A posteriori error estimations and convergence criteria in fast Fourier transform-based computational homogenization

Renaud Ferrier, Cédric Bellis

A posteriori error estimations and convergence criteria in FFT-based computational homogenization

Renaud Ferrier[*], Cédric Bellis[†]

January 23, 2023

Abstract: A stopping criterion for FFT-based iterative schemes in computational homogenization is proposed and investigated numerically. This criterion is based on the separate evaluation and comparison of the discretization and iteration errors on the computed fields. Some estimators for these errors are proposed and their performances are assessed on a set of 2D problems in the frameworks of both the classical FFT-based methods and these that use a modified version of the featured Green's operator. In particular, two novel strategies for estimating the discretization error are investigated: either using an image processing approach or transposing to the FFT-based setting the constitutive relation error that is well-established in the context of the finite element method. It is then shown that the resulting stopping criterion leads to a better control of the global error on the computed effective property compared to the classical criterion based on the residual of the iterative scheme alone.

Key Words: Error estimators, Fast Fourier Transform, Image Processing, Filtering, Constitutive Relation Error

1 Introduction

1.1 Context and motivations

Homogenization is commonly employed to model the macroscopic behavior of heterogeneous materials so as to perform efficient numerical computations on extended domains or structures that are made of different constituents at the microscopic scale. It revolves around the calculation or the numerical computation of an equivalent effective homogeneous material from the knowledge of the geometry of the microstructure and of the behavior of the different material phases. In the framework of periodic homogenization, the computation of such an effective model can be achieved via the resolution of an integral convolution equation commonly referred to as the periodic Lippmann-Schwinger equation.

In [20], it has been proposed to discretize the Lippmann-Schwinger equation in the space-based Fourier domain and use a fixed-point algorithm to solve the resulting system in an efficient Fast Fourier Transform (FFT)-based implementation. Since then, numerous variants of this approach have been proposed, either by modifying the discretization method, see e.g. [34, 5, 30, 12], or by using more efficient, i.e. faster, iterative algorithms, see e.g. [6, 36, 11].

Such variants of the FFT-based computational homogenization method amount in an iterative approach associated with a Cartesian grid-based discretization of the Lippmann-Schwinger equation. In this setting, the most commonly used indicator for quantifying the convergence of the solution is the residual of the numerical scheme, which can be interpreted as an indicator of the *iteration error*. If the question of the definition of this indicator has already been investigated in a number of studies, see e.g. [2] and the references therein, the notion of *discretization error estimation* has seldom been considered, see [21]. In this context, the present study aims at filling this gap, by proposing some estimators of the discretization error and a framework to use them in conjonction with an iteration error estimator in order to control the overall error on the computed fields or effective property.

The proposed procedure to estimate the discretization error is inspired by the ideas initially developed for the finite element method (FEM), see the monograph [14], as well as in the context of domain

[*]Mines Saint-Etienne, Univ Lyon, CNRS, UMR 5307 LGF, Centre SMS, Saint-Etienne, France
[†]Aix Marseille Univ, CNRS, Centrale Marseille, LMA, Marseille, France

decomposition [24]. Noticeably, the influence of the discretization in a finite volume method and of the incomplete resolution of the associated linear system has also been assessed in [10]. In the present study, to obtain some estimations of the discretization error, we investigate two types of approaches, which are classical for the FEM but still unexploited in the field of FFT-based computational homogenization. The first one draws from the so-called *ZZ1 method* [37] that can be assimilated, in the setting associated with regular computational grids, to smoothing methods used in the field of *image processing*, see also [19] in the present context. The second one is the so-called *constitutive relation error* [13]. To the best of our knowledge, this topic has never been investigated yet, except in a conference paper [4], and the present study constitutes a first step towards the formulation of robust a posteriori error estimators for Fourier-based computational homogenization platforms. Note that, the tools developed in the context of the FEM would also find a natural extension in this context if FE discretizations on regular grids coupled with FFT solvers were used, as in [25, 16].

In the remaining of this section, the homogenization problem for the model conductivity equation is presented, along with the general idea of the proposed approach to distinguish between the discretization and the iteration errors. In section 2, we present the numerical resolution method for the Lippmann-Schwinger equation as well as the test-case geometries considered for the numerical examples to come. Section 3 exposes the relationship between the residual of the discrete problem and the iteration error, but highlights numerically that this residual is not sufficient to provide a reliable indicator for the total error on the effective property. In section 4, three discretization error estimators are introduced and assessed numerically. Finally, Section 5 focuses on the proposition and the numerical evaluation of a stopping criterion based on the comparison between the estimated discretization and iteration errors, in 2D conductivity and 3D elasticity as well.

1.2 Problem setting

Consider a periodic medium characterized by the representative unit cell $\Omega \subset \mathbb{R}^d$ and a spatially-varying isotropic conductivity field $\gamma \in L^\infty_{\text{per}}(\Omega)$ with $\gamma > 0$. The latter constitutes a material parameter prototypical of linear constitutive laws, and a placeholder for possible extensions of the present work to anisotropic or elastic constituents. Our overall objective is to quantify the macroscopic behavior of the medium considered, through the computation of the effective tensor $\boldsymbol{\gamma}_{\text{eff}}$, which is the unique tensor defining an effective energy $W_{\text{eff}}(\overline{\boldsymbol{e}}) = \frac{1}{2}\overline{\boldsymbol{e}} \cdot \boldsymbol{\gamma}_{\text{eff}} \cdot \overline{\boldsymbol{e}}$ for all $\overline{\boldsymbol{e}} \in \mathbb{R}^d$, which satisfies

$$W_{\text{eff}}(\overline{\boldsymbol{e}}) = \min_{\boldsymbol{e}^* \in \boldsymbol{\mathcal{E}}_0} W(\overline{\boldsymbol{e}} + \boldsymbol{e}^*) \quad \text{with} \quad W(\boldsymbol{e}) = \frac{1}{2}\big\langle \boldsymbol{e}(\boldsymbol{x})\gamma(\boldsymbol{x})\boldsymbol{e}(\boldsymbol{x})\big\rangle, \tag{1}$$

where $\boldsymbol{e}(\boldsymbol{x}) = \overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}(\boldsymbol{x})$ and using the averaging operator $\langle\cdot\rangle$ defined as

$$\langle \boldsymbol{f} \rangle = \frac{1}{|\Omega|}\int_\Omega \boldsymbol{f}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{2}$$

In addition, we made use of the following functional space:

$$\boldsymbol{\mathcal{E}}_0 = \big\{\boldsymbol{e}^* \in \boldsymbol{L}^2_{\text{per}}(\Omega) \mid \exists\, w^* \in H^1_{\text{per}}(\Omega),\, \boldsymbol{e}^* = \boldsymbol{\nabla} w^*\big\},$$

where $\boldsymbol{L}^2_{\text{per}}(\Omega)$ is the subspace of tensor fields with components in $L^2_{\text{per}}(\Omega)$. In general $\boldsymbol{\gamma}_{\text{eff}}$ is a symmetric second-order tensor, i.e. it characterizes an anisotropic effective media. Yet, it can reduce to an isotropic one provided that the microstructure exhibits some symmetries.

Given $\overline{\boldsymbol{e}} \in \mathbb{R}^d$ and with $\tilde{\boldsymbol{e}}$ being the minimizer of (1), the Euler-Lagrange equations for the energy minimization problem (1) are equivalent to the following system of local equations:

$$\begin{cases} \boldsymbol{e}(\boldsymbol{x}) = \overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}(\boldsymbol{x}), \qquad \tilde{\boldsymbol{e}} = \boldsymbol{\nabla} u, \qquad u \text{ periodic on } \partial\Omega, \\ \boldsymbol{j}(\boldsymbol{x}) = \gamma(\boldsymbol{x})\boldsymbol{e}(\boldsymbol{x}), \\ \boldsymbol{\nabla}\cdot\boldsymbol{j}(\boldsymbol{x}) = 0, \qquad \boldsymbol{j}\cdot\boldsymbol{n} \text{ anti-periodic on } \partial\Omega, \end{cases} \tag{3}$$

with $\boldsymbol{n}$ being the unit outward normal on $\partial\Omega$. Physically, $\boldsymbol{e}$ is the electric field with prescribed mean value $\overline{\boldsymbol{e}}$ and $\boldsymbol{j}$ is the electric current or flux. In addition, the periodic fluctuation $\tilde{\boldsymbol{e}}$ is expressed as the

gradient of a scalar potential u sought in $H^1_{\text{per}}(\Omega)$, so that it satisfies the mean-free property $\langle \tilde{\boldsymbol{e}} \rangle = \boldsymbol{0}$.

We are interested in (1) as it is a prototypical problem in computational homogenization, which has a broad range of applicability as in conductivity, considered here, as well as in elasticity and electromagnetism [18]. The tensor $\boldsymbol{\gamma}_{\text{eff}}$ can be fully determined component-wise by solving Problem (3) at most d times using some linearly independent loadings $\overline{\boldsymbol{e}}$ and computing the scalar products of the corresponding solutions. Therefore, to be used in the forthcoming analysis, considering $\alpha \in L^\infty_{\text{per}}(\Omega)$ with $\alpha > 0$ we introduce the associated energetic scalar product and norm as

$$(\boldsymbol{f}_1, \boldsymbol{f}_2)_\alpha = \big\langle \boldsymbol{f}_1(\boldsymbol{x}) \cdot \alpha(\boldsymbol{x}) \boldsymbol{f}_2(\boldsymbol{x}) \big\rangle \qquad \text{and} \qquad \|\boldsymbol{f}\|_\alpha = (\boldsymbol{f}, \boldsymbol{f})_\alpha^{1/2}, \tag{4}$$

for all vectors $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \boldsymbol{L}^2_{\text{per}}(\Omega)$. As a consequence of this definition, the quadratic energy functional considered in (1) satisfies $W(\boldsymbol{f}) = \frac{1}{2}\|\boldsymbol{f}\|^2_\gamma$. Lastly, the standard scalar product and norm on $\boldsymbol{L}^2_{\text{per}}(\Omega)$ will be denoted as $(\cdot, \cdot)$ and $\|\cdot\|$, respectively.

1.3 Objectives

One considers approximating the solution $\tilde{\boldsymbol{e}}$ to (1), or equivalently u to (3), based on (i) the introduction of a subspace $\boldsymbol{\mathcal{E}}^h_0 \subset \boldsymbol{\mathcal{E}}_0$ constructed from a discretization of the domain using a *grid* size h, suitable basis functions and an integration scheme, and (ii) a given iterative scheme whose iteration number will be denoted by k, the combination of which yields an approximation $\tilde{\boldsymbol{e}}^h_k$. In this context, the overall objective of this study is to propose some convergence criteria for the effective property $\boldsymbol{\gamma}_{\text{eff}}$, which according to (1) amounts in quantifying the error $\delta^{\text{eff}}_{k,h}$ in the effective energy as

$$\delta^{\text{eff}}_{k,h} \overset{\text{def}}{=} \big|W(\boldsymbol{e}^h_k) - W(\boldsymbol{e})\big|^{1/2} \qquad \text{with} \qquad \boldsymbol{e}^h_k = \overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}^h_k, \tag{5}$$

for a given $\overline{\boldsymbol{e}} \in \mathbb{R}^d$. Upon introducing the limit $\boldsymbol{e}^h_\infty = \overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}^h_\infty$ of the iterative scheme considered (provided that it converges), the error above will be related in the following to the total error $\delta^{\text{tot}}_{k,h}$ on the fields, in the energetic norm, which satisfies

$$\delta^{\text{tot}}_{k,h} \overset{\text{def}}{=} \|\boldsymbol{e}^h_k - \boldsymbol{e}\|_\gamma \leqslant \underbrace{\|\boldsymbol{e}^h_k - \boldsymbol{e}^h_\infty\|_\gamma}_{\delta^h_k} + \underbrace{\|\boldsymbol{e}^h_\infty - \boldsymbol{e}\|_\gamma}_{\delta^h}, \tag{6}$$

where δ^h_k is the *iteration* error expressed in terms of the converged discrete solution $\boldsymbol{e}^h_\infty$ and δ^h is the *discretization* error expressed in terms of the continuous (exact) solution to (1). Note that the above errors and estimations are formulated using the *continuous* norm introduced in (4), an issue that we will return to in the ensuing analysis.

As neither the effective property $\boldsymbol{\gamma}_{\text{eff}}$ nor the continuous solution are known a priori, the errors introduced above are not accessible directly. Therefore, we aim at proposing some a posteriori estimations, denoted as Δ^h_k and Δ^h for the iteration and discretization errors respectively, which will be readily accessible during computations. Moreover, we also intend to use these error estimators to propose a stopping criterion for the iterative scheme.

Remark 1. *In this study, we only consider the iteration and discretization errors, i.e. we do not take into account a possible inaccurate representation of the geometry of the microstructure. In other words, the discrete geometry is somehow considered to be the exact one in the ensuing analysis.*

2 Solution methods

2.1 Green's operator and linear equation

Considering a homogeneous comparison medium with conductivity $\gamma_0 > 0$, we introduce the associated Green's operator $\boldsymbol{\Gamma}_0$ on $\boldsymbol{L}^2_{\text{per}}(\Omega)$ that is defined as

$$\boldsymbol{\Gamma}_0 : \boldsymbol{\tau} \mapsto \boldsymbol{e}^* = \boldsymbol{\Gamma}_0 \boldsymbol{\tau} \qquad \text{with} \qquad \boldsymbol{e}^* \in \boldsymbol{\mathcal{E}}_0 \text{ and } \boldsymbol{s} = (\gamma_0 \boldsymbol{e}^* - \boldsymbol{\tau}) \in \boldsymbol{\mathcal{S}}, \tag{7}$$

with the functional space $\boldsymbol{\mathcal{S}}$ being given by

$$\boldsymbol{\mathcal{S}} = \big\{\boldsymbol{s} \in \boldsymbol{L}^2_{\mathrm{per}}(\Omega) \mid \boldsymbol{\nabla}\cdot\boldsymbol{s}(\boldsymbol{x}) = 0,\ \boldsymbol{s}\cdot\boldsymbol{n} \text{ anti-periodic on } \partial\Omega\big\}.$$

With a slight abuse of notation, we have $\boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{E}}_0^{\perp}$ for the standard $\boldsymbol{L}^2_{\mathrm{per}}$-scalar product [18, 3]. Moreover, it is known that $\boldsymbol{\Gamma}_0\gamma_0$ is an orthogonal projector (self-adjoint) onto $\boldsymbol{\mathcal{E}}_0$ for the energetic scalar product $(\cdot,\cdot)_{\gamma_0}$. Therefore, for the standard inner product, $\boldsymbol{\Gamma}_0\gamma_0$ is an oblique projector (non-self-adjoint) onto $\boldsymbol{\mathcal{E}}_0$ and it is straightforward to show [30] that its adjoint is $\gamma_0\boldsymbol{\Gamma}_0$.

Now, the local problem (3) is equivalent to the following weak formulation: find $u \in H^1_{\mathrm{per}}(\Omega)$ such that

$$a(u,v) = \ell(v) \qquad \forall v \in H^1_{\mathrm{per}}(\Omega), \tag{8}$$

with a and ℓ being respectively the bilinear and linear forms defined on $H^1_{\mathrm{per}}(\Omega)$ as:

$$a(u,v) = \int_\Omega \gamma\boldsymbol{\nabla}u\cdot\boldsymbol{\nabla}v\,\mathrm{d}\boldsymbol{x} \qquad \text{and} \qquad \ell(v) = -\int_\Omega \gamma\,\overline{\boldsymbol{e}}\cdot\boldsymbol{\nabla}v\,\mathrm{d}\boldsymbol{x}.$$

The identity (8) is equivalent to

$$\big(\gamma(\overline{\boldsymbol{e}}+\tilde{\boldsymbol{e}}),\boldsymbol{e}^*\big) = 0 \qquad \forall \boldsymbol{e}^* = \boldsymbol{\nabla}v \in \boldsymbol{\mathcal{E}}_0, \tag{9}$$

so that $\boldsymbol{j} = \gamma(\overline{\boldsymbol{e}}+\tilde{\boldsymbol{e}})$ belongs to the subspace orthogonal to $\boldsymbol{\mathcal{E}}_0 = \mathrm{Im}(\boldsymbol{\Gamma}_0\gamma_0)$ for the standard $\boldsymbol{L}^2_{\mathrm{per}}$-scalar product. Therefore, owing to the properties of the adjoint operator, one has $\boldsymbol{j} \in \mathrm{Im}(\boldsymbol{\Gamma}_0\gamma_0)^{\perp} = \mathrm{Ker}(\gamma_0\boldsymbol{\Gamma}_0)$, which according to the above entails $\boldsymbol{\Gamma}_0\boldsymbol{j} = \boldsymbol{0}$. The latter identity can finally be recast as the following linear equation for $\tilde{\boldsymbol{e}} \in \boldsymbol{\mathcal{E}}_0$:

$$\boldsymbol{A}\tilde{\boldsymbol{e}} = \boldsymbol{b}, \qquad \text{with} \qquad \boldsymbol{A} = \boldsymbol{\Gamma}_0\gamma : \boldsymbol{\mathcal{E}}_0 \to \boldsymbol{\mathcal{E}}_0 \qquad \text{and} \qquad \boldsymbol{b} = -\boldsymbol{\Gamma}_0\gamma\overline{\boldsymbol{e}} \in \boldsymbol{\mathcal{E}}_0. \tag{10}$$

Note that this equation can equivalently be obtained from the first-order optimality condition for the minimization problem (1) by recognizing that the gradient of W in $\boldsymbol{\mathcal{E}}_0$ endowed with the energetic scalar product $(\cdot,\cdot)_{\gamma_0}$ writes as $\boldsymbol{\nabla}W(\boldsymbol{e}^*) = \boldsymbol{\Gamma}_0\gamma(\overline{\boldsymbol{e}}+\boldsymbol{e}^*)$ [11, 3].

