CZECH TECHNICAL UNIVERSITY IN PRAGUE

Faculty of Civil Engineering
Department of Mechanics

FFT-based method for homogenization of periodic media: Theory and applications

DOCTORAL THESIS

Mgr. Ing. Jaroslav Vondřejc

Ph.D. Programme: Civil Engineering
Branch of study: Mathematics in Civil Engineering

Supervisor: doc. Ing. Jan Zeman, Ph.D.

Prague, 2013

Czech Technical University in Prague
Faculty of Civil Engineering
Thákurova 7, 166 29 Prague 6

DECLARATION BY CANDIDATE

Author: Jaroslav Vondřejc
Title: FFT-based method for homogenization of periodic media: Theory and applications

I hereby declare that this thesis is my own work and effort and that it has not been submitted anywhere for any award. Where other sources of information have been used, they have been acknowledged.

In Prague, 2nd January 2013 Signature:

Abstract

This dissertation is devoted to an FFT-based homogenization scheme, a numerical method for the evaluation of the effective (homogenized) matrix of periodic linear heterogeneous materials. A problem is explained and demonstrated on a scalar problem modeling electric conduction, heat conduction, or diffusion.

Originally, FFT-based homogenization, that was proposed by Moulinec and Suquet in [32], is a numerical algorithm derived from Lippmann-Schwinger equation. Its equivalence to a corresponding weak formulation is shown; it eliminates a reference homogeneous material, a parameter of Lippmann-Schwinger equation. Next, Galerkin approximation with numerical integration is introduced to produce Moulinec-Suquet algorithm; trigonometric polynomials are taken as the trial space [47]. Convergence of approximate solutions to the solution of weak formulation is provided using a standard finite element approach together with approximation properties of trigonometric polynomials stated in [43].

Then, the solution of assembled non-symmetric linear system by Conjugate gradients, proposed by Zeman et al. in [57], is clarified.

Next, we study arbitrary accurate guaranteed bounds of homogenized matrix introduced by Dvořák in [12, 13] for a scalar problem and later independently by Wieçkowski in [55] for linear elasticity. This approach is also applicable for FFT-based homogenization. A general technique is proposed to allow for efficient calculation by FFT algorithm and to maintain the upper-lower bound structure. Dual formulation is employed to obtain lower bounds — for odd number of discretization points, the solution of dual formulation can be avoided. A general number of discretization points leads to a more complicated theory in both discretization and numerical treatment.

Finally, applications of FFT-based homogenization to real-world problems are demonstrated. The method is used to calculate homogenized matrix for cement paste, gypsum and aluminum alloy with local data obtained from nanoindentation. Next, it is employed as a part of two-step homogenization for a highly porous aluminium foam.

Keywords: homogenization, Fourier transform, FFT, discretization, finite element method, convergence, guaranteed bounds

Abstrakt

Tato práce je věnována homogenizační metodě založené na FFT, metodě počítající efektivní (homogenizovanou) matici periodického heterogenního materiálu. Problematika je vysvětlena a demonstrována na skalárním problému modelujícím elektrickou vodivost, tepelnou vodivost nebo difúzi.

Původně byla FFT homogenizace založena na řešení Lippmannovy-Schwingerovy rovnice a z něho navrženého numerického algoritmu podle Moulineca a Suqueta v [32].

Ukazujeme, že Lippmannova-Schwingerova rovnice je ekvivalentní příslušné slabé formulaci; to umožňuje eliminovat referenční homogenní materiálovou hodnotu, parametr rovnice. Dále jsme navrhli Galerkinovskou aproximaci s numerickou integrací tak, aby produkovala Moulinecův-Suquetův algoritmus; trigonometrické polynomy jsou použity jako konečně dimenzionální prostor [47]. Dokazujeme konvergenci přibližných řešení k řešení slabé formulace pomocí klasického přístupu z konečných prvků s využitím odhadů podle [43].

Dále je vysvětleno řešení vzniklé nesymetrické soustavy lineárních rovnic pomocí sdružených gradientů, které bylo navrženo Zemanem et al. v [57].

Věnujeme se libovolně přesným zaručeným mezím efektivní matice, jak byly navrženy Dvořákem v [12, 13] pro skalární problém a nezávisle Wieçkowskim v [55] pro lineární elasticitu. Tato metoda je využitelná i pro homogenizaci založenou na FFT. Navrhujeme obecnou metodu pro výpočet mezí tak, aby bylo možno využít efektivního algoritmu FFT a aby byla zachována struktura dolních a horních mezí. Duální formulace problému je využita pro dolní mez — ukazujeme, že pro lichý počet diskretizačních bodů se lze vyhnout řešení duální úlohy.

Nakonec je ukázána aplikace FFT homogenizace na lineární elasticitu. Metoda je využita k počítání efektivní matice pro cementovou pastu, sádru a hliníkovou slitinu s lokálními daty získanými z nanoindentace. Metoda je dále využita v dvoustupňové homogenizaci pro výpočet efektivních materiálových parametrů vysoce porézní hliníkové pěny.

Klíčová slova: homogenizace, Fourierova transformace, FFT, diskretizace, metoda konečných prvků, konvergence, zaručené meze

Acknowledgments

First of all, I would like to express my deepest gratitude and thanks to my supervisor doc. Ing. Jan Zeman, Ph.D for his ingenuous support and endless patience during the whole study and for the freedom that I have obtained in research directions. Special gratitude is to prof. RNDr. Ivo Marek, DrSc. for his support and critical comments in the fields of mathematical and numerical analysis.

I thank to co-authors of papers attached to the work, namely to already mentioned supervisors, Ing. Vlastimil Králík, doc. Ing. Jiří Němeček, Ph.D., and Ing. Jan Novák, Ph.D.

I also thank to prof. Ing. Milan Jirásek, DrSc. besides other things for his course Modeling of Localized Inelastic Deformation that I have held at the beginning of my Ph.D. studies. I also appreciate the courses, that I have taken at the Faculty of Mathematics and Physics at Charles University and I thank to their organizers: prof. RNDr. Vít Dolejší, Ph.D., DSc.; prof. RNDr. Miloslav Feistauer, DrSc., dr.h.c.; doc. RNDr. Jiří Felcman, CSc.; prof. RNDr. Jaroslav Haslinger, DrSc.; doc. RNDr. Petr Holický, CSc.; doc. Dr. Mgr. Petr Knobloch; doc. RNDr. Martin Kružík, Ph.D.; doc. Mgr. Milan Pokorný, Ph.D.; prof. Ing. Tomáš Roubíček, DrSc.; and doc. RNDr. Jan Zítko, CSc. The courses have provided me the opportunity to obtain the fundamental knowledge of mathematics that I have fully utilized during my works.

Most importantly, I would like to thank my wife Katka and daughter Justýnka for their patience and support during my studies and writing the works.

Last but not least, the financial support of this work provided by the Czech Science Foundation through project No. GAČR P105/12/0331, GAČR 103/09/1748, GAČR 103/09/P490 and by the Grant Agency of the Czech Technical University in Prague through project No. SGS 12/027/OHK1/1T/11, SGS10/124/OHK1/2T/11 is gratefully acknowledged.

"The mathematical facts worthy of being studied are those which, by their analogy with other facts, are capable of leading us to the knowledge of a physical law." (Henri Poincare)

"Empirical evidence can never establish mathematical existence — nor can the mathematician's demand for existence be dismissed by the physicist as useless rigor. Only a mathematical existence proof can ensure that the mathematical description of a physical phenomenon is meaningful." (Richard Courant)

Content

Part I
Integrating text

1 Motivation

In the recent decades, the computational mechanics — together with its tremendous development, increase in computer performance and hardware availability — has delivered extensive societal benefits. Simulations, replaced with computer models, have dramatically reduced the cost of the engineering design process, reduced design cycle times, and have made it possible to address scientific and engineering questions beyond the capabilities of experiments, [39, 2].

Nevertheless, the extensive engineering problems and their modeling, treated all in one with its complexity, promptly reach the limit of accessible computer capacity, particularly for heterogeneous materials with fully resolved microstructure. For example, concrete — the basic construction material in civil engineering structures — looks and behaves as a homogeneous substance at a macroscale, the scale of design object. However, a closer investigation into microscale discovers aggregates, sand, and binder; under another closer investigation, they are still heterogeneous. Concurrently, the heterogeneities are crucial, for example, in crack propagation problems; thus the design difficulty lies in a detail that is required. This justifies the development of new effective and reliable computational models, methods, and algorithms dealing with heterogeneities.

This work is dedicated to numerical homogenization of linear periodic materials or to the problem of finding effective (homogenized) material properties — electric conductivity, heat conductivity, or stiffness — by taking into account the heterogeneities at microscale. It consists of the calculation of microscale fields satisfying linear elliptic partial differential equation. Contrary to well-established h, p, hp-versions of Finite Element Method (FEM) based on weak formulation, see e.g. [1, 12, 13, 55, 58], they can be found by a method based on the Fourier Transform — Fast Fourier Transform homogenization (FFTH) — introduced in the form of numerical algorithm in [32]. The latter method is based on the solution of the Lippmann-Schwinger type of integral equation incorporating the Green function of an auxiliary homogeneous problem with a reference material property, the parameter of the method.

The method has become, next to analytical homogenizations [31, 8, 16] and already mentioned FEM, broadly used by engineers — especially for its simplicity in implementation, efficiency resulting from the application of the FFT algorithm, and direct usage of material geometry defined as pixel or voxel images. Although it is a rather special topic of interest, it has a wide area of applications; among others, recent relevant examples are listed: evaluation of effective properties for real structures [45, 54, 48] with data from nanoindentation [37, 38], microstructure modeling with visco-elastic [17, 53] and visco-plastic [21, 23, 44] material, conductivity with imperfect interfaces [20], micromechanical behavior of polycrystals [22, 21, 44], permeability of porous medium [30], non-local fracture (damage) models [24, 25], and microstructure reconstruction with Wang tilings [36].

In addition, for the last two decades, FFTH method has gone through extensive

investigation. The convergence of numerical scheme, based on the Neumann expansion, has been proposed in [33]. The accelerated schemes have emerged in [15, 49]. The algorithm has been extended for voided materials by augmented Lagrangians in [27, 28] and by application of the dual formulation in [29]; some of the methods were compared in [34]. Recently, the discretization of Lippmann-Schwinger equation, based on piecewise constant basis functions, with convergence results was suggested in [6, 7]. Another approach is based on approximation of the Green function with a FEM solution in [56].

Nevertheless, no rigorous theory — providing the resolvability of Lippmann-Schwinger equation for various parameters, comparing the Lippmann-Schwinger equation to variational formulation, establishing discretization and convergence of approximate solutions to the continuous one — has been published. It is the role of this work allowing the additional research within the standard mathematical instruments and contributing to reliability and confidence in engineering design, especially with guaranteed bounds of homogenized material properties.

Note that this dissertation is based on the compilation of six papers [57, 50, 37, 38, 51, 52] that are attached as Parts II–VII in the chronological order, see also Section 7 List of thesis paper. Paper 1 has been published in ISI journal, then Paper 2 was published as a peer-reviewed conference paper. Paper 3 is accepted for publication, while Paper 4 is in peer review stage. Finally, Papers 5 and 6, that are fundamental for the thesis, are to be sent for publication. Since each paper needs to be self-contained, the dissertation may contain repetitions.

2 State of the art

This section briefly overviews the state of the art of numerical homogenization of linear elliptic partial differential equation, namely the evaluation of homogenized matrix by FFT-based homogenization method. A scalar problem of electric conductivity, equally to diffusion or heat transfer, is chosen as a model case.

The variational formulation of homogenization problem is described in Section 2.1, followed by an alternative formulation with Lippmann-Schwinger equation including Moulinec-Suquet numerical solution [32] in Section 2.2 and the theory for guaranteed bounds [12, 13] in Section 2.3 are summarized.

For the later use in this dissertation, the following notation is introduced. Note that some articles can differ in some details, nevertheless, the most recent two [51, 52] should be consistent.

The letter d denotes the dimension of the problem, assuming $d = 2, 3$; the Greek letters α, β are reserved to indices relating dimension, thus ranging $1, \dots, d$ (the range is for simplicity often omitted).

The sets $\mathbb{C}^d$ and $\mathbb{R}^d$ are spaces of complex and real vectors with canonical basis $\{\boldsymbol{\epsilon}_\alpha\}$ and are equipped with the Lebesgue measure $\mathrm{d}\boldsymbol{x}$. We denote by $|\Omega|_d$ the d-dimensional Lebesgue measure of a measurable set $\Omega \subset \mathbb{R}^d$. The norm $\|\cdot\|_2$ on $\mathbb{C}^d$ is induced by scalar product $\big(\boldsymbol{u}, \boldsymbol{v}\big)_{\mathbb{C}^d} = \sum_\alpha u_\alpha \overline{v_\alpha}$ for $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^d$.

The set $\mathbb{R}^{d\times d}_{\mathrm{spd}}$ denotes the space of symmetric positive definite matrices of size $d \times d$ with norm $\|\boldsymbol{C}\|_2 = \max_{\boldsymbol{x}\in\mathbb{R}^d, \|\boldsymbol{x}\|=1} \|\boldsymbol{C}\boldsymbol{x}\|_2$ that equals to a largest eigenvalue.

A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is $\boldsymbol{Y}$-periodic (with period $\boldsymbol{Y} \in \mathbb{R}^d$) if $f(\boldsymbol{x} + \boldsymbol{Y} \odot \boldsymbol{k}) = f(\boldsymbol{x})$ for arbitrary $\boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{k} \in \mathbb{Z}^d$, where operator $\odot$ denotes element-wise

multiplication. The $\boldsymbol{Y}$-periodic functions are sufficient to define only on a periodic unit cell (PUC), set to $\mathcal{Y} := (-Y_\alpha, Y_\alpha)_{\alpha=1}^d \subset \mathbb{R}^d$. Two integrable functions which are almost everywhere equal are identified. The mean value of function $\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ over periodic unit cell $\mathcal{Y}$ is denoted as $\langle \boldsymbol{v} \rangle := \frac{1}{|\mathcal{Y}|_d} \int_{\mathcal{Y}} \boldsymbol{v}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \in \mathbb{R}^d$.

We define space $C_{\mathrm{per}}(\mathcal{Y};\mathbb{X})$ of continuous $\boldsymbol{Y}$-periodic functions $\mathbb{R}^d \mapsto \mathbb{X}$, where $\mathbb{X}$ is some finite dimensional vector space, e.g. $\mathbb{C}$, $\mathbb{R}$, $\mathbb{C}^d$, or $\mathbb{R}^d$. Vector valued functions, for $\mathbb{X} = \mathbb{C}^d$ or $\mathbb{X} = \mathbb{R}^d$, are denoted with small bold letters, e.g. $\boldsymbol{v}$ with components v_α.

