \mathbf{P}_0$ which is symmetric.*

As a consequence of the above property it holds $\mathbf{P}_0 \cdot \mathbf{G}^h = \mathbf{G}^h$ and thus, for all $\mathbf{f} \in \mathbb{R}_0^N$, one has:

$$(\mathbf{G}^h)^j \cdot \mathbf{f} = \underbrace{(\mathbf{P}_0 \cdot \mathbf{G}^h) \cdot (\mathbf{P}_0 \cdot \mathbf{G}^h) \ldots (\mathbf{P}_0 \cdot \mathbf{G}^h)}_{j} \cdot \mathbf{P}_0 \cdot \mathbf{f}$$

since $\mathbf{f} = \mathbf{P}_0 \cdot \mathbf{f}$. The terms $\mathbf{G}^h \cdot \mathbf{P}_0 = \mathbf{G}_0^h$ can be grouped together in the previous identity owing to the fact the ranges of $\mathbf{G}^h$ and $\mathbf{G}_0^h$ are both included in $\mathbb{R}_0^N$, which leads to the next property.

Property 3. *For all $\mathbf{f} \in \mathbb{R}_0^N$ and $j \geq 0$ one has $(\mathbf{G}^h)^j \cdot \mathbf{f} = (\mathbf{G}_0^h)^j \cdot \mathbf{f}$.*

Based on this property and since $\mathbf{f}_0^h = \mathbf{G}^h \cdot \overline{\mathbf{e}} \in \mathbb{R}_0^N$, then the matrix $\mathbf{G}^h$ in the Neumann series (34) can be replaced by its restriction $\mathbf{G}_0^h$. To summarize: unlike Γ_0, the continuous operator $\mathcal{G}$ is not self-adjoint in the whole space $L^2_{\text{per}}(0,\ell)$ and the associated matrix $\mathbf{G}^h$ in (30) is not symmetric. Moreover, as $\mathbf{G}^h$ is also not normal, since $(\mathbf{G}^h)^{\mathsf{t}} \cdot \mathbf{G}^h \neq \mathbf{G}^h \cdot (\mathbf{G}^h)^{\mathsf{t}}$, it is not diagonalizable by a unitary matrix. However, it is its restriction $\mathbf{G}_0^h$ to the space $\mathbb{R}_0^N$ of mean-free vectors that actually intervenes in the iterative construction of the discrete approximate solution $\mathbf{e}_k^h$. The matrix $\mathbf{G}_0^h$ being symmetric, it is diagonalizable and the remainder of this section focuses on the characterization of its eigenvalues and eigenvectors λ_m^h and $\mathbf{v}_m^h \in \mathbb{R}_0^N$ respectively, with $m = 0, \ldots, N-1$. Based on the definition of the material phases, see (21) and (22) in Section 2.3, one arrives at the main result of this section.

Proposition 1. *The spectrum of the diagonalizable matrix $\mathbf{G}_0^h$ can be described as follows:*

1. *In the case where $\gamma(x)$ and γ_0 are such that $\mathrm{g}_{n_p}^h \neq 0$ in all phases $p = 1, \ldots, \mathcal{P}$ then:*
   The matrix null space is such that $\dim(\mathcal{N}(\mathbf{G}_0^h)) = 1$ provided that $\langle 1/\mathbf{g}^h \rangle_h \neq 0$ and $\dim(\mathcal{N}(\mathbf{G}_0^h)) = 2$ when $\langle 1/\mathbf{g}^h \rangle_h = 0$. Moreover, the eigenpairs $(\lambda^h, \mathbf{v}^h)$ of $\mathbf{G}_0^h$ such that $\lambda^h \neq 0$ are given by:

(a) $\lambda^h = \mathrm{g}^h_{n_p}$ *with multiplicity* $(N_p - 1)$ *for all phase* $p \in \{1, \dots, \mathcal{P}\}$ *that satisfies* $N_p \geq 2$.
The corresponding eigenvectors satisfy $\langle \mathbf{v}^h \rangle_h = 0$ *with* $\mathbf{v}^h \neq \mathbf{0}$ *in* ϕ_p *while* $\mathbf{v}^h = \mathbf{0}$ *in all phases* ϕ_q *for* $q \neq p$.

(b) λ^h *is a zero of the rational fraction* $F(\lambda^h) = \sum_{p=1}^{\mathcal{P}} N_p (\mathrm{g}^h_{n_p} - \lambda^h)^{-1}$ *provided that* $\mathcal{P} \geq 2$.
The corresponding eigenvectors satisfy $\langle \mathbf{v}^h \rangle_h = 0$ *and they are constant in each phase.*

There are $\sum_{p=1}^{\mathcal{P}} (N_p - 1) = (N - \mathcal{P})$ *eigenvalues of type (a) and* $(\mathcal{P} - 1)$ *of type (b) counting multiplicity when* $\langle 1/\mathbf{g}^h \rangle_h \neq 0$ *and* $(\mathcal{P} - 2)$ *else.*

2. *In the case where* $\gamma(x)$ *and* γ_0 *are such that there exist a* reference *phase* $\phi_* \equiv \phi_{\mathcal{P}}$ *where* $\mathrm{g}^h_{n_*} = 0$ *then: The null space satisfies* $\dim(\mathcal{N}(\mathbf{G}^h_0)) = N_*$ *while the eigenpairs with non-zero eigenvalues are these of type (a) for all phases* $p \in \{1, \dots, \mathcal{P} - 1\}$ *satisfying* $N_p \geq 2$, *together with these of type (b) when* $\mathcal{P} \geq 2$. *This gives* $(N - N_* - \mathcal{P} + 1)$ *and* $(\mathcal{P} - 1)$ *eigenpairs of each type respectively.*

Proof. Consider the case 1 where $\mathrm{g}^h_n \neq 0$ for all n. Assuming $\mathbf{v}^h \in \mathcal{N}(\mathbf{G}^h_0)$, i.e. $\mathbf{G}^h_0 \cdot \mathbf{v}^h = \mathbf{0}$, entails:

$$\mathbf{Diag}[\mathbf{g}^h] \cdot \big(\mathbf{v}^h - \mathbf{A} \cdot \mathbf{v}^h\big) = \mathbf{A} \cdot \mathbf{Diag}[\mathbf{g}^h] \cdot \big(\mathbf{v}^h - \mathbf{A} \cdot \mathbf{v}^h\big) \tag{35}$$

where we make use of the definition (30). Based on the assumption that $\mathbf{g}^h$ is nowhere zero then the matrix $\mathbf{Diag}[\mathbf{g}^h]$ is invertible so that the above expression implies

$$\mathbf{v}^h = \mathbf{A} \cdot \mathbf{v}^h + \mathbf{Diag}[\mathbf{g}^h]^{-1} \cdot \mathbf{A} \cdot \mathbf{Diag}[\mathbf{g}^h] \cdot \big(\mathbf{v}^h - \mathbf{A} \cdot \mathbf{v}^h\big). \tag{36}$$

Averaging this identity by multiplication by the averaging matrix $\mathbf{A}$ yields the following necessary condition:

$$\left\langle \frac{1}{\mathbf{g}^h} \right\rangle_h \big\langle \mathbf{Diag}[\mathbf{g}^h] \cdot \big(\mathbf{v}^h - \mathbf{A} \cdot \mathbf{v}^h\big) \big\rangle_h = 0,$$

where, by an abuse of notation, the vector $1/\mathbf{g}^h$ is the vector with components $1/\mathrm{g}^h_n$. Therefore, if $\langle 1/\mathbf{g}^h \rangle_h \neq 0$ then one has necessarily $\big\langle \mathbf{Diag}[\mathbf{g}^h] \cdot \big(\mathbf{v}^h - \mathbf{A} \cdot \mathbf{v}^h\big) \big\rangle_h = 0$, which inserted back in (36) provides $\mathbf{v}^h = \mathbf{A} \cdot \mathbf{v}^h$. This means that the null space of $\mathbf{G}^h_0$ consists only in constant vectors, with components all equal, so that $\dim(\mathcal{N}(\mathbf{G}^h_0)) = 1$. If $\langle 1/\mathbf{g}^h \rangle_h = 0$, then (36) yields $\mathbf{v}^h$ in the form $\mathbf{v}^h = c_1 + c_2/\mathbf{g}^h$ with $c_1, c_2 \in \mathbb{R}$. Since $(1/\mathbf{g}^h) \in \mathbb{R}^N_0$ it establishes that $\mathbf{v}^h$ belongs to a subspace of dimension 2.

Now, for all $\lambda^h \neq 0$, the identity $\mathbf{G}^h_0 \cdot \mathbf{v}^h = \lambda^h \mathbf{v}^h$ implies $\mathbf{v}^h \in \mathcal{R}(\mathbf{G}^h_0)$ and so $\mathbf{A} \cdot \mathbf{v}^h = \mathbf{0}$ from Property 2. Then it holds:

$$(\mathbf{Diag}[\mathbf{g}^h] - \lambda^h \mathbf{I}) \cdot \mathbf{v}^h = \mathbf{A} \cdot \mathbf{Diag}[\mathbf{g}^h] \cdot \mathbf{v}^h \tag{37}$$

Therefore, the left-hand side term in (37) is constant and the two subcases below can be distinguished.

In the first subcase (a) it is assumed that there exists a phase ϕ_p for $p \in \{1, \dots, \mathcal{P}\}$, see definition (21), such that $\lambda^h = \mathrm{g}^h_{n_p}$. Then, expressing the identity (37) for the component n such that $x_n \in \phi_p$ yields $\langle \mathbf{Diag}[\mathbf{g}^h] \cdot \mathbf{v}^h \rangle_h = 0$, which in turn implies

$$(\mathrm{g}^h_{n_p} - \lambda^h) \mathrm{v}^h_{n_p} = 0 \qquad \text{for all } p = 1, \dots, \mathcal{P},$$

where we use the phase indexing. As a consequence, the eigenvector $\mathbf{v}^h$ vanishes in all phases ϕ_q when $q \neq p$. Therefore, it is non-zero only in ϕ_p. As $\langle \mathbf{v}^h \rangle_h = 0$, there are exactly $(N_p - 1)$ linearly independent vectors satisfying these constraints. Note that it is possible to construct such mean-free vectors for a given phase ϕ_p only if its dimension N_p is at least such that $N_p \geq 2$.

In the second subcase (b), one assumes that $\lambda^h \neq \mathrm{g}^h_{n_p}$ for all phase p, so that from (37) one gets

$$\mathrm{v}^h_{n_p} = \frac{c}{\mathrm{g}^h_{n_p} - \lambda^h} \qquad \text{for all } p = 1, \dots, \mathcal{P}, \tag{38}$$

with $c \in \mathbb{R}$. Averaging this identity and using again that $\langle \mathbf{v}^h \rangle_h = 0$ entail

$$F(\lambda^h) = 0 \qquad \text{where} \qquad F(\lambda^h) = \sum_{p=1}^{\mathcal{P}} \frac{N_p}{\mathrm{g}^h_{n_p} - \lambda^h},$$

since by averaging we have to account for N_p components $\mathrm{v}^h_{n_p}$ in (38) for each phase ϕ_p, see Section 2.3. The rational fraction F can be rewritten as

$$F(\lambda^h) = \left(\sum_{p=1}^{\mathcal{P}} \prod_{\substack{q=1 \\ q\neq p}}^{\mathcal{P}} N_p (\mathrm{g}^h_{n_q} - \lambda^h) \right) \left(\prod_{p=1}^{\mathcal{P}} (\mathrm{g}^h_{n_p} - \lambda^h) \right)^{-1}$$

so that it is seen that its numerator is a polynomial $Q(\lambda^h)$ of degree $(\mathcal{P}-1)$. Its term of degree zero, denoted as Q_0, is given by

$$Q_0 = \sum_{p=1}^{\mathcal{P}} \prod_{\substack{q=1 \\ q\neq p}}^{\mathcal{P}} N_p \, \mathrm{g}^h_{n_q} = \sum_{p=1}^{\mathcal{P}} \frac{N_p}{\mathrm{g}^h_{n_p}} \left(\prod_{q=1}^{\mathcal{P}} \mathrm{g}^h_{n_q} \right) = N \Big\langle \frac{1}{\mathbf{g}^h} \Big\rangle_h \left(\prod_{q=1}^{\mathcal{P}} \mathrm{g}^h_{n_q} \right). \tag{39}$$

As a consequence, for the case currently considered where $\mathrm{g}^h_n \neq 0$ for all n and under the additional assumption that $\langle 1/\mathbf{g}^h \rangle_h \neq 0$ then $Q_0 \neq 0$. Therefore, Q has $(\mathcal{P}-1)$ non-zero roots which constitute the remaining eigenvalues as long as the number $\mathcal{P}$ of phases is such that $\mathcal{P} \geq 2$. The corresponding eigenvectors can be constructed point-wise using (38). In the alternative situation where $\langle 1/\mathbf{g}^h \rangle_h = 0$ then $Q_0 = 0$ and one must study the term of degree 1 in the polynomial $Q(\lambda^h)$ which we denote as $Q_1 \lambda$. It can be checked that

$$Q_1 = -\sum_{p=1}^{\mathcal{P}} \sum_{\substack{q=1 \\ q\neq p}}^{\mathcal{P}} \prod_{\substack{r=1 \\ r\neq p \\ r\neq q}}^{\mathcal{P}} N_p \, \mathrm{g}^h_{n_r} = -\sum_{p=1}^{\mathcal{P}} \sum_{\substack{q=1 \\ q\neq p}}^{\mathcal{P}} \frac{N_p}{\mathrm{g}^h_{n_p} \mathrm{g}^h_{n_q}} \left(\prod_{r=1}^{\mathcal{P}} \mathrm{g}^h_{n_r} \right) = N \Big\langle \frac{1}{(\mathbf{g}^h)^2} \Big\rangle_h \left(\prod_{r=1}^{\mathcal{P}} \mathrm{g}^h_{n_r} \right),$$

where the last equality follows algebraically when using that $\langle 1/\mathbf{g}^h \rangle_h = 0$. Since $\langle 1/(\mathbf{g}^h)^2 \rangle_h \neq 0$ then the term Q_1 does not vanish. This finally implies that when $\langle 1/\mathbf{g}^h \rangle_h = 0$ there are exactly $(\mathcal{P}-2)$ non-zero roots to the polynomial $Q(\lambda^h)$ which are the sought eigenvalues.

Next, consider the case 2 where there exists a *reference* phase $\phi_* \equiv \phi_{\mathcal{P}}$, see (22). Any vector $\mathbf{v}^h \in \mathcal{N}(\mathbf{G}^h_0)$ satisfies (35) and this identity expressed for the component n such that $x_n \in \phi_*$ implies $\big\langle \mathbf{Diag}[\mathbf{g}^h] \cdot \big(\mathbf{v}^h - \mathbf{A} \cdot \mathbf{v}^h \big) \big\rangle_h = 0$. Inserted back in (35) and using the phase indexing entail

$$\mathrm{g}^h_{n_p} \big(\mathrm{v}^h_{n_p} - \langle \mathbf{v}^h \rangle_h \big) = 0 \qquad \text{for all } p = 1, \ldots, \mathcal{P}.$$

This identity imposes that $\mathbf{v}^h = \langle \mathbf{v}^h \rangle_h$ everywhere outside the *reference* phase ϕ_* while $(\mathbf{v}^h - \langle \mathbf{v}^h \rangle_h)$ can take any value in ϕ_*. According to the orthogonal decomposition (27), the number of linearly independent vector satisfying these constraints is equal $1 + (N_* - 1) = N_*$, i.e. the size of the *reference* phase, hence $\dim(\mathcal{N}(\mathbf{G}^h)) = N_*$.

Now, for the non-zero eigenvalues λ^h, the previous developments for the case 1 remain valid for the eigenvalues of type (a) at the exclusion of the *reference* phase. It provides a number $\sum_{p=1}^{\mathcal{P}-1}(N_p - 1) = (N - N_* - \mathcal{P} + 1)$ of eigenvalues. As of the eigenvalues of type (b) described previously, the corresponding developments still hold except that the zeroth order term Q_0 in (39) now reduces to

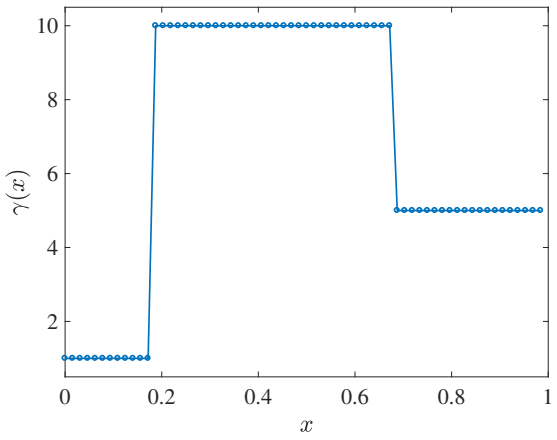
$$Q_0 = \sum_{p=1}^{\mathcal{P}} \prod_{\substack{q=1 \\ q\neq p}}^{\mathcal{P}} N_p \, \mathrm{g}^h_{n_q} = N_{\mathcal{P}} \prod_{q=1}^{\mathcal{P}-1} \mathrm{g}^h_{n_q}$$

owing to the indexing of the *reference* phase as $\phi_* \equiv \phi_{\mathcal{P}}$. By definition of the material phases, the above product cannot vanish. One can conclude that Q has exactly $(\mathcal{P}-1)$ non-zero roots when $\mathcal{P} \geq 2$ which gives the remaining sought eigenvalues. □

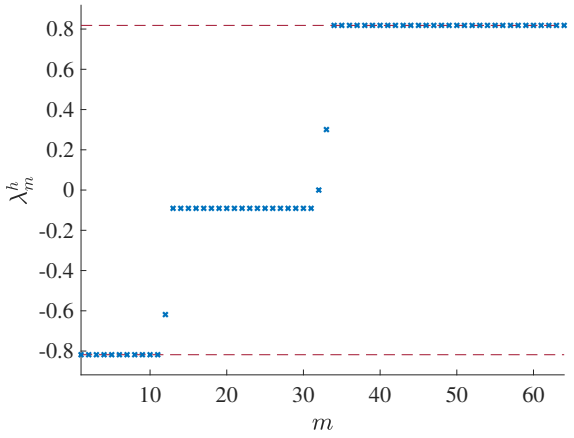
3.2 Numerical examples

We now illustrate Proposition 1 for different material distributions $\gamma(x)$. The period is chosen as $\ell = 1$ and it is discretized for a number $N \in \{64, 128, 256, 512\}$ of points. Moreover, the reference value γ_0 is set as $\gamma_0 = (\max_n(\gamma^h_n) + \min_n(\gamma^h_n))/2$, a choice which will be discussed in Section 4.3. The eigendecompositions presented hereafter are computed using Matlab, based on the Schur decomposition and the QR algorithm. Note that the computed eigenvectors are normalized using the energetic norm (25).