Finally, in the periodic setting considered, the Green's operator (7) can be expressed in closed-form using the Fourier transform $\mathscr{F}$, see Appendix A, as

$$\boldsymbol{\Gamma}_0\boldsymbol{\tau}(\boldsymbol{x}) = \mathscr{F}^{-1}\Big[\hat{\boldsymbol{\Gamma}}_0(\boldsymbol{\xi})\cdot\mathscr{F}[\boldsymbol{\tau}](\boldsymbol{\xi})\Big](\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in \Omega, \tag{11}$$

with the symmetric second-order tensor $\hat{\boldsymbol{\Gamma}}_0(\boldsymbol{\xi})$ being defined in the Fourier space by

$$\hat{\boldsymbol{\Gamma}}_0(\boldsymbol{0}) = \boldsymbol{0} \qquad \text{and} \qquad \hat{\boldsymbol{\Gamma}}_0(\boldsymbol{\xi}) = \frac{\boldsymbol{\xi}\otimes\boldsymbol{\xi}}{\gamma_0|\boldsymbol{\xi}|^2} \qquad \forall \boldsymbol{\xi} \in \mathcal{R}^*\backslash\{\boldsymbol{0}\}. \tag{12}$$

2.2 Discretization and numerical methods

2.2.1 Discrete problem

To transpose the continuous linear problem (10) into a discrete setting, we consider a regular grid of size h, with the corresponding set of interpolation points being $\{\boldsymbol{x}_{\boldsymbol{i}}\}$ where $\boldsymbol{i}$ a multi-index in dimension d, and the approximation subspace $\boldsymbol{\mathcal{T}}^h \subset \boldsymbol{L}^2_{\mathrm{per}}(\Omega)$ being generated by the trigonometric polynomials associated with the Discrete Fourier Transform (DFT) on the grid considered. Correspondingly, consider the discrete subspace $\boldsymbol{\mathcal{E}}_0^h = \boldsymbol{\mathcal{E}}_0 \cap \boldsymbol{\mathcal{T}}^h$ of mean-free gradient fields, relatively to the discrete version $\langle\cdot\rangle_h$ of the averaging operator (2) defined as

$$\langle\boldsymbol{f}\rangle_h = \sum_{\boldsymbol{i}} h^d \boldsymbol{f}(\boldsymbol{x}_{\boldsymbol{i}}), \tag{13}$$

a quadrature scheme that coincides with the trapezoidal rule [2]. The definition (13) also yields a discrete version $(\cdot,\cdot)_{h,\alpha}$ of the scalar product (4), as

$$(\boldsymbol{f}_1, \boldsymbol{f}_2)_{h,\alpha} = \sum_{\boldsymbol{i}} h^d \boldsymbol{f}_1(\boldsymbol{x_i}) \cdot \alpha(\boldsymbol{x_i}) \boldsymbol{f}_2(\boldsymbol{x_i})$$

whose associated norm $\|\cdot\|_{h,\alpha}$ defines the discrete energy functional

$$W_h(\boldsymbol{f}) = \frac{1}{2}\|\boldsymbol{f}\|^2_{h,\gamma}. \tag{14}$$

Remark 2. *In the ensuing analysis, we will have to deal with the continuous norms of some discrete fields. To do so, given a discrete field, say $\boldsymbol{f}^h$ in a discrete subspace such as $\boldsymbol{\mathcal{T}}^h$, then in the evaluation of its continuous norm $\|\cdot\|_\alpha$ we consider that $\boldsymbol{f}^h$ is extended as a piecewise-constant field on each element or pixel of the grid considered. By doing so it holds*

$$\|\boldsymbol{f}^h\|_\alpha = \|\boldsymbol{f}^h\|_{h,\alpha}.$$

This convention will allow us to evaluate quantities such as $\|\boldsymbol{f}^h - \boldsymbol{f}\|_\alpha$, where $\boldsymbol{f}$ is a continuous field. Note that we do not *interpolate $\boldsymbol{f}^h$ in the associated basis of continuous functions, which in the case of the Fourier interpolation would give continuous fields oscillating between sampling points that would in turn yield an artificially high value of $\|\boldsymbol{f}^h - \boldsymbol{f}\|_\alpha$.*

In addition, a DFT-based version $\boldsymbol{\Gamma}^h_0$ of the Green's operator in (10) is commonly defined by the straightforward transposition of (11–12) to the DFT setting. Let us underline that it has also been proposed [32, 5, 34, 33, 26] to modify the definition of the discrete Green's operator $\boldsymbol{\Gamma}^h_0$ by making use of staggered grids or introducing in (12) a filtering of the set of discrete spatial frequencies $\{\boldsymbol{\xi_i}\}$. The latter strategy is equivalent to approximating the continuous problem (3) using a finite-differences scheme, whose degree of approximation can be directly related to the chosen frequency filter [7]. In the present study, the considered modification of the Green's operator corresponds to the first degree central finite-differences scheme.

In this setting, one considers the corresponding discrete counterpart of the weak formulation (9) with the inexact integration scheme (13), which leads to the following equation for the sought discrete solution $\tilde{\boldsymbol{e}}^h \in \boldsymbol{\mathcal{E}}^h_0$:

$$\left(\gamma^h(\overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}^h), \boldsymbol{e}^{*h}\right)_h = 0 \qquad \forall \boldsymbol{e}^{*h} \in \boldsymbol{\mathcal{E}}^h_0, \tag{15}$$

where γ^h denotes the constitutive conductivity field evaluated at the DFT interpolation points, i.e. $\gamma^h = \{\gamma(\boldsymbol{x_i})\}$.

Remark 3. *As a consequence, the discrete conductivity γ^h is considered to be equal to the exact one γ at all the discretization points.*

Now, the equation (15) implies that $\boldsymbol{j}^h = \gamma^h(\overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}^h)$ belongs to the orthogonal subspace of $\boldsymbol{\mathcal{E}}^h_0$ for the (standard) discrete scalar product on $\boldsymbol{\mathcal{T}}^h$. Upon noting that the Helmholtz decomposition of $\boldsymbol{L}^2_{\text{per}}(\Omega)$ (as the direct sum of the mutually orthogonal subspaces of uniform, mean-free gradient and mean-free divergence-free fields) can be prolonged to $\boldsymbol{\mathcal{T}}^h$, it has been shown [28, 30] that $\boldsymbol{\Gamma}^h_0\gamma_0$ defines a projector onto $\boldsymbol{\mathcal{E}}^h_0$. Again, this operator is non-orthogonal for the standard scalar product while it is self-adjoint for the induced energetic one $(\cdot,\cdot)_{h,\gamma_0}$. Therefore, (15) leads to the following linear system:

$$\boldsymbol{A}^h\tilde{\boldsymbol{e}}^h = \boldsymbol{b}^h, \qquad \text{with} \qquad \boldsymbol{A}^h = \boldsymbol{\Gamma}^h_0\gamma^h : \boldsymbol{\mathcal{E}}^h_0 \to \boldsymbol{\mathcal{E}}^h_0 \qquad \text{and} \qquad \boldsymbol{b}^h = -\boldsymbol{\Gamma}^h_0\gamma^h\overline{\boldsymbol{e}} \in \boldsymbol{\mathcal{E}}^h_0. \tag{16}$$

This is the system under investigation and it should be noted that it will be handled as is, even if a modified version of the Green's operator is used.

Remark 4. *In the computational treatment of composite materials, it can be expected that the geometry of the microstructure cannot be represented exactly on a regular grid. The associated error on the geometry is not explicitly taken into account in this study. However, this error appears implicitly in the difference between the continuous and discrete integral forms, such as $\left(\cdot,\cdot\right)_\gamma$ and $\left(\cdot,\cdot\right)_{h,\gamma}$ respectively.*

2.2.2 Iterative schemes

To solve the linear system (16) numerically, we intend to use an iterative scheme that will yield an approximation $\tilde{\boldsymbol{e}}_k^h = \boldsymbol{\nabla} u_k^h$ of the discrete solution $\tilde{\boldsymbol{e}}^h$ at a given iterate k. Provided that the scheme considered converges, we formally get

$$\tilde{\boldsymbol{e}}_k^h \underset{k\infty}{\rightarrow} \tilde{\boldsymbol{e}}_\infty^h \equiv \tilde{\boldsymbol{e}}^h. \tag{17}$$

Here, two of the main approaches used in practice are adopted, namely a fixed-point scheme and the conjugate gradient method:

(i) Owing to the orthogonal decomposition of the discrete approximation space $\boldsymbol{\mathcal{T}}^h$ and using again the interpretation of $\boldsymbol{\Gamma}_0^h\gamma_0$ as a projector, we have $\boldsymbol{\Gamma}_0^h\gamma_0(\overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}^h) = \tilde{\boldsymbol{e}}^h$ so that (16) can be recast as

$$(\boldsymbol{I} + \boldsymbol{\Gamma}_0^h\delta\gamma^h)\boldsymbol{e}^h = \overline{\boldsymbol{e}}, \qquad \text{with} \qquad \boldsymbol{e}^h = \overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}^h$$

and where $\delta\gamma^h = \gamma^h - \gamma_0$. The equation above is a Lippmann-Schwinger integral equation for the total field $\boldsymbol{e}^h$, which can be inverted through a Neumann series expansion. This coincides with the fixed-point algorithm originally proposed [20]. The corresponding results shown hereafter will be labeled as FP or FPm provided that the Green's operator is used in its original or a modified version (as described in Section 2.2.1), respectively.

(ii) Alternatively, the linear system (16) can be solved using a conjugate gradient solver [36]. Note that, in general, care must be taken to use an energetic scalar product for which the operator $\boldsymbol{A}^h$ to be inverted is symmetric. Note that in the case of isotropic conductivity considered here some simplifications occurs. Accordingly, for any discrete field $\check{\boldsymbol{e}}^h \in \boldsymbol{\mathcal{E}}_0^h$, one introduces the corresponding residual:

$$\check{\boldsymbol{r}}^h(\check{\boldsymbol{e}}^h) = \boldsymbol{A}^h\check{\boldsymbol{e}}^h - \boldsymbol{b}^h, \tag{18}$$

which is minimized in a norm weighted by $(\boldsymbol{A}^h)^{-1}$. Indeed, the quadratic minimization problem associated with the linear system (16) satisfies

$$\tilde{\boldsymbol{e}}^h = \arg\min_{\check{\boldsymbol{e}}^h\in\boldsymbol{\mathcal{E}}_0^h} \left(\tfrac{1}{2}(\check{\boldsymbol{e}}^h, \boldsymbol{A}^h\check{\boldsymbol{e}}^h)_h - (\check{\boldsymbol{e}}^h, \boldsymbol{b}^h)_h\right) = \arg\min_{\check{\boldsymbol{e}}^h\in\boldsymbol{\mathcal{E}}_0^h} \|\check{\boldsymbol{r}}^h(\check{\boldsymbol{e}}^h)\|^2_{h,(\boldsymbol{A}^h)^{-1}}.$$

The associated results below will be labelled as CG or CGm depending on the version (original or modified) of the Green's operator employed.

Remark 5. *In both approaches, the computed iterates $\tilde{\boldsymbol{e}}_k^h$ are compatible fields in the sense of the Fourier-based discretization considered, i.e. $\tilde{\boldsymbol{e}}_k^h \in \boldsymbol{\mathcal{E}}_0^h = \boldsymbol{\mathcal{E}}_0 \cap \boldsymbol{\mathcal{T}}^h$, with $\boldsymbol{\mathcal{E}}_0$ being the subspace of mean-free gradient fields and $\boldsymbol{\mathcal{T}}^h$ the approximation subspace generated by trigonometric polynomials. The schemes are initialized by setting $\tilde{\boldsymbol{e}}_0^h = \boldsymbol{0}$.*

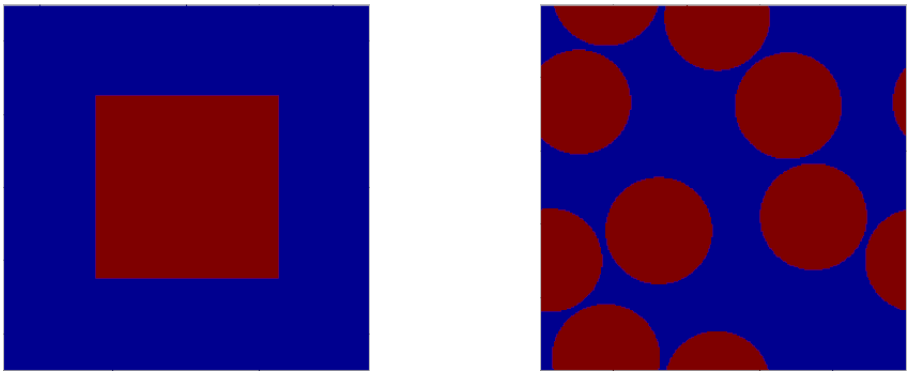
2.3 Test-cases

In this study, two distinct piecewise-homogeneous 2D material distributions are considered to illustrate the proposed approaches, see Figure 1. Different conductivity ratio ρ will be considered, with the definition $\rho = \gamma_\mathrm{i}/\gamma_\mathrm{m}$ where γ_i is the conductivity of the inclusion (or the set thereof) in red and γ_m this of of the matrix phase in blue.

The first geometry presents a regular pattern of square inclusions, with a surface fraction of 1/4, see Fig. 1a. In [22], the corresponding effective conductivity has been calculated analytically, which allows a reliable computation of the corresponding error by (5). The effective conductivity $\gamma_\mathrm{eff}^\mathrm{square}$ is found to be isotropic and writes as

$$\gamma_\mathrm{eff}^\mathrm{square} = \gamma_\mathrm{m}\sqrt{\frac{\gamma_\mathrm{m} + 3\gamma_\mathrm{i}}{3\gamma_\mathrm{m} + \gamma_\mathrm{i}}}.$$

This test-case presents the further advantage that the discretized geometry conforms strictly to the exact geometry. In addition, the continuous solution of the conductivity problem exhibits integrable



(a) Square inclusion. (b) Random circular inclusions.

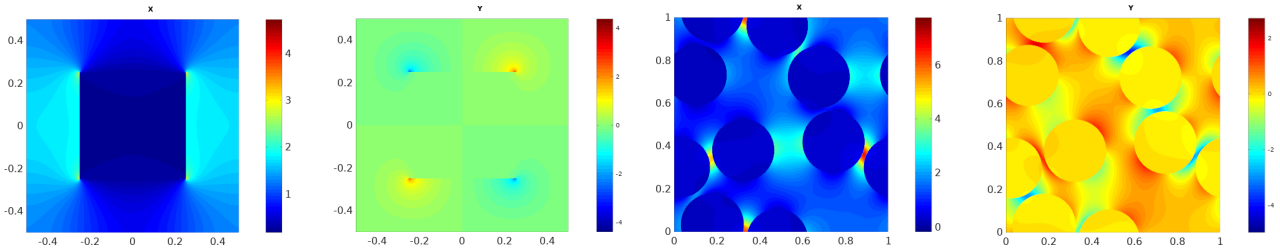
Figure 1: Geometries defining the test-cases considered.

singularities at the corners [22], which poses some difficulties in the computation of an approximation. Such singularities are typical of geometries with corners, which makes this problem interesting in the study of errors estimators.

The second geometry consists in a random pattern of circular inclusions, see Fig. 1b. The inclusions have a random radius satisfying $0.14 \leqslant r \leqslant 0.15$, the minimal distance between two inclusions is 0.01, and the phase fraction of the inclusions is equal to 1/4. The geometry in this case is not exactly rendered by the discretization, but the solution is regular (provided, as it is ensured here, a minimal exclusion distance is imposed between two inclusions). A reference numerical solution for this problem is computed using the finite element method on a mesh with P1 triangular elements and approximately 2^8 nodes on one side of the unit square. The mesh has 88337 nodes in total. The associated conductivity will be considered as the reference value.

To refer to the different configurations, we adopt a compact notation that indicates, in that order: the shape of the geometry (`squ` or `cir`), the phase ratio ρ, the number of pixels, and the solver used. For example, a computation labeled `[squ;rho=1E3;npix=2**6;FP]` is done on the square inclusion case, with a conductivity ratio equal to 10^3, a discretization of $2^6 \times 2^6$ pixels, and the fixed point method applied to the non-modified operator.

Finally, while the computation of the complete anisotropic effective conductivity $\boldsymbol{\gamma}_{\text{eff}}$ requires to solve Problem (3) twice with $\overline{\boldsymbol{e}}$ describing an orthonormal basis of $\mathbb{R}^2$, in the numerical experiments shown hereafter, only the loading $\overline{\boldsymbol{e}} = (1, 0)$ is considered. Accordingly, Figure 2 presents the maps of the components of the field $\tilde{\boldsymbol{e}}_k^h = \boldsymbol{\nabla} u_k^h$ computed for the two test-cases considered and obtained at convergence with the `CG` variant of the FFT-based method on a discretization of $2^{10} \times 2^{10}$ pixels.



(a) `[squ;rho=1E1;npix=2**10;CG]` (b) `[cir;rho=1E1;npix=2**10;CG]`

Figure 2: Components of the numerical solution $\tilde{\boldsymbol{e}}_k^h = \boldsymbol{\nabla} u_k^h$ for the two geometries considered and $\overline{\boldsymbol{e}} = (1, 0)$.

3 Towards a convergence criterion

3.1 Estimation of the iteration error

3.1.1 Relationship between the residual and the iteration error

As announced, since the converged solution $\tilde{\boldsymbol{e}}^h_\infty$ is not accessible a priori, an estimator Δ^h_k of the exact iteration error δ^h_k is needed. In the case of a gradient descent algorithm, such as the conjugate gradient (CG) considered here, it makes sense to directly use a norm of the residual $\boldsymbol{r}^h_k$ in (18) as a measure of convergence. Here, we consider the following energetic norm:

$$\Delta^h_k = \|\boldsymbol{r}^h_k\|_{h,\gamma} = \|\boldsymbol{\Gamma}^h_0 \gamma^h (\overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}^h_k)\|_{h,\gamma}. \tag{19}$$

This proposition is slightly different from the one in [3], where the norm $\|\boldsymbol{r}^h_k\|_{h,\gamma_0} = \|\boldsymbol{\nabla} W_h(\tilde{\boldsymbol{e}}^h_k)\|_{h,\gamma_0}$ is considered, with the featured gradient of W_h being computed in the approximation subspace $\boldsymbol{\mathcal{E}}^h_0$ endowed with the energetic scalar product $(\cdot,\cdot)_{h,\gamma_0}$. Yet, it should be noted that, provided that $\min \gamma^h \leqslant \gamma_0 \leqslant \max \gamma^h$ as it will be ensured later on, then the different energetic norms considered are equivalent as it holds

$$\frac{\min \gamma^h}{\gamma_0} \|\boldsymbol{f}\|^2_{h,\gamma_0} \leqslant \|\boldsymbol{f}\|^2_{h,\gamma} \leqslant \frac{\max \gamma^h}{\gamma_0} \|\boldsymbol{f}\|^2_{h,\gamma_0}.$$

Here, (19) is introduced to be consistent with the approximation of the discretization error in Section 4, and we generalize its use to all of the numerical schemes considered in the present study.