The spaces $L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{X})$ or $L^\infty_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^{d\times d}_{\mathrm{spd}})$ are composed of functions $\boldsymbol{v} : \mathbb{R}^d \mapsto \mathbb{X}$ or $\boldsymbol{A} : \mathbb{R}^d \mapsto \mathbb{R}^{d\times d}_{\mathrm{spd}}$ having $\boldsymbol{Y}$-periodic, measurable components v_α or $A_{\alpha\beta}$ and having finite norm, i.e. $\|\boldsymbol{v}\|_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{X})} < \infty$ or $\|\boldsymbol{A}\|_{L^\infty_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^{d\times d}_{\mathrm{spd}})} < \infty$; the first norm is generated by scalar product $\big(\boldsymbol{u},\boldsymbol{v}\big)_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{X})} = \frac{1}{|\mathcal{Y}|_d} \int_{\mathcal{Y}} \big(\boldsymbol{u}(\boldsymbol{x}),\boldsymbol{v}(\boldsymbol{x})\big)_{\mathbb{X}} \,\mathrm{d}\boldsymbol{x}$ while the second norm is defined as $\|\boldsymbol{A}\|_{L^\infty_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^{d\times d}_{\mathrm{spd}})} = \operatorname{esssup}_{\boldsymbol{x}\in\mathcal{Y}} \|\boldsymbol{A}(\boldsymbol{x})\|_2$. If there is no ambiguity, both the norms and the scalar products are denoted with subscript L^2_{per} or L^∞_{per} rather than $L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{X})$ or $L^\infty_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^{d\times d}_{\mathrm{spd}})$.

Next, we introduce the Helmholtz decomposition $L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) = \mathscr{U} \oplus^\perp \mathscr{E} \oplus^\perp \mathscr{J}$ to the spaces of constant, curl-free with zero mean, and divergence free with zero mean fields

$$\mathscr{U} = \{\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) : \boldsymbol{v}(\boldsymbol{x}) = \text{const.}\}, \tag{2.1a}$$

$$\mathscr{E} = \{\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) : \nabla \times \boldsymbol{v} = \boldsymbol{0}, \langle \boldsymbol{v} \rangle = \boldsymbol{0}\}, \tag{2.1b}$$

$$\mathscr{J} = \{\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) : \nabla \cdot \boldsymbol{v} = 0, \langle \boldsymbol{v} \rangle = \boldsymbol{0}\}, \tag{2.1c}$$

where differential operator $\nabla = (\frac{\partial}{\partial x_\alpha})_{\alpha=1}^d$ is meant in the distributional sense. For dimension $d \neq 3$, the curl-free condition in (2.1b) means $(\nabla \times \boldsymbol{v})_{\alpha\beta} := \frac{\partial v_\alpha}{\partial x_\beta} - \frac{\partial v_\beta}{\partial x_\alpha} = 0$. Since space $\mathscr{U}$ consists of constant functions, we identify the space $\mathscr{U}$ with $\mathbb{R}^d$; this validates the operations such as $\boldsymbol{E} + \boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ for $\boldsymbol{E} \in \mathbb{R}^d$, $\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ and $\boldsymbol{C}\boldsymbol{J} \in \mathbb{R}^d$ for $\boldsymbol{C} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{J} \in \mathscr{U}$.

2.1 Variational formulation

This section begins with a homogenization problem defining a homogenized or effective matrix for a scalar linear elliptic problem, particularly the problem of electric conductivity.

Here and in the sequel, $\boldsymbol{A} \in L^\infty_{\mathrm{per}}(\mathcal{Y},\mathbb{R}^{d\times d}_{\mathrm{spd}})$ denotes symmetric[1] and uniformly elliptic[2] material coefficients of electric conductivity, $\boldsymbol{e} \in \mathscr{E}$ and $\boldsymbol{\jmath} \in \mathscr{J}$ perturbation of electric field and electric current, and $\boldsymbol{E}, \boldsymbol{J} \in \mathscr{U}$ are their macroscopic counterparts.

Definition 2.1 (Homogenization problem)**.** *The primal and dual homogenization problem states: find homogenized matrix $\boldsymbol{A}_{\mathrm{eff}} \in \mathbb{R}^{d\times d}$ satisfying for arbitrary fixed*

[1] For almost all $\boldsymbol{x} \in \mathcal{Y}$, equality $\boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{A}(\boldsymbol{x})^T$ holds.

[2] There exists positive constant $c_A > 0$ such that for almost all $\boldsymbol{x} \in \mathcal{Y}$ and all nonzero $\boldsymbol{u} \in \mathbb{R}^d$, inequality $c_A \|u\|_2^2 \leq \big(\boldsymbol{A}(\boldsymbol{x})\boldsymbol{u},\boldsymbol{u}\big)_{\mathbb{R}^d}$ holds.

macroscopic quantities $\boldsymbol{E} \in \mathbb{R}^d$

$$\big(\boldsymbol{A}_{\text{eff}}\boldsymbol{E}, \boldsymbol{E}\big)_{\mathbb{R}^d} = \inf_{\boldsymbol{e}\in\mathscr{E}} \big(\boldsymbol{A}(\boldsymbol{E}+\boldsymbol{e}), \boldsymbol{E}+\boldsymbol{e}\big)_{L^2_{\text{per}}} \tag{2.2a}$$

$$\big(\boldsymbol{A}^{-1}_{\text{eff}}\boldsymbol{J}, \boldsymbol{J}\big)_{\mathbb{R}^d} = \inf_{\boldsymbol{\jmath}\in\mathscr{J}} \big(\boldsymbol{A}^{-1}(\boldsymbol{J}+\boldsymbol{\jmath}), \boldsymbol{J}+\boldsymbol{\jmath}\big)_{L^2_{\text{per}}} \tag{2.2b}$$

Remark 2.2. *For particular macroscopic load* $\boldsymbol{E}$*, a minimizer in previous definition exists and is unique according to the direct method in the calculus of variation since material coefficients* $\boldsymbol{A}$ *are bounded and uniformly elliptic, cf Rem. 2.5. Next, homogenized matrix* $\boldsymbol{A}_{\text{eff}}$ *coincide in both formulations, is symmetric and positive definite, see e.g. [3, 11].*

Remark 2.3. *Space* $\mathscr{E}$ *contains curl-free functions — such function* $\boldsymbol{e}$ *possess potential* U *such that* $\nabla U = \boldsymbol{e}$*. The reformulation of homogenization problem with potentials avoids a trouble with constructing finite element spaces with curl-free basis functions, however, FFT-based FEM naturally overcome this difficulty by employing a certain projection operators, see Lem. 4.2.*

Remark 2.4. *The previous homogenization formulas in Eq.* (2.2) *are obtained from an asymptotic behavior as* $\varepsilon \to 0$ *of a linear elliptic problem: for* $\varepsilon > 0$*, find* $u^\varepsilon \in H^1_0(\Omega) := \{v \in L^2(\Omega), v|_{\partial\Omega} = 0, \frac{\partial v_\alpha}{\partial x_\alpha} \in L^2(\Omega)\}$ *such that*[3]

$$\big(\boldsymbol{A}^\varepsilon \nabla u^\varepsilon, \nabla v\big)_{L^2(\Omega)} = F(v), \quad \forall v \in H^1_0(\Omega) \tag{2.3}$$

where Ω *is a bounded open set in* $\mathbb{R}^d$ *with the Lipschitz boundary* $\partial\Omega$*. Oscillating material coefficients are defined as* $\boldsymbol{A}^\varepsilon(\boldsymbol{x}) := \boldsymbol{A}\left(\frac{\boldsymbol{x}}{\varepsilon}\right)$ *for prescribed material coefficients* $\boldsymbol{A}$ *and linear functional* F *contains boundary conditions and source terms.*

Since the material coefficients are uniformly elliptic and bounded, the unique solutions u^ε *of weak formulation* (2.3) *exists, are uniformly bounded for* ε*, and thus weakly converges in* $H^1_0(\Omega)$ *to some function* $u_{\text{eff}} \in H^1_0(\Omega)$ *representing the averaged field of oscillating solutions* u^ε*.*

Various methods, see e.g. the notion of H-convergence in [11, 46] or formal method of asymptotic expansion [3], reveal that weak limit u_{eff} *satisfies variational formulation*

$$\big(\boldsymbol{A}_{\text{eff}} \nabla u_{\text{eff}}, \nabla v\big)_{L^2(\Omega)} = F(v), \quad \forall v \in H^1_0(\Omega) \tag{2.4}$$

with homogeneous material coefficients $\boldsymbol{A}_{\text{eff}}$ *obtained by homogenization formula* (2.2a)*.*

Hence the homogenized matrix represents a limit state, however, in reality it is a reliable value if the periodic unit cell is sufficiently small compared to the macroscale of a designed object.

The necessity for homogenization theories is observed from Eq. (2.3)*; a direct discretization, for some small* ε*, is inappropriate because of highly oscillating solutions — it requires huge number of degrees of freedom. The homogenization thus splits elliptic problem* (2.3) *to the evaluation of homogenized matrix* (2.2a) *and to the solution of Eq.* (2.4) *without the oscillatory part.*

[3] The values on boundary $v|_{\partial\Omega}$ are meant in the sense of trace operator, partial derivatives $\frac{\partial v_\alpha}{\partial x_\alpha}$ are meant in the sense of distributional derivatives.

In the sequel, an attention is focused on the primal homogenization problem in Def. 2.1; the dual homogenization problem is only utilized later in Sec. 4.3 for the guaranteed bounds of the homogenized matrix.

Remark 2.5. *Since material coefficients $\boldsymbol{A}$ are symmetric, the homogenization problem in Eq. 2.2a is equivalent to a weak formulation*[4]

$$\big(\boldsymbol{A}\tilde{\boldsymbol{e}}^{(\boldsymbol{E})},\boldsymbol{v}\big)_{L^2_{\mathrm{per}}} = -\big(\boldsymbol{AE},\boldsymbol{v}\big)_{L^2_{\mathrm{per}}},\quad \forall \boldsymbol{v}\in\mathscr{E} \tag{2.5}$$

describing the minimizer point. Since it has a linear structure, minimizer $\tilde{\boldsymbol{e}}^{E}\in\mathscr{E}$ for macroscopic field $E\in\mathbb{R}^d$ is obtained from unitary minimizers $\tilde{\boldsymbol{e}}^{(\alpha)}$, see the next definition, as

$$\tilde{\boldsymbol{e}}^{E} = \sum_{\alpha} E_\alpha \tilde{\boldsymbol{e}}^{(\alpha)}.$$

Definition 2.6 (auxiliary problems)**.** *We say that $\tilde{\boldsymbol{e}}^{(\alpha)}\in\mathscr{E}$ are unitary minimizers if they satisfy the weak formulations with unitary macroscopic loads $\boldsymbol{\epsilon}_\alpha$, i.e.*

$$\big(\boldsymbol{A}\tilde{\boldsymbol{e}}^{(\alpha)},\boldsymbol{v}\big)_{L^2_{\mathrm{per}}} = -\big(\boldsymbol{A}\boldsymbol{\epsilon}_\alpha,\boldsymbol{v}\big)_{L^2_{\mathrm{per}}},\quad \forall \boldsymbol{v}\in\mathscr{E}, \tag{2.6}$$

Unitary microscopic fields are noted without tilde, i.e. $\boldsymbol{e}^{(\alpha)} := \boldsymbol{\epsilon}_\alpha + \tilde{\boldsymbol{e}}^{(\alpha)} \in \mathscr{U}\oplus^{\perp}\mathscr{E}$.

Remark 2.7. *The unitary minimizers also serve to evaluate the components of the homogenized matrix; formula*

$$\boldsymbol{A}_{\mathrm{eff},\alpha\beta} = \big(\boldsymbol{A}(\boldsymbol{\epsilon}_\beta+\tilde{\boldsymbol{e}}^{(\beta)}),(\boldsymbol{\epsilon}_\alpha+\tilde{\boldsymbol{e}}^{(\alpha)})\big)_{L^2_{\mathrm{per}}}. \tag{2.7}$$

follows from the linear structure.

2.2 Formulation based on the Lippmann-Schwinger equation

This section is dedicated to an alternative formulation of the auxiliary problems in Def. 2.6, namely to the Lippmann-Schwinger equation incorporating the Green function for a reference homogeneous problem.

Definition 2.8 (Lippmann-Schwinger equation)**.** *Let parameter $\boldsymbol{A}^0\in\mathbb{R}^{d\times d}_{\mathrm{spd}}$ be a symmetric positive definite matrix. We say that $\boldsymbol{e}^{(\boldsymbol{E})}_{\mathrm{LS}}\in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ is the solution of Lippmann-Schwinger equation if it satisfies*

$$\boldsymbol{e}^{(\boldsymbol{E})}_{\mathrm{LS}}(\boldsymbol{x}) + \int_{\mathcal{Y}} \boldsymbol{\Gamma}^0(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{A}(\boldsymbol{y})-\boldsymbol{A}^0)\boldsymbol{e}^{(\boldsymbol{E})}_{\mathrm{LS}}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = \boldsymbol{E},\quad \textit{for almost all } \boldsymbol{x}\in\mathcal{Y} \tag{2.8}$$

where the convolution integral is defined with the help of the Fourier series

$$\int_{\mathcal{Y}} \boldsymbol{\Gamma}^0(\boldsymbol{x}-\boldsymbol{y})\boldsymbol{v}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} := \sum_{\boldsymbol{k}\in\mathbb{Z}^d\setminus\{\boldsymbol{0}\}} \frac{\boldsymbol{\xi}(\boldsymbol{k})\otimes\boldsymbol{\xi}(\boldsymbol{k})}{\big(\boldsymbol{A}^0\boldsymbol{\xi}(\boldsymbol{k}),\boldsymbol{\xi}(\boldsymbol{k})\big)_{\mathbb{R}^d}}\hat{\boldsymbol{v}}(\boldsymbol{k})\varphi_{\boldsymbol{k}}(\boldsymbol{x}) \tag{2.9}$$

[4]Left-hand side represents continuous symmetric uniformly elliptic bilinear form for $\tilde{\boldsymbol{e}}^{(\boldsymbol{E})}$ and $\boldsymbol{v}$. Right-hand side represents continuous linear functional for $\boldsymbol{v}$. The existence of an unique solution is provided by Lax-Milgram lemma or simply by Riesz representation theorem.

where $\boldsymbol{\xi}(\boldsymbol{k}) \in \mathbb{R}^d$ *is a vector with components* $\xi_\alpha(\boldsymbol{k}) = \frac{k_\alpha}{Y_\alpha}$*, operator* $\otimes$ *denotes the tensor product*[5]*, and* $\hat{\boldsymbol{v}}(\boldsymbol{k})$ *are the Fourier coefficients with components* $\hat{v}_\alpha(\boldsymbol{k}) := \big(v_\alpha, \varphi_{\boldsymbol{k}}\big)_{L^2_{\mathrm{per}}}$ *for trigonometric polynomial* $\varphi_{\boldsymbol{k}} = \exp(\mathrm{i}\pi \sum_\alpha \frac{x_\alpha k_\alpha}{Y_\alpha})$.