Test case 1: 3-phase laminate The material configuration is the 3-phase laminate of Fig. 1a for which $\gamma(x) \in \{1, 10, 5\}$ with the respective phases being of size $\{3/16, 1/2, 5/16\}$ respectively. Therefore, one gets $\langle 1/\gamma \rangle = 3/10$ according to which the corresponding exact solution $\mathbf{e}$ and effective property γ_{eff} can be computed. Moreover, for this material configuration, the corresponding numerical results shown in the figures 1b and 2 are in agreement with Proposition 1.

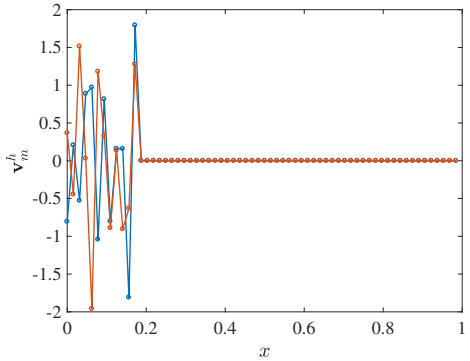


(a) Material distribution $\gamma(x)$

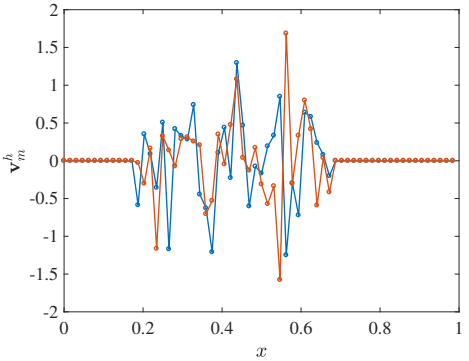


(b) Computed eigenvalues λ_m^h, for $m = 0, \ldots, N-1$, of the matrix $\mathbf{G}_0^h$. The horizontal dashed lines indicate the maximum and minimum values of the vector $\mathbf{g}^h$

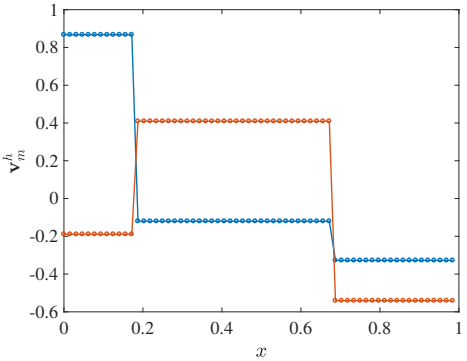
Figure 1: Test case 1: 3-phase laminate, shown here with $N = 64$.



(a) Eigenvectors of type (a) in phase 1 (eigenvalues $m = 5$ and $m = 10$).



(b) Eigenvectors of type (a) in phase 2 (eigenvalues $m = 40$ and $m = 50$).



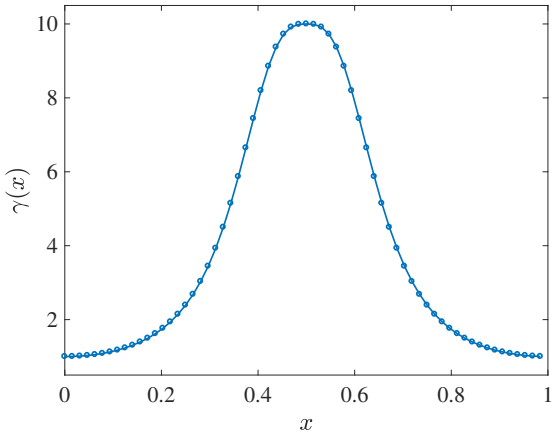
(c) Eigenvectors of type (b) (eigenvalues $m = 12$ and $m = 33$).

Figure 2: [3-phase laminate] Example of three sets of two eigenvectors $\mathbf{v}_m^h$ of the matrix $\mathbf{G}_0^h$ for the different types exposed in Proposition 1, shown here for $N = 64$.

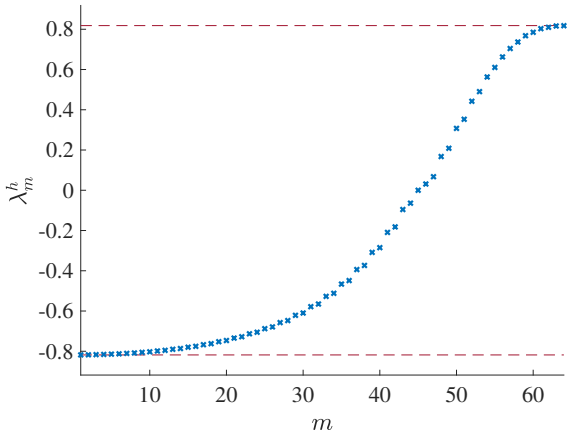
Test case 2: smooth distribution In a second example the material phase function $\gamma(x)$ is defined as the following function:

$$\gamma(x) = \frac{1}{\frac{1}{z} + \left(1 - \frac{1}{z}\right) \left|\cos(\pi x)\right|^r}, \tag{40}$$

shown in Figure 3a using the parameter values $r = 3$ and $z = 10$. It is a smooth distribution which inverse has a discontinuous r-th derivative when r is odd. The motivations for choosing such a function will be exposed in Section 5.2.



(a) Material distribution $\gamma(x)$



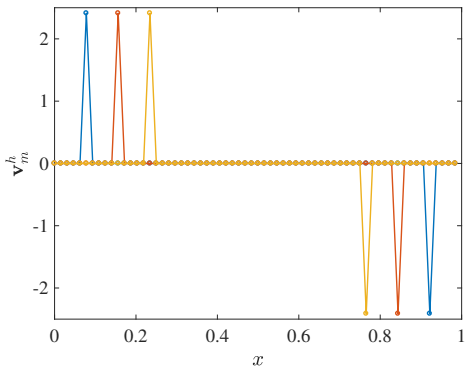
(b) Computed eigenvalues λ_m^h, for $m = 0, \ldots, N-1$, of the matrix $\mathbf{G}_0^h$. The horizontal dashed lines indicate the maximum and minimum values of the vector $\mathbf{g}^h$

Figure 3: Test case 2: smooth distribution, shown here with $N = 64$.

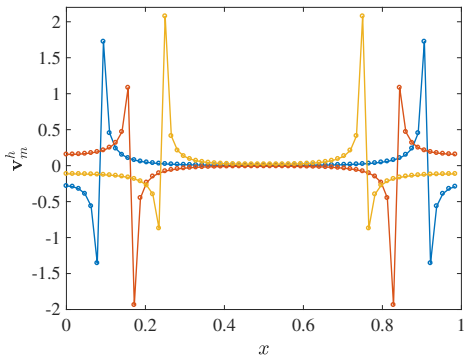
With $r = 3$ one obtains analytically the exact integral value:

$$\left\langle \frac{1}{\gamma} \right\rangle = \frac{1}{z} + \frac{4}{3\pi}\left(1 - \frac{1}{z}\right). \tag{41}$$

In this second example, the discrete distribution $\gamma(x_n)$ is symmetric, see Fig. 3a, so that it takes identical values at a number $(N-2)/2$ of two-points phases by the exclusion of the two extremal points $x_n = 0$ and $x_n = 1/2$. With Proposition 1 at hand, some examples of associated eigenvectors with eigenvalues of type (a) are shown in Figure 4a. There also exist eigenvalues of type (b) with some corresponding eigenvectors being shown in Fig. 4b.



(a) Eigenvectors of type (a) (eigenvalues $m = 10$, $m = 20$ and $m = 30$).



(b) Eigenvectors of type (b) (eigenvalues $m = 11$, $m = 21$ and $m = 31$).

Figure 4: [Smooth distribution] Example of two sets of three eigenvectors $\mathbf{v}_m^h$ of the matrix $\mathbf{G}_0^h$ for the different types exposed in Proposition 1, shown here for $N = 64$.

4 Convergence of the discrete Neumann series

4.1 Convergence of fields

In this section, the objective is to study the convergence properties of the discrete Neumann series (31) based on the results of Section 3. As the matrix $\mathbf{G}_0^h$ is diagonalizable it can be written as $\mathbf{G}_0^h = \frac{\gamma_0}{N}\mathbf{V}^h \cdot \mathbf{\Lambda}^h \cdot (\mathbf{V}^h)^{\mathrm{t}}$

with $\mathbf{\Lambda}^h$ and $\mathbf{V}^h$ being respectively the diagonal and the orthogonal matrices formed by the eigenvalues $\{\lambda_m^h\}_m$ and eigenvectors $\{\mathbf{v}_m^h\}_m$, or modes, which are described in Proposition 1. Note that the multiplicative factor γ_0/N is associated with a normalization of the eigenvectors $\mathbf{v}_m^h$ relatively to the energetic norm (25). From this decomposition and for $k \geq 1$, the equation (34) can be rewritten using the eigenbasis $\{(\lambda_m^h, \mathbf{v}_m^h)\}_{0 \leq m \leq N-1}$ of $\mathbf{G}_0^h$ as:

$$\mathbf{e}_k^h = \overline{\mathbf{e}} - \sum_{j=0}^{k-1} \sum_{m=0}^{N-1} (-\lambda_m^h)^j \left(\mathbf{v}_m^h, \mathbf{f}_0^h\right)_{\gamma_0} \mathbf{v}_m^h. \tag{42}$$

Moreover, provided that $|\lambda_m^h| < 1$ for all $m = 0, \ldots, N-1$, then this series converges when $k \to \infty$ to a limit $\mathbf{e}^h$ that reads:

$$\mathbf{e}^h = \overline{\mathbf{e}} - \sum_{m=0}^{N-1} \frac{1}{1+\lambda_m^h} \left(\mathbf{v}_m^h, \mathbf{f}_0^h\right)_{\gamma_0} \mathbf{v}_m^h, \tag{43}$$

an identity which constitutes a closed-form expression of the solution $\mathbf{e}^h$ to the discrete equation (30).

By definition of the matrix $\mathbf{G}^h$, the initialization vector $\mathbf{f}_0^h = \mathbf{G}^h \cdot \overline{\mathbf{e}}$ is constant in each phase. Therefore, in (42) and (43), the scalar product $\left(\mathbf{v}_m^h, \mathbf{f}_0^h\right)_{\gamma_0}$ is zero when the eigenvector $\mathbf{v}_m^h$ has zero mean in a given phase and is zero outside. In other words, the projection of $\mathbf{f}_0^h$ onto the space spanned by the eigenvectors $\mathbf{v}_m^h$ of type (a) is zero. This yields the proposition below.

Proposition 2. *The vector $\mathbf{e}_k^h$ computed at any iteration $k \geq 0$ and the discrete limit solution $\mathbf{e}^h$ when $k \to \infty$ belong to the union of the null space $\mathcal{N}(\mathbf{G}_0^h)$ and the space spanned by the eigenvectors of type (b) which are constant in each phase.*

With this result at hand, we now turn to the assessment of the properties of the convergence error. According to (42) and (43), the error $\boldsymbol{\epsilon}_k^h = \mathbf{e}_k^h - \mathbf{e}^h$ at step k can be expanded as

$$\boldsymbol{\epsilon}_k^h = \sum_{j=k}^{\infty} \sum_{m=0}^{N-1} (-\lambda_m^h)^j \left(\mathbf{v}_m^h, \mathbf{f}_0^h\right)_{\gamma_0} \mathbf{v}_m^h, \tag{44}$$

so that convergence rate of (34) is governed by the amplitude of the eigenvalues λ_m^h. The matrix $\mathbf{G}_0^h$ being diagonalizable, its spectral radius, defined as $\varrho(\mathbf{G}_0^h) = \max_m |\lambda_m^h|$, coincides with the matrix norm induced by (25), i.e. $\varrho(\mathbf{G}_0^h) = \|\mathbf{G}_0^h\|$. By the properties of the matrix norm one gets:

$$\|\mathbf{G}_0^h\| = \|\mathbf{P}_0 \cdot \mathbf{Diag}[\mathbf{g}^h] \cdot \mathbf{P}_0\| \leq \|\mathbf{P}_0\|^2 \, \|\mathbf{Diag}[\mathbf{g}^h]\| = \|\mathbf{Diag}[\mathbf{g}^h]\|,$$

since $\mathbf{P}_0$ being a projector it satisfies $\|\mathbf{P}_0\| = 1$. Therefore, recalling the definition (20) of the vector $\mathbf{g}^h$, one can obtain from the previous inequality the discrete version of the known upper bound on the norm of the continuous operator $\mathcal{G} = \Gamma_0 \gamma_0 (\delta\gamma/\gamma_0)$, see Equation (31) in Moulinec et al. (2018). In addition, from Proposition 1, it can be seen that any value $\mathrm{g}_{n_p}^h$ is an eigenvalue of $\mathbf{G}_0^h$ as long as the corresponding phase is composed of at least two points, i.e. $N_p \geq 2$. This leads to the next property.

Property 4. *The convergence rate of the Neumann series (31) is bounded by the spectral radius of the matrix $\mathbf{G}_0^h$, which satisfies $\varrho(\mathbf{G}_0^h) \leq \max_n |\delta\gamma(x_n)/\gamma_0|$. Moreover, if the phase ϕ_p such that $|\mathrm{g}_{n_p}^h| = \max_q |\mathrm{g}_{n_q}^h|$ satisfies $N_p \geq 2$ then $\varrho(\mathbf{G}_0^h) = |\delta\gamma(x_{n_p})/\gamma_0|$.*

This property implies that, without information on the material distribution $\gamma(x)$, the upper bound in Property 4 is optimal in the sense that there are some cases where it is attained. This bound could be tightened in some cases based on Proposition 1. Indeed, for any vector $\mathbf{f} \in \mathbb{R}_0^N$ one can identify the particular linear subset spanned by eigenvectors $\mathbf{v}_m^h$ to which it belongs and define the maximum eigenvalue $\varrho(\mathbf{G}_0^h, \mathbf{f})$, in absolute value, among the corresponding eigenvalues, i.e.

$$\varrho(\mathbf{G}_0^h, \mathbf{f}) = \max_m \left\{ |\lambda_m^h| \, : \, \left(\mathbf{v}_m^h, \mathbf{f}\right)_{\gamma_0} \neq 0 \right\},$$

with $\mathbf{f} = \sum_m (\mathbf{v}_m^h, \mathbf{f})_{\gamma_0} \mathbf{v}_m^h$ by definition. By bounding $\boldsymbol{\epsilon}_k^h$ in Equation (44) and owing to the expression of the remainder of a geometric series one obtains the following convergence result.