Considering the definitions (18) of the residual $\boldsymbol{r}^h_k$ and (17) of the discrete solution $\tilde{\boldsymbol{e}}^h$ obtained at convergence when $k \to \infty$, it holds

$$\boldsymbol{r}^h_k = \boldsymbol{\Gamma}^h_0 \gamma^h (\tilde{\boldsymbol{e}}^h_k - \tilde{\boldsymbol{e}}^h) = \boldsymbol{A}^h (\tilde{\boldsymbol{e}}^h_k - \tilde{\boldsymbol{e}}^h).$$

Now, the featured discrete operator $\boldsymbol{A}^h$ is symmetric, and hence diagonalizable, in the energetic norm $(\cdot,\cdot)_{h,\gamma}$. Indeed, for all $\boldsymbol{f}_1, \boldsymbol{f}_2$, it holds

$$(\boldsymbol{A}^h \boldsymbol{f}_1, \boldsymbol{f}_2)_{h,\gamma} = (\boldsymbol{\Gamma}^h_0 \gamma^h \boldsymbol{f}_1, \gamma^h \boldsymbol{f}_2)_h = (\gamma^h \boldsymbol{f}_1, \boldsymbol{\Gamma}^h_0 \gamma^h \boldsymbol{f}_2)_h = (\boldsymbol{f}_1, \boldsymbol{A}^h \boldsymbol{f}_2)_{h,\gamma},$$

where the second equality makes use of the reciprocity identity satisfied by $\boldsymbol{\Gamma}^h_0$ [3, Lemma 3]. As a consequence, upon introducing the lowest and largest eigenvalues of $\boldsymbol{A}^h$, denoted as $\lambda_{\min}$ and $\lambda_{\max}$ respectively, one gets

$$\lambda_{\min} \|\tilde{\boldsymbol{e}}^h_k - \tilde{\boldsymbol{e}}^h\|_{h,\gamma} \leqslant \|\boldsymbol{A}^h (\tilde{\boldsymbol{e}}^h_k - \tilde{\boldsymbol{e}}^h)\|_{h,\gamma} \leqslant \lambda_{\max} \|\tilde{\boldsymbol{e}}^h_k - \tilde{\boldsymbol{e}}^h\|_{h,\gamma},$$

which, owing to (19) and Remark 2, can be rewritten as

$$\lambda_{\min} \delta^h_k \leqslant \Delta^h_k \leqslant \lambda_{\max} \delta^h_k.$$

Finally, as is well known [17], the eigenvalues considered are bounded as

$$\frac{\min \gamma^h}{\gamma_0} \leqslant \lambda_{\min} \qquad \text{and} \qquad \lambda_{\max} \leqslant \frac{\max \gamma^h}{\gamma_0}, \tag{20}$$

and, a common choice for the reference medium that ensures convergence of the fixed-point scheme described in Section 2.2.2 is $\gamma_0 = \frac{1}{2}(\min \gamma^h + \max \gamma^h)$. Therefore, upon introducing the contrast $c = \max \gamma^h / \min \gamma^h \geqslant 1$ then one arrives at the following bounds:

$$\frac{2}{1+c} \delta^h_k \leqslant \Delta^h_k \leqslant \frac{2c}{1+c} \delta^h_k. \tag{21}$$

As a consequence, for very large contrasts then the estimated iteration error Δ^h_k can only be bounded as

$$0 \leqslant \Delta^h_k \leqslant 2\delta^h_k \qquad \text{when} \qquad c \gg 1,$$

whereas, conversely, for very small contrasts one has

$$\Delta^h_k \sim \delta^h_k \qquad \text{when} \qquad c \sim 1.$$

3.1.2 Numerical examples

In this section, we consider the case of random circular inclusions, for different values of the conductivity ratio, and a spacial discretization of $2^6 \times 2^6$ pixels. We illustrate numerically the obtained estimates (21) by considering the scaled ratio $(1+c)\Delta_k^h/\delta_k^h$.

This quantity is plotted against iterations k in Figure 3, for the different schemes considered and for various contrasts with $\rho = c$. Note that the contrast is inverted in the case of voids with the conductivity ratio between phases being set in this particular case as $\rho = 10^{-6} = c^{-1}$. To help distinguish between the different cases considered, the multiplying factor $(1+c)$ has been introduced in the representation of the ratio Δ_k^h/δ_k^h. On the downside, this choice artificially and excessively amplifies the relative variations of these quantities. In addition, both axes are in log-scale. Owing to (21), the lower bound is 2 independently of the contrast while the upper bound is equal to $2c$. Yet, on most of the numerical examples reported in Fig. 3 one observes that the quantity $(1+c)\Delta_k^h/\delta_k^h$ considered is about of the order of the contrast, i.e. the error ratio satisfies $\Delta_k^h/\delta_k^h = \mathcal{O}(1)$. Note that deviations from this behavior can be observed in some high-contrast cases. In addition, an asymptotic regime is met in all cases, i.e. the residual Δ_k^h stabilizes around a value proportional to the exact iteration error δ_k^h, typically in a few dozen iterations, as will be considered in the ensuing numerical experiments.

Clearly, the residual Δ_k^h is a reliable estimate of the exact iteration error δ_k^h when the contrast is low. This appears to also hold in the high contrast cases in examples considered, with the residual and error being of the same order of magnitude despite the loose bound (21). The residual will consequently be used as an estimate of the iteration error in the remainder of this article.

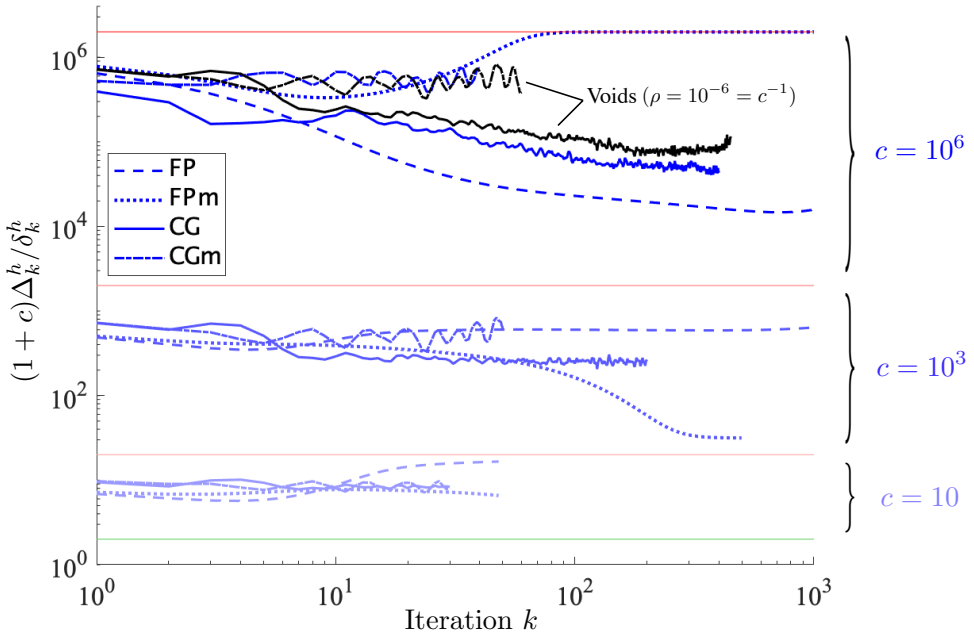


Figure 3: Ratio of the residual Δ_k^h by the exact iteration error δ_k^h, scaled by the contrast, in the random circular inclusions case for various contrasts and for the different schemes considered. The green line indicates the lower bound equal to 2 and the red lines corresponds to the upper bound $2c$ in (21).

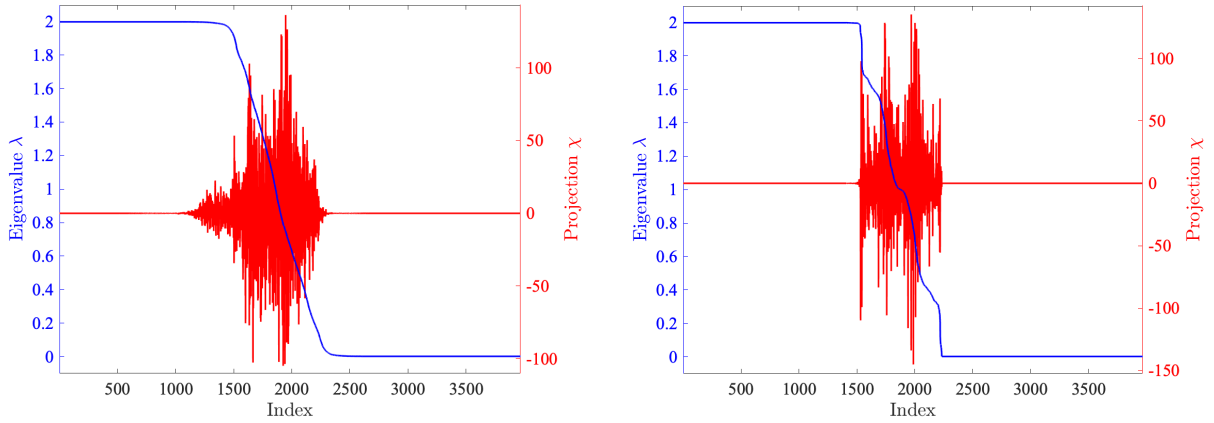
In Equation (21), the lower bound for Δ_k^h would be attained if the residual $(\tilde{\boldsymbol{e}}_k^h - \tilde{\boldsymbol{e}}^h)$ were only spanned by the eigenvectors of $\boldsymbol{A}^h$ associated with its lowest eigenvalue $\lambda_{\min}$. In order to visualize the projection of this residual onto the eigenvectors of $\boldsymbol{A}^h$, we build the matrix M that represents $\boldsymbol{A}^h$ in an energy-orthogonal basis $\{\boldsymbol{t}_i\}_{i=1,\dots,N}$ of $\boldsymbol{\mathcal{E}}_0^h = \boldsymbol{\mathcal{E}}_0 \cap \boldsymbol{\mathcal{T}}^h$, i.e. we have:

$$\begin{cases} (\boldsymbol{t}_i, \boldsymbol{t}_j)_{h,\gamma} = \delta_{ij} \\ \mathsf{M}_{ij} = \left(\boldsymbol{t}_i, \boldsymbol{A}^h \boldsymbol{t}_j\right)_{h,\gamma} \end{cases} \qquad \forall i,j \in \{1,\dots,N\}.$$

In this setting, we display on Figure 4 the numerically-computed eigenvalues λ_i of M in the case of a conductivity ratio $\rho = 10^3$. The vector χ is superposed on the same figure; it corresponds to the decomposition on the eigenbasis considered of the residual $(\tilde{\boldsymbol{e}}^h_{30} - \tilde{\boldsymbol{e}}^h)$ at the 30th fixed-point iteration, i.e. for all $i = 1, \dots, n$ we have

$$\chi_i = \sum_{j=1}^{n} \mathsf{U}_{ji} \left(\boldsymbol{t}_j, \tilde{\boldsymbol{e}}^h_{30} - \tilde{\boldsymbol{e}}^h \right)_{h,\gamma} \qquad \text{with} \qquad \begin{cases} \mathsf{U}^T \mathsf{M} \mathsf{U} = \Lambda, \\ \mathsf{U}^T \mathsf{U} = \mathsf{I}. \end{cases}$$

where I is the identity matrix and $\Lambda = \operatorname{diag}(\lambda_i)_{i=1,\dots,n}$. This is illustrated on Figure 4 for the fixed-point method using the standard or the modified operator, FP and FPm respectively. One can



(a) `[cir;rho=1E3;npix=2**6;FP]` (b) `[cir;rho=1E3;npix=2**6;FPm]`

Figure 4: Decomposition of the residual $(\tilde{\boldsymbol{e}}^h_{30} - \tilde{\boldsymbol{e}}^h)$ at the 30th fixed-point iteration on the discrete eigenvectors of $\boldsymbol{A}^h$.

observe a behavior that has already been highlighted [2]: there are two massively multiple eigenvalues (corresponding the bounds (20) and associated to the two distinct values of the local conductivity γ), that fill the majority of the spectrum, and on which the projection of the residual, and in fact of the solution itself, is zero. Rather, the solution can be decomposed on a smaller number of eigenvectors corresponding to intermediate eigenvalues. For both the non-modified and the modified operator, the overall contribution of each term χ_i to the residual are rather equally distributed among these eigenvalues. Accordingly, the ratio Δ^h_k / δ^h_k is expected to be weighted in between $\lambda_{\min}$ and $\lambda_{\max}$.

3.2 Examples of convergence behaviors in terms of error and residual

3.2.1 Error on the energy

Now that we have assessed the use of the residual Δ^h_k in (19) as a reliable indicator of the iteration error, one investigates numerically here its correlation with the error $\delta^{\text{eff}}_{k,h}$ in the effective energy (5) or with the total error $\delta^{\text{tot}}_{k,h}$ on the fields (6). It should be noted that the theoretical errors in (5) and (6) are defined relatively to the continuous energetic norm $\|\cdot\|_\gamma$. In this context, since $\tilde{\boldsymbol{e}} \in \boldsymbol{\mathcal{E}}_0$ is the solution of the weak formulation (9) of the problem, and introducing the solution $\tilde{\boldsymbol{e}}^h_\star$ to (9) in the approximation subspace $\boldsymbol{\mathcal{E}}^h_0 \subset \boldsymbol{\mathcal{E}}_0$ and with exact integration, we have the following Galerkin orthogonality relations:

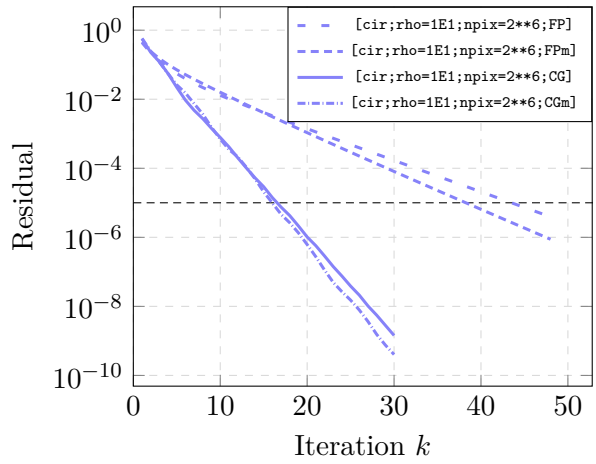
$$\big(\gamma(\overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}), \tilde{\boldsymbol{e}}\big) = 0 \qquad \text{and} \qquad \big(\gamma(\overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}), \tilde{\boldsymbol{e}}^h_\star\big) = 0, \tag{22}$$

which could formally be rewritten using the energetic scalar product $(\cdot,\cdot)_\gamma$. Moreover, considering the following identity

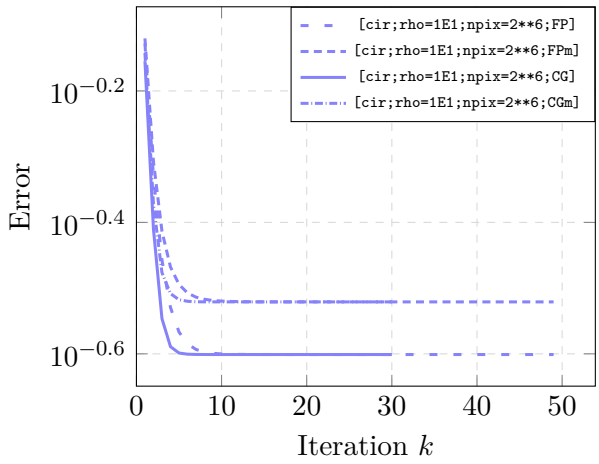
$$\|\tilde{\boldsymbol{e}} - \tilde{\boldsymbol{e}}^h_\star\|^2_\gamma = \|\overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}\|^2_\gamma + \|\overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}^h_\star\|^2_\gamma - 2(\overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}, \overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}^h_\star)_\gamma,$$

then the last right-hand side term simplifies as $(\overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}, \overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}^h_\star)_\gamma = \|\overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}\|^2_\gamma$ according to (22). Therefore, we finally get:

$$\frac{1}{2}\|\tilde{\boldsymbol{e}} - \tilde{\boldsymbol{e}}^h_\star\|^2_\gamma = W(\boldsymbol{e}^h_\star) - W(\boldsymbol{e}^h) \geqslant 0, \tag{23}$$



(a) Residual Δ_k^h and threshold at 10^{-5} (black dashed line)



(b) Error $d_{k,h}^{\text{eff}}$

Figure 5: Residual and error for the four methods considered.

where we defined $\boldsymbol{e}_\star^h = \overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}_\star^h$ and used that $W(\boldsymbol{e}_\star^h) = \frac{1}{2}\|\overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}_\star^h\|_\gamma^2$ and $W(\boldsymbol{e}^h) = \frac{1}{2}\|\overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}^h\|_\gamma^2$. In other words, the square error on the fields in energetic norm is equal to the error on the energy. Note that this result has already been discussed extensively [28, 30, 31, 29], in particular to construct guaranteed numerical bounds on the effective properties, but the argument is reproduced here for the reader's convenience.