Lippmann-Schwinger equation is formulated for microscopic field $\boldsymbol{e}^{(\boldsymbol{E})}_{\mathrm{LS}}$ contrary to weak formulation (2.5) written for the perturbation part $\tilde{\boldsymbol{e}}^{(\boldsymbol{E})} = \boldsymbol{e}^{(\boldsymbol{E})} - \boldsymbol{E}$; both of the solutions, if coincide, differ by constant $\boldsymbol{E}$ as $\langle \boldsymbol{e}^{(\boldsymbol{E})}_{\mathrm{LS}} \rangle = \boldsymbol{E}$ and $\langle \tilde{\boldsymbol{e}}^{(\boldsymbol{E})} \rangle = \boldsymbol{0}$.

Remark 2.9. *Lippmann-Schwinger equation is deduced from a strong formulation*[6]*: for prescribed macroscopic load* $E \in \mathbb{R}^d$*, find* $\tilde{\boldsymbol{e}}$ *with continuous partial derivatives satisfying*

$$\nabla \cdot [\boldsymbol{A}(E + \tilde{\boldsymbol{e}})] = 0, \qquad \nabla \times \tilde{\boldsymbol{e}} = 0, \qquad \langle \tilde{\boldsymbol{e}} \rangle = 0. \tag{2.10}$$

The problem is reformulated for a reference homogeneous material $\boldsymbol{A}^0 \in \mathbb{R}^{d\times d}_{\mathrm{spd}}$*, the parameter of Lippmann-Schwinger equation, to a homogeneous problem*

$$\nabla \cdot \Big[\boldsymbol{A}^0 \boldsymbol{e}(\boldsymbol{x})\Big] = \boldsymbol{f}(\boldsymbol{x})$$

where $\boldsymbol{f}(\boldsymbol{x}) = -\nabla \cdot \Big[\big(\boldsymbol{A}(\boldsymbol{x}) - \boldsymbol{A}^0\big)\boldsymbol{e}(\boldsymbol{x})\Big]$ *is called a divergence of polarization current and* $\boldsymbol{e} = \boldsymbol{E} + \tilde{\boldsymbol{e}}$ *is microscopic field with* $\tilde{\boldsymbol{e}}$ *satisfying the curl-free and zero mean condition. It is then transformed with the technique of the Fourier transform*[7] *to the system of algebraic equations, whose solution yields the Lippmann-Schwinger equation* (2.8)*.*

The resulting equation is already defined in a way to have a good sense and to be easily analyzed, see Sec. 4.1 and particularly Theorem 4.1. It will become apparent that the detailed derivation of Lippmann-Schwinger equation can be omitted; the only important part is the convolution integral providing projection on curl-free fields with zero mean, see Lem. 4.2.

Remark 2.10 (Solution of Lippmann-Schwinger equation)**.** *Lippmann-Schwinger equation* (2.8) *can be written in operator form* $(I + \mathcal{B})\boldsymbol{e}^{(\boldsymbol{E})}_{\mathrm{LS}} = \boldsymbol{E}$ *with the obvious definition of operators* I *and* $\mathcal{B}$*. The inverse of* $(I + \mathcal{B})$ *is then expressed using Neumann series expansion* $(I + \mathcal{B})^{-1} = \sum_{k=0}^{\infty} \mathcal{B}^k$ *as* $\|B\| < 1$ *for the special choice of parameter* $\boldsymbol{A}^0 := \frac{1}{2}(\|\boldsymbol{A}^{-1}\|_{L^\infty_{\mathrm{per}}} + \|\boldsymbol{A}\|_{L^\infty_{\mathrm{per}}})\boldsymbol{I}$*, see [28].*

Remark 2.11 (FFT-based method according to Moulinec and Suquet in [32])**.** *The solution of Lippmann-Schwinger equation obtained from Neumann series* $\boldsymbol{e}^{(\boldsymbol{E})}_{\mathrm{LS}} = \sum_{k=0}^{\infty} \mathcal{B}^k[\boldsymbol{E}]$ *is the limit in* L^2_{per} *norm as* $k \to \infty$ *of iteration algorithm* $\boldsymbol{e}_{(k+1)} = \mathcal{B}[\boldsymbol{e}_{(k)}] + \boldsymbol{E}$ *with initial value* $\boldsymbol{e}_{(0)} := E$*. The iteration algorithm serves as a foundation for the numerical algorithm proposed by Moulinec and Suquet in [32, 33]*

[5]The tensor product of two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ defines matrix $\boldsymbol{u} \otimes \boldsymbol{v} \in \mathbb{R}^{d\times d}$ with components $(\boldsymbol{u} \otimes \boldsymbol{v})_{\alpha\beta} = u_\alpha v_\beta$.

[6]The solution of the strong formulation also satisfies the weak formulation in Eq. (2.5); it is shown by the multiplication of a test function, by integration over PUC $\mathcal{Y}$, and by application of Green's theorem (boundary term vanishes because of periodicity).

[7]The most important properties are: derivative is transformed to the multiplication with a Fourier variable and the convolution of two functions is transformed to their multiplication. The inverse Fourier transform is provided with an analogical formula.

who approximated solution $\boldsymbol{e}$ with the function values at a regular grid corresponding to pixels or voxels in the images of periodic unit cell. The convolution integral of Lippmann-Schwinger equation is then replaced with the Discrete Fourier Transform[8]*, its inverse, and the multiplication by the integral kernel at the Fourier space; in a detail, the algorithm is described in Paper 1.*

2.3 Guaranteed bounds on homogenized matrix

The upper-lower bounds obtained from a posteriori error estimates were introduced by Dvořák [12, 13] for a scalar problem and independently by Wieçkowski [55] for linear elasticity. This section provides a summary of results in [12]. In this section, we work with some conforming approximations of unitary minimizers $\tilde{\boldsymbol{e}}^{(\alpha)}$ and $\tilde{\boldsymbol{\jmath}}^{(\alpha)}$, namely $\tilde{\boldsymbol{e}}^{(\alpha)}_{\boldsymbol{N}} = (\boldsymbol{e}^{(\alpha)}_{\boldsymbol{N}} - \boldsymbol{\epsilon}_\alpha) \in \mathscr{E}$ and $\tilde{\boldsymbol{\jmath}}^{(\alpha)}_{\boldsymbol{N}} = (\boldsymbol{\jmath}^{(\alpha)}_{\boldsymbol{N}} - \boldsymbol{\epsilon}_\alpha) \in \mathscr{J}$; parameter $\boldsymbol{N}$ represents the inverse of discretization size of FEM or the number of discretization points in the case of FFT-based method, for details see Sec. 4.2 or Papers 5 and 6. In what follows, relation $\boldsymbol{C} \preceq \boldsymbol{D}$ between symmetric and positive definite matrices $\boldsymbol{C}, \boldsymbol{D} \in \mathbb{R}^{d\times d}_{\text{spd}}$ stands for ordering in the sense of quadratic forms; it is equivalent to $\boldsymbol{E}\cdot\boldsymbol{C}\boldsymbol{E} \leq \boldsymbol{E}\cdot\boldsymbol{D}\boldsymbol{E}$ for all $\boldsymbol{E} \in \mathbb{R}^d$. Upper bound of homogenized matrix is obtained from the primal homogenization problem (2.2a). The replacement of minimizers with approximate minimizers leads to an increase in the value of the quadratic form with the homogenized matrix

$$\big(\boldsymbol{A}_{\text{eff}}\boldsymbol{E}, \boldsymbol{E}\big)_{\mathbb{R}^d} = \inf_{\boldsymbol{e}\in\mathscr{E}} \big(\boldsymbol{A}(\boldsymbol{E}+\boldsymbol{e}), \boldsymbol{E}+\boldsymbol{e}\big)_{L^2_{\text{per}}} \leq \big(\boldsymbol{A}(\boldsymbol{E}+\boldsymbol{e}^{(\boldsymbol{E})}_{\boldsymbol{N}}), \boldsymbol{E}+\boldsymbol{e}^{(\boldsymbol{E})}_{\boldsymbol{N}}\big)_{L^2_{\text{per}}}.$$

The last term then defines the upper bound of the homogenized matrix.

Definition 2.12. *We say that matrix $\overline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}} \in \mathbb{R}^{d\times d}$ defined as*

$$(\overline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}})_{\alpha\beta} = \big(\boldsymbol{A}(\boldsymbol{\epsilon}_\beta + \tilde{\boldsymbol{e}}^{(\beta)}_{\boldsymbol{N}}), \boldsymbol{\epsilon}_\alpha + \tilde{\boldsymbol{e}}^{(\alpha)}_{\boldsymbol{N}}\big)_{L^2_{\text{per}}}$$

is the upper bound on homogenized matrix $\boldsymbol{A}_{\text{eff}}$.

The upper bound of homogenized matrix $\boldsymbol{A}_{\text{eff}} \preceq \overline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}}$ can be completed with the upper bound of inverse homogenized matrix $\boldsymbol{A}^{-1}_{\text{eff}} \preceq \underline{\boldsymbol{A}}^{-1}_{\text{eff},\boldsymbol{N}}$ obtained from dual formulation (2.2b). Both relations lead to the upper-lower bound structure of the homogenized matrix

$$\underline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}} \preceq \boldsymbol{A}_{\text{eff}} \preceq \overline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}}. \tag{2.11}$$

Remark 2.13 (Element-wise upper-lower bounds)**.** *Upper-lower bounds* (2.11) *imply element-wise bounds, the bounds on components of the homogenized matrix.*

The mean of upper-lower bounds $\boldsymbol{A}_{\text{eff},\boldsymbol{N}} = \frac{1}{2}(\overline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}} + \underline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}})$ is a more reliable approximation of homogenized matrix with guaranteed error $\boldsymbol{D}_{\boldsymbol{N}} = \frac{1}{2}(\overline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}} - \underline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}})$. Following lemma provides the convergence[9] of the error to zero if the approximates minimizers converge to minimizers as $\min_\alpha N_\alpha \to \infty$.

[8] Numerically, it is realized with FFT algorithm, for details see [9].

[9] The trace of a matrix is a norm on the set of positive definite matrices, particularly on $\boldsymbol{D}_{\boldsymbol{N}}$ for $\boldsymbol{N} \in \mathbb{N}^d$.

Lemma 2.14 (Rate of convergence of homogenized properties)**.** *Guaranteed error* $\boldsymbol{D_N}$ *satisfies following inequality*

$$\operatorname{tr} \boldsymbol{D_N} \leq C_1 \sum_{\alpha} \|\tilde{\boldsymbol{e}}^{(\alpha)} - \tilde{\boldsymbol{e}}_{\boldsymbol{N}}^{(\alpha)}\|^2_{L^2_{\mathrm{per}}} + C_2 \sum_{\alpha} \|\tilde{\boldsymbol{j}}^{(\alpha)} - \tilde{\boldsymbol{j}}_{\boldsymbol{N}}^{(\alpha)}\|^2_{L^2_{\mathrm{per}}}$$

where constants C_1*,* C_2 *are independent of* $\boldsymbol{N}$*.*

3 Methodology

The goal of this dissertation consists of analysis of Lippmann-Schwinger equation (2.8) and particularly to Moulinec-Suquet algorithm [32], see Rem. 2.11. The relevant questions arising in the analysis can be summarized as:

1. Is there the unique solution of Lippmann-Schwinger equation for various parameters $\boldsymbol{A}^0$; the existence has been shown only for particular choice of parameter $\boldsymbol{A}^0$, see Rem. 2.10. Does the solution of Lippmann-Schwinger equation equal to the solution of weak formulation (2.5)? Do the solutions coincide for various parameters $\boldsymbol{A}^0$?

2. (a) Is there some consistent approximation of weak formulation (2.5) or of Lippmann-Schwinger equation (2.8) that is equivalent to Moulinec-Suquet numerical algorithm? Do approximate solutions, if exist, converge to the solution of weak formulation?

   (b) It has been observed by Zeman et al. in [57] that Moulinec-Suquet algorithm, Rem. 2.11, is equivalent to the solution of a system of linear equations. It appears that this non-symmetric system can be successfully solved using Conjugate gradient algorithm. What is the reason for that?

3. In [12, 13] and later independently in [55], the guaranteed bounds of the homogenized matrix were calculated using the standard Finite element methods. Is the approach applicable to the FFT-based homogenization?

The analysis of the method is provided with standard mathematical instruments and techniques. Namely, it is based on the following subjects with the list of literature: homogenization theory [11, 40, 18, 3], mathematical analysis [41, 42, 26], modern theory of partial differential equations [19, 5, 35], finite element method [10, 4], and trigonometric collocation method [43].

Numerical calculations stated in Sec. 4.4, 4.5 and attached papers were provided with the own software written in MATLAB® and PYTHON programming language; it is available at `http://mech.fsv.cvut.cz/~vondrejc/publications.php#SW`.

4 Results

This section provides the summary of the results obtained in six papers [57, 50, 37, 38, 51, 52], referenced as Papers 1-6, that are attached in the chronological order in Parts II–VII. The section is split into five subsections; The first three correspond to the tasks in Methodology Sec. 3 while the last two contain numerical examples and applications of FFT-based method to linear elasticity.

4.1 Weak formulation and Lippmann-Schwinger equation

This section start with theorem describing solvability of Lippmann-Schwinger equation, the main theorem about continuous formulation of homogenization problem.

Theorem 4.1 (Equivalence of weak formulation and Lippmann-Schwinger equation, Theorem 2.29 in Paper 5)**.** *Let* $\boldsymbol{A}^0 \in \mathbb{R}^{d\times d}_{\mathrm{spd}}$ *be symmetric and positive definite, then weak formulation* (2.5) *and Lippmann-Schwinger equation* (2.8) *are equivalent in the sense that the solution coincide in both formulations.*

The theorem shows that the unique solution of weak formulation (2.5) is the solution of Lippmann-Schwinger equation and contrary. It reveals the existence of the unique solution of Lippmann-Schwinger equation for various parameters $\boldsymbol{A}^0$.

The proof is based on a next lemma providing a projection on $\mathscr{E}$, the space of curl-free with zero mean fields that is both a trial space and the space of test functions in weak formulation (2.5).

Lemma 4.2 (Lem. 2.28 in Paper 5)**.** *Operator* $\mathcal{G}[\cdot] : L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) \to L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ *defined as*

$$\mathcal{G}[\boldsymbol{f}](\boldsymbol{x}) = \int_{\mathcal{Y}} \boldsymbol{\Gamma}^0(\boldsymbol{x}-\boldsymbol{y})\boldsymbol{A}^0\boldsymbol{f}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \tag{4.1}$$

is a projection *on* $\mathscr{E}$ *and orthogonal for* $\boldsymbol{A}^0 = \lambda\boldsymbol{I}$ *with* $\lambda > 0$.