Proposition 3. *For a given discretization parameter $h = \ell/N$, if the initialization vector $\mathbf{f}_0^h = \mathbf{G}^h \cdot \bar{\mathbf{e}}$ is such that $\varrho(\mathbf{G}_0^h, \mathbf{f}_0^h) < 1$ then the Neumann series (31) converges to the vector $\mathbf{e}^h$ whose components are given by:*

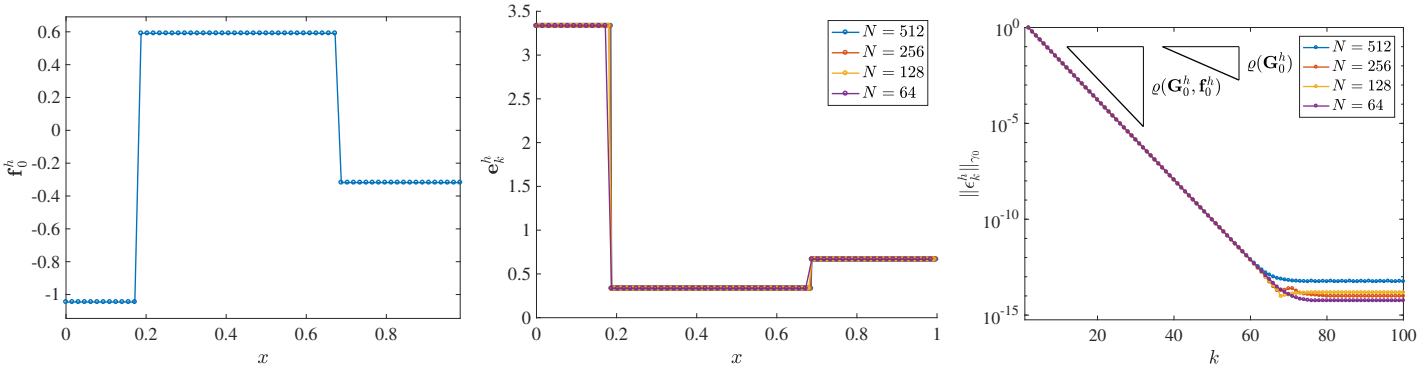
$$(\mathrm{e}^h)_n = \left(\frac{1}{N} \sum_{m=0}^{N-1} \frac{1}{\gamma(x_m)} \right)^{-1} \frac{\bar{e}}{\gamma(x_n)} \qquad \textit{for all } n = 0, \dots, N-1.$$

As such, $\mathbf{e}^h$ is the discrete counterpart of the analytical solution (6). Moreover, the error $\boldsymbol{\epsilon}_k^h = \mathbf{e}_k^h - \mathbf{e}^h$ at step k is bounded as:

$$\|\boldsymbol{\epsilon}_k^h\|_{\gamma_0} \leq \frac{\|\mathbf{f}_0^h\|_{\gamma_0}}{1 - \varrho(\mathbf{G}_0^h, \mathbf{f}_0^h)} \varrho(\mathbf{G}_0^h, \mathbf{f}_0^h)^k.$$

Since by definition it holds $\varrho(\mathbf{G}_0^h) \geq \varrho(\mathbf{G}_0^h, \mathbf{f})$ for all $\mathbf{f} \in \mathbb{R}_0^N$ then the above proposition shows that the convergence of the Neumann series can be faster than what can be expected when assessing only $\varrho(\mathbf{G}_0^h)$ jointly with the upper bound of Property 4. The convergence will be faster than that when a particular subspace of eigenvectors of $\mathbf{G}_0^h$ associated with eigenvalues that are strictly smaller than $\varrho(\mathbf{G}_0^h)$ in absolute value is activated during iterations. Such a result relies on the behavior of the initialization vector $\mathbf{f}_0^h$ relatively to the set of eigenvectors $\{\mathbf{v}_m^h\}_m$ of the matrix $\mathbf{G}_0^h$ associated with the laminate considered.

Test case 1: 3-phase laminate For this configuration, the upper bound of Property 4 is attained. Moreover, From the shape of the initialization vector $\mathbf{f}_0^h$ in Figure 5a and the properties of the eigenvectors illustrated in Fig. 2, the discrete local field vector $\mathbf{e}_k^h$ being computed by iterating in the particular subspace of eigenvectors that are constant in each phase, see Proposition 2, it does not exhibit spurious oscillations within the phases or at the interfaces between phases, see Fig. 5b. Moreover, the quantity $\varrho(\mathbf{G}_0^h, \mathbf{f}_0^h) = 0.6187$ is computed a priori using the eigendecomposition and the corresponding slope is reported in Figure 5c alongside with this associated with the spectral radius $\varrho(\mathbf{G}_0^h) = 0.8182$ for the discretization $N = 64$. In accordance with Proposition 3, it is seen that the convergence rate is proportional to the former.



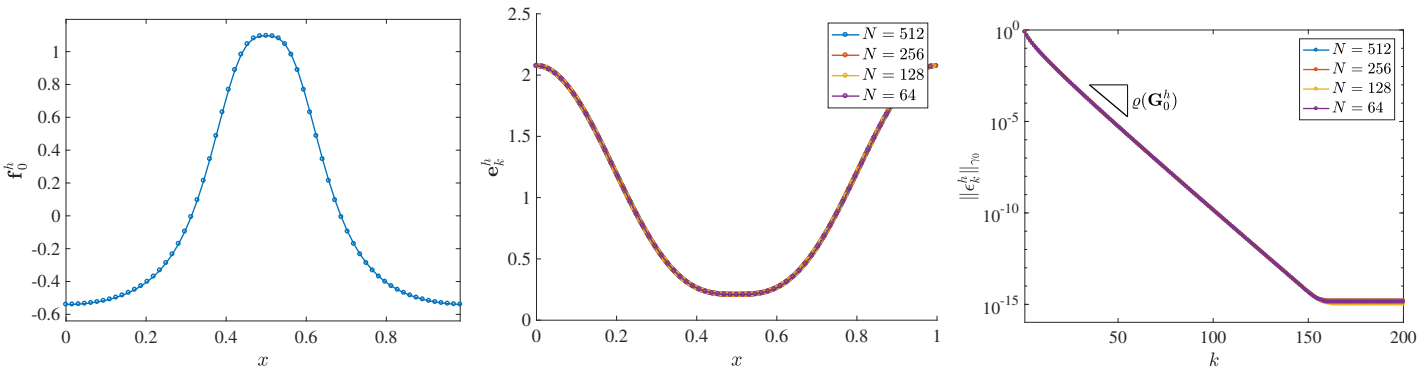
(a) Initialization vector $\mathbf{f}_0^h$ (b) Numerical solution $\mathbf{e}_k^h$ at $k = 100$ (c) Iteration error $\|\boldsymbol{\epsilon}_k^h\|_{\gamma_0}$ on the discrete solution

Figure 5: [3-phase laminate] Computation of the numerical solution by the discrete Neumann series for various discretizations. The reference slopes reported in panel (c) are computed from the eigenanalysis.

This example illustrates that the field $\mathbf{e}_k^h$ computed by the Neumann series converges up to machine precision to the exact discrete solution $\mathbf{e}^h$. Remarkably, this result does not depend on the discretization parameter. Lastly, the convergence behavior observed in Fig. 5c is faster than the upper bound provided in Property 4 and follows this of Proposition 3. Note that, when comparing the different resolutions, some numerical differences are observed at convergence and at a level comparable with the machine precision. To our opinion, this is only related to the numerical rounding error.

Test case 2: smooth distribution In this example the upper bound on the spectral radius in Property 4 is not attained and the numerical value of the corresponding gap is $\left(\max_q |\mathrm{g}_{n_q}^h| - \varrho(\mathbf{G}_0^h)\right) = 1.53 \cdot 10^{-4}$ for $N = 64$. In this example, the conclusions are similar to these of the previous one. However, it can be seen in Fig.

6c that the discrete field $\mathbf{e}_k^h$ converges at a rate equal to the spectral radius $\varrho(\mathbf{G}_0^h) = 0.8180$, which is computed a priori using the eigendecomposition. Indeed, for the configuration considered, the initialization vector $\mathbf{f}_0^h$ plotted in Figure 6a has a non-zero projection onto the eigenvector associated with the largest eigenvalue.



(a) Initialization vector $\mathbf{f}_0^h$ (b) Numerical solution $\mathbf{e}_k^h$ at $k = 200$ (c) Iteration error $\|\boldsymbol{\epsilon}_k^h\|_{\gamma_0}$ on the discrete solution

Figure 6: [Smooth distribution] Computation of the numerical solution by the discrete Neumann series for various discretizations. The slope corresponding to the spectral radius computed from the eigenanalysis is reported in panel (c).

Remark 2 (An additional remarkable example)**.** *Based on the above developments, it is possible to design an example for which the Neumann series converges in only one iteration. Indeed, let us consider a two phase laminate where* $\gamma(x) = \gamma_p$ *for* $p = 1, 2$ *and* $\gamma_1 = 1$, $\gamma_2 = 10$ *with the corresponding phase being of size* $c_1 = 1/4$ *and* $c_2 = 3/4$ *respectively. For this example, we would like to address the particular situation described in Proposition 1 where* $\dim(\mathcal{N}(\mathbf{G}_0^h)) = 2$*. This occurs when* $\langle 1/\mathbf{g}^h \rangle_h = 0$*, i.e. when* $\gamma_0 = c_1\gamma_2 + c_2\gamma_1$*. For such a choice of the reference conductivity, the spectrum consists of eigenvectors of type (a) which play no role in the construction of the solution and two vectors that span the kernel of the matrix* $\mathcal{N}(\mathbf{G}_0^h)$*. For these vectors, the radius of convergence of the Neumann series is zero by definition, so that the latter converges at the first iterate. This astonishing property is illustrated in Figure 7 where the norm of the error on the discrete solution is plotted as a function of the iteration number* k *for different values of the reference conductivity* γ_0 *for the discretization* $N = 512$*. As expected, convergence in only one iteration is observed when choosing* $\gamma_0 = c_1\gamma_2 + c_2\gamma_1$*. Note that this one-step convergence, independently of the phase conductivity values, has been described in Vinogradov and Milton (2008) for the particular case* $c_1 = c_2 = 1/2$*.*

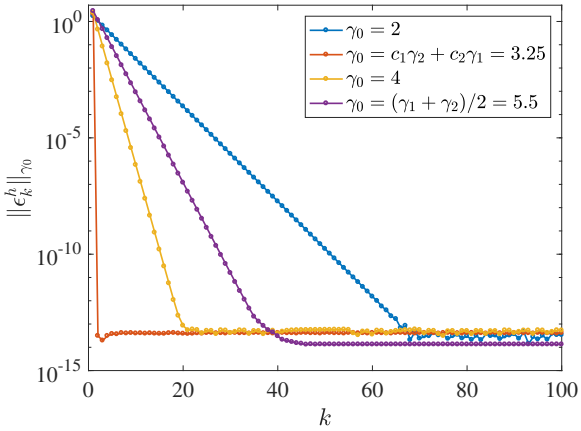


Figure 7: [2-phase laminate] A remarkable convergence result for a 2-phase laminate with $\gamma_1 = 1$ and $\gamma_2 = 10$. Norm $\|\boldsymbol{\epsilon}_k^h\|_{\gamma_0}$ of the iteration error on the discrete solution for different values of the reference conductivity γ_0 and for the discretization $N = 512$.

4.2 Convergence of the effective property

In this section, the objective is to investigate the convergence rate of the discrete effective property $(\gamma_{\text{eff}})_k^h$ in (32), whose limit is denoted as $(\gamma_{\text{eff}})^h$ if it converges when $k \to \infty$. Quadratic upper and lower bounds are discussed in this section. Such bounds are classic in optimization, see e.g. Nesterov (2004), but they are derived below for the reader's convenience. As a starting point, owing to the identity (29) and since $\gamma \in L^\infty_{\text{per}}(0,\ell)$ with $\gamma(x) \geq \kappa > 0$ then $\mathbf{f} \mapsto \boldsymbol{\nabla}\mathrm{W}^h(\mathbf{f})$ is both Lipschitz continuous and coercive on $\mathbb{R}^N_0$, i.e. there exist two constants $M \geq m > 0$ such that for all $\mathbf{f}_1, \mathbf{f}_2 \in \mathbb{R}^N_0$:

$$\|\boldsymbol{\nabla}\mathrm{W}^h(\mathbf{f}_1) - \boldsymbol{\nabla}\mathrm{W}^h(\mathbf{f}_2)\|^2_{\gamma_0} \leq M\|\mathbf{f}_1 - \mathbf{f}_2\|^2_{\gamma_0} \qquad \text{and} \qquad \big(\boldsymbol{\nabla}\mathrm{W}^h(\mathbf{f}_1) - \boldsymbol{\nabla}\mathrm{W}^h(\mathbf{f}_2), \mathbf{f}_1 - \mathbf{f}_2\big)_{\gamma_0} \geq m\|\mathbf{f}_1 - \mathbf{f}_2\|^2_{\gamma_0}. \quad (45)$$

The Lipschitz and coercivity constants, M and m respectively, satisfy:

$$M = \frac{\max_n(\gamma_n^h)}{\gamma_0} \leq \frac{\max_x(\gamma(x))}{\gamma_0} \qquad \text{and} \qquad m = \frac{\min_n(\gamma_n^h)}{\gamma_0} \geq \frac{\min_x(\gamma(x))}{\gamma_0}. \quad (46)$$

From the first inequality in (45) one gets that $\mathbf{f} \mapsto \frac{M}{2}\|\mathbf{f}\|^2_{\gamma_0} - \mathrm{W}^h(\mathbf{f})$ is convex on $\mathbb{R}^N_0$, from which one obtains the following quadratic upper bound:

$$\mathrm{W}^h(\mathbf{f}_1) \leq \mathrm{W}^h(\mathbf{f}_2) + \big(\boldsymbol{\nabla}\mathrm{W}^h(\mathbf{f}_2), \mathbf{f}_1 - \mathbf{f}_2\big)_{\gamma_0} + \frac{M}{2}\|\mathbf{f}_1 - \mathbf{f}_2\|^2_{\gamma_0} \qquad \forall \mathbf{f}_1, \mathbf{f}_2 \in \mathbb{R}^N_0. \quad (47)$$

Upon setting $\mathbf{f}_1 = \tilde{\mathbf{e}}_k^h = \mathbf{e}_k^h - \overline{\mathbf{e}}$ and $\mathbf{f}_2 = \tilde{\mathbf{e}}^h = \mathbf{e}^h - \overline{\mathbf{e}}$ in this identity one gets

$$\mathrm{W}^h(\tilde{\mathbf{e}}_k^h) - \mathrm{W}^h(\tilde{\mathbf{e}}^h) \leq \frac{M}{2}\|\tilde{\mathbf{e}}_k^h - \tilde{\mathbf{e}}^h\|^2_{\gamma_0}. \quad (48)$$

since, by definition, the vector $\tilde{\mathbf{e}}^h$ being the solution to the variational principle (28) it satisfies the optimality condition $\boldsymbol{\nabla}\mathrm{W}^h(\tilde{\mathbf{e}}^h) = \mathbf{0}$. Moreover, from (47) again one has

$$\mathrm{W}^h(\tilde{\mathbf{e}}^h) \leq \min_{\mathbf{e}' \in \mathbb{R}^N_0} \mathrm{W}^h(\mathbf{e}') \leq \min_{\mathbf{e}' \in \mathbb{R}^N_0} \Big\{ \mathrm{W}^h(\mathbf{f}_2) + \big(\boldsymbol{\nabla}\mathrm{W}^h(\mathbf{f}_2), \mathbf{e}' - \mathbf{f}_2\big)_{\gamma_0} + \frac{M}{2}\|\mathbf{e}' - \mathbf{f}_2\|^2_{\gamma_0} \Big\} \qquad \forall \mathbf{f}_2 \in \mathbb{R}^N_0. \quad (49)$$

The Euler-Lagrange equation associated with this minimization problem reads:

$$\big(\boldsymbol{\nabla}\mathrm{W}^h(\mathbf{f}_2) + M(\mathbf{e}' - \mathbf{f}_2), \tilde{\mathbf{e}}'\big)_{\gamma_0} = 0 \qquad \forall \tilde{\mathbf{e}}' \in \mathbb{R}^N_0.$$

As all the terms in this identity belongs to $\mathbb{R}^N_0$, the above condition implies that the minimizer in (49) is given by:

$$\mathbf{e}' = \mathbf{f}_2 - \frac{1}{M}\boldsymbol{\nabla}\mathrm{W}^h(\mathbf{f}_2),$$

which inserted back in (49) yields the following inequality for all $\mathbf{f}_2 \in \mathbb{R}^N_0$:

$$\mathrm{W}^h(\mathbf{f}_2) - \mathrm{W}^h(\tilde{\mathbf{e}}^h) \geq \frac{1}{2M}\|\boldsymbol{\nabla}\mathrm{W}^h(\mathbf{f}_2)\|^2_{\gamma_0}. \quad (50)$$

Finally, using (50) with $\mathbf{f}_2 = \tilde{\mathbf{e}}_k^h$ and combining that inequality with (48) yield

$$\frac{1}{2M}\|\boldsymbol{\nabla}\mathrm{W}^h(\tilde{\mathbf{e}}_k^h)\|^2_{\gamma_0} \leq \mathrm{W}^h(\tilde{\mathbf{e}}_k^h) - \mathrm{W}^h(\tilde{\mathbf{e}}^h) \leq \frac{M}{2}\|\boldsymbol{\epsilon}_k^h\|^2_{\gamma_0} \quad (51)$$

since the error $\boldsymbol{\epsilon}_k^h$ on the fields satisfies $\boldsymbol{\epsilon}_k^h = \mathbf{e}_k^h - \mathbf{e}^h = \tilde{\mathbf{e}}_k^h - \tilde{\mathbf{e}}^h$.

By using the coercivity property of (45) one can show that the functional $\mathbf{f} \mapsto \mathrm{W}^h(\mathbf{f}) - \frac{m}{2}\|\mathbf{f}\|^2_{\gamma_0}$ is convex on $\mathbb{R}^N_0$, from which a quadratic lower bound on W^h analogous to (47) can be obtained. The former can in turn be used to obtain:

$$\frac{m}{2}\|\boldsymbol{\epsilon}_k^h\|^2_{\gamma_0} \leq \mathrm{W}^h(\tilde{\mathbf{e}}_k^h) - \mathrm{W}^h(\tilde{\mathbf{e}}^h) \leq \frac{1}{2m}\|\boldsymbol{\nabla}\mathrm{W}^h(\tilde{\mathbf{e}}_k^h)\|^2_{\gamma_0}. \quad (52)$$

Therefore, by combining the inequalities (51) and (52) together with Proposition 3 one obtains the following result.