Noticeably, the above developments are relative to the solution $\tilde{\boldsymbol{e}}_\star^h$ that satisfies the weak formulation (9) in the approximation subspace $\boldsymbol{\mathcal{E}}_0^h$ and with the integration being performed exactly, which is a crucial argument in the previous developments. The approximate solution $\tilde{\boldsymbol{e}}^h \in \boldsymbol{\mathcal{E}}_0^h$ we deal with however, is solution of the weak formulation (15) where the integration is performed numerically according to the inexact trapezoidal rule (13). As a consequence, an inequality such as (23) is not guaranteed anymore for the numerical solution $\tilde{\boldsymbol{e}}^h$ considered, so that according to (14):

$$\text{the inequality } W_h(\boldsymbol{e}^h) \geqslant W(\boldsymbol{e}) \text{ cannot be ensured.} \tag{24}$$

Note that in the previous expression, the discrete and exact solutions are each associated with the corresponding definition of the energy. Owing to (24), the effective property estimated using the numerical schemes considered can become smaller than the exact one along iterations, which has already been observed in numerous studies. The impact on the development of error estimators is the impossibility to guarantee strict bounds in such a setting, as discussed later on.

In addition, we show in the remainder of this section that the iteration error estimate Δ_k^h is not sufficient in itself and that a reliable discretization error estimate is also needed. To illustrate this and based on (24), we consider next a numerically consistent version $d_{k,h}^{\text{eff}}$ of $\delta_{k,h}^{\text{eff}}$ in (5) as

$$d_{k,h}^{\text{eff}} = \left|W_h(\boldsymbol{e}_k^h) - W(\boldsymbol{e})\right|^{1/2}. \tag{25}$$

Therefore, the convergence of the four variants of the iterative FFT-based homogenization method are investigated numerically next on the set of test-cases considered and with the reference energy term $W(\boldsymbol{e})$ in (25) being either known analytically or computed using the finite elements method on a reference fine grid, see Section 2.3.

3.2.2 Numerical examples

Let us first consider a discretization of $2^6 \times 2^6$ pixels and a conductivity ratio $\rho = 10$. The error and residual convergence history for the random circular inclusions case are displayed on Figure 5. One observes that, for such a moderate contrast, the modified Green operator leads to a slightly faster convergence of the residual Δ_k^h, but to a higher asymptotic error $d_{k,h}^{\text{eff}}$. This was already noticed in

[21]. As expected, the Conjugate Gradient, when associated with the modified operator or not, leads to a faster convergence, a fact that does not affect in the end the asymptotic error compared to the fixed-point scheme. This last result is due to the fact that the underlying discrete linear problem, and thus its solution, does not depend on the resolution method. With these results at hand, one considers the convergence criterion $\Delta_k^h \leqslant 10^{-\beta}\Delta_0^h$, with $\beta = 5$ being chosen here as a typical value, and report the iteration number and the error reached when the threshold is met. These values are then reported on Figure 6. The same is done for the four schemes considered, on the two test-cases as well as for a high conductivity ratio $\rho = 10^3$.

At all points displayed on the plot, the residual Δ_k^h is the same (except for the non-converged points). However, the error $d_{k,h}^{\text{eff}}$ is different for each of them. This shows that the accuracy of the current iterate cannot be deduced from the residual alone.

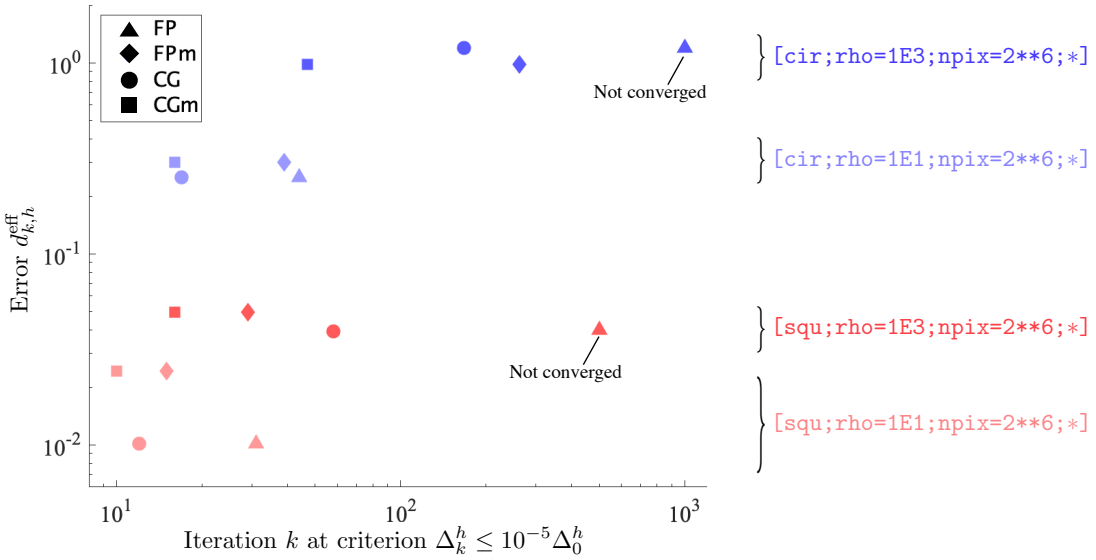


Figure 6: Error $d_{k,h}^{\text{eff}}$ on the energy vs iteration number when the criterion $\Delta_k^h \leqslant 10^{-5}\Delta_0^h$ is met, for the different schemes and test cases considered.

Remark 6. *One finds here a result compatible with what was obtained in [34] with a slightly different modification. The convergence of the residual is much faster when using the modified operator. The asymptotic error is better with the modified operator in the case of circular inclusions and high contrast, and worse in the square inclusion cases. This can be explained because the modification tends to filter out the singularities, which is favorable in the circular case, and not in the square case.*

3.3 Principle of a stopping criterion

Based on the previous discussion, we now expose the principle of a stopping criterion that would be based on estimations of both the iteration and the convergence errors. Considering (6), the triangle inequality entails $\delta^h - \delta_k^h \leqslant \delta_{k,h}^{\text{tot}} \leqslant \delta_k^h + \delta^h$, which in turn implies:

$$|\delta_{k,h}^{\text{tot}} - \delta^h| \leqslant \delta_k^h. \tag{26}$$

As a consequence, if the iteration error δ_k^h becomes much smaller than the discretization error δ^h then the total error $\delta_{k,h}^{\text{tot}}$ cannot be expected to decrease noticeably any more. The bound (26) actually traduces our implicit assumption that, since the numerical schemes considered are expected to converge, the total error would tend to be mostly governed by the discretization error. This invites us to consider a stopping criterion relying on a comparison of δ_k^h and δ^h (or rather their estimations Δ_k^h and Δ^h). Note that a similar idea has been used in the context of domain decomposition [24] .

Let us introduce a parameter $\beta > 0$ such that the iterations are to be stopped when

$$\delta_k^h \leqslant 10^{-\beta}\delta^h. \tag{27}$$

Let us now suppose that the error estimators Δ_k^h and Δ^h are such that there exist some positive parameters m_{d}^h, M_{d}^h, m_{i}^h and M_{i}^h such that the following bounds hold:

$$m_{\mathrm{i}}^h\Delta_k^h \leqslant \delta_k^h \leqslant M_{\mathrm{i}}^h\Delta_k^h \qquad \text{and} \qquad m_{\mathrm{d}}^h\Delta^h \leqslant \delta^h \leqslant M_{\mathrm{d}}^h\Delta^h.$$

Note that, owing to (21), one could choose $m_{\mathrm{i}}^h = (1+c)/2c$ and $M_{\mathrm{i}}^h = (1+c)/2$. In this context, a sufficient condition ensuring (27) writes as the following uniform bound on k:

$$M_{\mathrm{i}}^h\Delta_k^h \leqslant 10^{-\beta}m_{\mathrm{d}}^h\Delta^h,$$

which readily provides a convergence criterion that will be discussed and illustrated numerically in Section 5.

4 Estimation of the discretization error

4.1 Preliminaries

At convergence when $k \to \infty$ the obtained solution $\tilde{\boldsymbol{e}}^h = \boldsymbol{\nabla} u^h$, see (17), satisfies the discrete weak formulation (15) and our objective is now to compute an estimation Δ^h to the corresponding discretization error δ^h in (6) relatively to the exact solution $\tilde{\boldsymbol{e}} = \boldsymbol{\nabla} u$. As δ^h involves both a discrete quantity and a continuous one, the evaluation of the featured continuous norm has to be consistent. As already discussed in Remark 2, computing this norm from the interpolation of $\tilde{\boldsymbol{e}}^h$ as a continuous Fourier series would yield an exaggeratedly high error due to the spurious oscillations of the former between sampling points. Therefore, we consider a piecewise-constant extension of the sampled version of $\tilde{\boldsymbol{e}}^h$ when appropriate, while keeping the same notation for simplicity.

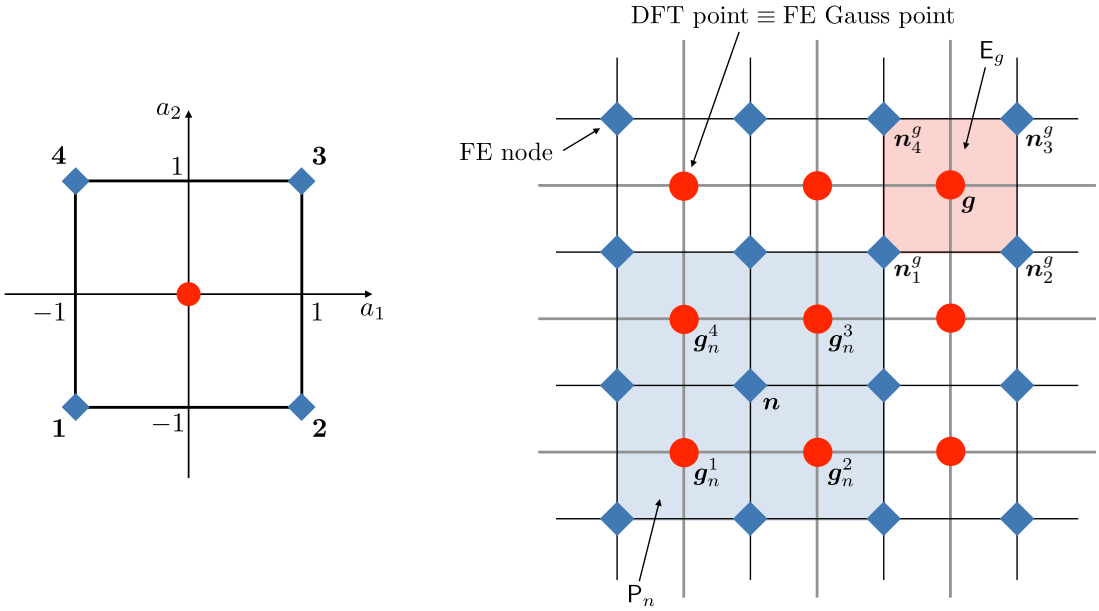
The discretization error estimators developed in this section are inspired by methods that have proven their reliability in the framework of the Finite Element (FE) method. In this context and focusing on 2D problems without loss of generality, the cornerstone of the proposed approach is the interpretation of the DFT grid as a periodic and structured mesh of square Q1-Lagrange finite elements with reduced integration [9], i.e. elements having 4 nodes and 1 Gauss point at the center, see Figure 7a. In addition, we consider that the DFT interpolation points are at the centre of the pixels, which is only a display convention (note for example that it is not the one used by default in Matlab), and we identify these interpolation points as the Gauss points of the periodic FE mesh. The nodes of the FE mesh are thus at the corners of the pixels. This will allow a relatively simple transport of the discrete flux $\boldsymbol{j}^h = \gamma^h(\overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}^h)$ computed at the Gauss points, and piecewise-constant over each $Q1$ element, to a FE version $\mathbf{j}^h$ available at the nodes.

For a given Gauss point $\boldsymbol{g}$, we consider the surrounding element E_g whose associated set of nodes is defined as $\mathcal{N}(\mathsf{E}_g) = \{\boldsymbol{n}_j^g,\, j = 1,\dots,4\}$, see Fig. 7b. A point $\boldsymbol{a} = (a_1, a_2)$ in the square reference unit element of Fig. 7a is mapped to a point $\boldsymbol{x} \in \mathsf{E}_g$ using translations and dilatations of the former as

$$\boldsymbol{x}(\boldsymbol{a}) = \sum_{k=1}^{4} N_k(\boldsymbol{a})\boldsymbol{x}_k^g \qquad \text{with} \qquad \begin{cases} N_1(\boldsymbol{a}) = \dfrac{1}{4}(1-a_1)(1-a_2) \\ N_2(\boldsymbol{a}) = \dfrac{1}{4}(1+a_1)(1-a_2) \\ N_3(\boldsymbol{a}) = \dfrac{1}{4}(1+a_1)(1+a_2) \\ N_4(\boldsymbol{a}) = \dfrac{1}{4}(1-a_1)(1+a_2), \end{cases} \tag{28}$$

with the bilinear shape functions N_k in (28) being also used to define an isoparametric FE interpolation of a vectorial field $\mathbf{f}^h$ at a point $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{a}) \in \mathsf{E}_g$ as

$$\mathbf{f}^h(\boldsymbol{x}) = \sum_{k=1}^{4} N_k(\boldsymbol{a})\mathbf{f}^h(\boldsymbol{x}_k^g). \tag{29}$$



(a) Reference element

(b) FE mesh and DFT grid

Figure 7: Nodes of the FE mesh (blue diamonds) and interpolation points of the DFT grid (red dots). A given DFT interpolation point $\boldsymbol{g}$ at the center of a pixel is seen as the FE Gauss point of the element E_g that coincides geometrically with that pixel.

For further use, for each node $\boldsymbol{n}$ we define the surrounding patch P_n as the set of 4 elements that share this node. The set of Gauss points that belong to this patch is given by $\mathcal{G}(\mathsf{P}_n) = \{\boldsymbol{g}_n^j,\, j = 1, \dots, 4\}$, see Fig. 7b. Moreover, for each Gauss point $\boldsymbol{g}$, we also define A_g as the set of 9 adjacent elements (in 2D) that share at least one node with E_g (including E_g itself). Finally, the set of Gauss points that belong to these elements are denoted by $\mathcal{G}(\mathsf{A}_g)$.

The common idea for the three proposed error estimation approaches is that from the available flux $\boldsymbol{j}^h$, that results from a Fourier-based computation of the solution to the discrete problem (16), one constructs a suitably modified flux $\boldsymbol{j}_*^h$ and use the energetic distance $\Delta^h = \|\boldsymbol{j}^h - \boldsymbol{j}_*^h\|_{h,\gamma^{-1}}$ as an estimation of the discretization error δ^h. In all approaches we will make use of an intermediate FE flux quantity $\mathbf{j}^h$ computed at the nodes following some principles that are described below. Note that the justification that makes the field $\boldsymbol{j}_*^h$ relevant is not the same for the third method and the two others.

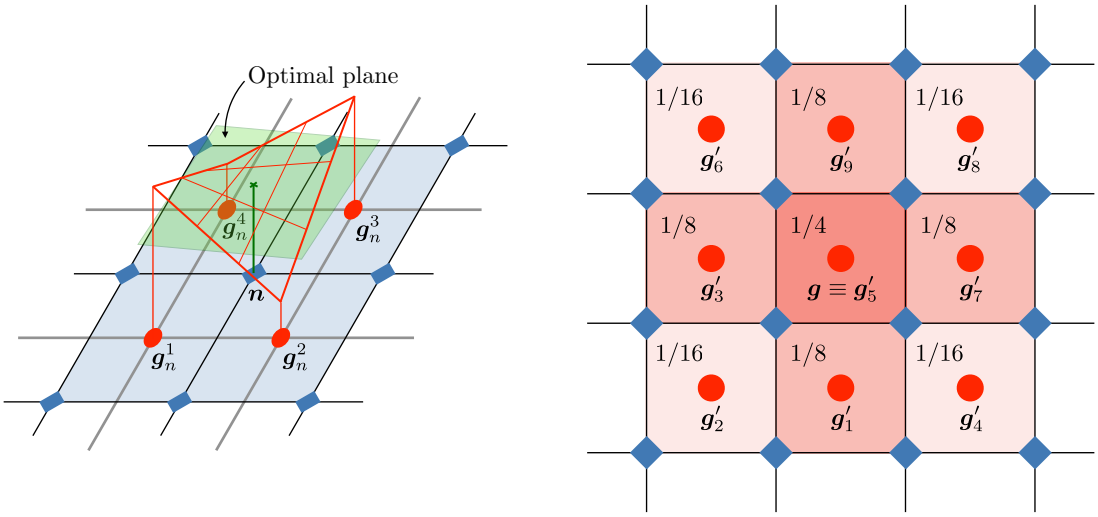
Remark 7. *For the methods described hereafter, it has been chosen to construct the modified flux $\boldsymbol{j}_*^h$ at a given Gauss point from the values of $\boldsymbol{j}^h$ at the nearest neighbors only. In some cases, for example for highly refined discretizations, this window can be made larger and the proposed methods be easily extended to such configurations. Note also that periodicity must be taken into account for the computational treatment involving the Gauss points of the pixels at the boundary of the image.*

Remark 8. *The mean and median filters are commonly used in image processing. These two error estimators can be interpreted as follows: each component of the field j^h is considered as a noisy image (due to the non-physical oscillations), and this image is smoothed out to obtain j_*^h. Note that an image-filtering based approach has also been proposed in [19] to improve FFT-based computations in homogenization.*

4.2 Weighted mean filter

In this method, we aim at constructing a modified flux $\boldsymbol{j}_*^h$ that is aimed at being closer to the exact solution $\boldsymbol{j} = \gamma(\overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}})$ in an appropriate norm. As the Fourier-based discrete solution $\boldsymbol{j}^h$ may typically exhibit unphysical oscillations, often seen as aliasing effects or Gibbs phenomena (see for example Fig. 13), we propose here to construct $\boldsymbol{j}_*^h$ as a smooth version of $\boldsymbol{j}^h$. To do so, we draw from the

rather simple but seminal approach proposed in [37], where an improved version of the flux $\boldsymbol{j}_*^h$ is computed by a nodal averaging process based on the FE shape functions (28–29).



(a) Patch P_n for node $\boldsymbol{n}$ and interpolation of the flux from the associated Gauss points based on the introduction of an optimal fitting plane.

(b) Weights $\alpha_{g,g'}$ associated with the adjacent elements A_g surrounding a given Gauss point $\boldsymbol{g}$.

Figure 8: Weighted mean filter process.