The proof is based on the expression of convolution integral in the Fourier space, cf. Def. 2.8. It is shown that the operator is continuous and real valued. Since matrix $\frac{\boldsymbol{\xi}(\boldsymbol{k})\otimes\boldsymbol{\xi}(\boldsymbol{k})}{\boldsymbol{A}^0\boldsymbol{\xi}(\boldsymbol{k})\cdot\boldsymbol{\xi}(\boldsymbol{k})}\boldsymbol{A}^0$ is simply a projection, it shows that operator $\mathcal{G}$ is. Besides, we show that it is the projection on $\mathscr{E}$ with the help of a potential that exists in this case as the periodic unit cell is simply connected.

Proof outline of Theorem 4.1. The proof is outlined for a special choice of parameter $\boldsymbol{A}^0$ equal to the identity matrix $\boldsymbol{A}^0 = \boldsymbol{I}$. First, we show that the solution of the weak formulation is the solution of Lippmann-Schwinger equation. We enlarge space $\mathscr{E}$ in the weak formulation by adding the projection operator

$$\big(\boldsymbol{A}\tilde{\boldsymbol{e}}, \mathcal{G}[\boldsymbol{v}]\big)_{L^2_{\mathrm{per}}} = -\big(\boldsymbol{A}\boldsymbol{E}, \mathcal{G}[\boldsymbol{v}]\big)_{L^2_{\mathrm{per}}}, \quad \forall \boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d).$$

In this special case, operator $\mathcal{G}$ is self-adjoint leading to

$$\big(\mathcal{G}[\boldsymbol{A}\tilde{\boldsymbol{e}}], \boldsymbol{v}\big)_{L^2_{\mathrm{per}}} = -\big(\mathcal{G}[\boldsymbol{A}\boldsymbol{E}], \boldsymbol{v}\big)_{L^2_{\mathrm{per}}}, \quad \forall \boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d).$$

Finally, a following equation is deduced

$$\mathcal{G}[\boldsymbol{A}\tilde{\boldsymbol{e}}] = -\mathcal{G}[\boldsymbol{A}\boldsymbol{E}], \quad \text{for almost all } \boldsymbol{x} \in \mathcal{Y}$$

by removing the testing functions and the rest is the consequence of elementary algebra and properties of projection $\mathcal{G}$.

The opposite implication is done analogically by splitting the solution $\boldsymbol{e}^{(\boldsymbol{E})}_{\mathrm{LS}}$ with projection operator $\mathcal{G}$. The general case for $\boldsymbol{A}^0 \in \mathbb{R}^{d\times d}_{\mathrm{spd}}$ follows similar ideas. □

4.2 Discretization via trigonometric polynomials

This section is dedicated to the discretization of the homogenization problem (2.2) in a way to produce the identical linear system like in Moulinec-Suquet algorithm [32], cf. Rem. 2.11.

The space of trigonometric polynomials is used as a finite dimensional space for discretization. Relating definitions and properties are summarized in Sec. 4.2.1 according to [47, 43], however, a slight modification is performed for the non-odd number of discretization points to assure conforming approximations that are required for guaranteed bounds of the homogenized matrix, see Sec. 4.3 or Papers 5-6 for more details.

Then, approximate solutions are defined through Galerkin approximation (GA) and Galerkin approximation with numerical integration (GAwNI). Together with their convergence to the minimizers and the numerical solution of the linear system, it is described in Sec. 4.2.2 and in a detail in Papers 5 and 6.

Notation 4.3. *In the sequel, let* $\boldsymbol{N} \in \mathbb{N}^d$ *be reserved for a number of discretization points with the number of degrees of freedom* $|\boldsymbol{N}|_\Pi := \prod_\alpha N_\alpha$*; if* N_α *is odd (even) for all* α *we talk about odd (even) number of discretization points, otherwise about non-odd ones. The reduced and full index set state for*

$$\mathbb{Z}^d_{\boldsymbol{N}} = \left\{ \boldsymbol{k} \in \mathbb{Z}^d : |k_\alpha| < \frac{N_\alpha}{2} \right\}, \qquad \underline{\mathbb{Z}}^d_{\boldsymbol{N}} = \left\{ \boldsymbol{k} \in \mathbb{Z}^d : -\frac{N_\alpha}{2} \le k_\alpha < \frac{N_\alpha}{2} \right\}. \tag{4.2}$$

A multi-index notation is employed, in which $\mathbb{R}^{\boldsymbol{N}}$ *represents* $\mathbb{R}^{N_1 \times \cdots \times N_d}$*. Set* $\mathbb{X}_{\boldsymbol{N}}$ *represents the space of vectors* v *with components* $\mathsf{v}^{\boldsymbol{n}}_\alpha$ *and* $\mathbb{X}^2_{\boldsymbol{N}}$ *the space of matrices* A *with components* $\mathsf{A}^{\boldsymbol{nm}}_{\alpha\beta}$ *for* α, β *and* $\boldsymbol{n}, \boldsymbol{m} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}$*. Next,* $\mathbf{v}^{\boldsymbol{n}} \in \mathbb{R}^d$ *for* $\boldsymbol{n} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}$ *and* $\mathbf{v}_\alpha \in \mathbb{R}^{\boldsymbol{N}}$ *for* α *represent subvectors of* v *with components* $\mathsf{v}^{\boldsymbol{n}}_\alpha$*; analogically the submatrices* $\mathbf{A}^{\boldsymbol{nm}} \in \mathbb{R}^{d\times d}$ *and* $\mathbf{A}_{\alpha\beta} \in \mathbb{R}^{\boldsymbol{N}\times\boldsymbol{N}}$ *can be defined. A scalar product on* $\mathbb{X}_{\boldsymbol{N}}$ *is defined as* $\big(\mathsf{u}, \mathsf{v}\big)_{\mathbb{X}_{\boldsymbol{N}}} := \sum_\alpha \sum_{\boldsymbol{n}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \mathsf{u}^{\boldsymbol{n}}_\alpha \mathsf{v}^{\boldsymbol{n}}_\alpha$ *and matrix* A *by vector* v *multiplication as* $(\mathsf{Av})^{\boldsymbol{n}}_\alpha := \sum_\beta \sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \mathsf{A}^{\boldsymbol{nm}}_{\alpha\beta} \mathsf{v}^{\boldsymbol{m}}_\beta$*. Matrix* A *is symmetric positive definite if* $\mathsf{A}^{\boldsymbol{mn}}_{\alpha\beta} = \mathsf{A}^{\boldsymbol{nm}}_{\beta\alpha}$ *holds for all components and* $\big(\mathsf{Av}, \mathsf{v}\big)_{\mathbb{X}_{\boldsymbol{N}}} > 0$ *applies for arbitrary* $\mathsf{v} \in \mathbb{X}_{\boldsymbol{N}}$*. We use a serif font for vectors* v *and matrices* A *to distinguish from vectors* $\boldsymbol{E} \in \mathbb{R}^d$ *and matrices* $\boldsymbol{A}_{\mathrm{eff}} \in \mathbb{R}^{d\times d}$ *and from vector valued functions* $\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y}; \mathbb{R}^d)$*. In order to distinguish vectors and matrices for different number of discretization points* $\boldsymbol{N}$*, we write them with subscript, i.e.* $\mathsf{v}_{\boldsymbol{N}}$ *and* $\mathsf{A}_{\boldsymbol{N}}$*.*

Operator $\oplus^\perp$ *denotes the direct sum of mutually orthogonal subspaces, e.g.* $\mathbb{R}^d = \boldsymbol{\epsilon}_1 \oplus^\perp \boldsymbol{\epsilon}_2 \oplus^\perp \ldots \oplus^\perp \boldsymbol{\epsilon}_d$*.*

4.2.1 Trigonometric polynomials

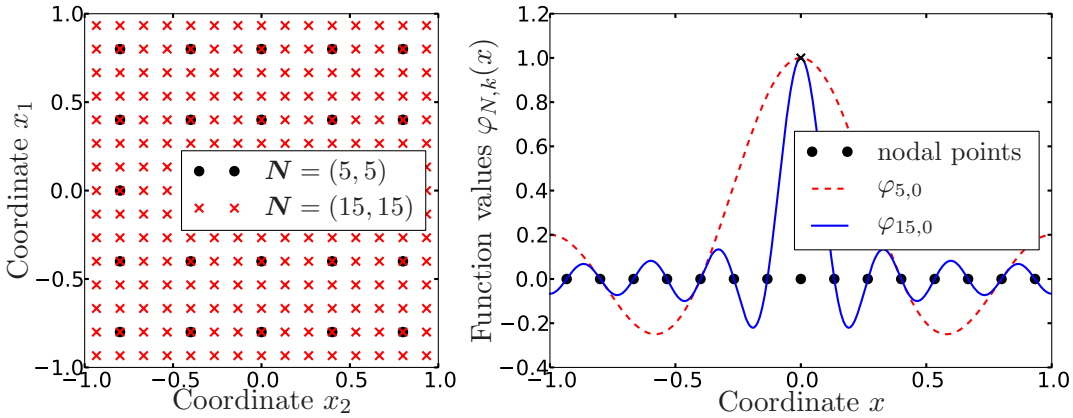
This section presents the trigonometric polynomials with their properties; it is also well described in Paper 5 for the odd number of discretization points and in Paper 6 for the general number of discretization points, see also [43].

First, the trigonometric polynomials can be expressed as the linear combination of the Fourier coefficients and the Fourier basis functions, see Eq. (4.3) and (4.5). Second, they can be expressed as the linear combination of function values at nodal

points and shape basis functions, see Def. 4.5 and Eq. (4.6); Fig. 1 shows the examples of the nodal points and the shape basis function. Both formulations are related through the Discrete Fourier Transform (DFT) see Rem. 4.12.

Dirac delta property of shape basis functions $\varphi_{\boldsymbol{N},\boldsymbol{m}}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{n}}) = \delta_{\boldsymbol{mn}}$ ensures the uniqueness of trigonometric polynomials representation and allows us to define an interpolation operator through the nodal points, see Def. 4.8 and 4.5.

Both definitions of trigonometric polynomials $\mathscr{T}_{\boldsymbol{N}}$ and $\tilde{\mathscr{T}}_{\boldsymbol{N}}$ in Def. 4.7 coincide if the number of discretization points $\boldsymbol{N}$ is odd for all components, cf. Rem. 4.13.



(a) nodal values at PUC (b) shape basis functions

Figure 1: Nodal values for $d = 2$ and shape basis function for $d = 1$

Notation 4.4 (DFT). *For $\boldsymbol{N} \in \mathbb{N}^d$ we define, up to constant, unitary matrices $\mathsf{F}_{\boldsymbol{N}}, \mathsf{F}_{\boldsymbol{N}}^{-1} \in \mathbb{C}^{d\times d\times \boldsymbol{N}\times \boldsymbol{N}}$ of the Discrete Fourier transform (DFT) and its inverse (iDFT) as*

$$\mathsf{F}_{\boldsymbol{N}} = \frac{1}{|\boldsymbol{N}|_\Pi}\left(\delta_{\alpha\beta}\omega_{\boldsymbol{N}}^{-\boldsymbol{mn}}\right)_{\alpha,\beta=1,\dots,d}^{\boldsymbol{m},\boldsymbol{n}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} \qquad \mathsf{F}_{\boldsymbol{N}}^{-1} = \left(\delta_{\alpha\beta}\omega_{\boldsymbol{N}}^{\boldsymbol{mn}}\right)_{\alpha,\beta=1,\dots,d}^{\boldsymbol{m},\boldsymbol{n}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d}.$$

where $\delta_{\alpha\beta}$ is Kronecker delta and $\omega_{\boldsymbol{N}}^{\boldsymbol{mn}} = \exp\left(2\pi\mathrm{i}\sum_{\alpha=1}^d \frac{m_\alpha n_\alpha}{N_\alpha}\right)$ with $\boldsymbol{m},\boldsymbol{n} \in \mathbb{Z}^d$.

Definition 4.5 (nodal points, basis functions). *Let $\boldsymbol{N} \in \mathbb{N}^d$. For $\boldsymbol{n} \in \underline{\mathbb{Z}}_{\boldsymbol{N}}^d$, we define nodal points of the periodic unit cell $\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{n}} = \sum_\alpha \frac{2Y_\alpha n_\alpha}{N_\alpha}\boldsymbol{\epsilon}_\alpha$ and the Fourier and shape basis functions*

$$\varphi_{\boldsymbol{n}}(\boldsymbol{x}) = \exp\left(\pi\mathrm{i}\sum_\alpha \frac{n_\alpha x_\alpha}{Y_\alpha}\right), \qquad \varphi_{\boldsymbol{N},\boldsymbol{n}}(\boldsymbol{x}) = \frac{1}{|\boldsymbol{N}|_\Pi}\sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} \omega_{\boldsymbol{N}}^{-\boldsymbol{mn}}\varphi_{\boldsymbol{m}}(\boldsymbol{x}). \tag{4.3}$$

Lemma 4.6 (Properties of $\varphi_{\boldsymbol{m}}$ and $\varphi_{\boldsymbol{N},\boldsymbol{m}}$). *Let $\boldsymbol{m},\boldsymbol{n} \in \underline{\mathbb{Z}}_{\boldsymbol{N}}^d$, then*

$$\left(\varphi_{\boldsymbol{m}},\varphi_{\boldsymbol{n}}\right)_{L^2_{\mathrm{per}}} = \delta_{\boldsymbol{mn}} \qquad \varphi_{\boldsymbol{n}}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{m}}) = \omega_{\boldsymbol{N}}^{\boldsymbol{mn}} \tag{4.4a}$$

$$\varphi_{\boldsymbol{N},\boldsymbol{m}}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{n}}) = \delta_{\boldsymbol{mn}} \qquad \left(\varphi_{\boldsymbol{N},\boldsymbol{m}},\varphi_{\boldsymbol{N},\boldsymbol{n}}\right)_{L^2_{\mathrm{per}}} = \frac{\delta_{\boldsymbol{mn}}}{|\boldsymbol{N}|_\Pi} \tag{4.4b}$$