Proposition 4. *For a given discretization parameter $h = \ell/N$, if the initialization vector $\mathbf{f}_0^h = \mathbf{G}^h \cdot \bar{\mathbf{e}}$ is such that $\varrho(\mathbf{G}_0^h, \mathbf{f}_0^h) < 1$ then $(\gamma_{\text{eff}})_k^h$ in (32) converges to the quantity $(\gamma_{\text{eff}})^h$ given by*

$$(\gamma_{\text{eff}})^h = \left(\frac{1}{N} \sum_{n=0}^{N-1} \frac{1}{\gamma(x_n)} \right)^{-1}$$

and there exist two constants $C \geq c > 0$, which depend on γ, γ_0 and $\bar{e}$, such that the error in effective property is bounded as

$$c\, \varrho(\mathbf{G}_0^h, \mathbf{f}_0^h)^{2k} \leq (\gamma_{\text{eff}})_k^h - (\gamma_{\text{eff}})^h \leq C \varrho(\mathbf{G}_0^h, \mathbf{f}_0^h)^{2k}.$$

Moreover, there also exist two constants $C' \geq c' > 0$ such that the above error satisfies:

$$c' \|\boldsymbol{\nabla} \mathrm{W}^h(\tilde{\mathbf{e}}_k^h)\|_{\gamma_0}^2 \leq (\gamma_{\text{eff}})_k^h - (\gamma_{\text{eff}})^h \leq C' \|\boldsymbol{\nabla} \mathrm{W}^h(\tilde{\mathbf{e}}_k^h)\|_{\gamma_0}^2.$$

As shown in the above proposition, the error associated with the computation of the effective property converges to zero twice as fast as (in logarithmic scale) the norm of the error $\|\boldsymbol{\epsilon}_k^h\|_{\gamma_0}$ associated with the fields themselves. Indeed the convergence rate is proportional to $\varrho(\mathbf{G}_0^h, \mathbf{f}_0^h)^{2k}$ for the former and $\varrho(\mathbf{G}_0^h, \mathbf{f}_0^h)^{k}$ for the latter. It is worth noting that this is due to the fact that the effective property $(\gamma_{\text{eff}})_k^h$ is computed based on the energetic definition (32). Alternatively, one can consider the effective property $(\check{\gamma}_{\text{eff}})_k^h$ defined as:

$$(\check{\gamma}_{\text{eff}})_k^h = \frac{\langle \mathbf{j}_k^h \rangle_h}{\bar{e}} \qquad \text{with} \qquad \mathbf{j}_k^h = \mathbf{Diag}[\gamma^h] \cdot \mathbf{e}_k^h \tag{53}$$

which converges to a limit $(\check{\gamma}_{\text{eff}})^h$ when it does converge. Owing to Property 2 and Proposition 3, it is clear that the limits of (53) and (32) coincide, i.e. $(\check{\gamma}_{\text{eff}})^h = (\gamma_{\text{eff}})^h$, but the convergence error associated with (53) is bounded as:

$$\left|(\check{\gamma}_{\text{eff}})_k^h - (\gamma_{\text{eff}})^h\right| \leq \check{C} \varrho(\mathbf{G}_0^h, \mathbf{f}_0^h)^k$$

where $\check{C} > 0$ is a constant that depends on γ, γ_0 and $\bar{e}$. This entails that, if the effective property is defined by (53), then its convergence rate can only be bounded by this of the norm of the error on the fields. It implies convergence at a rate possibly slower than this associated with the energetic definition (32).

Lastly, Proposition 4 confirms that the norm $\|\boldsymbol{\nabla} \mathrm{W}^h(\tilde{\mathbf{e}}_k^h)\|_{\gamma_0}$ yields a reliable stopping criterion to assess numerically the convergence of the Neumann series (31) as of the computation of the effective property associated with the discretization considered. Owing to (52), this criterion is also relevant to the convergence of the local fields. Note finally that, if one is interested in a quantitative comparison between the convergence rates associated with different choices of the reference medium, i.e. for different values of γ_0, then the use of the energetic norm makes it necessary in practice to consider a *normalized* version of the stopping criterion.

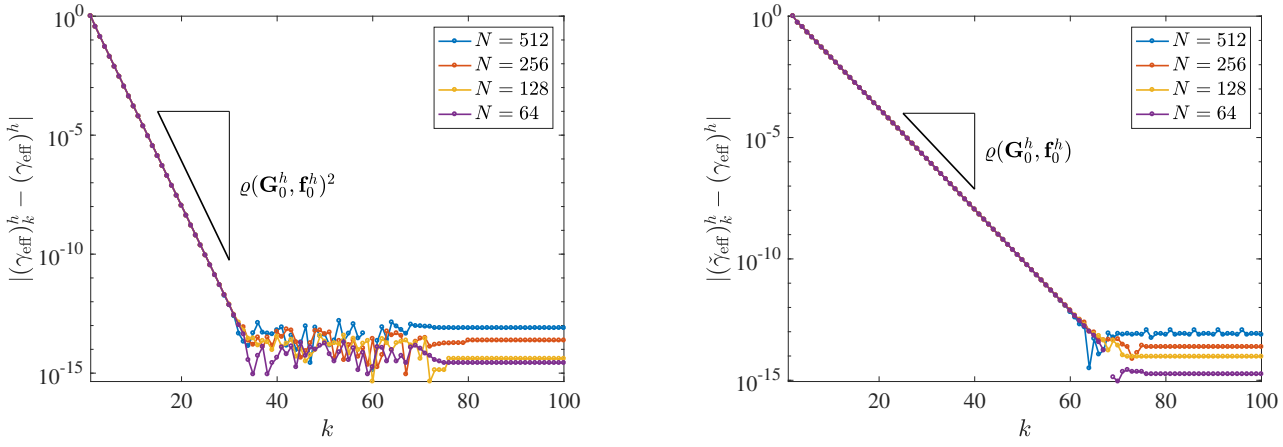


Figure 8: [3-phase laminate] Error on the effective property measured as functions of the iteration number k. *(left)* Quantity $(\gamma_{\text{eff}})_k^h$ from energetic definition (32) and *(right)* quantity $(\check{\gamma}_{\text{eff}})_k^h$ using the alternative definition (53). The reference slopes indicated are computed from the eigenanalysis.

Test case 1: 3-phase laminate As seen in Fig. 5c, the Neumann series converges to the exact discrete solution up to machine precision, in terms of the discrete field $\mathbf{e}_k^h$, and this convergence extends to the computation of the effective property $(\gamma_{\text{eff}})_k^h$ as well, see Fig. 8. In Figure 8, the reported slopes corresponding to $\varrho(\mathbf{G}_0^h, \mathbf{f}_0^h)^2$ and $\varrho(\mathbf{G}_0^h, \mathbf{f}_0^h)$ are computed a priori, based on the eigenanalysis, and the observed evolutions of the curves confirm that their convergence rates are these described in Proposition 4 and in the ensuing discussion.

Test case 2: smooth distribution For the second test case, the numerical results illustrate again the convergence of the effective property $(\gamma_{\text{eff}})_k^h$ to the discrete solution up to machine precision. In the right panel of Figure 9 it is seen that, using the definition (53), the convergence is not monotonic, whereas it appears to be so in the left panel, where $(\gamma_{\text{eff}})_k^h$ is computed based on the energetic definition (32). Again, the convergence rate is quadratic for this latter computation and related to the upper bound associated with the spectral radius $\varrho(\mathbf{G}_0^h)$.

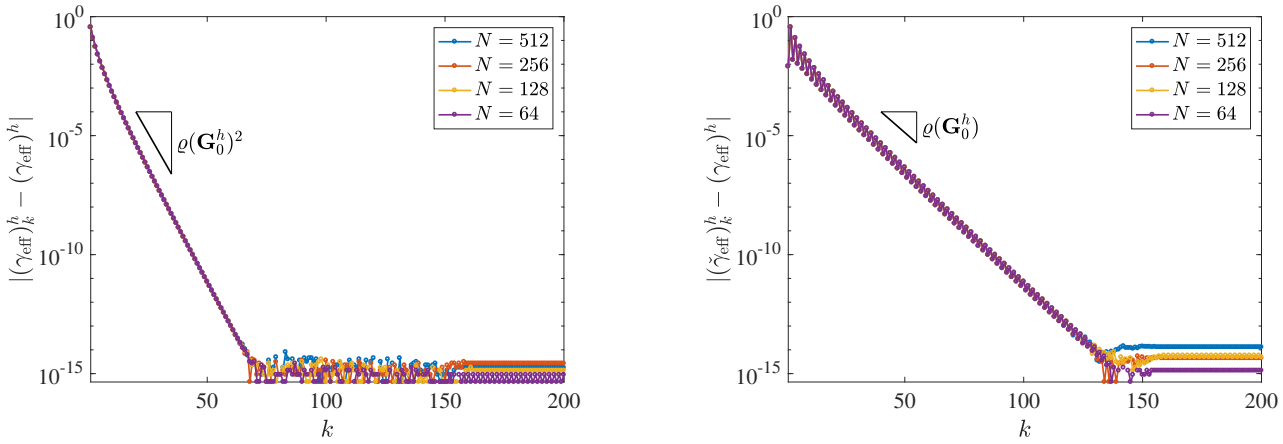


Figure 9: [Smooth distribution] Error on the effective property measured as functions of the iteration number k. *(left)* Quantity $(\gamma_{\text{eff}})_k^h$ from energetic definition (32) and *(right)* quantity $(\check{\gamma}_{\text{eff}})_k^h$ using the alternative definition (53). Reference slopes associated with the spectral radius are indicated.

4.3 Monotonicity properties

As discussed in Section 2.3, the iterations associated with the Neumann series (31) are equivalent to the gradient-descent scheme with fixed step (15) for the discrete minimum energy principle (28). In this context, this section focuses on the monotonicity properties associated with such a scheme.

First, upon setting $\mathbf{f}_1 = \tilde{\mathbf{e}}_{k+1}^h$ and $\mathbf{f}_2 = \tilde{\mathbf{e}}_k^h$ in (47), then one has

$$\mathrm{W}^h(\tilde{\mathbf{e}}_{k+1}^h) \leq \mathrm{W}^h(\tilde{\mathbf{e}}_k^h) - t_k\Big(1 - \frac{Mt_k}{2}\Big)\|\boldsymbol{\nabla}\mathrm{W}^h(\tilde{\mathbf{e}}_k^h)\|_{\gamma_0}^2,$$

where we made use of the relation $\tilde{\mathbf{e}}_{k+1}^h = \tilde{\mathbf{e}}_k^h - t_k\boldsymbol{\nabla}\mathrm{W}^h(\tilde{\mathbf{e}}_k^h)$ that holds between two successive iterates in terms of a generic fixed step $t_k > 0$. As a consequence, the discrete energy W^h is guaranteed to decrease monotonically at each iterate provided that $t_k \leq 2/M$, where the Lipschitz constant M is given by (46). With the Neumann series (31) corresponding to $t_k = 1$, the next property summarizes this discussion.

Property 5. *The energy* $\mathrm{W}^h(\tilde{\mathbf{e}}_k^h)$ *associated with the iterates of the Neumann series* (31) *is a* strictly *decreasing function of* k *as long as the reference medium is such that* $\gamma_0 > \max_n(\gamma_n^h)/2$. *This is satisfied in particular when* $\gamma_0 = (\max_n(\gamma_n^h) + \min_n(\gamma_n^h))/2$, *and in which case one has:*

$$\mathrm{W}^h(\tilde{\mathbf{e}}_{k+1}^h) - \mathrm{W}^h(\tilde{\mathbf{e}}_k^h) \leq -\big(1 + \max_n(\gamma_n^h)/\min_n(\gamma_n^h)\big)^{-1}\|\boldsymbol{\nabla}\mathrm{W}^h(\tilde{\mathbf{e}}_k^h)\|_{\gamma_0}^2.$$

Property 5 concerns the monotonicity property of the energy functional. Hereafter, we discuss the fact that the iterated vector $\tilde{\mathbf{e}}_k^h$ is itself characterized by a monotonic convergence. This relies on the coercivity property in (45), which entails that the functional $\mathrm{K} : \mathbf{f} \mapsto \mathrm{W}^h(\mathbf{f}) - \frac{m}{2}\|\mathbf{f}\|_{\gamma_0}^2$ is convex on $\mathbb{R}_0^N$ as seen previously. In

turn, this implies that $\mathbf{f} \mapsto \boldsymbol{\nabla}\mathrm{K}(\mathbf{f})$ is Lipschitz continuous with constant $(M-m)$. As a consequence, $\boldsymbol{\nabla}\mathrm{K}$ is co-coercive, i.e. it satisfies

$$\big(\boldsymbol{\nabla}\mathrm{K}(\mathbf{f}_1)-\boldsymbol{\nabla}\mathrm{K}(\mathbf{f}_2),\mathbf{f}_1-\mathbf{f}_2\big)_{\gamma_0} \geq \frac{1}{M-m}\|\boldsymbol{\nabla}\mathrm{K}(\mathbf{f}_1)-\boldsymbol{\nabla}\mathrm{K}(\mathbf{f}_2)\|^2_{\gamma_0} \qquad \forall \mathbf{f}_1,\mathbf{f}_2 \in \mathbb{R}^N_0.$$

By making the expression for $\boldsymbol{\nabla}\mathrm{K}$ explicit into this inequality one finally obtains

$$\big(\boldsymbol{\nabla}\mathrm{W}^h(\mathbf{f}_1)-\boldsymbol{\nabla}\mathrm{W}^h(\mathbf{f}_2),\mathbf{f}_1-\mathbf{f}_2\big)_{\gamma_0} \geq \frac{mM}{M+m}\|\mathbf{f}_1-\mathbf{f}_2\|^2_{\gamma_0}+\frac{1}{M+m}\|\boldsymbol{\nabla}\mathrm{W}^h(\mathbf{f}_1)-\boldsymbol{\nabla}\mathrm{W}^h(\mathbf{f}_2)\|^2_{\gamma_0} \qquad \forall \mathbf{f}_1,\mathbf{f}_2 \in \mathbb{R}^N_0. \quad (54)$$

With this result at hand, let consider the residual, with respect to the solution $\tilde{\mathbf{e}}^h$ of the minimum energy principle (28), of the gradient-descent scheme at step $k+1$, i.e.

$$\begin{aligned}\|\tilde{\mathbf{e}}^h_{k+1}-\tilde{\mathbf{e}}^h\|^2_{\gamma_0} &= \|\tilde{\mathbf{e}}^h_k-\tilde{\mathbf{e}}^h-t_k\boldsymbol{\nabla}\mathrm{W}^h(\tilde{\mathbf{e}}^h_k)\|^2_{\gamma_0}\\ &= \|\tilde{\mathbf{e}}^h_k-\tilde{\mathbf{e}}^h\|^2_{\gamma_0}+t_k^2\|\boldsymbol{\nabla}\mathrm{W}^h(\tilde{\mathbf{e}}^h_k)\|^2_{\gamma_0}-2t_k\big(\boldsymbol{\nabla}\mathrm{W}^h(\tilde{\mathbf{e}}^h_k),\tilde{\mathbf{e}}^h_k-\tilde{\mathbf{e}}^h\big)_{\gamma_0}\end{aligned}$$

By using the inequality (54) with $\mathbf{f}_1=\tilde{\mathbf{e}}^h_k$ and $\mathbf{f}_2=\tilde{\mathbf{e}}^h$, which satisfies the optimality condition $\boldsymbol{\nabla}\mathrm{W}^h(\tilde{\mathbf{e}}^h)=\mathbf{0}$, then the previous inequality finally entails:

$$\|\tilde{\mathbf{e}}^h_{k+1}-\tilde{\mathbf{e}}^h\|^2_{\gamma_0} \leq \Big(1-\frac{2mMt_k}{M+m}\Big)\|\tilde{\mathbf{e}}^h_k-\tilde{\mathbf{e}}^h\|^2_{\gamma_0}+t_k\Big(t_k-\frac{2}{M+m}\Big)\|\boldsymbol{\nabla}\mathrm{W}^h(\tilde{\mathbf{e}}^h_k)\|^2_{\gamma_0}. \quad (55)$$

As a consequence of this inequality, when the step satisfies $0<t_k\leq 2/(M+m)$ then the field error $\boldsymbol{\epsilon}^h_k=\tilde{\mathbf{e}}^h_k-\tilde{\mathbf{e}}^h=\mathbf{e}^h_k-\mathbf{e}^h$ satisfies $\|\boldsymbol{\epsilon}^h_{k+1}\|^2_{\gamma_0}<\|\boldsymbol{\epsilon}^h_k\|^2_{\gamma_0}$. Moreover, given that the multiplicative factor in the first right-hand side term of (55) satisfies

$$1-\frac{2mMt_k}{M+m}=1-\frac{\max_n(\gamma^h_n)\min_n(\gamma^h_n)}{\max_n(\gamma^h_n)+\min_n(\gamma^h_n)}\frac{t_k}{\gamma_0}$$

then two different strategies can be adopted to optimize the convergence rate of the gradient-descent scheme with fixed step (15): (i) Maximize t_k while the reference medium γ_0 is chosen arbitrarily. This leads to the optimal value of the fixed step $t_k=2\gamma_0/(\max_n(\gamma^h_n)+\min_n(\gamma^h_n))$. (ii) Set $t_k=1$ and minimize γ_0, which leads to the optimal value of the reference medium conductivity $\gamma_0=(\max_n(\gamma^h_n)+\min_n(\gamma^h_n))/2$. As the case $t_k=1$ is the one corresponding to the Neumann series, one obtains the next property.