This method has two steps. The first one consists in computing $\mathbf{j}^h(\boldsymbol{n})$ for each FE node $\boldsymbol{n}$ from the field values at the neighboring Gauss points, i.e. $\boldsymbol{j}^h(\boldsymbol{g}_n^j)$ for $\boldsymbol{g}_n^j \in \mathcal{G}(\mathsf{P}_n)$. In order to do this, we define $\mathbf{j}^h(\boldsymbol{n})$ as a point on an optimal hyper-surface that minimizes the least-squares distances to the values at the chosen Gauss points. Here, this hyper-surface is simply defined as a plane, see Fig. 8a, which amounts to use the interpolating bilinear Lagrange polynomials (28) with the reference square element mapped to the Gauss points $\boldsymbol{g}_n^j \in \mathcal{G}(\mathsf{P}_n)$. This polynomial is then evaluated at $\boldsymbol{n}$, the central point, so that we get

$$\mathbf{j}^h(\boldsymbol{n}) = \frac{1}{4} \sum_{\boldsymbol{g}_n^j \in \mathcal{G}(\mathsf{P}_n)} \boldsymbol{j}^h(\boldsymbol{g}_n^j). \tag{30}$$

In the second step, the obtained values are then transported from the nodes back to the Gauss points by evaluation of the shape functions (28). Once again, the use of square elements implies that, for each element E_g the interpolated values at the central Gauss point $\boldsymbol{g}$ is the mean among the four nodal values, i.e.

$$\boldsymbol{j}_*^h(\boldsymbol{g}) = \frac{1}{4} \sum_{\boldsymbol{n}_j^g \in \mathcal{N}(\mathsf{E}_{g'})} \mathbf{j}^h(\boldsymbol{n}_j^g). \tag{31}$$

All in all, combining (30) and (31), leads to a resulting smoothed field computed at a given $\boldsymbol{g}$ from a weighted mean over the values at the Gauss points $\boldsymbol{g}'$ of the adjacent elements, see Fig. 8b, as

$$\boldsymbol{j}_*^h(\boldsymbol{g}) = \sum_{\boldsymbol{g}' \in \mathcal{G}(\mathsf{A}_g)} \alpha_{g,g'} \boldsymbol{j}^h(\boldsymbol{g}') \tag{32}$$

where for any two Gauss point $\boldsymbol{g}$, $\boldsymbol{g}'$, we denote by $\mu_{g,g'}$ the number of nodes shared by E_g and $\mathsf{E}_{g'}$ and define

$$\alpha_{g,g'} = \frac{\mu_{g,g'}}{16} \qquad \text{with} \qquad \sum_{\boldsymbol{g}' \in \mathcal{G}(\mathsf{A}_g)} \alpha_{g,g'} = 1$$

4.3 Weighted median filter

Computing the weighted average (32) can be interpreted as solving for each Gauss point $\boldsymbol{g}$, i.e. for each pixel, the following local weighted quadratic minimization problem:

$$\boldsymbol{j}_*^h(\boldsymbol{g}) = \arg \min_{\boldsymbol{j}_* \in \mathbb{R}^2} \frac{1}{2} \sum_{\boldsymbol{g}' \in \mathcal{G}(\mathsf{A}_g)} \alpha_{g,g'} \|\boldsymbol{j}^h(\boldsymbol{g}') - \boldsymbol{j}_*\|_2^2 \tag{33}$$

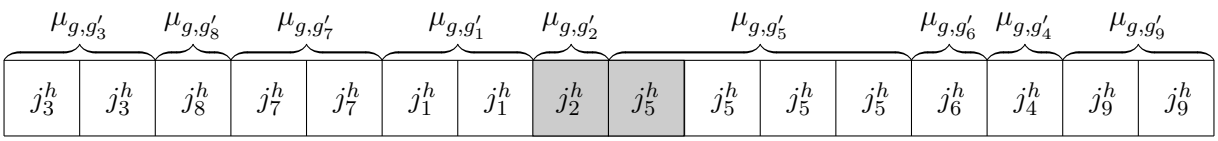


Figure 9: Sorting process for the computation of the weighted median filter (arbitrary example). Each component of $j_\ell^h = j^h(\boldsymbol{g}'_\ell)_\zeta$ for $\boldsymbol{g}'_\ell \in \mathcal{G}(\mathsf{A}_g)$, with $\zeta = x$ and y and using the notations of Fig. 8b, are sorted in ascending order and one of the median values (greyed) is chosen as a minimizer of (34).

where the featured norm $\|\cdot\|_2$ is the standard Euclidean norm. As it is known that the resulting field $\boldsymbol{j}_*^h$ will not respect the possible discontinuities of the original flux, this process may be prone to overestimation of the error. This has been underlined in [27], where an alternative error estimation strategy is proposed, which however cannot be used in our context as it does not directly provides an estimation of $\|\tilde{\boldsymbol{e}} - \tilde{\boldsymbol{e}}^h\|_\gamma$. We rather consider here an alternate weighted median filter that is meant to allow for such possible discontinuities. As mentioned in Remark 8, this type of filtering is commonly employed in signal processing, see for example [8, 35], but it has never been used to our knowledge in the present context. To do so, the minimization problem (33) is modified using the L^1-norm and rewritten component-wise as

$$j_*^h(\boldsymbol{g})_\zeta = \arg\min_{j_* \in \mathbb{R}} \sum_{\boldsymbol{g}' \in \mathcal{G}(\mathsf{A}_g)} \alpha_{g,g'} \left| j^h(\boldsymbol{g}')_\zeta - j_* \right| \qquad \text{with} \qquad \zeta = x \ \text{ or } y. \tag{34}$$

These minimization problems are then solved independently at each point $\boldsymbol{g}$ and for each sought component (x and y in 2D) of the vector $\boldsymbol{j}_*^h(\boldsymbol{g})$. Each one can then be reformulated as two (or three in 3D) unidimensional minimization problems, see (34), for cost functionals that are linear by parts and whose minimum might not be unique. In that case, one of the minimizers is arbitrarily chosen.

Two methods can be used in practice to find a minimizer. The first one consists in evaluating each cost functional in (34) for each of the $j^h(\boldsymbol{g}')_\zeta$, with $\zeta = x$ and y, i.e. at the *corners* of such piecewise linear functions, and storing the value for which each of them is minimal. The second one uses the fact that the minimizer in (34) is the weighted median of the set of vector values $\{j^h(\boldsymbol{g}')_\zeta, \boldsymbol{g}' \in \mathcal{G}(\mathsf{A}_g)\}$. To compute the latter, one can sort the elements of this set in ascending order while duplicating them $\mu_{g,g'}$ times, and pick the median value (here the 8th or 9th as $\sum_{g'} \mu_{g,g'}$ is even). For the sake of the example an arbitrary sorted set is represented in Fig. 9 with reference to the notations of Fig. 8b. Note that, in the case where the filter window is chosen to be larger, see Remark 7 then fast algorithms can be used instead of the above sorting method [8].

4.4 Constitutive Relation Error

The third and last method that will be tested on our problem is the Constitutive Relation Error method (CRE) [13]. This method builds an estimation of the error that, provided a few hypotheses are verified, is proven to be higher than the true error. Its usual limitation is that it is more computationally costly than competing methods. However, as it is shown in this section, in the case where all the elements of the mesh are identical, the cost of the method is similar as the one of the ZZ1 method.

4.4.1 Principle of the method

Let us use the field $\tilde{\boldsymbol{e}}_\star^h$ defined in Section 3.2.1 as the solution to the discrete Lippmann-Schwinger problem with exact integration, with associated total field $\boldsymbol{e}_\star^h = \overline{\boldsymbol{e}} + \tilde{\boldsymbol{e}}_\star^h$. We now aim at estimating the discretization error $\delta^h = \|\tilde{\boldsymbol{e}} - \tilde{\boldsymbol{e}}_\star^h\|_\gamma = \|\gamma(\boldsymbol{e} - \boldsymbol{e}_\star^h)\|_{\gamma^{-1}}$.

For all $\boldsymbol{j}_*^h \in \boldsymbol{\mathcal{S}}$ we have:

$$\begin{aligned} \|\boldsymbol{j}_*^h - \gamma \boldsymbol{e}_\star^h\|_{\gamma^{-1}}^2 &= \|\boldsymbol{j}_*^h - \gamma \boldsymbol{e} + \gamma(\boldsymbol{e} - \boldsymbol{e}_\star^h)\|_{\gamma^{-1}}^2 \\ &= \|\boldsymbol{j}_*^h - \gamma \boldsymbol{e}\|_{\gamma^{-1}}^2 + \|\gamma(\boldsymbol{e} - \boldsymbol{e}_\star^h)\|_{\gamma^{-1}}^2 + 2\big(\boldsymbol{j}_*^h - \gamma \boldsymbol{e}, \gamma(\boldsymbol{e} - \boldsymbol{e}_\star^h)\big)_{\gamma^{-1}}. \end{aligned}$$

Note that it is the exact conductivity γ that enters the above identities consistently with Remark 3. As $(\boldsymbol{e}-\boldsymbol{e}^h_\star)\in\boldsymbol{\mathcal{E}}_0$ and $(\boldsymbol{j}^h_*-\gamma\boldsymbol{e})\in\boldsymbol{\mathcal{S}}$, and since these functional spaces are orthogonal in the sense of the standard scalar product on $\boldsymbol{L}^2_{\rm per}(\Omega)$ (see Section 2.1), then one has:

$$\big(\boldsymbol{j}^h_*-\gamma\boldsymbol{e},\gamma(\boldsymbol{e}-\boldsymbol{e}^h_\star)\big)_{\gamma^{-1}}=\big(\boldsymbol{j}^h_*-\gamma\boldsymbol{e},\boldsymbol{e}-\boldsymbol{e}^h_\star\big)=0, \tag{35}$$

and then

$$\|\boldsymbol{j}^h_*-\gamma\boldsymbol{e}^h_\star\|^2_{\gamma^{-1}}=\|\boldsymbol{j}^h_*-\gamma\boldsymbol{e}\|^2_{\gamma^{-1}}+\|\gamma(\boldsymbol{e}-\boldsymbol{e}^h_\star)\|^2_{\gamma^{-1}}. \tag{36}$$

As a conclusion, for any $\boldsymbol{j}^h_*\in\boldsymbol{\mathcal{S}}$, the quantity $\|\boldsymbol{j}^h_*-\gamma\boldsymbol{e}^h_\star\|_{\gamma^{-1}}$ is an estimator of the discretization error $\delta^h=\|\gamma(\boldsymbol{e}-\boldsymbol{e}^h_\star)\|_{\gamma^{-1}}$, that is all the more accurate as $\|\boldsymbol{j}^h_*-\gamma\boldsymbol{e}\|_{\gamma^{-1}}$ is small, i.e. when $\boldsymbol{j}^h_*$ gets close to the exact flux $\boldsymbol{j}=\gamma\boldsymbol{e}$. What is more, this estimator is guaranteed to be an upper bound of the true discretization error.

We recall however that, due to inexact integration, the orthogonality relations (35) are not satisfied despite the fact that the fields $\boldsymbol{e}^h_k$ computed are compatible, see Remark 5. In other words, the inexact integration amounts to prolongating $\boldsymbol{e}^h$ by a piecewise-constant field that will not coincide with $\boldsymbol{e}^h_\star$ (see Remark 2). As a result, we reach a conclusion similar to (24), in that the above bound on the discretization error cannot be ensured for $\boldsymbol{e}^h$. Yet, we will make use of the derivation above as a guideline to devise an estimator that we discuss next, i.e. we will apply the CRE procedure, designed to estimate $\|\tilde{\boldsymbol{e}}-\tilde{\boldsymbol{e}}^h_\star\|_\gamma$, to construct an estimate of $\|\tilde{\boldsymbol{e}}-\tilde{\boldsymbol{e}}^h\|_\gamma$. The relevance of this approach will be assessed on a number of numerical examples.

The flux-equilibration method [15] will be used to compute a modified field $\boldsymbol{j}^h_*$ from the available numerical solution. It consists in two steps:
– Firstly, determine an equilibrated normal flux F^h_* at the boundary of each element E_g of the F.E. mesh from the original flux $\boldsymbol{j}^h=\gamma^h\boldsymbol{e}^h$ available at the Gauss points.
– Secondly, compute $\boldsymbol{j}^h_*$ from a F.E. approximation on a refined mesh of the solution $\boldsymbol{j}_*$ to the following continuous subproblem:

$$\begin{cases}\boldsymbol{j}_*(\boldsymbol{x})=\gamma^g\boldsymbol{\nabla}\boldsymbol{u}_*(\boldsymbol{x}) & \boldsymbol{x}\in\mathsf{E}_g\\ \boldsymbol{\nabla}\cdot\boldsymbol{j}_*(\boldsymbol{x})=0 & \boldsymbol{x}\in\mathsf{E}_g\\ \boldsymbol{j}_*(\boldsymbol{x})\cdot\boldsymbol{\nu}_g(\boldsymbol{x})=F^h_*(\boldsymbol{x}) & \boldsymbol{x}\in\partial\mathsf{E}_g\end{cases} \tag{37}$$

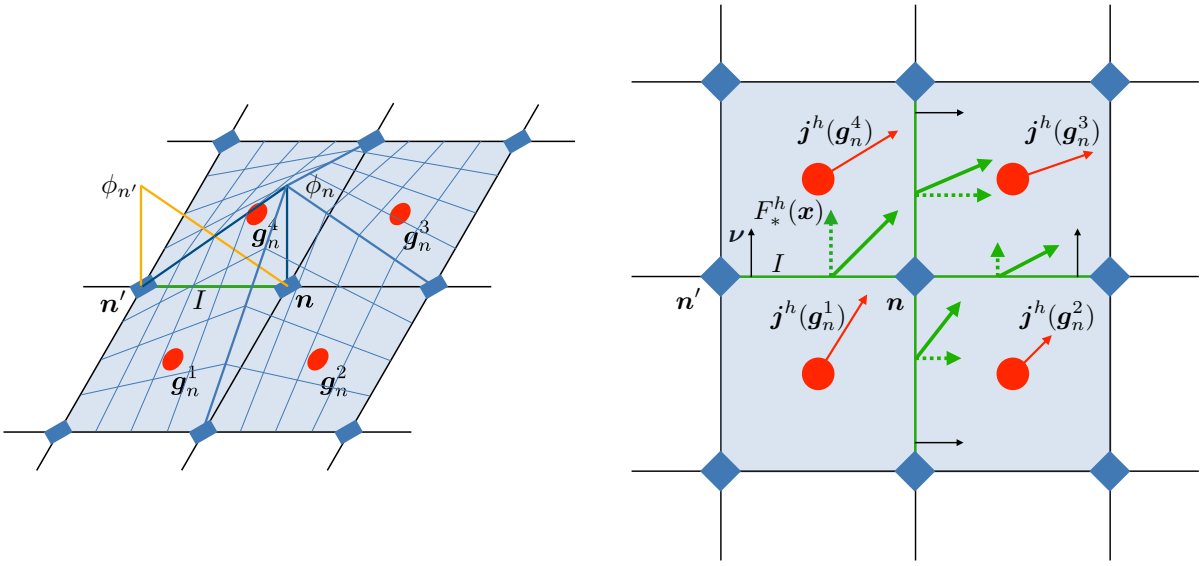
where γ^g is the *homogeneous* conductivity value sampled at the Gauss point $\boldsymbol{g}$ and $\boldsymbol{\nu}_g$ is the unit outward normal on $\partial\mathsf{E}_g$. The fact that the constant value γ^g of γ^h in E_g is used in (37) pertains to the assumption that the latter constitutes the exact conductivity, see remarks 1 and 3 (the geometrical error is not accounted for in the present work). In addition, this choice allows the factorization by the material property, which in turn makes the resolution of (37) numerically efficient. If an exact conductivity field $\gamma(\boldsymbol{x})$, different from its discrete counterpart γ^h, were known then it should be used in (37) to get a consistent *non-homogeneous* material distribution in the sub-discretization of E_g.
Doing so, the field $\boldsymbol{j}^h_*$ respects the discrete equilibrium equation on the whole domain, and its proximity to $\gamma\boldsymbol{e}$ depends on the relevance of the boundary fluxes.

4.4.2 Computation of the boundary fluxes

Given a node $\boldsymbol{n}$, together with the associated patch P_n, let ϕ_n denote the associated global shape function constructed from (28). Considering the inter-element interface I connecting $\boldsymbol{n}$ to another node $\boldsymbol{n}'$, see Fig. 10a, the objective is then to compute an equilibrated boundary flux $F^h_*(\boldsymbol{x})$, spatially varying along I and which must be consistent with the available values of the flux $\boldsymbol{j}^h(\boldsymbol{g}^j_n)$ at the Gauss points $\boldsymbol{g}^j_n\in\mathcal{G}(\mathsf{P}_n)$, see Fig. 10b.

To do so, one considers the prolongation equation [13]. It amounts in equating the virtual works associated with $\boldsymbol{j}^h$ and $\boldsymbol{j}_*$ in any element $\mathsf{E}\in\mathsf{P}_n$ and for all test function ϕ_m relative to a node $\boldsymbol{m}$ of this element, i.e. we set

$$\begin{aligned}\int_{\mathsf{E}}\boldsymbol{j}^h(\boldsymbol{x})\cdot\boldsymbol{\nabla}\phi_m(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}&=\int_{\mathsf{E}}\boldsymbol{j}_*(\boldsymbol{x})\cdot\boldsymbol{\nabla}\phi_m(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\\&=\int_{\partial\mathsf{E}}F^h_*(\boldsymbol{x})\phi_m(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\end{aligned} \tag{38}$$



(a) Global shape function ϕ_n at node $\boldsymbol{n}$ (blue) and interface I (green) connecting to the node $\boldsymbol{n}'$ with shape function $\phi_{n'}$ (whose trace on I is in yellow).

(b) Computation of the normal fluxes $F_*^h(\boldsymbol{x})$ along each interface I (dashed green) with normal $\boldsymbol{\nu}$ (black) from the fluxes $\boldsymbol{j}^h(\boldsymbol{g}_n^j)$ at the Gauss points (red).