Definition 4.7 (Trigonometric polynomials)**.** *For $\boldsymbol{N} \in \mathbb{N}^d$, we define the spaces of trigonometric polynomials $\mathscr{T}_{\boldsymbol{N}}$, $\tilde{\mathscr{T}}_{\boldsymbol{N}}$ and their vector valued versions $\mathscr{T}^d_{\boldsymbol{N}}$,$\tilde{\mathscr{T}}^d_{\boldsymbol{N}}$ as*

$$\mathscr{T}_{\boldsymbol{N}} = \Big\{ \sum_{\boldsymbol{n}\in\mathbb{Z}^d_{\boldsymbol{N}}} \hat{\mathsf{v}}^{\boldsymbol{n}} \varphi_{\boldsymbol{n}} : \hat{\mathsf{v}}^{\boldsymbol{n}} \in \mathbb{C}, \hat{\mathsf{v}}^{\boldsymbol{n}} = \overline{(\hat{\mathsf{v}}^{-\boldsymbol{n}})} \Big\}, \qquad \mathscr{T}^d_{\boldsymbol{N}} = \big\{ \boldsymbol{v} : v_\alpha \in \mathscr{T}_{\boldsymbol{N}} \big\}. \tag{4.5}$$

$$\tilde{\mathscr{T}}_{\boldsymbol{N}} = \Big\{ \sum_{\boldsymbol{n}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \mathsf{v}^{\boldsymbol{n}} \varphi_{\boldsymbol{N},\boldsymbol{n}} : \mathsf{v}^{\boldsymbol{n}} \in \mathbb{R} \Big\}, \qquad \tilde{\mathscr{T}}^d_{\boldsymbol{N}} = \big\{ \boldsymbol{v} : v_\alpha \in \tilde{\mathscr{T}}_{\boldsymbol{N}} \big\}. \tag{4.6}$$

where the index sets are given by

$$\mathbb{Z}^d_{\boldsymbol{N}} = \Big\{ \boldsymbol{k} \in \mathbb{Z}^d : |k_\alpha| < \frac{N_\alpha}{2} \Big\}, \qquad \underline{\mathbb{Z}}^d_{\boldsymbol{N}} = \Big\{ \boldsymbol{k} \in \mathbb{Z}^d : -\frac{N_\alpha}{2} \le k_\alpha < \frac{N_\alpha}{2} \Big\}. \tag{4.7}$$

Definition 4.8 (Interpolation projection)**.** *We define interpolation operator $Q_{\boldsymbol{N}} : C_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) \to L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{C}^d)$ as*

$$Q_{\boldsymbol{N}}[f] = \sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} f(\boldsymbol{x}^{\boldsymbol{m}}_{\boldsymbol{N}}) \varphi_{\boldsymbol{N},\boldsymbol{m}}.$$

Lemma 4.9. *Interpolation operator $Q_{\boldsymbol{N}}$ is projection and its image is $\tilde{\mathscr{T}}^d_{\boldsymbol{N}}$.*

Definition 4.10. *The operator $\mathcal{I}_{\boldsymbol{N}} : \tilde{\mathscr{T}}^d_{\boldsymbol{N}} \to \mathbb{R}^{d\times\boldsymbol{N}}$ stocks the values of trigonometric polynomials at nodal points to a vector $\mathcal{I}_{\boldsymbol{N}}[\boldsymbol{v}_{\boldsymbol{N}}] = \big(\boldsymbol{v}_{\boldsymbol{N},\alpha}(\boldsymbol{x}^{\boldsymbol{n}}_{\boldsymbol{N}})\big)^{\boldsymbol{n}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}}_{\alpha=1,\dots,d}$.*

Lemma 4.11. *The operator $\mathcal{I}_{\boldsymbol{N}}$ from previous definition is isomorphism.*

Remark 4.12 (Connection of representations)**.** *The trigonometric polynomial $\boldsymbol{v}_{\boldsymbol{N}} \in \tilde{\mathscr{T}}^d_{\boldsymbol{N}}$ can be uniquely expressed using both the Fourier coefficients and the function values at nodal points*

$$\boldsymbol{v}_{\boldsymbol{N}} = \sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \boldsymbol{v}_{\boldsymbol{N}}(\boldsymbol{x}^{\boldsymbol{m}}_{\boldsymbol{N}}) \varphi_{\boldsymbol{N},\boldsymbol{m}} = \sum_{\boldsymbol{n}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \hat{\boldsymbol{v}}_{\boldsymbol{N}}(\boldsymbol{n}) \varphi_{\boldsymbol{n}} \tag{4.8}$$

with connection through the DFT $\hat{\mathbf{v}}_{\boldsymbol{N}} = \mathsf{F}_{\boldsymbol{N}} \mathbf{v}_{\boldsymbol{N}}$, where $\mathbf{v}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}[\boldsymbol{v}_{\boldsymbol{N}}]$ and $\hat{\mathbf{v}}_{\boldsymbol{N}} = \big(\hat{v}_{\boldsymbol{N},\alpha}(\boldsymbol{m})\big)^{\boldsymbol{m}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}}_{\alpha=1,\dots,d}$. Thus, space $\tilde{\mathscr{T}}^d_{\boldsymbol{N}}$ can be possibly characterized with the Fourier coefficients as $\tilde{\mathscr{T}}^d_{\boldsymbol{N}} = \{\sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \hat{\mathbf{v}}^{\boldsymbol{m}}_{\boldsymbol{N}} \varphi_{\boldsymbol{m}} : \hat{\mathbf{v}}_{\boldsymbol{N}} \in \mathsf{F}_{\boldsymbol{N}}(\mathbb{R}^{d\times\boldsymbol{N}})\}$.

Remark 4.13. *The trigonometric polynomials are real valued if the Fourier coefficients obey conjugate symmetry $\hat{\boldsymbol{v}}(\boldsymbol{n}) = \overline{\hat{\boldsymbol{v}}(-\boldsymbol{n})}$, $\boldsymbol{n} \in \mathbb{Z}^d$; from definition, it is valid for the trigonometric polynomials $\mathscr{T}^d_{\boldsymbol{N}} \subset L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$.*

The peculiar situation occurs for the space $\tilde{\mathscr{T}}^d_{\boldsymbol{N}}$. If $\boldsymbol{N}$ is odd, both spaces coincide $\mathscr{T}^d_{\boldsymbol{N}} = \tilde{\mathscr{T}}^d_{\boldsymbol{N}}$ as the index sets do $\mathbb{Z}^d_{\boldsymbol{N}} = \underline{\mathbb{Z}}^d_{\boldsymbol{N}}$; generally, the inclusion $\mathscr{T}^d_{\boldsymbol{N}} \subseteq \tilde{\mathscr{T}}^d_{\boldsymbol{N}}$ holds. Unfortunately, the space $\tilde{\mathscr{T}}^d_{\boldsymbol{N}}$ fails to be real valued $\tilde{\mathscr{T}}^d_{\boldsymbol{N}} \not\subseteq L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ because the Fourier coefficients with frequencies $\boldsymbol{n} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}} \setminus \mathbb{Z}^d_{\boldsymbol{N}}$ miss opposite counterpart with frequencies $-\boldsymbol{n}$.

4.2.2 Galerkin approximation with numerical integration

Approximate solutions are gained from Galerkin approximation (GA) and Galerkin approximation with numerical integration (GAwNI); both methods can be defined either for weak formulation (2.5) or minimization formulation (2.2). In all cases, formulations are based on trial space $\mathscr{E}$ or $\mathscr{J}$ in the primal or dual formulation resp. The following remark clarifies their finite dimensional relatives.

Remark 4.14. *We confine the description to the problem for the odd number of discretization points $\boldsymbol{N}$ for which trigonometric polynomials $\tilde{\mathscr{T}}^d_{\boldsymbol{N}}$ and $\mathscr{T}^d_{\boldsymbol{N}}$ coincide; thus it is possible to directly follow the results in Paper 5. The case of general number of discretization points can be found in Section 3.2 in Paper 6.*

Hence, we define a conforming finite dimensional spaces as $\mathscr{E}_{\boldsymbol{N}} := \mathscr{T}^d_{\boldsymbol{N}} \bigcap \mathscr{E}$ and $\mathscr{J}_{\boldsymbol{N}} := \mathscr{T}^d_{\boldsymbol{N}} \bigcap \mathscr{J}$ that satisfy the finite dimensional Helmholtz decomposition

$$\tilde{\mathscr{T}}^d_{\boldsymbol{N}} = \mathscr{U} \oplus^{\perp} \mathscr{E}_{\boldsymbol{N}} \oplus^{\perp} \mathscr{J}_{\boldsymbol{N}}.$$

Fully discrete spaces defined as $\mathbb{U}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}[\mathscr{U}]$, $\mathbb{E}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}[\mathscr{E}_{\boldsymbol{N}}]$, $\mathbb{J}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}[\mathscr{J}_{\boldsymbol{N}}]$, with isomorphism from Def. 4.10, are their associates and thus satisfy an analogue

$$\mathcal{I}_{\boldsymbol{N}}[\tilde{\mathscr{T}}^d_{\boldsymbol{N}}] = \mathbb{R}^{d\times \boldsymbol{N}} = \mathbb{U}_{\boldsymbol{N}} \oplus^{\perp} \mathbb{E}_{\boldsymbol{N}} \oplus^{\perp} \mathbb{J}_{\boldsymbol{N}}. \tag{4.9}$$

Definition 4.15 (Galerkin approximation, Def. 3.20 in Paper 5)**.** *Galerkin approximation of the auxiliary problems in Def. 2.6 states: for unitary macroscopic load $\boldsymbol{\epsilon}_\alpha$, find minimizer $\tilde{\boldsymbol{e}}^{(\alpha)}_{\boldsymbol{N}}$ satisfying*

$$\big(\boldsymbol{A}\tilde{\boldsymbol{e}}^{(\alpha)}_{\boldsymbol{N}}, \boldsymbol{v}_{\boldsymbol{N}}\big)_{L^2_{\mathrm{per}}} = -\big(\boldsymbol{A}\boldsymbol{\epsilon}_\alpha, \boldsymbol{v}_{\boldsymbol{N}}\big)_{L^2_{\mathrm{per}}}, \quad \forall \boldsymbol{v}_{\boldsymbol{N}} \in \mathscr{E}_{\boldsymbol{N}}. \tag{4.10}$$

Remark 4.16 (Solution and convergence)**.** *The unique approximate solutions of Galerkin approximation are provided by Lax-Milgram lemma as material coefficients $\boldsymbol{A}$ are uniformly elliptic and bounded. The convergence of the approximate solutions to the minimizers is provided by Cea's lemma*

$$\|\boldsymbol{e}^{(\alpha)} - \boldsymbol{e}^{(\alpha)}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}} \leq C \inf_{\boldsymbol{v}_{\boldsymbol{N}} \in \mathscr{E}_{\boldsymbol{N}}} \|\boldsymbol{e}^{(\alpha)} - \boldsymbol{v}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}} \tag{4.11}$$

together with density of trigonometric polynomials in space L^2_{per}. Although, the convergence can be arbitrarily slow, more regular minimizers permit the order of convergence, see Sec. 3.4 in Paper 5.

Approximate solutions from Galerkin approximation behaves meaningfully, unfortunately, the linear systems determined by GA cannot be obtained in the closed form for general coefficients $\boldsymbol{A}$. Thus, the need for numerical integration arises.

Definition 4.17 (GAwNI)**.** *Let the material coefficients $\boldsymbol{A}$ be continuous. Galerkin approximation with numerical integration of the primal homogenization problem, Def. 2.1, states as: for arbitrary fixed macroscopic load $\boldsymbol{E} \in \mathbb{R}^d$, find discrete homogenized matrix $\boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff},\boldsymbol{N}} \in \mathbb{R}^{d\times d}$ satisfying*

$$\big(\boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff},\boldsymbol{N}}\boldsymbol{E}, \boldsymbol{E}\big)_{\mathbb{R}^d} = \inf_{\boldsymbol{e}_{\boldsymbol{N}} \in \mathscr{E}_{\boldsymbol{N}}} \big(Q_{\boldsymbol{N}}[\boldsymbol{A}(\boldsymbol{E} + \boldsymbol{e}_{\boldsymbol{N}})], \boldsymbol{E} + \boldsymbol{e}_{\boldsymbol{N}}\big)_{L^2_{\mathrm{per}}}. \tag{4.12}$$

It is shown in Lem. 3.28 in Paper 5 and in Lem. 3.22 in Paper 6, that GAwNI is equivalent to the fully discrete formulation.

Definition 4.18 (Fully discrete formulation of GAwNI)**.** *Find* $\boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff},\boldsymbol{N}} \in \mathbb{R}^{d\times d}$ *such that for arbitrary fixed* $\boldsymbol{E} \in \mathbb{R}^d$ *holds*

$$\left(\boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff},\boldsymbol{N}}\boldsymbol{E},\boldsymbol{E}\right)_{\mathbb{R}^d} = \frac{1}{|\boldsymbol{N}|_\Pi}\inf_{\mathbf{e}_{\boldsymbol{N}}\in\mathbb{E}_{\boldsymbol{N}}}\left(\mathbf{A}_{\boldsymbol{N}}(\mathbf{E}_{\boldsymbol{N}}+\mathbf{e}_{\boldsymbol{N}}),\mathbf{E}_{\boldsymbol{N}}+\mathbf{e}_{\boldsymbol{N}}\right)_{\mathbb{R}^{d\times\boldsymbol{N}}} \tag{4.13}$$

where $\mathbf{E}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}[\boldsymbol{E}] \in \mathbb{U}_{\boldsymbol{N}}$ *and* $\mathbf{A}_{\boldsymbol{N}} \in \mathbb{R}^{d\times d\times\boldsymbol{N}\times\boldsymbol{N}}$ *with components assembled as* $\mathbf{A}^{\boldsymbol{mn}}_{\boldsymbol{N},\alpha\beta} = \boldsymbol{A}_{\alpha\beta}(\boldsymbol{x}^{\boldsymbol{m}}_{\boldsymbol{N}})\delta_{\boldsymbol{mn}}$ *for* α,β *and* $\boldsymbol{m},\boldsymbol{n} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}$.