Property 6. *The residual error $\|\boldsymbol{\epsilon}^h_k\|_{\gamma_0}$ on the field computed by the Neumann series (31) is a* strictly *decreasing function of k as long as $\gamma_0 \geq (\max_n(\gamma^h_n)+\min_n(\gamma^h_n))/2$. In the optimal case where $\gamma_0=(\max_n(\gamma^h_n)+\min_n(\gamma^h_n))/2$ then one has:*

$$\|\boldsymbol{\epsilon}^h_{k+1}\|_{\gamma_0} \leq \beta\|\boldsymbol{\epsilon}^h_k\|_{\gamma_0} \qquad \textit{with} \qquad \beta=\frac{\max_n(\gamma^h_n)-\min_n(\gamma^h_n)}{\max_n(\gamma^h_n)+\min_n(\gamma^h_n)}.$$

Property 6 shows that the Neumann series converges towards the *discrete* solution as a geometric series with ratio β. However, it yields an upper-bound on the converge rate that coincides exactly with this of Property 4, as it can be checked. Therefore, this property provides a key information on the monotonicity of the convergence of the iterates while the convergence rate itself is better estimated in Proposition 3 based on the eigendecomposition of the iterated matrix $\mathbf{G}^h_0$.

Test cases The numerical results associated with the test cases 1 and 2 are in agreement with the properties 5 and 6. First, the choice of the reference medium $\gamma_0=(\max_n(\gamma^h_n)+\min_n(\gamma^h_n))/2$ is justified by Property 6 and both the field error and the energy-based effective property are then observed to converge monotonically up to the machine precision, see the figures 5c, 6c and the left panel of the figures 8 and 9 respectively. It is also remarkable that, for the alternative definition (53) of the effective property, the associated behavior in the case of the second material distribution (40) is not monotonic, see the right panel of Figure 9.

Remark 3. *In the developments of this section, it is implicitly assumed that $\min_n(\gamma^h_n)>0$, a requirement which is self-evident for the 1D material setting considered. In 2D and 3D such an assumption will however be violated in the case of porous materials. In such a case, the results of Section 4.2 have to be revised too, since the energy functional W^h is no longer coercive with $m=0$ in (45).*

Remark 4. *To conclude, it should be noted that the results of the sections 4.2 and 4.3 are classic in optimization, see e.g. Nesterov (2004) and they stem from the properties (45) of the energy functional W^h. While the results of these sections are investigated here within a discretized setting and in the case of conductive laminates, it is straightforward to obtain them at the continuous level too and for any constitutive relations provided that the gradient $\boldsymbol{\nabla}W$ of the corresponding energy functional is both coercive and Lipschitz continuous.*

5 Convergence of the discrete solution

5.1 Fourier-based analysis

Now that the convergence of the discrete Neumann series has been investigated, this section focuses on the convergence of the discrete solution with regard to the evaluation $\mathbf{e}$ of the exact continuous solution at the grid points. The propositions 3 and 4 show that evaluating the error $\|\mathbf{e}^h - \mathbf{e}\|_{\gamma_0}$ in (33) boils down to assessing the following approximation error:

$$\epsilon^h = |I^h - I| \qquad \text{where} \qquad I^h = \frac{1}{N}\sum_{n=0}^{N-1} \frac{1}{\gamma(x_n)} \qquad \text{and} \qquad I = \left\langle \frac{1}{\gamma} \right\rangle \tag{56}$$

owing to the definition (23) of the numerical integration scheme using the trapezoidal rule. This quadrature is *exact* for functions that are piecewise constant or linear on the discretization grid. It is also known to be such that the numerical integration error ϵ^h decreases at worst as $\mathcal{O}(1/N^2)$ if γ^{-1} is twice differentiable. However, depending on the smoothness properties of γ^{-1} the convergence can be much faster, see e.g. Boyd (2000), even exponential if γ^{-1} is analytic Trefethen and Weideman (2014). In fact, this convergence rate can be driven by the decaying behavior of the Fourier transform of the integrand, which can be shown using simple arguments that we produce hereafter for the reader's convenience. Upon setting $\rho = \gamma^{-1}$ then $\rho \in L^2_{\text{per}}(0,\ell)$ and one introduces its Fourier series $S[\rho]$ as

$$S[\rho](x) = \sum_{\nu\in\mathbb{Z}/\ell} \hat{\rho}(\nu) e^{2\mathrm{i}\pi\nu x} \qquad \text{with} \qquad \hat{\rho}(\nu) = \frac{1}{\ell}\int_0^\ell \rho(x) e^{-2\mathrm{i}\pi\nu x}\,\mathrm{d}x, \tag{57}$$

so that $\langle\rho\rangle = \hat{\rho}(0)$. Assuming that $S[\rho]$ converges pointwisely to ρ at the discretization points x_n, then the approximation I^h of the integral I of ρ over the interval $(0,\ell)$ using the trapezoidal rule satisfies

$$I^h = \frac{1}{N}\sum_{n=0}^{N-1} \rho(x_n) = \frac{1}{N}\sum_{n=0}^{N-1}\sum_{\nu\in\mathbb{Z}/\ell} \hat{\rho}(\nu) e^{2\mathrm{i}\pi\nu x_n} = \frac{1}{N}\sum_{\nu\in\mathbb{Z}/\ell} \hat{\rho}(\nu) \sum_{n=0}^{N-1} e^{2\mathrm{i}\pi\nu n\frac{\ell}{N}},$$

which can be rewritten as

$$I^h = I + \frac{1}{N}\sum_{\substack{\nu\in\mathbb{Z}/\ell\\ \nu\neq 0}} \hat{\rho}(\nu) \sum_{n=0}^{N-1} e^{2\mathrm{i}\pi\nu n\frac{\ell}{N}}. \tag{58}$$

In the last term, the second sum is zero unless $\nu = mN/\ell$ with $m \in \mathbb{Z}$, in which case it is equal to N. This is the phenomenon of aliasing, which is characterized by the fact that the functions $x \mapsto e^{2\mathrm{i}\pi m N x/\ell}$ are indistinguishable from the constant function 1 on the grid considered so that, in (58), they are accountable for the integration error in the trapezoidal rule. As a consequence, one arrives at the property below.

Property 7. *The discretization error* (56) *satisfies* $\epsilon^h \leq \left|\sum_{m=1}^{\infty}\big(\hat{\rho}(m/h) + \hat{\rho}(-m/h)\big)\right|$ *with* $h = \ell/N$.

Property of continuous function $\rho(x)$	Behavior of Fourier coefficient $\hat{\rho}(\nu)$
Piecewise continuous and of bounded variation	$\mathcal{O}\big(\lvert\nu\rvert^{-1}\big)$
$\rho \in C^{r-1}_{\text{per}}(0,\ell)$ and $\rho^{(r)} \in L^1_{\text{per}}(0,\ell)$	$\mathcal{O}\big(\lvert\nu\rvert^{-r}\big)$
$\rho \in C^{r-1}_{\text{per}}(0,\ell)$ and $\rho^{(r)}$ of bounded variation	$\mathcal{O}\big(\lvert\nu\rvert^{-(r+1)}\big)$
$C^{\infty}_{\text{per}}(0,\ell)$	$\mathcal{O}\big(\lvert\nu\rvert^{-r}\big)$ for all $r \geq 0$

Table 1: Examples of convergence rates of Fourier coefficients.

This property implies that, if the integrand in (23) were a trigonometric polynomial of the form $e^{\pm 2\mathrm{i}\pi n x/\ell}$ with $n \geq 0$, then the trapezoidal rule would be *exact* for all $N > n$. This result seems in contradiction with the

Nyquist criterion of sampling a band-limited function at twice its cut-off frequency. Yet, this apparent paradox has been extensively discussed and resolved in Trefethen and Weideman (2014).

According to Property 7, the error ϵ^h converges to zero at a rate that is governed by the decay of the Fourier coefficients $\hat{\rho}(\pm m/h)$ for $m \geq 1$. The properties of these coefficients are well-known for a broad class of functions, see e.g. Gottlieb and Orszag (1977); Boyd (2000), and some classic results are collected in Table 1 to help evaluate the discretization error.

Aliasing and the Gibbs phenomenon For the 1D problem considered and according to the preceding developments, the aliasing effects are entirely accountable for the error between the discrete solution $\mathbf{e}^h$ and the theoretical one $\mathbf{e}$, and likewise for $(\gamma_{\text{eff}})^h$ with respect to γ_{eff}. To prove this, the starting point is the assumption that the Fourier series (57) converges pointwisely to ρ at the grid points x_n. Excluding some very atypical functions, this assumption is valid as long as the spatial discretization is compatible with the function ρ. For example, if ρ were discontinuous then, in a standard discretization, the discontinuity would be placed in-between two grid points so that the *infinite* Fourier series $S[\rho]$ would converge to ρ at these points.

In this context, if one considers the M-th partial sum $S_M[\rho]$ of the Fourier series of ρ, i.e. a *finite* truncated version of (57), then $S_M[\rho](x)$ can exhibit a local oscillatory behavior near discontinuities. This is the Gibbs phenomenon that depends both on the discretization and truncation parameters N and M respectively. It can occur in particular when using the Discrete Fourier Transform (DFT) while interpolating on refined grids, i.e. evaluating the discrete fields in-between grid points. That being said, neither the DFT nor truncated Fourier series are used in the present study and the Gibbs phenomenon plays no role in it.

5.2 Numerical examples

Test case 1: 3-phase laminate Considering the material distribution of Figure 1a, the corresponding function $1/\gamma$ is integrated exactly by the trapezoidal rule. The errors quantifying the discrepancy between the discrete iterated solution and the theoretical one are plotted in Figure 10. As expected, these errors do not depend on the discretization parameter N and convergence to the exact solution is obtained up to machine precision, see also Fig. 11.

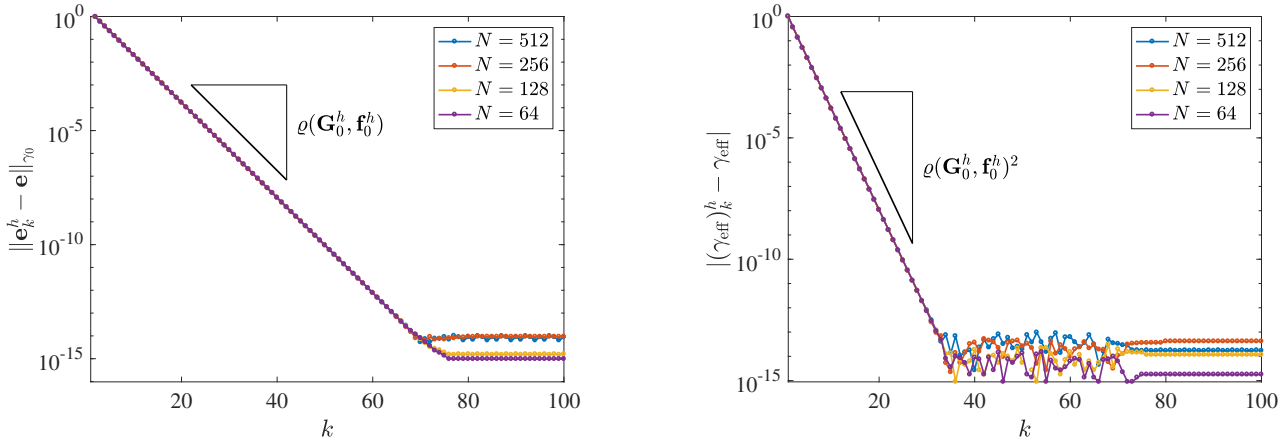


Figure 10: [3-phase laminate] Numerical errors to the analytical solution as functions of the iteration number k and when varying the discretization as $N \in \{64, 128, 256, 512\}$. *(left)* Global field error $\|\mathbf{e}_k^h - \mathbf{e}\|_{\gamma_0}$ and *(right)* global error on the effective property using its energetic definition (32).

Test case 2: Smooth distribution In the second example, the material distribution defined by (40) and shown in Fig. 3a is chosen so that, when r is odd, the function $\rho = \gamma^{-1}$ belongs to $C^{r-1}_{\text{per}}(0, \ell)$ but not to $C^r_{\text{per}}(0, \ell)$ due to a singularity of the derivative $\rho^{(r)}$ at $x = 1/2$. However, $\rho^{(r)}$ is of bounded variation. As a consequence, from the analysis of Section 5.1 and the properties reported in Table 1, the discretization error associated with the trapezoidal rule (23) is expected to decay as $\mathcal{O}(1/N^{r+1})$. Global errors are computed from the identity (41) and shown in Figure 12 as functions of the iteration number k and for various values of the discretization parameter N. Unlike in the example of the 3-phase laminate, a dependence on N is observed as

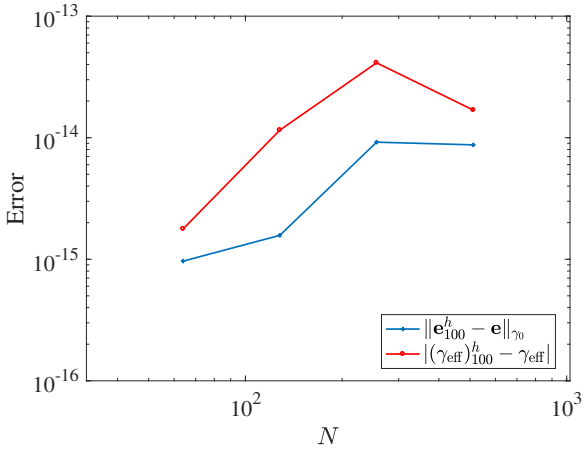


Figure 11: [3-phase laminate] Errors values obtained at $k = 100$ from Figure 10 as functions of the discretization parameter $N \in \{64, 128, 256, 512\}$.

expected. Reminding that one considers (40) with $r = 3$, Figure 13 shows the associated errors obtained at convergence, i.e. at $k = 200$, along with the slopes corresponding to $1/N^2$ and $1/N^4$ for comparison. A global convergence rate as $\mathcal{O}\big(1/N^4\big)$ is obtained, which is in agreement with the preceding Fourier-based analysis.

Note finally that the right panel of Figure 12 shows the evolution of the error between the energy-based effective property $(\gamma_{\text{eff}})^h_k$ and the exact *continuous* solution γ_{eff}. The absolute value being taken, this error appears to be not monotonic. This is not however in contradiction with the results of Section 4.3 that pertain to the *discrete* solution and state that the discrete energy $\mathrm{W}^h(\tilde{\mathbf{e}}^h_k)$ decreases monotonically to its limit $\mathrm{W}^h(\tilde{\mathbf{e}}^h)$.

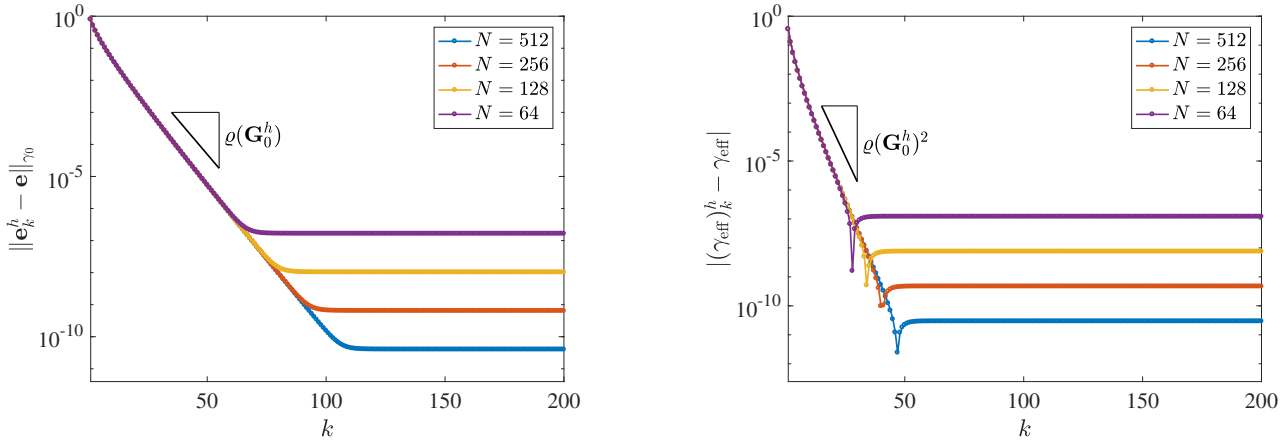


Figure 12: [Smooth distribution] Numerical errors to the analytical solution as functions of the iteration number k and when varying the discretization as $N \in \{64, 128, 256, 512\}$. *(left)* Global field error $\|\mathbf{e}^h_k - \mathbf{e}\|_{\gamma_0}$ and *(right)* global error on the effective property using its energetic definition (32).

6 Discussion

The preceding study focuses on conductive laminated composites. The associated homogenization problem is solved using the stationary iterative scheme of Moulinec and Suquet (1998), which main properties are investigated analytically in a discrete setting. The latter are illustrated numerically on two test cases. This allows to shed light on a number of interesting features that we summarize below. In this context, the question of extending these results to other configurations is also discussed qualitatively in this section on a set of additional numerical examples for 2D and 3D microstructures but without going through a detailed analysis as before.

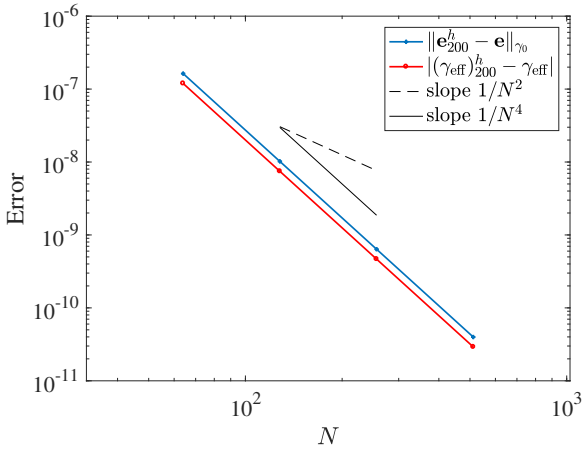


Figure 13: [Smooth distribution] Errors values obtained at $k = 200$ from Figure 12 as functions of the discretization parameter $N \in \{64, 128, 256, 512\}$. The slopes corresponding to $1/N^2$ and $1/N^4$ are shown for comparison.