Figure 10: Representation and computation of the boundary fluxes.

where the second equality makes use of (37). Since the set of shape functions $\{\phi_n\}_n$ satisfies the property of partition of unity $\sum_n \phi_n(\boldsymbol{x}) = 1$, then (38) ensures that $F_*^h(\boldsymbol{x})$ is equilibrated over E, i.e. it holds

$$\int_{\partial \mathsf{E}} F_*^h(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0.$$

Now, at a given boundary I of an element E, to enforce the continuity of the flux across elements we write $F_*^h(\boldsymbol{x}) = \eta_{\mathsf{E}}^I\,\mathsf{F}_I^h(\boldsymbol{x})$ with the factor $\eta_{\mathsf{E}}^I = \pm 1$ being chosen so as to ensure that an outgoing flux is the opposite of an entering flux on each adjacent elements. This is done consistently with the orientation of the normal $\boldsymbol{\nu}$ in Figure 10b.

In addition, consistently with the discretization (28) considered, we define $\mathsf{F}_I^h(\boldsymbol{x})$ as a linear function of $\boldsymbol{x} \in I$, so that it can be associated with two scalar degrees of freedom denoted as $\mathsf{F}_{I,n}^h$ and $\mathsf{F}_{I,n'}^h$, and which we defined as the following projection

$$\mathsf{F}_{I,n}^h = \int_I \mathsf{F}_I^h(\boldsymbol{x})\phi_n(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \tag{39}$$

and likewise for $\mathsf{F}_{I,n'}^h$ using $\phi_{n'}$. As a consequence for the mesh considered, introducing the set B_n of four interfaces between the elements of P_n, there are four degrees of freedom associated to each node $\boldsymbol{n}$, which we gather into the following vector:

$$\mathsf{F}_n^h = \left\{\mathsf{F}_{I,n}^h,\ I \in \mathsf{B}_n\right\} \in \mathbb{R}^4.$$

Therefore, owing to (38) and upon choosing the test function relative to the node $\boldsymbol{m} = \boldsymbol{n}$ we get

$$\sum_{I \in \mathsf{B}_n} \eta_{\mathsf{E}}^I\,\mathsf{F}_{I,n}^h = \int_{\mathsf{E}} \boldsymbol{j}^h(\boldsymbol{x}) \cdot \boldsymbol{\nabla}\phi_n(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \tag{40}$$

Given that the available flux $\boldsymbol{j}^h$ is piecewise constant in each of the four elements $\mathsf{E} \in \mathsf{P}_n$ (see Section 4.1), then (40) leads to the following linear system:

$$\mathsf{A}\,\mathsf{F}_n^h = \mathsf{b}_n$$

$$\text{with} \quad \mathsf{A} = \begin{pmatrix} -1 & 0 & 0 & -1 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix} \quad \text{and} \quad \mathsf{b}_n = \begin{pmatrix} -j^h(\boldsymbol{g}_n^1)_x - j^h(\boldsymbol{g}_n^1)_y \\ j^h(\boldsymbol{g}_n^2)_x - j^h(\boldsymbol{g}_n^2)_y \\ j^h(\boldsymbol{g}_n^3)_x + j^h(\boldsymbol{g}_n^3)_y \\ -j^h(\boldsymbol{g}_n^4)_x + j^h(\boldsymbol{g}_n^4)_y \end{pmatrix}. \tag{41}$$

However, the matrix A is singular and its null space is of dimension 1. To overcome this, we use its Moore-Penrose pseudo-inverse [1], denoted by A^+. Moreover, as proposed in [15], the extra degree of freedom (in the null space of A) is determined by minimizing the distance between the components of F_n^h and these of an average vector $\langle \mathsf{F}_n^h \rangle$ defined on each interface I from the mean value of the projections onto the normal $\boldsymbol{\nu}$ of the adjacent fluxes, i.e. $\frac{1}{2}\big(\boldsymbol{j}^h(\boldsymbol{g}_n^1) + \boldsymbol{j}^h(\boldsymbol{g}_n^4)\big) \cdot \boldsymbol{\nu}$ for example. All in all, F_n^h is computed from the available components of the flux by solving the following equation:

$$\mathsf{F}_n^h = \big(\mathsf{A}^+\mathsf{B}_1 + \mathsf{k}\mathsf{k}^T\mathsf{B}_2\big)\mathsf{j}_n^h$$

$$\text{with} \quad \mathsf{j}_n^h = \begin{pmatrix} j^h(\boldsymbol{g}_n^1)_x & j^h(\boldsymbol{g}_n^1)_y & j^h(\boldsymbol{g}_n^2)_x & j^h(\boldsymbol{g}_n^2)_y & j^h(\boldsymbol{g}_n^3)_x & j^h(\boldsymbol{g}_n^3)_y & j^h(\boldsymbol{g}_n^4)_x & j^h(\boldsymbol{g}_n^4)_y \end{pmatrix}^T,$$

$$\text{span}(\mathsf{k}) = \text{null}(\mathsf{A}), \quad \mathsf{B}_1 = \begin{pmatrix} -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}, \quad \mathsf{B}_2 = \frac{1}{2}\begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The matrices B_1 and B_2 build respectively b_n and the components of $\langle \mathsf{F}_n^h \rangle$ from j_n^h. As the geometries of the element patches are all the same, the matrices A, B_1, B_2 and vector k do not depend on the specific node $\boldsymbol{n}$. As a consequence, the matrix operator $\big(\mathsf{A}^+\mathsf{B}_1 + \mathsf{k}\mathsf{k}^T\mathsf{B}_2\big)$ is assembled once for all in a pre-processing step and it is then applied to a multi-vector constituted from the set $\big\{\mathsf{j}_n^h\big\}_n$, which is indexed by the nodes $\boldsymbol{n}$ of the given mesh. The boundary fluxes $\big\{\mathsf{F}_n^h\big\}_n$ are then available after such a computation.

Remark 9. *In usual applications of this approach in the context of finite elements [13, 15], one can show that b_n is orthogonal to k. As a result, one can write the problem as the minimization of $\|\mathsf{F}_n^h - \langle \mathsf{F}_n^h \rangle\|$ under the constraint of Equation (41). It can be shown that the idea proposed here is mathematically equivalent to solving this minimization problem (by remarking that (41) only imposes the projection of F_n^h in the subspace orthogonal to k), with the advantage of being also usable when there is no solution to (41).*

4.4.3 Computation of a self-equilibrated flux

For each Gauss point $\boldsymbol{g}$, we now aim at computing a self-equilibrated flux $\boldsymbol{j}_*^h$ for the associated element (or pixel) E_g that respects the boundary fluxes computed at the previous step. We introduce the vector $\mathsf{F}_g^h = \{(\mathsf{F}_g^h)_{jk},\, j = 1, \ldots, 4,\, k = 1, 2\} \in \mathbb{R}^8$ that gathers the relevant degrees of freedom in $\mathsf{F}_{n_j^g}^h$ for all nodes $\boldsymbol{n}_j^g \in \mathcal{N}(\mathsf{E}_g)$, i.e. two associated with each element edge, see Fig. 11.

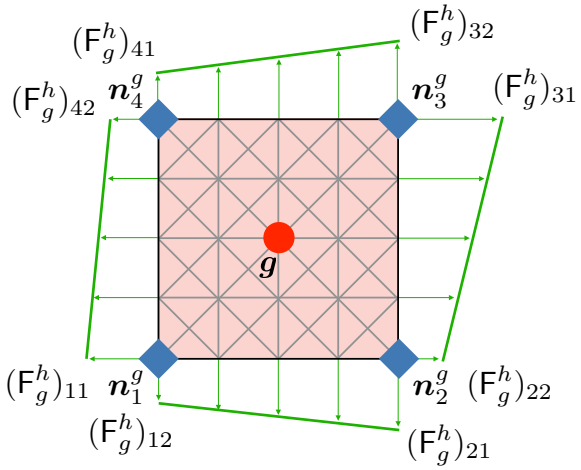


Figure 11: Finite element-based computation of a self-equilibrated flux $\boldsymbol{j}_*^h$ on a given element E_g from the associated boundary fluxes computed previously and using a refined mesh.

Considering the continuous interpolation operator $\mathcal{M}$ that yields piecewise linear normal fluxes $F_*^h(\boldsymbol{x})$ from its projected components F_g^h, see (39), we consider the solution $\boldsymbol{j}_*$ to (37) with the

boundary condition

$$\boldsymbol{j}_*(\boldsymbol{x}) \cdot \boldsymbol{\nu}_g(\boldsymbol{x}) = \mathcal{M}\big(\mathsf{F}_g^h\big)(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in \partial \mathsf{E}_g.$$

The Neumann problem (37) is then discretized and solved using the finite element method on a mesh that is refined compared to the original pixel-based discretization, and using a primal formulation. Noticeably, a p-refinement strategy could be used at this step but h-refinement is simple and efficient here since one only needs to build a single mesh as all the elements we are dealing with are identical. In addition, h-refinement is preferred here as it would allow a direct mapping to refined FFT grids if a multi-grid approach were used to improve the computations. This leads to an approximated flux that is considered to be *sufficiently self-equilibrated* provided the mesh of this subproblem is fine enough [15]. Given $h' < h$ the associated mesh size, we denote by $\mathsf{K}^{h'}$ the corresponding *stiffness* matrix and by $\mathsf{M}^{h'}$ the discretized version of the operator $\mathcal{M}$ and define $\mathsf{u}_*^{h'}$ as the solution to

$$\mathsf{K}^{h'} \mathsf{u}_*^{h'} = \mathsf{M}^{h'} \mathsf{F}_g^h. \tag{42}$$

Due to the absence of Dirichlet boundary conditions, the matrix $\mathsf{K}^{h'}$ is singular and its null space corresponds to the fields that are uniform in E_g. Therefore, we make use of its Moore-Penrose pseudo-inverse $\mathsf{K}^{h'+}$. At this stage a self-equilibrated flux is available on a refined mesh.

Finally, in order to compute the energy of the difference between $\boldsymbol{j}_*^h$ and the original field $\boldsymbol{j}^h$, we consider a discrete operator G that computes a final modified flux $\boldsymbol{j}_*^h$ at the original Gauss point $\boldsymbol{g}$ from the refined finite element solution $\mathsf{u}_*^{h'}$, i.e. $\boldsymbol{j}_*^h(\boldsymbol{g}) = \mathsf{G}\,\mathsf{u}_*^{h'}$. Different strategies can be adopted to do so and the operator chosen here simply amounts in averaging locally the fluxes computed at the Gauss points of the finite elements of E_g that are adjacent to $\boldsymbol{g}$.

As the number of degrees of freedom involved in (42) is relatively small and since all pixels can be discretized using the same refined mesh, then the finite element system (42) can be inverted in a pre-processing step and its inverse stored in matrix form. Therefore, the operator that gives the value $\boldsymbol{j}_*^h(\boldsymbol{g})$ at the Gauss point from the components of the boundary flux F_g^h reads

$$\boldsymbol{j}_*^h(\boldsymbol{g}) = \mathsf{G}\,\mathsf{K}^{h'+}\,\mathsf{M}^{h'}\,\mathsf{F}_g^h.$$

Again, the key-point of the proposed approach is that, all pixels having the same geometry, the matrices G and $\mathsf{K}^{h'}$ only depend on $\boldsymbol{g}$ through a pre-factor γ^g. Therefore, $\Big(\mathsf{G}\,\mathsf{K}^{h'+}\Big)$ and $\mathsf{M}^{h'}$ do not depend on $\boldsymbol{g}$ and the computation above can be directly performed on a multi-vector $\big\{\mathsf{F}_g^h\big\}_g$ that is a mere reordering of the nodal degrees of freedom $\big\{\mathsf{F}_n^h\big\}_n$ of the boundary fluxes. This is slightly different in the case of elasticity, see Remark 12.

4.5 Error maps

The three proposed smoothing methods are now illustrated numerically on the two test cases. First, the Obnosov square inclusion geometry with a conductivity ratio $\rho = 10$ is considered and a numerical solution $\boldsymbol{j}^h$ is computed on a coarse discretization of $2^6 \times 2^6$ pixels using $k = 40$ iterations of the method FP. Then we show on Figure 12 the different modified fields $\boldsymbol{j}_*^h$ computed using the mean, median and CRE methods, which can thus be compared to the field computed originally and to a reference field. The latter corresponds to a computation on a finer grid of $2^{10} \times 2^{10}$ pixels and averaged locally to be shown on the coarse grid.

One can observe that the mean smoothing method does not respect the discontinuities of the flux at the top and bottom boundaries of the inclusion. The CRE and the median methods seem to allow for these discontinuities. In addition the median smoothing does not preserve the corner singularities on the square inclusion. Figure 14a presents the corresponding maps of the local estimated error $(\boldsymbol{j}^h - \boldsymbol{j}_*^h)/\sqrt{\gamma}$. It appears that the over-smoothing done by the mean method leads to a large over-estimation of the error in the discontinuity regions. The median method over-estimates the error nearby the singularities while, on this test-case, the CRE method performs at best. On Figure 13, we present the reconstructed fluxes for the random circles test-case with the local errors being displayed on Figure 14b.

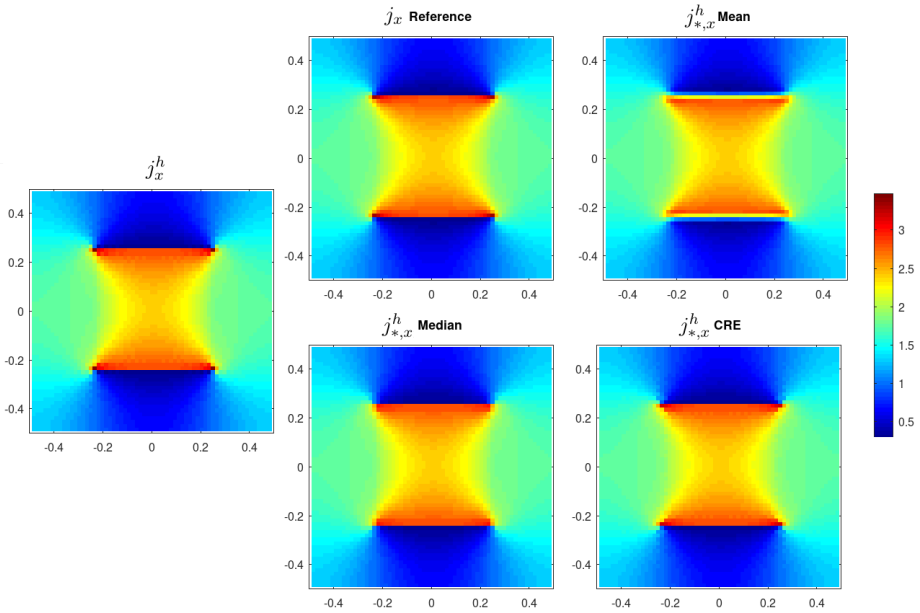


Figure 12: `[squ;rho=1E1;npix=2**6;FP]`: Discrete flux j_x^h, reference flux j_x and reconstructed fluxes $j_{*,x}^h$ with the different methods.

4.6 Performance of the global error estimate

In this section, we investigate the accuracy of the three different error estimators on 16 different settings. They are tested on both the square and random circles inclusions geometries, with both $2^6 \times 2^6$ and $2^8 \times 2^8$ pixels, with both the modified and original operators, and with both $\rho = 10$ and $\rho = 10^3$. The fixed-point algorithm was used but the specific choice of solver does not matter here as only the error at convergence is studied (a number $k = 40$ of iterations were used when $\rho = 10$, and $k = 500$ when $\rho = 10^3$). The results are displayed on Figure 15 for all combinations of the numerical parameter that lead to a true error $\delta^h \leqslant 1$. Therefore, in these figures we have discarded the points for which the true error is above 1, which are here obtained in some of the high-contrast cases. In such cases our estimations are not relevant.

The mean filter based error estimator gives mostly always the worst estimation. It must be noticed however that this estimator is the easiest to develop, and is marginally less CPU costly than the CRE one. When the original Green's operator is used (Figure 15a), the CRE method appears to lead to the most reliable estimator in the cases of low contrast (although not a guaranteed bound, see Section 4.4.1). However, in the cases of high contrast, this estimator does not perform much better than the mean filter. Finally, the median filter-based method appears to lead to the best estimator in the high contrasts cases, while being acceptable in the low contrast cases considered.

When the modified operator is used (Figure 15b), the median filter and CRE-based error estimators turn out to have quite equivalent accuracies, and perform reasonably well on the test cases considered. It must be noticed that, for the investigated configurations, these estimators tend to yield values lower than the exact one, which is often considered to be undesirable in error estimation. Yet, the theoretical bound of Section 4.4.1 has no reason to be satisfied when using the modified operator. In addition and as previously discussed, because of inexact integration, the schemes considered violate the identity (36), which conventionally yields a guaranteed upper bound with the CRE method. This explains that we can get estimated errors below the true ones even with this method. The use of exact numerical integration would allow to recover guaranteed bounds.

5 Convergence criterion: numerical results

In this section, we assess the relevance of a stopping criterion based on the comparison between the estimations of the iteration error and of the discretization one, as discussed in Section 3.3. Here the

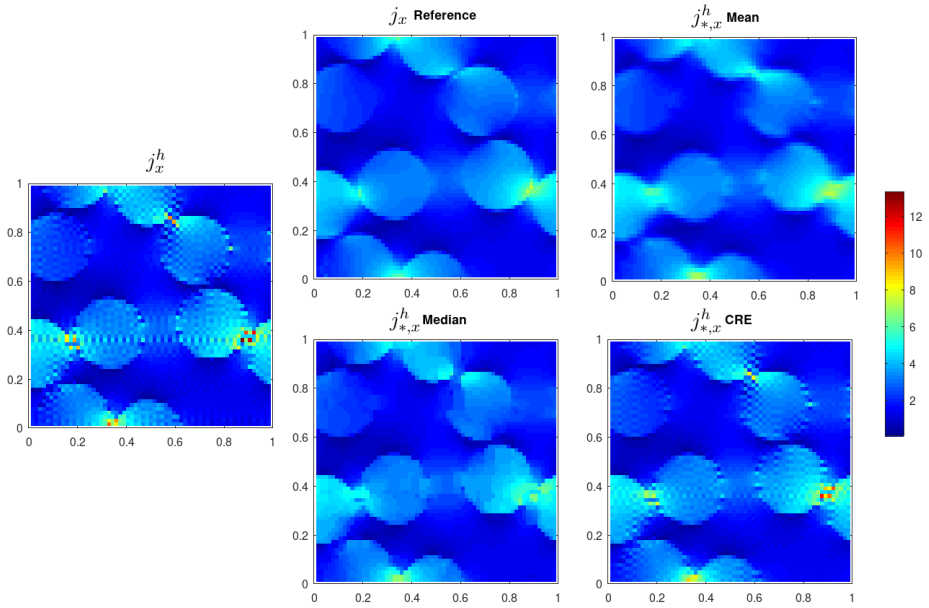


Figure 13: `[cir;rho=1E1;npix=2**6;FP]`: Discrete flux j_x^h, reference flux j_x and reconstructed fluxes $j_{*,x}^h$ with the different methods.

discretization is considered to be given and fixed.
Two convergence criteria, denoted by R_β and E_β^{method} are considered:

(i) R_β consists in stopping the algorithm at the first iteration for which $\Delta_k^h \leqslant 10^{-\beta}\ \Delta_0^h$ where Δ_0^h is the initial residual. This criterion corresponds to the one most employed in the literature.