Remark 4.19 (Solution of GAwNI by Conjugate gradients)**.** *In Papers 2 and 5, it is explained that the minimizers of fully discrete formulation can be obtained by Conjugate gradients that minimize a quadratic functional on a subspace* $\mathbb{E}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}[\mathscr{E}_{\boldsymbol{N}}]$. *It is equivalent to the solution of linear system* $\mathbf{Cx} = \mathbf{b}$ *defined for particular* α *as*

$$\underbrace{\mathbf{G}^1_0\mathbf{A}_{\boldsymbol{N}}}_{\mathbf{C}}\underbrace{\mathbf{e}^{(\alpha)}_{\boldsymbol{N}}}_{\mathbf{x}} = \underbrace{-\mathbf{G}^1_0\mathbf{A}_{\boldsymbol{N}}\mathbf{E}^{(\alpha)}_{\boldsymbol{N}}}_{\mathbf{b}}$$

for the initial approximation $\mathbf{x}_{(0)} \in \mathbb{E}_{\boldsymbol{N}}$ *from the appropriate subspace. Matrix* $\mathbf{G}^1_0$ *is an orthogonal projection on* $\mathbb{E}_{\boldsymbol{N}}$. *It follows from continuous projection* (4.2) *on* $\mathscr{E}$, *see Sec. 3.2 in Paper 6 providing the scheme of subspaces for both of the primal and the dual formulations.*

Remark 4.20 (Convergence of discrete minimizers)**.** *In Paper 5, the convergence of the approximate solutions of GAwNI to the minimizers of homogenization problem* (2.2) *is provided for sufficiently regular material coefficients* $\boldsymbol{A}$. *It incorporates first Strang's lemma — the standard approach in the finite element method — and the estimates of interpolation operator, Def.* (4.8)*, provided in [43] for one and two-dimensional setting, see also Sec. 3.2 in Paper 5 for d-dimensional setting. Numerical examples confirming the rates of convergence can be found in Paper 6 in Sec. 4.2.*

Remark 4.21. *Regularity of material coefficients* $\boldsymbol{A}$ *principally influences the rates of convergence. For rough coefficients, arbitrary slow convergence can be observed for GA. However, GAwNI is even difficult to define as interpolation operator* $Q_{\boldsymbol{N}}$ *requires function values — it requires some type of continuity. In Paper 5, we remedy this trouble by smoothing of the material coefficients.*

4.3 Guaranteed bounds by FFT-based homogenization

This section summarizes results obtained in Paper 6 for guaranteed bounds on the homogenized matrix, see summary in Sec. 4.3, that were introduced by Dvořák in [12, 13] for a scalar problem and later independently by Wieçkowski [55] for linear elasticity.

The upper-lower bounds are based on the primal and dual formulations of homogenization problem in Def. 2.1. In Sec. 4.3.1, the connection of the primal and dual formulations is investigated in fully discrete setting. Then in Sec. 4.3.2, the method for an effective evaluation of the upper-lower bounds is provided.

Galerkin approximation with numerical integration of the homogenization problem in Def. (2.1) leads to a fully discrete homogenization problem in both primal and dual setting

$$\big(\tilde{\boldsymbol{A}}^{\mathrm{FFTH}}_{\mathrm{eff},\boldsymbol{N}}\boldsymbol{E},\boldsymbol{E}\big)_{\mathbb{R}^d} = \frac{1}{|\boldsymbol{N}|_\Pi}\inf_{\mathbf{e}_{\boldsymbol{N}}\in\mathbb{E}_{\boldsymbol{N}}}\big(\mathbf{A}_{\boldsymbol{N}}(\mathbf{E}_{\boldsymbol{N}}+\mathbf{e}_{\boldsymbol{N}}),\mathbf{E}_{\boldsymbol{N}}+\mathbf{e}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}}, \tag{4.14a}$$

$$\big((\tilde{\boldsymbol{A}}^{\mathrm{FFTH}}_{\mathrm{eff,D},\boldsymbol{N}})^{-1}\boldsymbol{J},\boldsymbol{J}\big)_{\mathbb{R}^d} = \frac{1}{|\boldsymbol{N}|_\Pi}\inf_{\mathbf{j}_{\boldsymbol{N}}\in\mathbb{J}_{\boldsymbol{N}}}\big(\mathbf{A}^{-1}_{\boldsymbol{N}}(\mathbf{J}_{\boldsymbol{N}}+\mathbf{j}_{\boldsymbol{N}}),\mathbf{J}_{\boldsymbol{N}}+\mathbf{j}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}}, \tag{4.14b}$$

compare to Def. 4.18. Minimizers $\boldsymbol{e}^{(\alpha)}_{\boldsymbol{N}} := \mathcal{I}^{-1}_{\boldsymbol{N}}[\mathbf{e}^{(\alpha)}_{\boldsymbol{N}}]$ and $\boldsymbol{\jmath}^{(\alpha)}_{\boldsymbol{N}} := \mathcal{I}^{-1}_{\boldsymbol{N}}[\mathbf{j}^{(\alpha)}_{\boldsymbol{N}}]$ are then used to evaluate the upper-lower bounds of the homogenized matrix.

4.3.1 Connection of primal and dual formulation

The connection between formulations in (4.14) is summarized for odd number of discretization points.

Theorem 4.22 (Primal-dual formulation for odd number of discretization points, Corollary 3.27 in Paper 6)**.** *Let the number of discretization points $\boldsymbol{N}\in\mathbb{N}^d$ be odd. Then, the fully discrete formulations satisfy:*

1. *Both primal and dual homogenized matrices coincide $\boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff},\boldsymbol{N}} = \boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff,D},\boldsymbol{N}}$.*

2. *Primal and dual discrete minimizers $\tilde{\mathbf{e}}^{(\alpha)}_{\boldsymbol{N}}$ and $\tilde{\mathbf{j}}^{(\alpha)}_{\boldsymbol{N}}$ are related as*

$$\boldsymbol{\epsilon}_\beta + \tilde{\mathbf{j}}^{(\beta)}_{\boldsymbol{N}} = \mathbf{A}_{\boldsymbol{N}}\sum_\alpha E_\alpha(\boldsymbol{\epsilon}_\alpha + \tilde{\mathbf{e}}^{(\alpha)}_{\boldsymbol{N}}) \tag{4.15}$$

where $\boldsymbol{E} = (\boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff},\boldsymbol{N}})^{-1}\boldsymbol{\epsilon}_\beta$.

This enables to avoid the solution of the dual formulation in order to obtain the dual minimizers. The proof is the consequence of perturbation duality theorem [14], that is mentioned for a discrete setting in Lem. 3.26 in Paper 6, and discrete version of the Helmholtz decomposition (4.9).

Unfortunately, the previous theorem fails to hold for the general number of discretization points as the discrete version of the Helmholtz decomposition (4.9) is no longer valid. A general theorem is provided in Cor. 3.27 in Paper 6.

4.3.2 Calculation of bounds

The calculation of the upper-lower bounds of the homogenized matrix consists of the integral evaluation of type $\big(\boldsymbol{A}\boldsymbol{e}^{(\alpha)}_{\boldsymbol{N}},\boldsymbol{e}^{(\beta)}_{\boldsymbol{N}}\big)_{L^2_{\mathrm{per}}}$ occurring in Def. 2.12.Generally, the integral cannot be evaluated in a closed form because of non-specific material coefficients. The idea is to adjust the material coefficients to calculate the integrals accurately and efficiently and simultaneously keep the upper-lower bounds structure.

For an easier orientation among various homogenized matrices, we refer to their scheme in Fig. 2. The matrices $\boldsymbol{A}_{\mathrm{eff}}$, $\underline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}$, $\overline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}$, $\boldsymbol{A}_{\mathrm{eff},\boldsymbol{N}}$, and $\boldsymbol{D}_{\boldsymbol{N}}$ already introduced in Sec. 4.3 are in no relation to matrices $\boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff},\boldsymbol{N}}$, $\boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff,D},\boldsymbol{N}}$ from fully discrete formulation (4.14) because of the variational crime caused by numerical integration in Def. 4.17.

$$
\begin{array}{ccccccccccc}
 & & & & \tilde{\boldsymbol{A}}^{\mathrm{lin},M}_{\mathrm{eff,D},\boldsymbol{N}} & & & & \tilde{\boldsymbol{A}}^{\mathrm{lin},M}_{\mathrm{eff},\boldsymbol{N}} & & \\
 & & & & \wr & & & & \wr & & \\
0 & \preceq & \underline{\underline{\boldsymbol{A}}}^{\mathrm{con},M}_{\mathrm{eff},\boldsymbol{N}} & \preceq & \underline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}} & \preceq & \boldsymbol{A}_{\mathrm{eff}} & \preceq & \overline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}} & \preceq & \overline{\overline{\boldsymbol{A}}}^{\mathrm{con},M}_{\mathrm{eff},\boldsymbol{N}} \\
 & & & & \| & & & & \| & & \\
 & & & & \boldsymbol{A}_{\mathrm{eff},\boldsymbol{N}} - \boldsymbol{D}_{\boldsymbol{N}} & \preceq & \boldsymbol{A}_{\mathrm{eff},\boldsymbol{N}} & \preceq & \boldsymbol{A}_{\mathrm{eff},\boldsymbol{N}} + \boldsymbol{D}_{\boldsymbol{N}} & & \\
 & & & & \nleq & & & & \nleq & & \\
 & & & & \boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff,D},\boldsymbol{N}} & \underset{\preceq}{\text{equality if } \boldsymbol{N} \text{ is odd}} & & & \boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff},\boldsymbol{N}} & &
\end{array}
$$

Figure 2: The overview of homogenized material bounds

In Paper 6 in Sec. 3.5, we introduce the approximation of upper-lower bounds $\tilde{\boldsymbol{A}}^{\mathrm{lin},\boldsymbol{M}}_{\mathrm{eff,D},\boldsymbol{N}}$, $\tilde{\boldsymbol{A}}^{\mathrm{lin},\boldsymbol{M}}_{\mathrm{eff},\boldsymbol{N}}$ based on piecewise bilinear material coefficients. Next, piecewise constant material coefficients defined in a way to guaranty bounds produce upper-lower bounds of homogenized matrix, i.e. $\underline{\underline{A}}^{\mathrm{con},\boldsymbol{M}}_{\mathrm{eff},\boldsymbol{N}}$, $\overline{\overline{A}}^{\mathrm{con},\boldsymbol{M}}_{\mathrm{eff},\boldsymbol{N}}$. The next lemma shows sufficient condition to guaranty bounds with adjusted material coefficients.

Lemma 4.23 (Sufficient condition for the upper-lower bound structure, Lem. 3.31 in Paper 6)**.** *Let $\boldsymbol{A} \in L^\infty_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^{d\times d})$ be material coefficients and $\overline{\boldsymbol{A}}, \underline{\boldsymbol{A}} \in L^\infty_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^{d\times d})$ its upper and lower approximations satisfying*

$$\underline{\boldsymbol{A}}(\boldsymbol{x}) \preceq \boldsymbol{A}(\boldsymbol{x}) \preceq \overline{\boldsymbol{A}}(\boldsymbol{x}), \quad \text{for almost all } \boldsymbol{x} \in \mathcal{Y}. \tag{4.16}$$

Let $\tilde{\boldsymbol{e}}^{(\alpha)}_{\boldsymbol{N}} \in \mathscr{E}_{\boldsymbol{N}}$ and $\tilde{\boldsymbol{j}}^{(\alpha)}_{\boldsymbol{N}} \in \mathscr{J}_{\boldsymbol{N}}$ be unitary minimizers for material coefficients $\boldsymbol{A}$, cf. Def. 2.6. Then matrices $\overline{\overline{\boldsymbol{A}}}_{\mathrm{eff}}, \underline{\underline{\boldsymbol{A}}}_{\mathrm{eff}} \in \mathbb{R}^{d\times d}$, defined as

$$(\overline{\overline{\boldsymbol{A}}}_{\mathrm{eff},\boldsymbol{N}})_{\alpha\beta} = \big(\overline{\boldsymbol{A}}(\boldsymbol{\epsilon}_\beta + \tilde{\boldsymbol{e}}^{(\beta)}_{\boldsymbol{N}}), \boldsymbol{\epsilon}_\alpha + \tilde{\boldsymbol{e}}^{(\alpha)}_{\boldsymbol{N}}\big)_{L^2_{\mathrm{per}}}, \tag{4.17a}$$

$$(\underline{\underline{\boldsymbol{A}}}^{-1}_{\mathrm{eff},\boldsymbol{N}})_{\alpha\beta} = \big(\underline{\boldsymbol{A}}^{-1}(\boldsymbol{\epsilon}_\beta + \tilde{\boldsymbol{j}}^{(\beta)}_{\boldsymbol{N}}), \boldsymbol{\epsilon}_\alpha + \tilde{\boldsymbol{j}}^{(\alpha)}_{\boldsymbol{N}}\big)_{L^2_{\mathrm{per}}}, \tag{4.17b}$$

comply with the upper-lower bound structure, i.e.

$$\underline{\underline{\boldsymbol{A}}}_{\mathrm{eff},\boldsymbol{N}} \preceq \underline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}} \preceq \boldsymbol{A}_{\mathrm{eff}} \preceq \overline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}} \preceq \overline{\overline{\boldsymbol{A}}}_{\mathrm{eff},\boldsymbol{N}}.$$

A special situation occurs when the material coefficients are expressed as the linear combination of some functions placed at nodal points of regular grid. Then the integrals required for evaluating upper-lower bounds can be calculated by FFT algorithm, see Eq. (4.19) in next lemma and Lem. 3.33 in Paper 6.

Lemma 4.24 (Calculation of homogenized matrices, Lem. 3.32 in Paper 6)**.** *Let $\boldsymbol{u}_{\boldsymbol{N}}, \boldsymbol{v}_{\boldsymbol{N}} \in \mathscr{T}^d_{\boldsymbol{N}}$ be trigonometric polynomials and $\boldsymbol{A}_{\boldsymbol{M}} \in L^\infty_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^{d\times d}_{\mathrm{spd}})$ for $\boldsymbol{M} \in \mathbb{N}^d$ be function explicitly expressed as*

$$\boldsymbol{A}_{\boldsymbol{M}}(\boldsymbol{x}) = \sum_{\boldsymbol{n}\in\mathbb{Z}^d_{\boldsymbol{M}}} \psi(\boldsymbol{x} + \boldsymbol{x}^{\boldsymbol{n}}_{\boldsymbol{M}})\mathbf{A}^{\boldsymbol{n}}, \quad \boldsymbol{x} \in \mathcal{Y}$$

where $\psi \in L^\infty_{\mathrm{per}}(\mathcal{Y};\mathbb{R})$ is some basis function and $\mathbf{A} \in \mathbb{R}^{d\times d\times \boldsymbol{M}}$. Then the integrals of the type occurring in Eq. (4.17) can be calculated as

$$\big(\boldsymbol{A}_{\boldsymbol{M}}\boldsymbol{u}_{\boldsymbol{N}}, \boldsymbol{v}_{\boldsymbol{N}}\big)_{L^2_{\mathrm{per}}} = \frac{1}{|\mathcal{Y}|_d} \sum_{\alpha,\beta} \sum_{\boldsymbol{m}\in\mathbb{Z}^d_{2\boldsymbol{N}}} w(\boldsymbol{m}) \widehat{\mathsf{u}}\widehat{\mathsf{v}}^{\boldsymbol{m}}_{\boldsymbol{N},\alpha,\beta} \hat{\mathsf{A}}^{\boldsymbol{m}}_{\alpha\beta} \tag{4.18}$$

where the integration weight $w(\boldsymbol{m})$ *is defined as* $w(\boldsymbol{m}) := \int_{\mathcal{Y}} \psi(\boldsymbol{x})\varphi_{\boldsymbol{m}}(\boldsymbol{x})$ *and factors* $\widehat{\mathsf{uv}}^{\boldsymbol{m}}_{\boldsymbol{N},\beta,\alpha}$, $\hat{\mathsf{A}}^{\boldsymbol{m}}_{\alpha\beta}$ *are defined as*

$$\widehat{\mathsf{uv}}^{\boldsymbol{m}}_{\boldsymbol{N},\beta,\alpha} = \frac{1}{2|\boldsymbol{N}|_\Pi} \sum_{\boldsymbol{k}\in\underline{\mathbb{Z}}^d_{2\boldsymbol{N}}} \boldsymbol{u}_{\boldsymbol{N},\beta}(\boldsymbol{x}^{\boldsymbol{k}}_{2\boldsymbol{N}})\boldsymbol{v}_{\boldsymbol{N},\alpha}(\boldsymbol{x}^{\boldsymbol{k}}_{2\boldsymbol{N}})\omega^{-\boldsymbol{mk}}_{2\boldsymbol{N}} \tag{4.19a}$$

$$\hat{\mathsf{A}}^{\boldsymbol{m}}_{\alpha\beta} = \sum_{\boldsymbol{n}\in\underline{\mathbb{Z}}^d_{\boldsymbol{M}}} \mathsf{A}^{\boldsymbol{n}}_{\alpha\beta}\omega^{-\boldsymbol{mn}}_{\boldsymbol{M}} \tag{4.19b}$$