6.1 Highlights of the study on laminated composites

1. **Eigendecomposition.**

   Considering the Neumann series (31), the eigendecomposition of the iterated operator is investigated in Section 3. This analysis is performed in a particular context with three main characteristics: (i) The eigendecomposition problem is addressed in a discrete setting in $\mathbb{R}^N$, by focusing on the matrix $\mathbf{G}^h$ in (30), which is the discrete counterpart of the operator $\mathcal{G} = \Gamma_0 \, \delta\gamma$. (ii) The continuous Green's operator Γ_0 reduces to the rather simple expression (17) for laminated composites. (iii) The space $\mathbb{R}^N$ can be decomposed as $\mathbb{R}^N = \mathbb{R} \oplus \mathbb{R}^N_0$ using the orthogonal matrix decomposition (27), with $\mathbb{R}$ and $\mathbb{R}^N_0$ being respectively the discrete versions of the spaces $\mathscr{S}$ in (4) and $\mathscr{E}_0$ in (2). Extending the study of Section 3 to other configurations can be achieved by revisiting one or more of these points.

   For the purposes of the discussion, let us consider here a completely general configuration with a composite material being characterized by a constitutive tensor denoted as $\boldsymbol{\gamma}$, while $\boldsymbol{\gamma}_0$ is a uniform reference tensor. The continuous Green's operator Γ_0 being defined through $\boldsymbol{\Gamma}_0$ in (10), the featured functional spaces $\mathscr{E}_0$ and $\mathscr{S}$ have to be redefined in a more general framework as admissibility spaces of mean-free gradients and diverge-free fields respectively, see Milton (2002). In this context, $\mathscr{S}$ being the polar space of $\mathscr{E}_0$, see Bellis and Suquet (2018), then the orthogonal subspace $\mathscr{E}_0^\perp$ is a space much larger than before as, now, it no longer only includes the uniform fields. This also pertains to the discrete versions of these functional spaces and, in particular, the averaging matrix $\mathbf{A}$ must be replaced in (27) by the projection matrix $\mathbf{P}_0^\perp$ associated with $\mathscr{E}_0^\perp$. As the orthogonal decomposition (27) plays a key role in the proof of Proposition 1, generalizing the latter would imply the study of a larger number of subcases and the variety of eigenvectors is expected to be enriched. In this process, the particular form (17) of the Green's operator would no longer holds so that the eigenvectors have to be constructed based on the definition (10) and either the discrete version of the local equations (1) or a discrete Fourier transform-based expression of $\boldsymbol{\Gamma}_0$. To conclude this paragraph, let us mention that characterizing the spectrum of the *continuous* operator $\boldsymbol{\Gamma}_0(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)$ would be a task much more involved Reed and Simon (1980) than diagonalizing the symmetric matrix $\mathbf{G}^h_0$ and it requires using the self-adjointness property of the operator $\boldsymbol{\Gamma}_0\boldsymbol{\gamma}_0$ that holds in a suitable Hilbert space, see Bellis and Suquet (2018).

2. **Case-dependent convergence rate of the Neumann series.**

   Property 4 recalls an upper bound on the convergence rate of the Neumann series (31) that is governed by the spectral radius $\varrho(\mathbf{G}^h_0)$. In turn, $\varrho(\mathbf{G}^h_0)$ is bounded by the supremum norm of the material property vector (20), which therefore yields the well-known bound on the convergence rate of (31) in terms of the material contrast only, see Property 4. This is a conservative bound that is independent of the microstructure but it is optimal in the sense that it is attained for some microstructures. Finally, let us mention that the reference conductivity value γ_0 can be chosen so as to minimize this upper bound. This optimization can be done without information on the microstructure but the bound could be further improved if geometrical information on the microstructure is actually used, see Moulinec et al. (2018).

In this context and based on the eigendecomposition of the matrix featured in the discrete Neumann series (31), one also establishes in Proposition 2 that the vector $\mathbf{e}_k^h$ computed at each iteration and its limit $\mathbf{e}^h$ are both expressed in term of a particular subset of eigenvectors. This allows to show in Proposition 3 that the actual convergence rate of the Neumann series is potentially smaller than $\varrho(\mathbf{G}_0^h)$ and governed by the largest eigenvalue associated with the subset of eigenvectors that is activated during iterations. This results is illustrated numerically on the two test cases considered.

3. **Quadratic convergence of the energy-based effective property.**
With these results on the fields $\mathbf{e}_k^h$ at hand, we turn to the convergence properties of the energy-based effective conductivity $(\gamma_{\text{eff}})_k^h$ in (32). For the material configuration considered, the discrete energy functional is coercive and Lipschitz, see (45). Based on these inequalities, Proposition 4 establishes that $(\gamma_{\text{eff}})_k^h$ converges to its limit $(\gamma_{\text{eff}})^h$ at a quadratic rate compared to this governing $\mathbf{e}_k^h$. Moreover, it is observed that this convergence is monotonic for the example considered. These properties are all the more remarkable as they are not shared with the alternative definition (53) of the effective conductivity $(\check{\gamma}_{\text{eff}})_k^h$. Indeed, although the latter converges to $(\gamma_{\text{eff}})^h$ as well, it does so at a rate only proportional to this of $\mathbf{e}_k^h$.
Finally, as already pointed out in Remark 4, the quadratic convergence of the discrete energy-based effective property is a classic result from optimization Nesterov (2004) and the starting point to derive the associated estimates are the inequalities (45). In this derivation, the fact that the composite considered is a laminate is not used. Therefore, these results are also valid for other microstructures and in any dimension provided that (45) holds. Moreover, as the corresponding coercive and Lipschitz properties also hold for the energy functional W in (7), the developments of Section 4.2 can easily be extended at the continuous level. Note finally, that such properties do not make use of the linearity of the constitutive relations per se. As a consequence, they could also be established in the case of non-linear composites as long as energy bounds such as (45) are valid.

4. **Stopping criterion.**
In Proposition 4, it is also established that the convergence error $|(\gamma_{\text{eff}})_k^h - (\gamma_{\text{eff}})^h|$ in the effective property, can be bounded using $\|\boldsymbol{\nabla}\mathrm{W}^h(\tilde{\mathbf{e}}_k^h)\|_{\gamma_0}^2$. Note, that from the identity (29), one has

$$\|\boldsymbol{\nabla}\mathrm{W}^h(\tilde{\mathbf{e}}_k^h)\|_{\gamma_0} = \left\| \frac{1}{\gamma_0}\mathbf{P}_0 \cdot \mathbf{j}_k^h \right\|_{\gamma_0} \qquad \text{with} \qquad \mathbf{j}_k^h = \mathbf{Diag}[\gamma^h] \cdot (\overline{\mathbf{e}} + \tilde{\mathbf{e}}_k^h). \tag{59}$$

As the matrix $\mathbf{P}_0/\gamma_0$ is the discretized counterpart of the Green's operator Γ_0, one arrives here at a convergence criterion that has already been discussed in a number of studies, see Moulinec et al. (2018) and the references therein. In (59), the featured norm is the energetic norm defined by γ_0. Although the choice of the norm is not critical for the 1D conductivity problem considered here, see Remark 1, it is actually of prime importance for other material configurations, see Moulinec et al. (2018) and Bellis and Suquet (2018). Finally, as an added value to these latter references, let us point out that Proposition 4 provides, not only a stopping criterion for the discrete Neumann series, but both an upper and a lower bound on the error in the effective property.

5. **Monotonicity properties**
Revisiting the Neumann series (31) as the gradient-descent scheme with fixed step (15) allows to highlight a number of monotonicity properties. These properties stem from the Lipschitz continuity and the coercivity of the gradient of the discrete energy functional $\mathbf{W}^h$, see (45). In this context and using an approach which is standard in the field of optimization, a sufficient condition on the reference medium conductivity γ_0 is obtained to ensure that $\mathbf{W}^h$ decreases monotonically. In addition, the residual error $\|\boldsymbol{\epsilon}_k^h\|_{\gamma_0}$ on the field is also shown to converge monotonically under a slightly more constraining condition. Lastly, the choice $\gamma_0 = (\max_n(\gamma_n^h) + \min_n(\gamma_n^h))/2$, proposed in Moulinec and Suquet (1998) and which is used in the present study, is shown to be the optimal value that satisfies these conditions. It is straightforward to extend these results in 2D or 3D and other material configurations. Care must be exercised however if the composite considered includes some cavities, a case which would be characterized by a lack of coercivity of the energy functional. Extensions to continuous formulations are also straightforward.

6. **Discretization error estimation.**
With the convergence properties of the Neumann series being assessed, one focuses in a second step on the estimation of the discretization error, i.e. the evaluation of the second right-hand side term in the error estimate (33). In this study on laminated composites, the discretization scheme is entirely encapsulated in

the definition of the discrete averaging operator $\langle\cdot\rangle_h$. Numerically, such integrals are computed using the trapezoidal rule (23) due to its link to the discrete Fourier transform and therefore by consistency with the FFT-based computational homogenization methods. It should be noted, however, that the present study on laminates relies directly on the particular form (17) of the Green's operator and does not make use of the discrete Fourier transform. In this setting, it is shown that the discretization error can be driven by the decay rate of the Fourier coefficients $\hat{\rho}(\nu)$ of the resistivity field $\rho = \gamma^{-1}$, see Property 7. Recalling this result, which may seem astonishing, the scheme considered can reach high levels of accuracy for smooth material distributions, while the overall numerical error can be estimated for a broad class of functions

7. **Assessment of aliasing and Gibbs phenomena.**
Lastly and in connection with discussions in the FFT-based computational homogenization literature, the possible effects of the aliasing and Gibbs phenomena are assessed. In the particular context of laminated composites, where the action of the Green's operator Γ_0 is directly computed in the physical space using (17), it is shown that the discretization error is entirely due to aliasing effects, see Property 7. Obviously, as the computation of Γ_0 does not rely neither on the discrete Fourier transform nor on truncated Fourier series, the Gibbs phenomenon is not at play here. In particular, it is shown in Proposition 2 that the computed local field $\mathbf{e}_k^h$ is constant in each phase at any iteration k. For other material configurations however, Fourier-based computational homogenization methods make an intensive use of the FFT. In this context, potential undesirable fields oscillations near material discontinuities and the possible manifestations of the Gibbs phenomenon have then been pointed out in W. H. Müller (1996); C. M. Brown and W. W. Dreyer and W. H. Müller (2002); Willot et al. (2014); Schneider et al. (2016). In order to reduce these spurious effects, the strategy of using modified Green operators has been proposed, see e.g. W. H. Müller (1996); Brisard and Dormieux (2012); Willot (2015); Schneider et al. (2016).

6.2 Extension to other configurations: numerical examples

In this section, we illustrate on a set of additional examples some of the key features of the stationary iterative scheme of Moulinec and Suquet (1998) that have been highlighted previously in the case of laminated composites. While the objective is not to go through the same type of detailed analysis, the examples of this section, which pertain to other material configurations, may serve as an illustrative guidance for future studies.

6.2.1 2D conductivity: the Obnosov problem

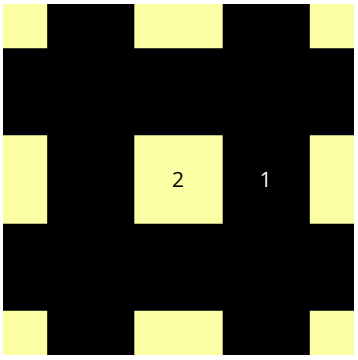


Figure 14: Double-periodic array of square inclusions with volume fraction 0.25.

In a first example, we consider the 2D isotropic conductivity problem that has been investigated analytically in Obnosov (1992, 1999). It consists of a double-periodic array of unit square inclusions with volume fraction 0.25, see Figure 14, where the conductivity field is piecewise constant with $\gamma(\boldsymbol{x}) = \gamma_p$ in each phase ϕ_p for $p = 1, 2$. For this problem, the exact field $\boldsymbol{e}(\boldsymbol{x})$ solution of (1) and the effective conductivity γ_{eff} can be found analytically, see Obnosov (1992, 1999), and they are provided in Appendix B for the reader's convenience. As previously, the notation $\mathbf{e}$ is used to denote the map of the numerical values of the continuous solution $\boldsymbol{e}$ at the 2D grid points. While this analytical *continous* solution is available, we do not have at our disposal neither the closed-form expression of the *discrete* solution $\mathbf{e}^h$, which is defined as the limit of the discrete Neumann series, nor the associated expression $(\gamma_{\text{eff}})^h$ of the effective property.

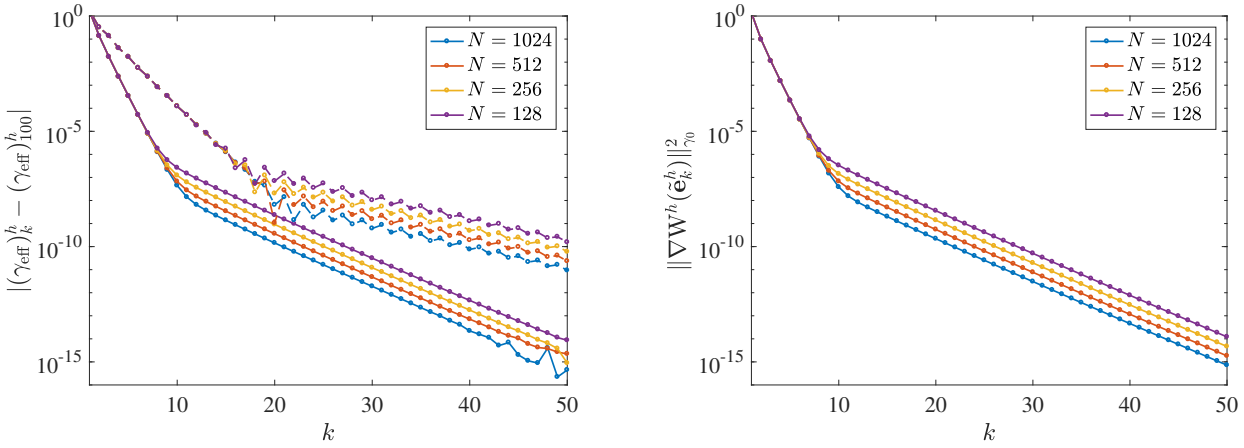


Figure 15: [2D conductivity] *(left)* Error on the effective property at each iteration relatively to the value computed at iterate $k = 100$ for the corresponding discretization. Solid lines correspond to the energetic definition (32) and dashed lines to the alternative definition (53), i.e. $(\gamma_{\text{eff}})_k^h$ being replaced by $(\check{\gamma}_{\text{eff}})_k^h$. *(right)* Convergence indicator for the discrete Neumann series.

The numerical results presented here are obtained using a Fourier-based implementation of the stationary iterative scheme (16) in the case $\gamma_1 = 1$, $\gamma_2 = 10$ and with a discretization of the cell $[-2;2] \times [-2;2]$ of Figure 21a using $N \times N$ pixels. The formalism described previously is formally extended in 2D with γ^h being the discrete conductivity map, $\gamma_0 = (\gamma_1 + \gamma_2)/2$ the reference isotropic conductivity and $\boldsymbol{\Gamma}_0^h$ the corresponding discrete Green's operator that is computed using discrete Fourier series, taking into account all of the frequencies associated with the discretization considered, and the FFT algorithm. In this setting, Figure 15 illustrates the convergence of the Neumann series, which is ensured by the choice of γ_0, through the behavior of the error on the effective property. For a given discretization, the value computed at each iteration is compared to the numerical value obtained at the iterate $k = 100$, which we consider to be the discrete limit value in the absence of the explicit form of the discrete solution $(\gamma_{\text{eff}})^h$. As in the case of 1D laminates, convergence is achieved up to machine precision. However, for a given discretization parameter N, two distinct slopes are observed, i.e. the Neumann series is governed by two different convergence rates. The error seems to be independent of N in the first stage and correlated to it in the second stage. When the energy-based effective property $(\gamma_{\text{eff}})_k^h$ is replaced by the alternative definition $(\check{\gamma}_{\text{eff}})_k^h$ using (53) then the convergence rate decreases in accordance with the previous results, see dashed lines compared to solid ones.

Drawing from the previous eigenanalysis, and for the sake of the discussion, consider that the iteration error $\boldsymbol{\epsilon}_k^h$ in (44) for the discrete solution has the following form in the 2D case:

$$\boldsymbol{\epsilon}_k^h = \sum_{j=k}^{\infty} \left(\sum_{m=1}^{N_1} (-\lambda_m^h)^j \left(\mathbf{v}_m^h, \mathbf{f}_0^h\right)_{\gamma_0} \mathbf{v}_m^h + \sum_{m=1}^{N_2} (-\mu_m^h)^j \left(\mathbf{w}_m^h, \mathbf{f}_0^h\right)_{\gamma_0} \mathbf{w}_m^h \right), \tag{60}$$

with $N_1 + N_2 = 2N^2$ being the total number of components in the field $\mathbf{e}_k^h$, while $\{\lambda_m^h, \mu_m^h\}$ and $\{\mathbf{v}_m^h, \mathbf{w}_m^h\}$ are respectively the sets of eigenvalues and eigenvectors of the iterated matrix for the 2D configuration considered. Having $(\mathbf{w}_m^h, \mathbf{f}_0^h)_{\gamma_0} \ll (\mathbf{v}_m^h, \mathbf{f}_0^h)_{\gamma_0}$ with in the mean time $|\lambda_m^h| < |\mu_m^h|$ would be compatible with a two-stage convergence behavior as this observed in Figure 15. Indeed, this would mean that the eigenvectors $\mathbf{v}_m^h$, or modes, are predominant in the discrete solution $\mathbf{e}_k^h$ and they converge at a faster rate compared to some other minor modes $\mathbf{w}_m^h$. The slower convergence of the latter being accountable for the iteration error in the second stage. This can only be justified rigorously by performing the eigenanalysis of the 2D Obnosov problem.