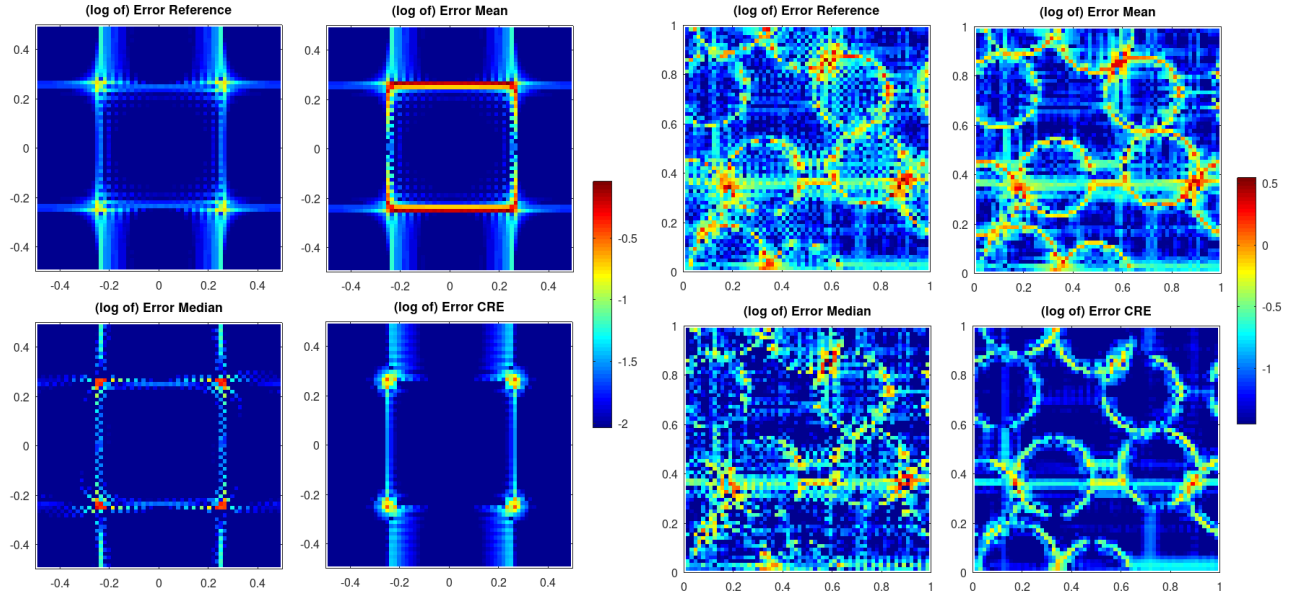
(ii) E_β^{method} consists in stopping the algorithm at the first iteration for which $\Delta_k^h \leqslant 10^{-\beta}\ \Delta^h$. In other words, it consists in waiting until the estimated iteration error has no more significant effect on the global error by being a threshold lower than the estimated discretization error. This is similar to the criterion employed in [10], which couples estimations of the discretization error and of the algebraic error associated with an inexact solution of the linear system in a finite volume method. Here, criterion E_β^{med} uses the median filter to estimate Δ^h, while E_β^{cre} uses the CRE method. Noticeably, we will here make use of the approaches described in Section 4 to compute some estimations Δ^h of the discretization error, not in the limit $\boldsymbol{e}^h = \lim_{k\to\infty} \boldsymbol{e}_k^h$ but for the specific values of k considered. In practice, it was noticed that at the very first iterations, the estimator Δ^h was much smaller than the residual Δ_k^h, which prevented the stopping criterion to be met accidentally due to the use of $\boldsymbol{e}_k^h$ instead of $\boldsymbol{e}^h$ in the computation of Δ^h.

Once the chosen stopping criterion is met then the computation is simply stopped. At this point, the numerical results could be further improved by performing a new computation on a refined grid and using the last computation as an initial guess, see e.g. [6]. This is however beyond the scope of this work.

5.1 2D conductivity test-cases

In the case of the square inclusion, Figure 16 plots the *true* error $\delta_{k,h}^{\text{eff}}$ on the effective conductivity and the associated number of iterations. Here, $\delta_{k,h}^{\text{eff}}$ is computed from the knowledge of the analytical effective conductivity [22]. The conductivity ratio is $\rho = 10^3$, and we use the variant of the method denoted by CG because of its faster convergence. Different discretizations are investigated.

On this test-case, the criteria E_2^{med} and E_2^{cre} yields similar numbers of iterations. However, as for a high contrast the CRE method over-estimated the discretization error (see Section 4.6), the corresponding stopping criterion requires less iterations in that case. The proposed criteria give errors that are relatively close to the standard criterion R_5 but they require different numbers of iterations.



(a) `[squ;rho=1E1;npix=2**6;FP]` (b) `[cir;rho=1E1;npix=2**6;FP]`

Figure 14: Reference and error estimators (logarithmic plot with scaling by the conductivity).

As a consequence, we introduce a new indicator, that we will refer to as *indicator of improvement*, in order to understand which stopping criterion is the most efficient. This indicator, denoted by $\mathcal{I}_k^h$, aims at measuring how much the error on the effective property could have been reduced if the iterative procedure had not been stopped. It is defined as the ratio between the iteration error and discretization error but uses the reference effective property and a numerical solution $\boldsymbol{e}_\infty^h$ obtained after a sufficiently large number of iterations as

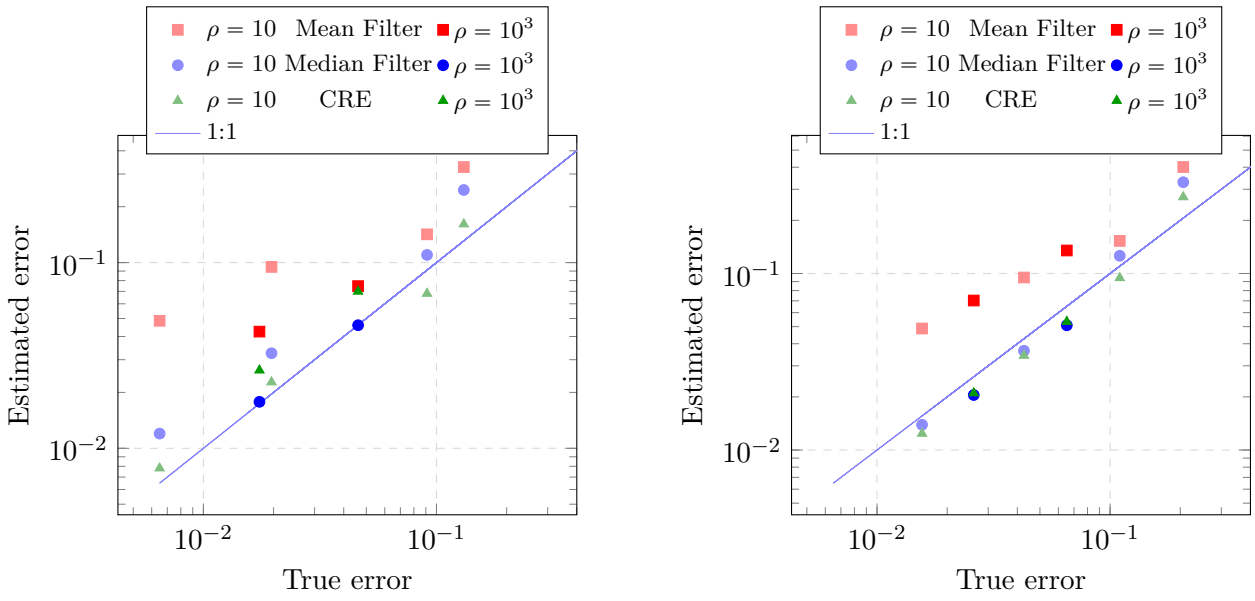
$$\mathcal{I}_k^h = \frac{|W_h(\boldsymbol{e}_k^h) - W_h(\boldsymbol{e}_\infty^h)|^{1/2}}{|W(\boldsymbol{e}) - W_h(\boldsymbol{e}_\infty^h)|^{1/2}}.$$

A *good* criterion would stop at an iteration k that ensures that the error on the effective property (ie. on the energy) cannot be decreased noticeably anymore, disregarding the discretization, geometry or conductivity ratio. This can be ensured when $\mathcal{I}_k^h = \alpha$, with $\alpha \ll 1$ being a parameter that tunes how small $\mathcal{I}_k^h$ is wanted to be, and that does not vary with the discretization, geometry nor conductivity ratio. Typically, we choose $\alpha = 0.1$

Remark 10. *Criterion E_β^{method} is based on the discretization error estimators developped in Section 4, that estimate the energetic error on the field $\boldsymbol{e}$, defined in Equation* (6)*, rather than the error on the effective property, or equivalently on the energy which is used to evaluate $\mathcal{I}_k^h$. According to* (23)*, if exact integration were used, both errors would be equal.*

On Figure 17a, we plot the evolution of the indicator of improvement $\mathcal{I}_k^h$ for the different stopping criteria considered, as functions of the discretization. One can notice that the criteria of type E_β^{method}, which are based on the proposed error estimators, tend to make $\mathcal{I}_k^h$ independent to the discretization, while the conventional residual-based criterion R_5 makes it increase with the discretization. This means that, for a coarse grid, the latter requires too many iterations, while for a finer grid, it does not allow to benefit from the full precision offered by the discretization.

In the case of random circular inclusions, the trends on the error and number of iterations are similar to the case of a square inclusion (see Figure 16) but they are not shown here. However, on Figure 17b, we display the evolution of the indicator of improvement $\mathcal{I}_k^h$ with the discretization for the former test-case. One notices once again that this quantity is stable when the stopping criteria E_β^{method} are used, while it increases with the number of degrees of freedom with the conventional criterion R_5. In addition, the comparison between Figures 17a and 17b shows that the levels of $\mathcal{I}_k^h$ have the same magnitude for both test-cases, and according to our computations on the test-cases



(a) `[squ/cir;rho=1E1/1E3;npix=2**6/2**8;FP]` (b) `[squ/cir;rho=1E1/1E3;npix=2**6/2**8;FPm]`

Figure 15: Estimated error Δ^h vs true error δ^h for different configurations and choice of numerical parameters.

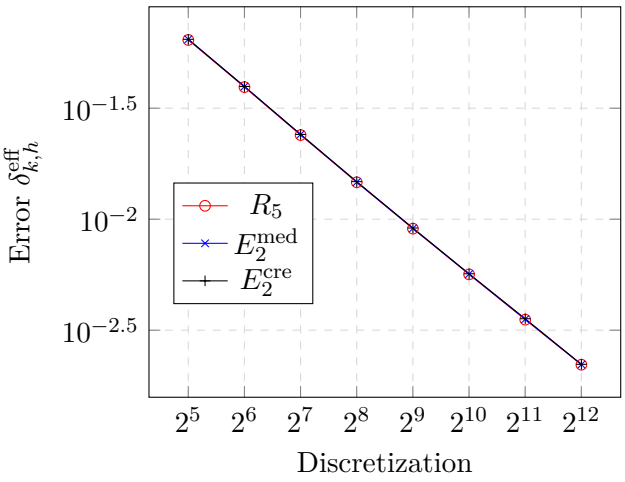
considered, this magnitude seems to be correlated with β. This implies that the latter parameter can be used to control the significance of the discretization in the total error on the effective property.

Figure 18 displays the evolution of the indicator $\mathcal{I}_k^h$ against the number of pixels and for the same two geometries but at a lower conductivity ratio $\rho = 10$. For the criteria E_β^{method}, the behavior of this indicator in the case of random circular inclusions is satisfactory, with all stopping criteria leading to rather stable values of $\mathcal{I}_k^h$, while the criterion R_5 leads to an unstable indicator, which increases with the number of pixels. The square inclusion test-case leads to slightly different behaviors: the residual-based stopping criterion R_5 makes $\mathcal{I}_k^h$ increase for more than two orders of magnitude when the number of pixels increases; the proposed criteria E_β^{method} also lead to an increase of $\mathcal{I}_k^h$ but it is however smaller. This shows that, even when the proposed stopping criteria do not exhibit an optimal behavior, they however lead to a less unstable $\mathcal{I}_k^h$.

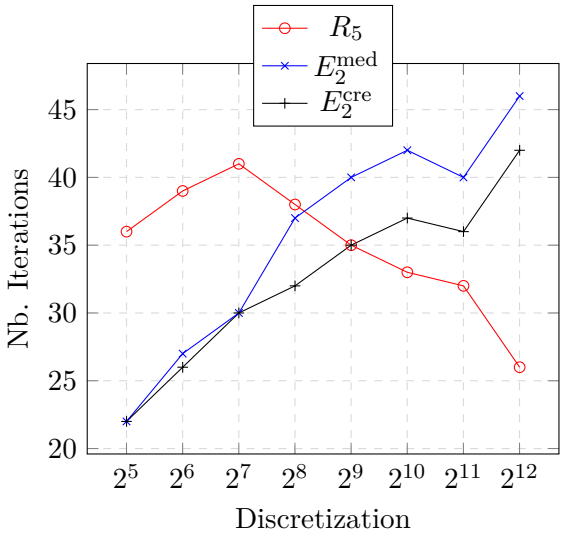
Remark 11. *The relatively sub-optimal behavior of the proposed stopping criteria on the test-case* `[squ;rho=1E1;npix=*;CG]` *of Figure 18a can be explained as follows: the discretization error estimators do not match the error on the effective property (see Remark 10) and the discrepancy seems to slightly increase with the discretization. In addition, on this test-case, all stopping criteria require only a few iterations to be met (typically 8 or 9), so that a variation of a few iterations leads to potentially significantly different values of* $\mathcal{I}_k^h$.

5.2 3D elasticity test-cases

The proposed procedure can be applied to the case of 3D linear elasticity. Adapting the proposed errors estimators is straightforward although more demanding both in terms of implementation and hardware. We consider a periodic microstructure of hard ellipsoids (Young modulus $E = 200\,\text{GPa}$) in a soft matrix ($E = 20\,\text{GPa}$ for $\rho = 10$ or $E = 200$ MPa for $\rho = 10^3$). The ellipsoids are allowed to interpenetrate. The Poisson coefficient is homogeneous and set to $\nu = 0.3$. The geometry of the inclusions is displayed on Figure 19a for an example discretization with $\left(2^6\right)^3 = 262\,144$ voxels. The number of voxels used in the computations will vary between $\left(2^4\right)^3 = 4\,096$ and $\left(2^9\right)^3 = 134\,217\,728$. For this computation, no Finite Element solution was computed, and the reference solution is obtained with an overkill FFT computation with $\left(2^{10}\right)^3 = 1\,073\,741\,824$ voxels. Note that the use of this overkill solution as a reference for the finest computation being evaluated is arguable. The stress σ_{xx}^h at



(a) Relative errors



(b) Nb. Iterations

Figure 16: Error and required number of iterations for different convergence criteria as functions of the discretization and for the different stopping criteria. `[squ;rho=1E3;npix=*;CG]`

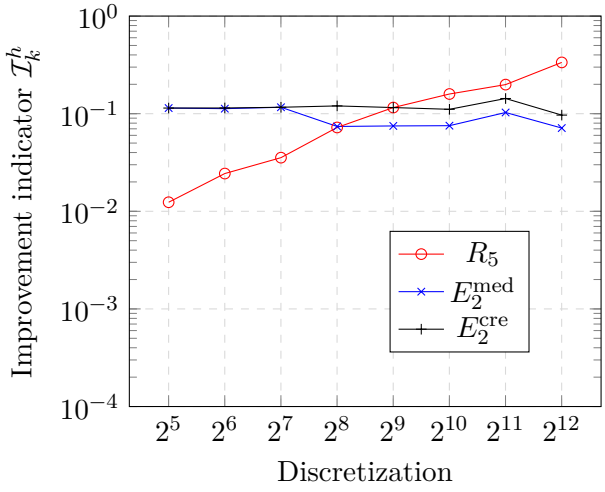
convergence (with $k = 40$ iterations) and its transformed versions with the median and CRE methods are displayed on Figures 19b, 19c and 19d.

Remark 12. *For the CRE estimator, a heterogeneous Poisson coefficient would require to invert several rigidity matrices* $\mathsf{K}_\nu^{h'}$ *at the step of Section 4.4.3 (one for each different value of* ν*), and to apply the operator* $\mathsf{G}_\nu \mathsf{K}_\nu^{h'+} \mathsf{M}^{h'}$ *independently to the voxels of each inclusion. However, this should have only a limited impact on the computational cost and implementation difficulty of the method provided there is only a small number of different homogeneous inclusions.*

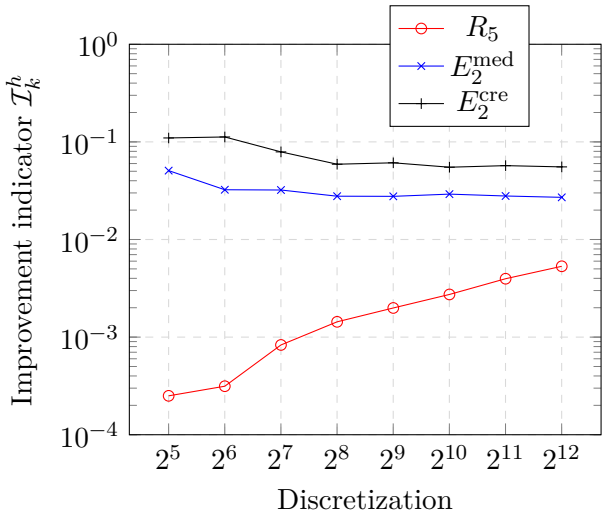
We display on Figure 20 the evolution of the improvement indicator $\mathcal{I}_k^h$ (ratio between the iteration and total errors on the energy) with the number of voxels for the different stopping criteria. The results obtained for these numerical experiments in 3D elasticity are similar to those of Section 5.1 in that both stopping criteria E_β^{med} and E_β^{cre} lead to a ratio between iteration error and total error that is more stable than R_β and thus are more reliable.