Investigation of previous lemma reveals that functions ψ are chosen to have analytical expression of integral weights $w(\boldsymbol{m}) := \int_{\mathcal{Y}} \psi(\boldsymbol{x})\varphi_{\boldsymbol{m}}(\boldsymbol{x})$. A reasonable choice is a constant and a bilinear function

$$\chi_{\boldsymbol{N}}(\boldsymbol{x}) = \begin{cases} 1, & |x_\alpha| < \frac{Y_\alpha}{M_\alpha} \text{ for all } \alpha \\ 0, & \text{otherwise} \end{cases}, \quad \operatorname{tri}_{\boldsymbol{N}}(\boldsymbol{x}) = \prod_\alpha \max\{1 - |\frac{x_\alpha M_\alpha}{2Y_\alpha}|, 0\},$$

leading to the weights

$$w^0_{\boldsymbol{N}}(\boldsymbol{m}) := \prod_\alpha \frac{2Y_\alpha}{M_\alpha} \operatorname{sinc}\left(\frac{m_\alpha}{M_\alpha}\right) \qquad w^1_{\boldsymbol{N}}(\boldsymbol{m}) := \prod_\alpha \frac{2Y_\alpha}{M_\alpha} \operatorname{sinc}^2\left(\frac{m_\alpha}{M_\alpha}\right)$$

with function $\operatorname{sinc}(x) := \begin{cases} 1, & x = 0 \\ \frac{\sin(\pi x)}{\pi x}, & x \neq 0 \end{cases}$. Another example of the function with the analytical expression of the weight is a circle function with the weight based on the Bessel function.

More details about calculation of upper-lower bounds can be found in Sec. 3.5 in Paper 6.

4.4 Numerical experiments

In this section, numerical experiments presented in Papers 1,2, and 6 are summarized. Sec. 4.4.1 is dedicated to the acceleration of the FFT-based homogenization by Conjugate gradients, while Sec. 4.4.2 is dedicated to the guaranteed bounds of the homogenized matrix.

4.4.1 Acceleration by Conjugate gradients

In this section, the theoretical result about the acceleration of the original FFT-based homogenization by Conjugate gradients is validated. We show that Conjugate gradients are independent on reference conductivity $\boldsymbol{A}^0$, the parameter of Lippmann-Schwinger equation. Moreover, both methods, the original and accelerated one, are compared in terms of computational times.

From Paper 2, a three-dimensional electric conduction in a cubic periodic unit cell $\mathcal{Y} = \prod_{\alpha=1}^3 (-\frac{1}{2}, \frac{1}{2})$ is chosen as a model problem. The conductivity parameters are defined as

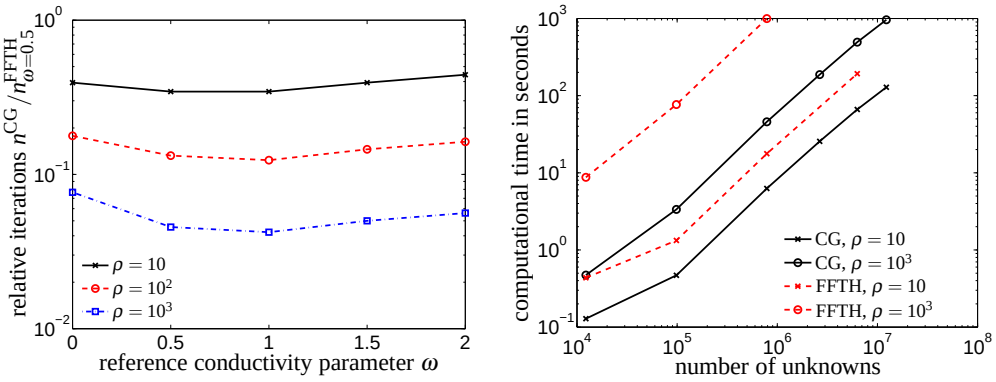
$$\boldsymbol{A}(\boldsymbol{x}) = \begin{cases} \rho\boldsymbol{I}, & \|\boldsymbol{x}\|_2 < (\frac{3}{16\pi})^{\frac{1}{3}} \\ \begin{pmatrix} 1 & 0.2 & 0.2 \\ 0.2 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{pmatrix}, & \text{otherwise} \end{cases}$$

where $\rho > 0$ denotes the contrast of phase conductivities; it represents a two-phase medium with spherical inclusions of 25% volume fraction. We consider the macroscopic field $\boldsymbol{E} := \boldsymbol{\epsilon}_1$ and discretize the unit cell with $\boldsymbol{N} = [n, n, n]$ nodes[10]. The conductivity of the homogeneous reference medium $\boldsymbol{A}^0 \in \mathbb{R}^{d\times d}$ is parametrized as

$$\boldsymbol{A}^0 = \lambda \boldsymbol{I}, \qquad \lambda = 1 - \omega + \rho\omega, \tag{4.20}$$

where $\omega \approx 0.5$ delivers the optimal convergence of the original Moulinec-Suquet Fast-Fourier Transform-based Homogenization (FFTH) algorithm [32].

We first investigate the sensitivity of Conjugate Gradients (CG) to the choice of the reference medium. The results appear in Fig. 3(a) plotting the relative number of iterations for CG against the conductivity of the reference medium parametrized by ω, recall Eq. (4.20). As expected, CG solver achieve a significant improvement over FFTH method as it requires about 40% iterations of FFTH for a mildly-contrasted composite down to 4% for $\varrho = 10^3$. The minor differences visible especially for $\rho = 10^3$ can be therefore attributed to accumulation of round-off errors.



(a) (b)

Figure 3: (a) Relative number of iterations as a function of the reference medium parameter ω and (b) computational time as a function of the number of unknowns.

In Fig. 3(b), we present the total computational time[11] as a function of the number of degrees of freedom and the phase ratio ρ. The results confirm that the computational times scales linearly with the increasing number of degrees of freedom for both schemes for fixed phase ratio ρ [57]. The ratio of the computational time for CG and FFTH algorithms remains almost constant, which indicates that the cost of a single iteration of CG and FFTH method is comparable.

In addition, the memory requirements of both schemes are also comparable. This aspect emphasized the major advantage of the short-recurrence CG-based scheme over alternative schemes for non-symmetric systems, such as GMRES. Finally, we note that finer discretizations can be treated by a straightforward parallel implementation.

[10] In particular, n was taken consequently as $16, 32, 64, 128$ and 160 leading up to $3 \cdot 160^3 \doteq 12.2 \times 10^6$ unknowns

[11] The problem was solved with a MATLAB® in-house code on a machine Intel® Core™2 Duo 3 GHz CPU, 3.28 GB computing memory with Debian linux 5.0 operating system.

4.4.2 Guaranteed bounds

In this section, the properties of the upper-lower bounds of the homogenized matrices are validated. We consider 2-dimensional model problem from Paper 6 with material coefficients defined on a periodic unit cell $\mathcal{Y} = \Pi_{\alpha=1}^{2}(-1,1) \subset \mathbb{R}^2$ as

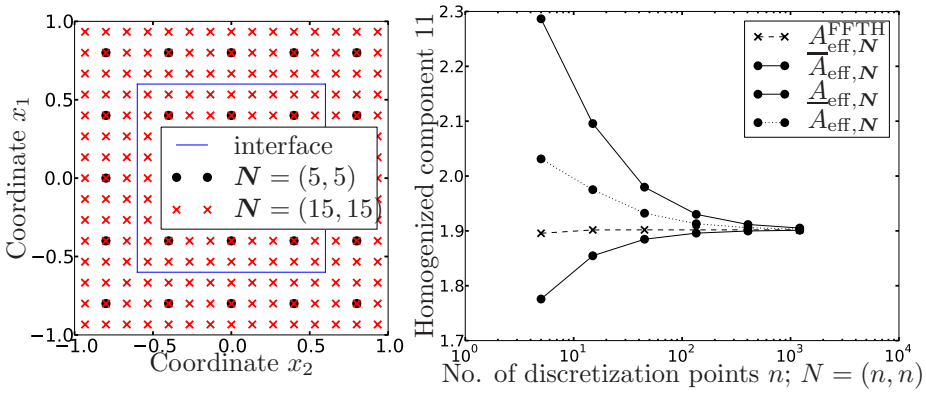
$$\boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{I}[1 + 10f(\boldsymbol{x})], \quad \boldsymbol{x} \in \mathcal{Y},$$

where $\boldsymbol{I} \in \mathbb{R}^{d\times d}$ is identity matrix and $f : \mathcal{Y} \to \mathbb{R}$ is a scalar nonnegative function defined explicitly as

$$f(\boldsymbol{x}) = \begin{cases} 1, & \|\boldsymbol{x}\|_\infty < \frac{3}{4} \\ 0, & \text{otherwise} \end{cases}.$$

The problem is discretized with odd number of discretization points $\boldsymbol{N} = (n, n)$ where $n \in \{5, 15, 45, 135, 405, 1215\}$.

Fig. 4(a) shows the periodic unit cell with the interface between phases and the nodal points sets, $\{\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{n}} \in \mathcal{Y} : \boldsymbol{n} \in \underline{\mathbb{Z}}_{\boldsymbol{N}}^{d}\}$, for particular $\boldsymbol{N}$. Next in Fig. 4(b), we demonstrate the properties of the homogenized matrices for their particular diagonal component. The inequality $\underline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}} \preceq \overline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}}$ stated in (2.11) is satisfied and the error, difference between them, is approaching zero supporting Lem. 2.14.



(a) PUC with nodal points (b) bounds of homogenized coefficient

Figure 4: Periodic unit cell with nodal points and guaranteed bounds of homogenized material coefficient

In this case, approximate homogenized matrix $\boldsymbol{A}_{\text{eff},\boldsymbol{N}}^{\text{FFTH}}$ resembles the real homogenized coefficients $\boldsymbol{A}_{\text{eff}}$ properly even for small $\boldsymbol{N}$ compared to the mean value $\boldsymbol{A}_{\text{eff},\boldsymbol{N}} = \frac{1}{2}(\underline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}} + \overline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}})$ that overestimates. It is the consequence of good approximation of inclusion topology by nodal points, see Fig. 4(a); the interface lies exactly between the nodes. Generally, the approximate homogenized matrix $\boldsymbol{A}_{\text{eff},\boldsymbol{N}}^{\text{FFTH}}$ can be located either over or under the bounds — for a more detail, see Sec. 4.1 in Paper 6.

4.5 Applications

In this section, applications of FFT-based homogenization to linear elasticity are discussed according to Papers 3 and 4.

Linear elasticity is certainly more complicated than the scalar problem presented so far, however, the similar structure is observed. The complexity consists of more complicated material law between second order tensors, strain and stress, through fourth order stiffness tensor. Moreover, Green function and derived projection operator to admissible fields — analogue to (2.9) and (4.1) — are more complicated as it naturally splits into two parts, one corresponding to volumetric fields and second to deviatoric fields.

Paper 3 deals with FFT-based homogenization in the framework of representative volume element to produce homogenized stiffness matrix of cement paste, gypsum, and aluminum alloy. A grid nanoindentation is used for the determination of phase properties in grid points at microscale; the direct utilization of grid data simplifies the numerical evaluation and contributes to its efficiency. The method is compared to some simple analytical homogenization procedures with material phases obtained by a statistical deconvolution.

Paper 4 describes a technique to homogenize highly porous aluminium foam — porosity causes the violation of uniform ellipticity. Although, some modifications of FFT-based homogenization were proposed [27, 29] to overcome this difficulty, no rigorous convergence theory has been provided yet. Thus, the FFT-based homogenization is used as a part of two step homogenization, particularly, it is incorporated at a lower scale composed of aluminum melt with admixtures. A higher scale containing significant volume fraction of air voids exceeding 90% is homogenized using two-dimensional Finite element method.

5 Conclusions

The most important results addressing the questions in Sec. 3 are summarized:

1. It has been shown that Lippmann-Schwinger equation is equivalent, in the sense the unique solution coincide, to the corresponding weak formulation for an arbitrary parameter $\boldsymbol{A}^0 \in \mathbb{R}^{d\times d}_{\mathrm{spd}}$.

2. The discretization of weak formulation (2.5) has been proposed to reproduce the original Moulinec-Suquet numerical algorithm, Rem 2.11; it can be newly considered as the Finite element method with trigonometric polynomials basis functions. Convergence of approximate solutions to continous one has been proven. Moreover, the successful application of Conjugate gradients to non-symmetric linear system has been explained.

3. The method for guaranteed bounds of the homogenized matrix has been used for FFT-based homogenization; FFT algorithm can be utilized for the evaluation. The solution of the dual formulation, that is required for lower bound, can be avoided for the odd number of discretization points.

The presented results clarify FFT-based homogenization in a way to be further analyzed by standard mathematical instruments. Main areas of investigation and employment of results are:

- preconditioning of linear system coming from discretization,

- providing theory for linear elasticity,
- providing theory for porous materials,
- generalization to another physical problems (Stokes problem),
- validation of other homogenization methods with the help of reliable guaranteed bounds.

6 References

[1] N. S. Bakhvalov and A. V. Knyazev, *Efficient computation of averaged characteristics of composites of a periodic structure of essentially different materials*, Soviet mathematics - Doklady **42** (1991), no. 1, 57–62.

[2] T. Belytschko, T.J.R. Hughes, N. Patankar, C.T. Herakovich, and E.C. Bakis, *Research directions in computational and composite mechanics*, Tech. report, United States National Committee on Theoretical and Applied Mechanics, 2007.

[3] A. Bensoussan, G. Papanicolau, and J.L. Lions, *Asymptotic analysis for periodic structures*, vol. 5, North Holland, 1978.