To get rid of the comparison with the numerical value at the iterate $k = 100$, one can use instead the quantity $\|\boldsymbol{\nabla} \mathrm{W}^h(\tilde{\mathbf{e}}_k^h)\|_{\gamma_0}^2$ as a convergence indicator for the discrete Neumann series. Owing to (25) and (29), this indicator is computed in the 2D setting as

$$\|\boldsymbol{\nabla} \mathrm{W}^h(\tilde{\mathbf{e}}_k^h)\|_{\gamma_0}^2 = \left\langle \boldsymbol{\Gamma}_0^h \mathbf{j}_k^h \cdot \gamma_0 \boldsymbol{\Gamma}_0^h \mathbf{j}_k^h \right\rangle_h \qquad \text{with} \qquad \mathbf{j}_k^h = \gamma^h \left(\overline{\mathbf{e}} + \tilde{\mathbf{e}}_k^h\right), \tag{61}$$

in the energetic norm weighted by γ_0. As expected, there is an excellent agreement between this indicator, shown in the right panel of Fig. 15, and the convergence error on the effective property obtained at the discrete

level (left panel). This is all the more interesting as the indicator (61) can be fully computed numerically from the quantities available at the current iterate.

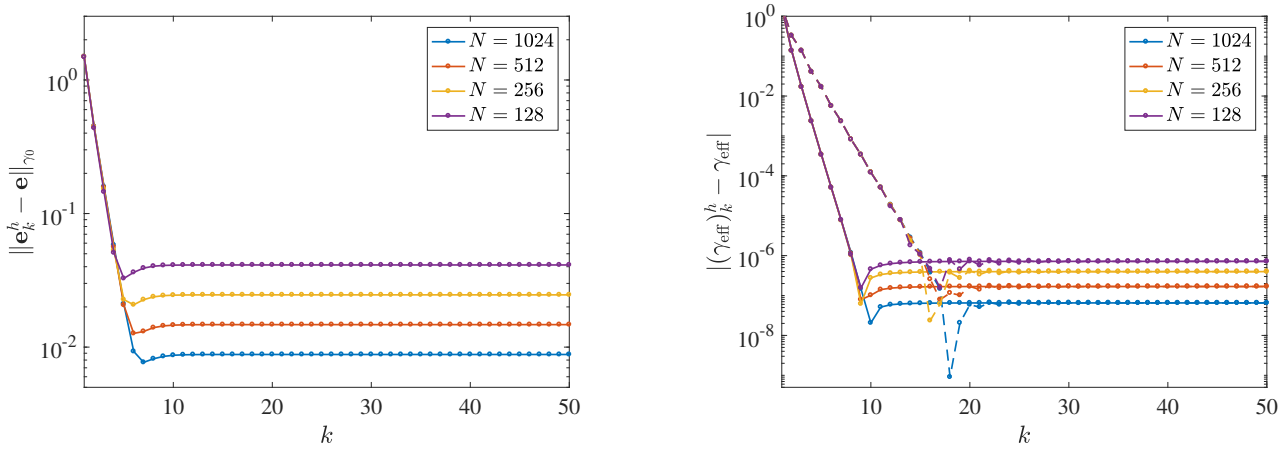


Figure 16: [2D conductivity] Numerical errors to the exact analytical solution as functions of the iteration number k and when varying the discretization as $N \in \{128, 256, 512, 1024\}$. *(left)* Global field error $\|\mathbf{e}_k^h - \mathbf{e}\|_{\gamma_0}$ and *(right)* global error on the effective property. Solid lines correspond to the energetic definition (32) and dashed lines to the alternative definition (53), i.e. $(\gamma_{\text{eff}})_k^h$ being replaced by $(\check{\gamma}_{\text{eff}})_k^h$.

Turning to the assessment of the discretization error, comparisons with the exact analytical solution of Appendix B are shown in Figure 16. The energetic norm of the field error is plotted in the left panel of Fig. 16 and comparisons with the exact effective property γ_{eff} are shown in the right panel for both definitions (32) and (53). The corresponding final errors at the iterate $k = 50$ are reported in Figure 17. In these results the discretization plays a key role similar to this it has for the smooth material distribution of the test case 2, see Section 5.2 and figures 12 and 13. Errors rapidly reach a plateau whose height is correlated to the discretization parameter N. Note that these asymptotic values are attained within the first convergence stage where the convergence rate appears to be independent of N. Moreover, for a given value N, the effective property obtained asymptotically is the same for both definitions $(\gamma_{\text{eff}})_k^h$ and $(\check{\gamma}_{\text{eff}})_k^h$ but it is attained with different convergence rates, as previously discussed. Overall, these errors with respect to the analytical solutions $\mathbf{e}$ and γ_{eff} could be due to the integrable corner singularities contained in the exact field solution, see Obnosov (1992, 1999). Indeed, the global convergence of the discrete solution $\mathbf{e}_k^h$ is expected to be penalized due to its behavior within the corner regions, see the local error maps of Figure 18.

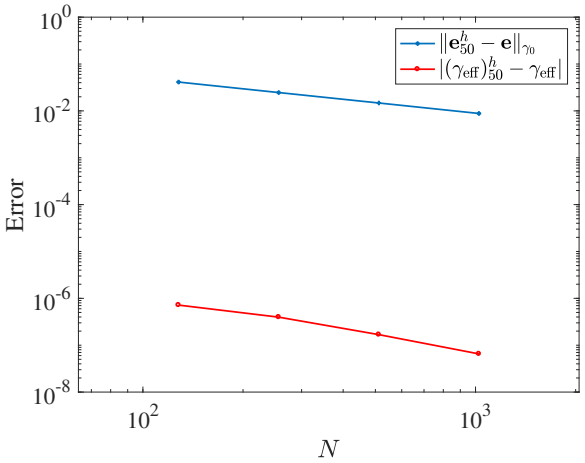
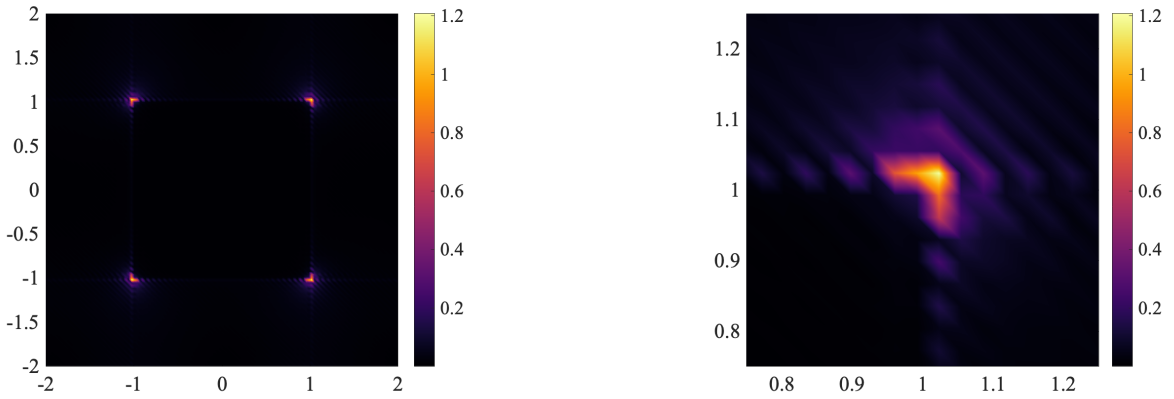


Figure 17: [2D conductivity] Errors values obtained at $k = 50$ from Figure 16 as functions of the discretization parameter $N \in \{128, 256, 512, 1024\}$.



(a) $|\mathbf{e}_k^h(\boldsymbol{x}_n) - \mathbf{e}(\boldsymbol{x}_n)|_{\gamma_0}$ (b) $|\mathbf{e}_k^h(\boldsymbol{x}_n) - \mathbf{e}(\boldsymbol{x}_n)|_{\gamma_0}$ close-up in corner region

Figure 18: [2D conductivity] Maps of local numerical error of the discrete solution $\mathbf{e}_k^h$ with respect to the theoretical solution $\mathbf{e}$, computed at the iteration $k = 100$ and for the discretization parameter $N = 128$.

6.2.2 3D elasticity

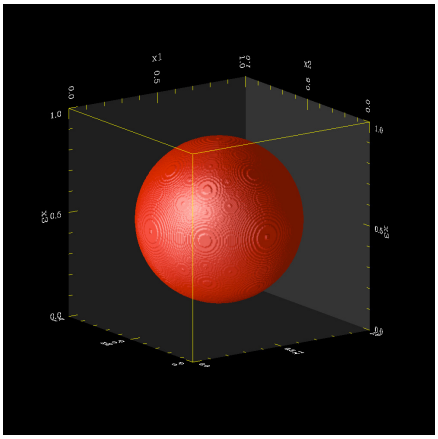


Figure 19: Triple-periodic array of spherical elastic inclusions with volume fraction 0.25 discretized using N^3 voxels.

In a second example we consider a triple-periodic array of spherical elastic inclusions with Young modulus $E_2 = 10$ and Poisson's ratio $\nu_2 = 0.2$ embedded in an isotropic elastic matrix with parameters $E_1 = 1$ and $\nu_1 = \nu_2$. The volume fraction of the inclusions is 0.25. Correspondingly, this material configuration defines a fourth-order and isotropic elasticity tensor $\boldsymbol{L}(\boldsymbol{x})$. Accordingly, the local problem (1) has to be recast using $\boldsymbol{L}(\boldsymbol{x})$ and a second-order strain tensor $\boldsymbol{\varepsilon}(\boldsymbol{x})$ instead of $\gamma(x)$ and $\boldsymbol{e}(x)$ respectively. Moreover, one introduces a reference homogenous tensor $\boldsymbol{L}_0$. Numerically, a unit cubic cell is considered and discretized using $N \times N \times N$ voxels, see Figure 19. For this example, we do not have at our disposal neither the discrete solutions ε^h, $(\mathbf{L}_{\text{eff}})^h$ obtained at convergence of the discrete Neumann series, nor the exact continuous solutions $\boldsymbol{\varepsilon}$, $\boldsymbol{L}_{\text{eff}}$ of the problem. Therefore, as in the previous example, for each discretization the convergence of the discrete Neumann series is assessed by comparing numerical quantities computed at each iteration k relatively to their corresponding values at the iteration $k = 100$ at the same discretization. More precisely, for this elastic case, one considers the discrete macroscopic energy defined as the quantity

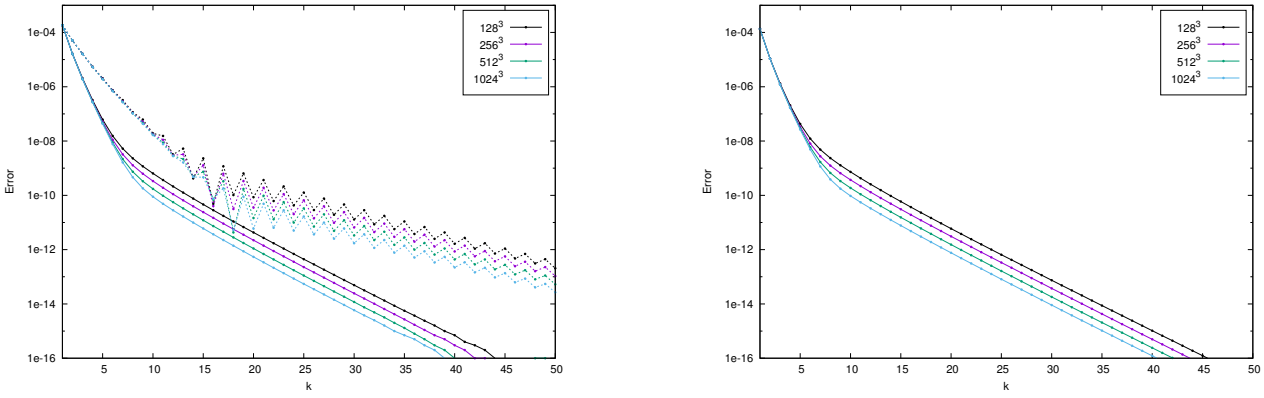
$$\mathrm{W}^h(\tilde{\varepsilon}_k^h) = \frac{1}{2}\left\langle \left(\bar{\varepsilon} + \tilde{\varepsilon}_k^h\right) : \mathbf{L}^h : \left(\bar{\varepsilon} + \tilde{\varepsilon}_k^h\right) \right\rangle_h, \tag{62}$$

which is expressed as a function of the fluctuating part $\tilde{\varepsilon}_k^h$ of the discrete strain field computed at each iteration of the scheme (16) and that features the imposed macroscopic strain $\bar{\varepsilon}$, the discretized elasticity tensor $\mathbf{L}^h$ and

uses a 3D discrete averaging operator $\langle\cdot\rangle_h$ and, by an abuse of notation, a doubly contracted product ":". As such, this definition is consistent with (32). For comparison, and with consistency with the alternative definition (53) considered previously, one introduces a second definition of the discrete energy as

$$\check{\mathrm{W}}^h(\tilde{\varepsilon}_k^h) = \frac{1}{2}\big\langle \big(\overline{\varepsilon} + \tilde{\varepsilon}_k^h\big)\big\rangle_h : \big\langle \mathbf{L}^h : \big(\overline{\varepsilon} + \tilde{\varepsilon}_k^h\big)\big\rangle_h. \tag{63}$$

Based on Hill's lemma, which pertains to the continuous solution of the problem (1), the energies computed according to (62) and (63) are expected to be equal at convergence of the discrete scheme. However, as in Section 4.2, the evolutions of these quantities with iterations provide us with information on the convergence of the discrete Neumann series.



(a) $|\mathrm{W}^h(\tilde{\varepsilon}_k^h) - \mathrm{W}^h(\tilde{\varepsilon}_{100}^h)|$ and $|\check{\mathrm{W}}^h(\tilde{\varepsilon}_k^h) - \check{\mathrm{W}}^h(\tilde{\varepsilon}_{100}^h)|$ (b) $\|\boldsymbol{\nabla}\mathrm{W}^h(\tilde{\varepsilon}_k^h)\|^2_{\gamma_0}$

Figure 20: [3D elasticity] *(left)* Error on the macroscopic energy computed at each iteration relatively to the values at iterate $k = 100$, for the corresponding discretization, using either the definition (62) or the alternative definition (63), i.e. $\mathrm{W}^h(\tilde{\varepsilon}_k^h)$ being replaced by $\check{\mathrm{W}}^h(\tilde{\varepsilon}_k^h)$. *(right)* Convergence indicator for the discrete Neumann series.

The errors on the quantities (62) and (63), relatively to their values computed at the iteration $k = 100$ for each discretization, are shown in Figure 20a. These results illustrate that, again, convergence of the Neumann series up to machine precision is achieved and the convergence rate using the definition (62) is quadratic compared to (63). Note that, as in the 2D conductivity problem above, two regimes are observed, with the same type of dependence on the discretization parameter N.

Lastly, to avoid assessing the convergence of the discrete Neumann series by comparison with values obtained at the iteration $k = 100$, one considers instead in Fig. 20b the indicator function $\|\boldsymbol{\nabla}\mathrm{W}^h(\tilde{\varepsilon}_k^h)\|^2_{\gamma_0}$ that can be fully computed from quantities available at the current iterate. In the elastic case, by consistency with (25) and (29), this indicator is computed as

$$\|\boldsymbol{\nabla}\mathrm{W}^h(\tilde{\varepsilon}_k^h)\|^2_{\gamma_0} = \big\langle \boldsymbol{\Gamma}_0^h\sigma_k^h : \mathbf{L}_0 : \boldsymbol{\Gamma}_0^h\sigma_k^h\big\rangle_h \qquad \text{with} \qquad \sigma_k^h = \mathbf{L}^h : \big(\overline{\varepsilon} + \tilde{\varepsilon}_k^h\big), \tag{64}$$

and where $\boldsymbol{\Gamma}_0^h$ denotes the discrete elastic Green's operator associated with $\mathbf{L}_0$, which is computed using the complete discrete Fourier series of the discretization considered and the FFT algorithm. In (64), one makes use of an energetic norm weighted by $\mathbf{L}_0$. As in Section 6.2.1, there is a good agreement between the results in Figure 20b and those in Fig. 20a associated with the definition (62) of the energy, in the sense that the evolutions of the corresponding curves are comparable. Therefore, the indicator function (64) can be used as a stopping criterion as of the convergence of the discrete Neumann series.

7 Conclusions

This study focuses on the conductivity problem in laminated periodic composites. Its aim is to examine the convergence properties of the field and effective property that are computed through the inversion of the Lippman-Schwinger equation by a Neumann series. For the configuration considered, the conventional Fourier-based

formulation is avoided as the Green's operator is obtained in closed-form in the physical space. In this context, both the iteration error and the discretization error are evaluated within a discrete setting.