6 Conclusion

The present study addresses the issue of the computation of a global convergence criterion for iterative solvers in the FFT-based computational homogenization of periodic materials. The main idea is to separate and estimate the contributions to the error of the iterative scheme and of the discretization.



(a) `[squ;rho=1E3;npix=*;CG]`



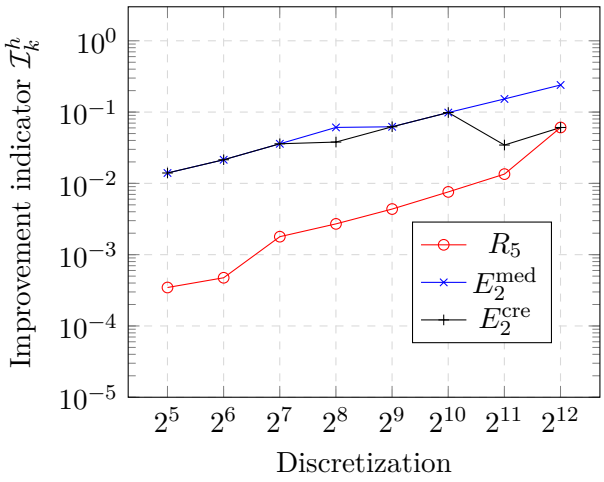
(b) `[cir;rho=1E3;npix=*;CG]`

Figure 17: Indicator of improvement $\mathcal{I}_k^h$ as a function of the discretization and for the different stopping criteria.

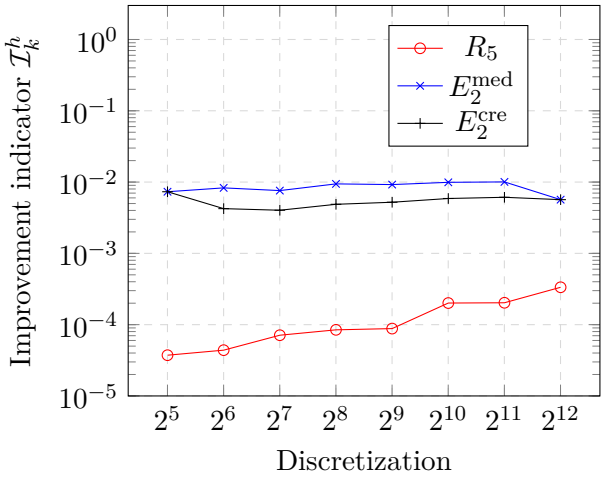
Some associated estimators have been proposed and their performances have been assessed on two prototypical 2D conductivity test-cases. From this study, we can conclude that:

- The residual of the discrete problem is not sufficient for a reliable estimation of the total error on the effective property.
- Provided that the conductivity ratio is not too high (up to 10^3 for the configurations considered), this residual is nonetheless a suitable estimator of the iteration error.
- Two methods have proven to be useful for estimating the discretization error. The first one is the *median filter*, which works best with rather high conductivity ratios ($\sim 10^3$), while the second one, the *constitutive relation error*, appears in our computations to be the best choice for moderate conductivity ratios (~ 10).
- A parameter-controlled stopping criterion has been introduced and tested: it consists in interrupting the iterative scheme when the estimated iteration error is *much* smaller than the estimated discretization error, the sense of much being given by the featured parameter.

Finally, we based the evaluation of the stopping criteria on an *improvement indicator* that measures how much the error on the effective property could have been reduced if the iterative scheme had not been stopped. In the test cases considered, the proposed stopping criteria appear to be superior in several ways to the conventional criterion solely based on the residual:



(a) `[squ;rho=1E1;npix=*;CG]`



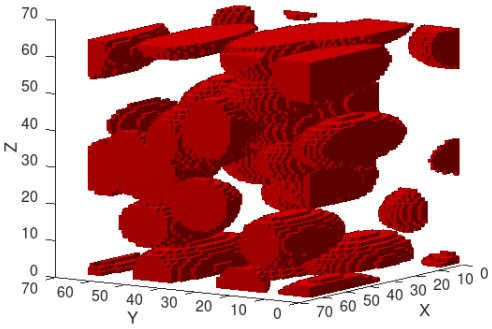
(b) `[cir;rho=1E1;npix=*;CG]`

Figure 18: Indicator of improvement $\mathcal{I}_k^h$ as a function of the discretization for the different stopping criteria.

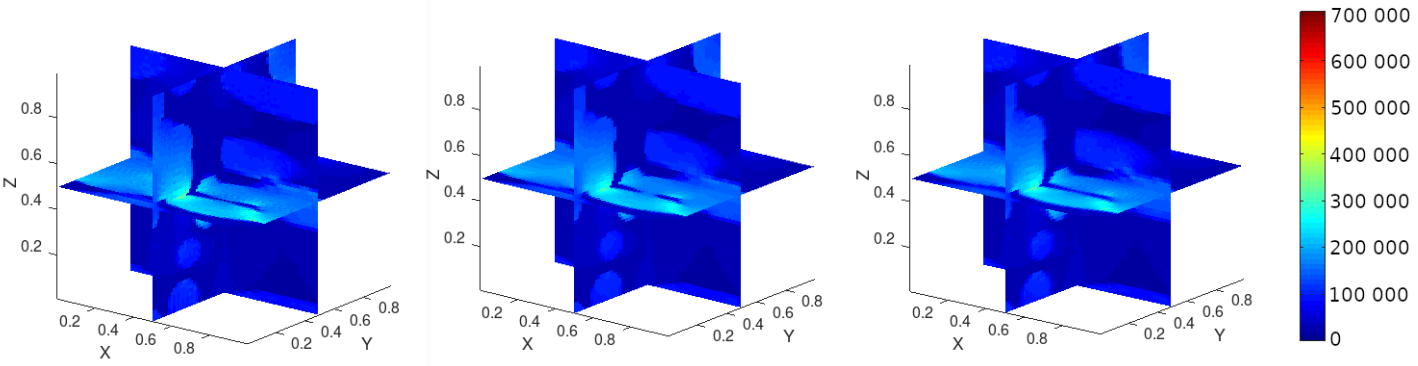
- The residual-based criterion tends to require less iterations when the discretization becomes finer. As a consequence, the improvement indicator increases with the number of pixels. On the contrary, our stopping criteria tends to require more iterations when the discretization is finer, which leads to a more stable improvement indicator so that full advantage can be taken of the discretization.
- The proposed stopping criteria involve a parameter β, which value seems to control the improvement indicator associated with the computations. If confirmed, this would mean that, when choosing the value of this parameter, one could adjust the share of the iteration error in the total error on the effective property.

At this stage, a number of perspectives emerge for this work:

- The estimation of the discretization error has a non-negligible cost in the computation (despite the fact that the regular grid allows for significant speedups). For this reason, it will be essential to compute it only when necessary, i.e. possibly not at every iterations, which calls for an adapted strategy.
- The *constitutive relation method* appears to be well suited in the cases of small contrast. It could be interesting to improve it in cases of strong contrast, in particular by using a more accurate method of flux reconstruction. One could for example use the approach proposed in [23].
- There are three valuable theoretical results that do not hold due to the fact that the discretization methods we work with are inexactly integrated Galerkin methods. These results are (i) the



(a) Ellipsoidal inclusions with $\left(2^6\right)^3$ voxels



(b) σ^h_{xx} (c) σ^h_{xx} median (d) σ^h_{xx} CRE

Figure 19: Geometry and stress field σ^h_{xx} after $k = 40$ iterations.

equality between the energy error on the field and the error on the effective property, (ii) the bound between the computed and the exact effective properties and (iii) the bound between the exact error and the estimator provided by the CRE method. As a consequence, it would be of great interest to extend the methods proposed in the present paper to exactly integrated Galerkin schemes [30, 5] or FEM-based discretizations coupled with FFT solvers [25, 16].

A Fourier transforms

Consider the unit cell Ω filling the space $\mathbb{R}^d$ by translation along d vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_d$. The lattice $\mathcal{R}$ generated by these vectors is defined as

$$\mathcal{R} = \left\{ \boldsymbol{Y} \mid \boldsymbol{Y} = \sum_{j=1}^{d} n_j \boldsymbol{Y}_j,\ n_j \in \mathbb{Z} \right\}.$$

Let $\mathcal{R}^*$ denote the reciprocal lattice of $\mathcal{R}$ generated by the vectors

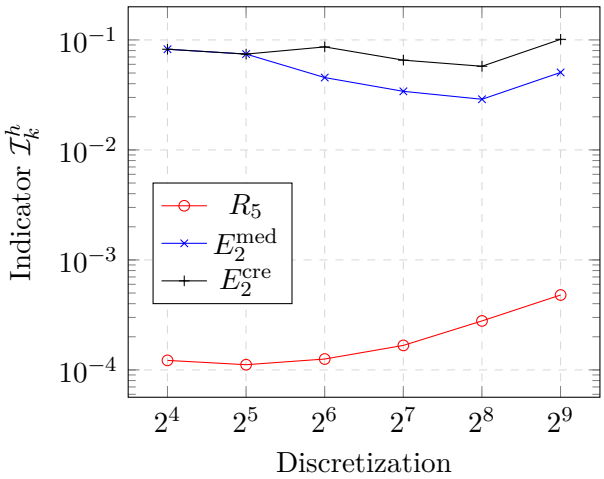
$$\boldsymbol{Y}_i^* = \frac{2\pi}{|\Omega|} \boldsymbol{Y}_j \wedge \boldsymbol{Y}_k,$$

where (i, j, k) is a direct circular permutation. The Fourier transform f of f is defined on $\mathcal{R}^*$ as:

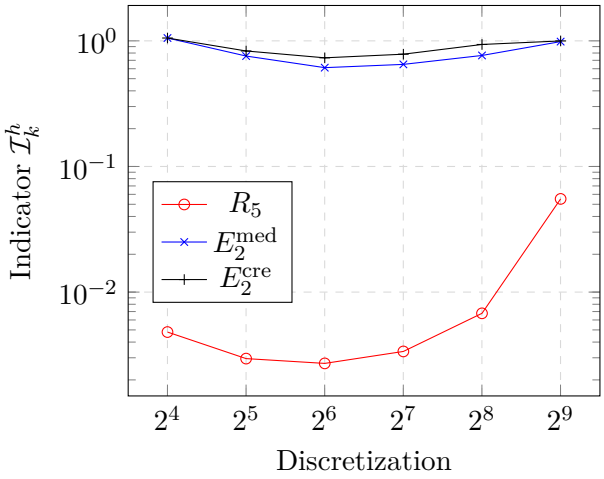
$$\hat{f}(\boldsymbol{\xi}) = \mathscr{F}[f](\boldsymbol{\xi}) = \frac{1}{|\Omega|} \int_\Omega f(\boldsymbol{x}) e^{-\mathrm{i}\boldsymbol{\xi}\cdot\boldsymbol{x}} \, \mathrm{d}\boldsymbol{x}, \quad \text{where} \quad \mathrm{i} = \sqrt{-1}.$$

The periodic function f in $L^2_{\mathrm{per}}(\Omega)$ can be reconstructed from its Fourier transform by

$$f(\boldsymbol{x}) = \mathscr{F}^{-1}[\hat{f}](\boldsymbol{x}) = \sum_{\boldsymbol{\xi} \in \mathcal{R}^*} \hat{f}(\boldsymbol{\xi}) e^{\mathrm{i}\boldsymbol{\xi}\cdot\boldsymbol{x}}.$$



(a) $\rho = 10$



(b) $\rho = 10^3$

Figure 20: Indicator of improvement $\mathcal{I}_k^h$ as a function of the discretization and for the different stopping criteria considering the 3D elasticity test-cases.

References

[1] J. C. A. Barata and M. S. Hussein. The moore–penrose pseudoinverse: A tutorial review of the theory. *Brazilian Journal of Physics*, 42(1):146–165, 2012.

[2] C. Bellis, H. Moulinec, and P. Suquet. Eigendecomposition-based convergence analysis of the Neumann series for laminated composites and discretization error estimation. *International Journal for Numerical Methods in Engineering*, 121(2):201–232, 2020.

[3] C. Bellis and P. Suquet. Geometric variational principles for computational homogenization. *Journal of Elasticity*, 137(2):119–149, 2019.

[4] S. Brisard and L. Chamoin. Constitutive relation error for FFT-based methods. In *ECCOMAS Congress 2016 Proceedings*. VII European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS Congress2016), Greece, 2016.

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

[6] D. J. Eyre and G. W. Milton. A fast numerical scheme for computing the response of composites using grid refinement. *The European Physical Journal Applied Physics*, 6(1):41–47, 1999.

[7] B. Fornberg. The pseudospectral method: Comparisons with finite differences for the elastic wave equation. *Geophysics*, 52(4):483–501, 1987.

[8] T. Huang, G. Yang, and G. Tang. A fast two-dimensional median filtering algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(1):13–18, 1979.

[9] T. J. R. Hughes. *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*. Prentice-Hall, Inc., 1987.

[10] P. Jiránek, Z. Strakoš, and M. Vohralík. A posteriori error estimates including algebraic error and stopping criteria for iterative solvers. *SIAM Journal on Scientific Computing*, 32(3):1567–1590, 2010.

[11] M. Kabel, T. Böhlke, and M. Schneider. Efficient fixed point and Newton–Krylov solvers for FFT-based homogenization of elasticity at large deformations. *Computational Mechanics*, 54(6):1497–1514, 2014.

[12] J. Kochmann, K. Manjunatha, C. Gierden, S. Wulfinghoff, B. Svendsen, and S. Reese. A simple and flexible model order reduction method for fft-based homogenization problems using a sparse sampling technique. *Computer Methods in Applied Mechanics and Engineering*, 347:622–638, 2019.

[13] P. Ladeveze and D. Leguillon. Error estimate procedure in the finite element method and applications. *SIAM Journal on Numerical Analysis*, 20(3):485–509, 1983.

[14] P. Ladevèze and J.-P. Pelle. *Mastering Calculations in Linear and Nonlinear Mechanics*. Springer, New York, 2005.

[15] P. Ladevèze and P. Rougeot. New advances on a posteriori error on constitutive relation in f.e. analysis. *Computer Methods in Applied Mechanics and Engineering*, 150(1-4):239–249, 1997.

[16] M. Leuschner and F. Fritzen. Fourier-accelerated nodal solvers (fans) for homogenization problems. *Comput. Mech.*, 62, 2018.

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

[18] G. W. Milton. *The Theory of Composites*. Cambridge university press, 2002.

[19] L. Morin, R. Brenner, K. Derrien, and K. Dorhmi. Periodic smoothing splines for fft-based solvers. *Computer Methods in Applied Mechanics and Engineering*, 373:113549, 2021.

[20] H. Moulinec and P. Suquet. A numerical method for computing the overall response of nonlinear composites with complex microstructure. *Computer methods in applied mechanics and engineering*, 157(1-2):69–94, 1998.

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

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

[23] A. Parret-Fréaud, V. Rey, P. Gosselet, and C. Rey. Improved recovery of admissible stress in domain decomposition methods—application to heterogeneous structures and new error bounds for feti-dp. *International Journal for Numerical Methods in Engineering*, 111(1):69–87, 2017.

[24] V. Rey, C. Rey, and P. Gosselet. A strict error bound with separated contributions of the discretization and of the iterative solver in non-overlapping domain decomposition methods. *Computer Methods in Applied Mechanics and Engineering*, 270:293–303, 2014.

[25] M. Schneider, D. Merkert, and M. Kabel. Fft-based homogenization for microstructures discretized by linear hexahedral elements. *International Journal for Numerical Methods in Engineering*, 109(10), 2017.

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

[27] R. M. Sydenstricker, A. L. Coutinho, M. A. Martins, L. Landau, and J. L. Alves. A posteriori error estimate for stress analysis of homogeneous and heterogeneous materials: An engineering approach. *Finite elements in analysis and design*, 42(3):171–188, 2005.

[28] J. Vondřejc. *FFT-based method for homogenization of periodic media: Theory and applications*. PhD thesis, Czech Technical University, 2013.

[29] J. Vondřejc. Improved guaranteed computable bounds on homogenized properties of periodic media by the Fourier–Galerkin method with exact integration. *International Journal for Numerical Methods in Engineering*, 107(13):1106–1135, 2016.

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

[31] J. Vondřejc, J. Zeman, and I. Marek. Guaranteed upper–lower bounds on homogenized properties by FFT-based Galerkin method. *Computer Methods in Applied Mechanics and Engineering*, 297:258–291, 2015.

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

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

[34] F. Willot, B. Abdallah, and Y.-P. Pellegrini. Fourier-based schemes with modified Green operator for computing the electrical response of heterogeneous media with accurate local fields. *International Journal for Numerical Methods in Engineering*, 98(7):518–533, 2014.

[35] L. Yin, R. Yang, M. Gabbouj, and Y. Neuvo. Weighted median filters: a tutorial. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 43(3):157–192, 1996.

[36] J. Zeman, J. Vondřejc, J. Novák, and I. Marek. Accelerating a FFT-based solver for numerical homogenization of periodic media by conjugate gradients. *Journal of Computational Physics*, 229(21):8065–8071, 2010.

[37] O. C. Zienkiewicz and J. Z. Zhu. A simple error estimator and adaptive procedure for practical engineering analysis. *International journal for numerical methods in engineering*, 24(2):337–357, 1987.