The main points of this study are as follows. (1) The eigendecomposition of the matrix iterated within the Neumann series, which provides a detailed description of the eigenvectors, or modes, that are used in the construction of the solution, together with the associated eigenvalues. This allows next (2) to quantify precisely the convergence rate of the Neumann series for the specific material distribution characterizing the laminate considered. In connection to this result, one discusses (3) quadratic upper and lower bounds on the energy-based effective property, which provide a convergence result for the latter and lead to (4) a stopping criterion for the iterative scheme, in a form that has already been discussed in the literature. (5) The monotonicity properties of both the energy functional and the residual error on the field are also discussed in relation with the choice of the reference medium conductivity γ_0. With these results at hand, the next step is (6) to evaluate the discretization error, which is achieved by assessing the spectral properties of the conductivity field, thus showing how the global accuracy of the scheme considered is related to the smoothness of the spatial distribution of the material considered. Although the present study does not make use of the Fourier transform, a connection with the latter is established at this point through the study of the trapezoidal quadrature rule. This also leads to (7) the assessment of the role of aliasing in the discretization error and to confirm that the Gibbs phenomenon is not at play in the performed computations.

This study constitutes a first step towards studying a priori global error estimates for iterative schemes in computational homogenization. The configuration considered is rather simple but the intent here is to provide some illustrations of the phenomena at play in the construction of the approximated solution and to quantify them precisely through a semi-analytical analysis. In a more general context, a systematic derivation of a priori global error estimates is much needed. Extending the eigenanalysis of the present study to other 2D and 3D configurations would be a next step in this direction. For such configurations, Fourier-based implementations would be obvious choices and their performance could be evaluated using error analysis tools that have been developed for spectral methods.

Acknowledgments

The Authors are grateful to J.-C. Michel for fruitful discussions and to Y. Obnosov for his help on the analytical solution reproduced in Appendix B. The Authors have received funding from Excellence Initiative of Aix-Marseille University - A*MIDEX, a French "Investissements d'Avenir" program in the framework of the Labex MEC.

References

Abramowitz, M. and Stegun, I., editors (1974). *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Dover Publications.

Bellis, C. and Suquet, P. (2018). Geometric variational principles for computational homogenization. *Journal of Elasticity*.

Boyd, J. P. (2000). *Chebyshev and Fourier Spectral Methods*. Dover, New-York.

Brisard, S. and Dormieux, L. (2012). Combining Galerkin approximation techniques with the principle of Hashin and Shtrikman to derive a new FFT-based numerical method for the homogenization of composites. *Comput Methods Appl Mech Eng.*, 217–220:197–212.

C. M. Brown and W. W. Dreyer and W. H. Müller (2002). Discrete fourier transforms and their application to stress-strain problems in composite mechanics: a convergence study. *Proc R Soc Lond A.*, 458:1–21.

Eyre, D. J. and Milton, G. W. (1999). A fast numerical scheme for computing the response of composites using grid refinement. *Eur Phys J Appl Phys.*, 6:41–47.

Gélébart, L. and Mondon-Cancel, R. (2013). Non linear extension of FFT-based methods accelerated by conjugate gradients to evaluate the mechanical behavior of composite materials. *Computational Materials Science*, 77:430–439.

Gottlieb, D. and Orszag, S. (1977). *Numerical Analysis of Spectral Methods*. Society for Industrial and Applied Mathematics.

Kabel, M., Böhlke, T., and Schneider, M. (2014). Efficient fixed point and Newton-Krylov solvers for FFT-based homogenization of elasticity at large deformations. *Comput. Mech.*, 54:1497–1514.

Kröner, E. (1972). *Statistical Continuum Mechanics*, volume 92. CISM Lecture Notes, Springer-Verlag.

Michel, J.-C., Moulinec, H., and Suquet, P. (2001). A computational scheme for linear and non-linear composites with arbitrary phase contrast. *Int J Numer Methods Eng.*, 52:139–160.

Milton, G. W. (2002). *The Theory of Composites*. Cambridge university press.

Mishra, N., Vondřejc, J., and Zeman, J. (2016). A comparative study on low-memory iterative solvers for FFT-based homogenization of periodic media. *J Comput Phys.*, 321:151–168.

Monchiet, V. and Bonnet, G. (2012). A polarization-based FFT iterative scheme for computing the effective properties of elastic composites with arbitrary contrast. *Int J Numer Methods Eng.*, 89:1419–1436.

Moulinec, H. and Suquet, P. (1994). A fast numerical method for computing the linear and nonlinear properties of composites. *C. R. Acad. Sc. Paris, II*, 318:1417–1423.

Moulinec, H. and Suquet, P. (1998). A numerical method for computing the overall response of nonlinear composites with complex microstructure. *Computer Methods in Applied Mechanics and Engineering*, 157(1):69 – 94.

Moulinec, H., Suquet, P., and Milton, G. W. (2018). Convergence of iterative methods based on neumann series for composite materials: Theory and practice. *International Journal for Numerical Methods in Engineering*, 114(10):1103–1130.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization*. Springer US.

Obnosov, Y. V. (1992). Solution of a boundary value problem of $\mathbb{R}$-linear conjugation with piecewise-constant coefficients. *Izv. VUZ Mat.*, 36(4):39–48. Russian Math. (Izv. VUZ), 36 (1992), 39–47.

Obnosov, Y. V. (1999). Periodic heterogeneous structures: New explicit solutions and effective characteristics of refraction of an imposed field. *SIAM Journal on Applied Mathematics*, 59(4):1267–1287.

Reed, M. and Simon, B. (1980). *Methods of modern mathematical physics. I : Functional Analysis*. Academic Press.

Schneider, M. (2017). An FFT-based fast gradient method for elastic and inelastic unit cell homogenization problems. *Computer Methods in Applied Mechanics and Engineering*, 315:846–866.

Schneider, M., Ospald, F., and Kabel, M. (2016). Computational homogenization of elasticity on a staggered grid. *Int J Numer Methods Eng.*, 105:693–720.

Trefethen, L. and Weideman, J. (2014). The exponentially convergent trapezoidal rule. *SIAM Review*, 56:385–458.

Vinogradov, V. and Milton, G. W. (2008). An accelerated FFT algorithm for thermoelastic and non-linear composites. *International Journal for Numerical Methods in Engineering*, 76(11):1678–1695.

Vondřejc, J., Zeman, J., and Marek, I. (2014). An fft-based galerkin method for homogenization of periodic media. *Computers & Mathematics with Applications*, 68(3):156 – 173.

Vondřejc, J., Zeman, J., and Marek, I. (2015). Guaranteed upper-lower bounds on homogenized properties by fft-based galerkin method. *Computer Methods in Applied Mechanics and Engineering*, 297:258 – 291.

W. H. Müller (1996). Mathematical versus experimental stress analysis of inhomogeneities in solids. *J Phys IV*, 6:139–148.

Willis, J. (1981). Variational and related methods for the overall properties of composites. *Advances in Applied Mechanics*, 21:1–78.

Willot, F. (2015). Fourier-based schemes for computing the mechanical response of composites with accurate local fields. *Comptes Rendus Mécanique*, 343(3):232–245.

Willot, F., Abdallah, B., and Pellegrini, Y. P. (2014). Fourier-based schemes with modified green operator for computing the electrical response of heterogeneous media with accurate local fields. *Comput Methods Appl Mech Eng.*, 98:518–533.

Zeman, J., Vondřejc, J., Novák, J., and Marek, I. (2010). Accelerating a FFT-based solver for numerical homogenization of periodic media by conjugate gradients. *J Comput Phys.*, 229:8065–8071.

A Mathematical definitions

A.1 Functional spaces

For $x = \boldsymbol{x} \cdot \boldsymbol{d}$ with $\boldsymbol{x}$, $\boldsymbol{d} \in \mathbb{R}^D$ and $p \in \{1, 2\}$, define spaces of periodic scalar functions and vector fields as:

$$
\begin{aligned}
&L^\infty_{\rm per}(0,\ell) = \{f \ : \ \exists C \geq 0 \text{ such that } |f(x)| \leq C \text{ a.e. } x \in \mathbb{R},\ f(x+\ell) = f(x) \text{ a.e. } x \in \mathbb{R}\}\,,\\
&L^p_{\rm per}(0,\ell) = \{f \in L^p_{\rm loc}(\mathbb{R}) \ : \ f(x+\ell) = f(x) \text{ a.e. } x \in \mathbb{R}\}\,,\\
&L^p_{\rm per}(0,\ell)^D = \big\{\boldsymbol{f} = (f_j)|_{1\leq j\leq D} \ : \ f_j \ \in L^p_{\rm per}(0,\ell)\big\}\,,\\
&H^1_{\rm per}(0,\ell) = \big\{f \in H^1_{\rm loc}(\mathbb{R}^d),\ f \in L^2_{\rm per}(0,\ell),\ \partial_{x_j} f \in L^2_{\rm per}(0,\ell),\ 1 \leq j \leq D\big\}\,.
\end{aligned}
$$

A.2 Matrices

For all vector $\mathbf{f} \in \mathbb{R}^N$, we introduce the following notation

$$
\mathbf{Diag}[\mathbf{f}] = \begin{bmatrix} \mathrm{f}_0 & 0 & \dots & 0 \\ 0 & \mathrm{f}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \mathrm{f}_{N-1} \end{bmatrix}.
$$

Moreover, one defines the discrete averaging matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ as

$$
\mathbf{A} = \frac{1}{N}\begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix} \tag{65}
$$

Since $\mathbf{A}^2 = \mathbf{A}$, the matrix $\mathbf{A}$ is the orthogonal projector onto $\mathbb{R}$, i.e. the space of constant vectors. The orthogonal projection matrix $\mathbf{P}_0$ onto $\mathbb{R}^N_0$ in (26) is given by:

$$
\mathbf{P}_0 = \mathbf{I} - \mathbf{A} \tag{66}
$$

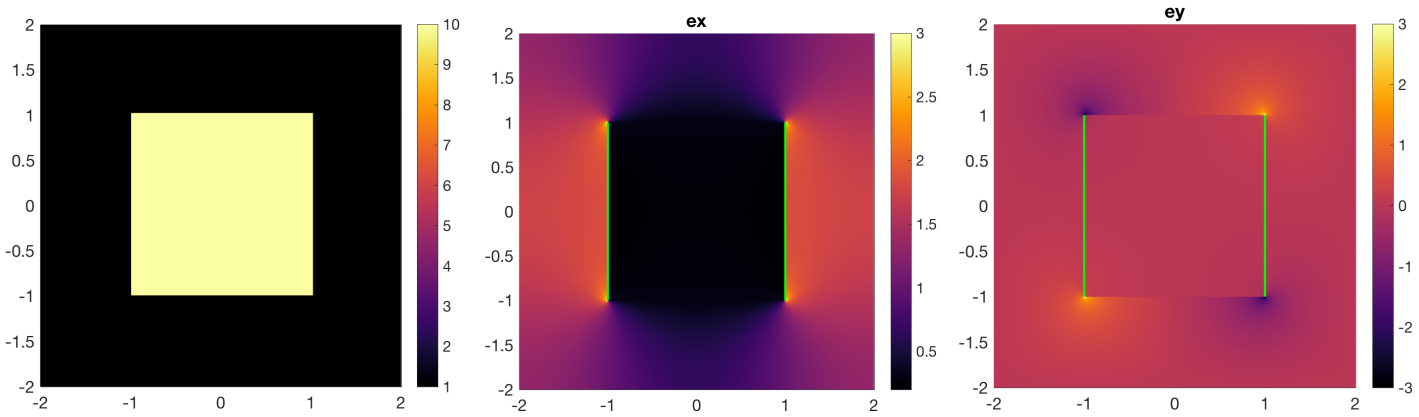
with $\mathbf{I}$ being the identity matrix. It can indeed by checked that $(\mathbf{P}_0)^2 = \mathbf{P}_0$ and that its range satisfies $\mathcal{R}(\mathbf{P}_0) = \mathbb{R}^N_0$. Moreover, one has $\mathbf{A}\cdot\mathbf{P}_0 = \mathbf{P}_0\cdot\mathbf{A} = \mathbf{0}$ so that one can formally use the notation $\mathbf{A} = \mathbf{P}_0^\perp$. Lastly, it should be reminded that from Property 1, the matrix $\mathbf{P}_0$ is the discretized version of the orthogonal projection operator $\mathcal{P}_0 = \Gamma_0\gamma_0$ from $L^2_{\rm per,0}(0,\ell)$ onto itself.

B Exact solution for the Obnosov problem

We consider here the exact field solution of the 2D conductivity problem of Section 6.2.1 for a configuration where the inclusions, i.e. phase 2, are unit squares. The explicit solution for a double-periodic array of rectangular inclusions has been derived in Obnosov (1992, 1999) and we reproduce it below for the particular configuration considered for the reader's convenience. Reference to Abramowitz and Stegun (1974) can be made for details on the special functions that are used below.

First, the effective conductivity is isotropic and given by:

$$\gamma_{\text{eff}} = \gamma_1 \sqrt{\frac{\gamma_1 + 3\gamma_2}{\gamma_2 + 3\gamma_1}}. \tag{67}$$



(a) $\gamma(\boldsymbol{x})$ (b) $e_x(\boldsymbol{x})$ (c) $e_y(\boldsymbol{x})$

Figure 21: Physical configuration considered and the corresponding exact field solution when $\overline{\boldsymbol{e}} = (1, 0)$. In the second and third panels the vertical green lines indicate branch cuts. The color scale has been adjusted to enhance the readability of the figures.

Next, introducing the complex variable $z = x + \mathrm{i}y$ for $\boldsymbol{x} = (x, y) \in \mathbb{R}^2$, let $v(z) = \gamma(x, y)\big(e_x(x, y) - \mathrm{i}e_y(x, y)\big)$ be a complex-valued field where e_x and e_y are the components of $\boldsymbol{e}$ in the canonical basis of $\mathbb{R}^2$ and $\overline{e}_x$, $\overline{e}_y$ are these of the imposed macroscopic intensity $\overline{\boldsymbol{e}}$. The field v is given analytically in each phase by

$$\begin{aligned} v(z) &= \Lambda_1 e^{\mathrm{i}\pi\alpha}\chi(z) + \Lambda_2 e^{-\mathrm{i}\pi\alpha}\chi(z)^{-1} && \text{in } \phi_1, \\ v(z) &= -(1+\Delta)\big[\Lambda_1 e^{-3\mathrm{i}\pi\alpha}\chi(z) + \Lambda_2 e^{3\mathrm{i}\pi\alpha}\chi(z)^{-1}\big] && \text{in } \phi_2. \end{aligned}$$

The parameter in the above identities are given by:

$$\Delta = \frac{\gamma_1^{-1} - \gamma_2^{-1}}{\gamma_1^{-1} + \gamma_2^{-1}}, \qquad \lambda = \frac{2}{\pi}\arcsin\Big(\frac{|\Delta|}{2}\Big), \qquad \alpha = \frac{1}{4}\big(\lambda + \operatorname{sign}(\Delta)\big).$$

Moreover, one defines

$$\Lambda_1 = \frac{\gamma_{\text{eff}}\theta\big(\overline{e}_x - \overline{e}_y \operatorname{sign}(\Delta)\big)}{\sqrt{2+\Delta}}, \qquad \Lambda_2 = \frac{\gamma_{\text{eff}}\theta\big(\overline{e}_x + \overline{e}_y \operatorname{sign}(\Delta)\big)}{\sqrt{2+\Delta}},$$

where the denominators are based on the corresponding expressions in Obnosov (1992) because of a typo in Obnosov (1999). The effective conductivity γ_{eff} is given by (67) and one has

$$\theta = 2\,\Gamma(1/4)^2 \left(\sqrt{4-\Delta^2}\,\Gamma\Big(\frac{1-\lambda}{4}\Big)\,\Gamma\Big(\frac{1+\lambda}{4}\Big)\right)^{-1},$$

which is expressed using the standard gamma function that, in this appendix only, is denoted as Γ. In addition, one has

$$\chi(z) = \left(\frac{(1+\mathrm{i})\,\mathrm{dn}(Kz/2\,|\,1/2)^2 - 1}{\sqrt{2}\,\mathrm{dn}(Kz/2\,|\,1/2)^2 - (1+\mathrm{i})/\sqrt{2}}\right)^{\lambda}$$

where $K = \Gamma(1/4)^2/(4\sqrt{\pi})$ and $\mathrm{dn}(\tilde{z}|m)$ is the Jacobian elliptic delta function that, for the complex argument $\tilde{z} = \tilde{x} + \mathrm{i}\tilde{y}$ and parameter m, is given by

$$\mathrm{dn}(\tilde{z}|m) = \frac{\mathrm{dn}(\tilde{x}|m)\,\mathrm{cn}(\tilde{y}|m_1)\,\mathrm{dn}(\tilde{y}|m_1) - \mathrm{i}m\,\mathrm{sn}(\tilde{x}|m)\,\mathrm{cn}(\tilde{x}|m)\,\mathrm{sn}(\tilde{y}|m_1)}{\mathrm{cn}(\tilde{y}|m_1)^2 + m\,\mathrm{sn}(\tilde{x}|m)^2\,\mathrm{sn}(\tilde{y}|m_1)^2}$$

with cn, sn and dn being the Jacobian elliptic functions for real arguments and $m_1 = 1 - m$. Note that the function $\chi(z)$ is single-valued with branch cuts along the vertical segments of $\partial\phi_2$. In the homogeneous case $\gamma_1 = \gamma_2$, calculating the limit of the above identities yields the correct homogeneous solution. In the case where $\gamma_1 = 1$, $\gamma_2 = 10$ and $\overline{\boldsymbol{e}} = (1, 0)$, this exact analytical solution is plotted Figure 21.