[4] D. Braess, *Finite elements: Theory, fast solvers, and applications in solid mechanics*, Cambridge University Press, 2001.

[5] H. Brézis, *Functional analysis, Sobolev spaces and partial differential equations*, Universitext Series, Springer, 2010.

[6] S. Brisard and L. Dormieux, *FFT-based methods for the mechanics of composites: A general variational framework*, Computational Materials Science **49** (2010), no. 3, 663–671.

[7] ———, *Combining Galerkin approximation techniques with the principle of Hashin and Shtrikman to derive a new FFT-based numerical method for the homogenization of composites*, Computer Methods in Applied Mechanics and Engineering (2012).

[8] B. Budiansky, *On the elastic moduli of some heterogeneous materials*, Journal of the Mechanics and Physics of Solids **13** (1965), no. 4, 223–227.

[9] C. S. Burrus and T. W. Parks, *DFT/FFT and convolution algorithms*, Wiley, New York, 1985.

[10] P.G. Ciarlet, *Handbook of numerical analysis: Finite element methods*, vol. II, Elsevier Science Publishers B.V. (North-Holland), 1991.

[11] D. Cioranescu and P. Donato, *An introduction to homogenization*, Oxford Lecture Series in Mathematics and Its Applications, Oxford University Press, 1999.

[12] J. Dvořák, *Optimization of composite materials*, Master's thesis, The Charles University in Prague, June 1993.

[13] ———, *A reliable numerical method for computing homogenized coefficients*, Tech. report, available at CiteSeerX `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.45.1190`, 1995.

[14] I. Ekeland and R. Temam, *Convex analysis and variational problems*, SIAM, 1976.

[15] 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** (1999), no. 1, 41–47.

[16] R. Hill, *A self-consistent mechanics of composite materials*, Journal of the Mechanics and Physics of Solids **13** (1965), no. 4, 213–222.

[17] H. Hoang-Duc and G. Bonnet, *Effective properties of viscoelastic heterogeneous periodic media: an approximate solution accounting for the distribution of heterogeneities.*, Mechanics of Materials (2012).

[18] V.V. Jikov, S.M. Kozlov, and O.A. Oleinik, *Homogenization of differential operators and integral functionals*, Springer-Verlag, 1994.

[19] L.C.Evans, *Partial differential equations*, vol. 19, American Mathematical Society, 2000.

[20] H. Le Quang, T.L. Phan, and G. Bonnet, *Effective thermal conductivity of periodic composites with highly conducting imperfect interfaces*, International Journal of Thermal Sciences **50** (2011), no. 8, 1428–1444.

[21] R.A. Lebensohn, A.K. Kanjarla, and P. Eisenlohr, *An elasto-viscoplastic formulation based on fast fourier transforms for the prediction of micromechanical fields in polycrystalline materials*, International Journal of Plasticity (2011).

[22] R.A. Lebensohn, A.D. Rollett, and P. Suquet, *Fast fourier transform-based modeling for the determination of micromechanical fields in polycrystals*, JOM Journal of the Minerals, Metals and Materials Society **63** (2011), no. 3, 13–18.

[23] S.B. Lee, RA Lebensohn, and A.D. Rollett, *Modeling the viscoplastic micromechanical response of two-phase materials using Fast Fourier Transforms*, International Journal of Plasticity **27** (2011), no. 5, 707–727.

[24] J. Li, S. Meng, X. Tian, F. Song, and C. Jiang, *A non-local fracture model for composite laminates and numerical simulations by using the FFT method*, Composites Part B: Engineering (2011).

[25] J. Li, X.X. Tian, and R. Abdelmoula, *A damage model for crack prediction in brittle and quasi-brittle materials solved by the FFT method*, International journal of fracture (2012), 1–12.

[26] J. Lukeš and J. Malý, *Measure and integral*, Matfyzpress Prague, 1995.

[27] J. C. Michel, H. Moulinec, and P. Suquet, *A computational method based on augmented Lagrangians and fast Fourier transforms for composites with high contrast*, CMES-Computer Modeling in Engineering & Sciences **1** (2000), no. 2, 79–88.

[28] ———, *A computational scheme for linear and non-linear composites with arbitrary phase contrast*, International Journal for Numerical Methods in Engineering **52** (2001), no. 1–2, 139–160.

[29] V. Monchiet and G. Bonnet, *A polarization-based FFT iterative scheme for computing the effective properties of elastic composites with arbitrary contrast*, International Journal for Numerical Methods in Engineering **89** (2012), no. 11, 1419–1436.

[30] V. Monchiet, G. Bonnet, and G. Lauriat, *A FFT-based method to compute the permeability induced by a stokes slip flow through a porous medium*, Comptes Rendus Mécanique **337** (2009), no. 4, 192–197.

[31] T. Mori and K. Tanaka, *Average stress in matrix and average elastic energy of materials with misfitting inclusions*, Acta metallurgica **21** (1973), no. 5, 571–574.

[32] H. Moulinec and P. Suquet, *A fast numerical method for computing the linear and nonlinear mechanical properties of composites*, Comptes rendus de l'Académie des sciences. Série II, Mécanique, physique, chimie, astronomie **318** (1994), no. 11, 1417–1423.

[33] ———, *A numerical method for computing the overall response of nonlinear composites with complex microstructure*, Computer Methods in Applied Mechanics and Engineering **157** (1997), no. 1–2, 69–94.

[34] ———, *Comparison of FFT-based methods for computing the response of composites with highly contrasted mechanical properties*, Physica B: Condensed Matter **338** (2003), no. 1–4, 58–60.

[35] J. Nečas, *Direct methods in the theory of elliptic equations*, Springer, 2012.

[36] J. Novák, A. Kučerová, and J. Zeman, *Microstructural enrichment functions based on stochastic wang tilings*, arXiv preprint arXiv:1110.4183 (2011).

[37] J. Němeček, V. Králík, and J. Vondřejc, *Micromechanical analysis of heterogeneous structural materials*, Cement and Concrete Composites (2012).

[38] J. Němeček, V. Králík, and J. Vondřejc, *A two-scale micromechanical model for aluminium foam based on results from nanoindentation*, (2012), Submitted.

[39] J.T. Oden, T. Belytschko, I. Babuška, and T.J.R. Hughes, *Research directions in computational mechanics*, Computer Methods in Applied Mechanics and Engineering **192** (2003), no. 7, 913–922.

[40] O.A. Oleinik, A.S. Shamaev, and G.A. Yosifian, *Mathematical problems in elasticity and homogenization*, Studies in Mathematics and Its Applications, North-Holland, 1992.

[41] W. Rudin, *Functional analysis*, McGraw-Hill, Inc., New York, 1991.

[42] ———, *Real and complex analysis*, Tata McGraw-Hill Education, 2006.

[43] J. Saranen and G. Vainikko, *Periodic integral and pseudodifferential equations with numerical approximation*, Springer Monographs Mathematics, 2000.

[44] P. Suquet, H. Moulinec, O. Castelnau, M. Montagnat, N. Lahellec, F. Grennerat, P. Duval, and R. Brenner, *Multi-scale modeling of the mechanical behavior of polycrystalline ice under transient creep*, Procedia IUTAM **3** (2012), 64–78.

[45] J. Sýkora, J. Zeman, and M. Šejnoha, *Selected topics in homogenization of transport processes in historical masonry structures*, arXiv preprint arXiv:1204.6199 (2012).

[46] L. Tartar, *The general theory of homogenization: a personalized introduction*, vol. 7, Springer, 2009.

[47] G. Vainikko, *Fast solvers of the Lippmann-Schwinger equation*, Direct and Inverse Problems of Mathematical Physics (R. P. Gilbert, J. Kajiwara, and Y. S. Xu, eds.), International Society for Analysis, Applications and Computation, vol. 5, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 423–440.

[48] R. Valenta and M. Šejnoha, *Two-step homogenization of asphalt mixtures*, Bulletin of Applied Mechanics **4** (2008), no. 14, 61–66.

[49] V. Vinogradov and G. W. Milton, *An accelerated FFT algorithm for thermoelastic and non-linear composites*, International Journal for Numerical Methods in Engineering **76** (2008), no. 11, 1678–1695.

[50] J. Vondřejc, J. Zeman, and I. Marek, *Analysis of a Fast Fourier Transform based method for modeling of heterogeneous materials*, Lecture Notes in Computer Science **7116** (2012), 512–522.

[51] ———, *FFT-based finite element method for homogenization*, (2013), In preparation.

[52] ———, *Guaranteed bounds of effective material properties using FFT-based FEM*, (2013), In preparation.

[53] V. Šmilauer and Z.P. Bažant, *Identification of viscoelastic C-S-H behavior in mature cement paste by FFT-based homogenization method*, Cement and Concrete Research **40** (2010), no. 2, 197–207.

[54] V. Šmilauer and Z. Bittnar, *Microstructure-based micromechanical prediction of elastic properties in hydrating cement paste*, Cement and Concrete Research **36** (2006), no. 9, 1708–1718.

[55] Z. Wieçkowski, *Dual finite element methods in mechanics of composite materials*, Journal of Theoretical and Applied Mechanics **2** (1995), no. 33, 233–252.

[56] J. Yvonnet, *A fast method for solving microstructural problems defined by digital images: a space lippmann–schwinger scheme*, International Journal for Numerical Methods in Engineering (2012).

[57] 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** (2010), no. 21, 8065–8071.

[58] J. Zeman and M. Šejnoha, *Numerical evaluation of effective elastic properties of graphite fiber tow impregnated by polymer matrix*, Journal of the Mechanics and Physics of Solids **49** (2001), no. 1, 69–90.

7 List of thesis papers

Paper 1 : ISI journal paper

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** (2010), no. 21, 8065–8071.

Paper 2 : Conference paper in SCOPUS

J. Vondřejc, J. Zeman, and I. Marek, *Analysis of a Fast Fourier Transform based method for modeling of heterogeneous materials*, Lecture Notes in Computer Science **7116** (2012), 512–522.

Paper 3 : ISI journal paper (In Press):

J. Němeček, V. Králík, and J. Vondřejc, *Micromechanical analysis of heterogeneous structural materials*, Cement and Concrete Composites (2012).

Paper 4 : Peer review in ISI journal

J. Němeček, V. Králík, and J. Vondřejc, *A two-scale micromechanical model for aluminium foam based on results from nanoindentation*, (2012).

Paper 5 : In preparation for ISI journal

J. Vondřejc, J. Zeman, and I. Marek, *FFT-based finite element method homogenization*, (2013), In preparation.

Paper 6 : In preparation for ISI journal

J. Vondřejc, J. Zeman, and I. Marek, *Guaranteed bounds of effective material properties using FFT-based FEM*, (2013), In preparation.

Part II
Paper 1

Authors:
Jan Zeman, Jaroslav Vondřejc, Jan Novák, and Ivo Marek

Title:
Accelerating a FFT-based solver for numerical homogenization of periodic media by conjugate gradients

Source:
JOURNAL OF COMPUTATIONAL PHYSICS

Volume:
229

Issue:
21

Pages:
8065–8071

DOI:
10.1016/j.jcp.2010.07.010

Accession number:
WOS:000282118500001

Published:
20th October 2010

Journal of Computational Physics 229 (2010) 8065–8071

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

Short note

Accelerating a FFT-based solver for numerical homogenization of periodic media by conjugate gradients

Jan Zeman [a,*], Jaroslav Vondřejc [a], Jan Novák [b], Ivo Marek [c]

[a] Department of Mechanics, Faculty of Civil Engineering, Czech Technical University in Prague, Thákurova 7, 166 29 Prague 6, Czech Republic
[b] Centre for Integrated Design of Advanced Structures, Faculty of Civil Engineering, Czech Technical University in Prague, Thákurova 7, 166 29 Prague 6, Czech Republic
[c] Department of Mathematics, Faculty of Civil Engineering, Czech Technical University in Prague, Thákurova 7, 166 29 Prague 6, Czech Republic

ARTICLE INFO

Article history:
Received 7 April 2010
Received in revised form 30 June 2010
Accepted 9 July 2010
Available online 17 July 2010

Keywords:
Numerical homogenization
FFT-based solvers
Trigonometric collocation method
Conjugate gradient solvers

ABSTRACT

In this short note, we present a new technique to accelerate the convergence of a FFT-based solver for numerical homogenization of complex periodic media proposed by Moulinec and Suquet [1]. The approach proceeds from discretization of the governing integral equation by the trigonometric collocation method due to Vainikko [2], to give a linear system which can be efficiently solved by conjugate gradient methods. Computational experiments confirm robustness of the algorithm with respect to its internal parameters and demonstrate significant increase of the convergence rate for problems with high-contrast coefficients at a low overhead per iteration.

1. Introduction

A majority of computational homogenization algorithms rely on the solution of the unit cell problem, which concerns the determination of local fields in a representative sample of a heterogeneous material under periodic boundary conditions. Currently, the most efficient numerical solvers of this problem are based on discretization of integral equations. In the case of particulate composites with smooth bounded inclusions embedded in a matrix phase, the problem can be reduced to internal interfaces and solved with remarkable accuracy and efficiency by the fast multipole method, see [3, and references therein]. An alternative method has been proposed by Moulinec and Suquet [1] to treat problems with general microstructures supplied in the form of digital images. The algorithm is based on the Neumann series expansion of the inverse to an operator arising in the associated Lippmann–Schwinger equation and exploits the Fast Fourier Transform (FFT) to evaluate the action of the operator efficiently.

The major disadvantage of the FFT-based method consists in its poor convergence for composites exhibiting large jumps in material coefficients. To overcome this difficulty, Eyre and Milton proposed in [4] an accelerated scheme derived from a modified integral equation treated by means of the series expansion approach. In addition, Michel et al. [5] introduced an equivalent saddle-point formulation solved by the Augmented Lagrangian method. As clearly demonstrated in a numerical study by Moulinec and Suquet [6], both methods converge considerably faster than the original variant; the number of iterations is proportional to the square root of the phase contrast instead of the linear increase for the basic scheme. However,

* Corresponding author. Tel.: +420 2 2435 4482; fax: +420 2 2431 0775.
E-mail addresses: zemanj@cml.fsv.cvut.cz (J. Zeman), vondrejc@gmail.com (J. Vondřejc), novakj@cml.fsv.cvut.cz (J. Novák), marek@mbox.ms.cuni.cz (I. Marek).
URL: http://mech.fsv.cvut.cz/~zemanj (J. Zeman).

0021-9991/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2010.07.010

this comes at the expense of increased computational cost per iteration and the sensitivity of the Augmented Lagrangian algorithm to the setting of its internal parameters.

In this short note, we introduce yet another approach to improve the convergence of the original FFT-based scheme [1] based on the trigonometric collocation method [7] and its application to the Helmholtz equation as introduced by Vainikko [2]. We observe that the discretization results in a system of linear equations with a structured dense matrix, for which a matrix–vector product can be computed efficiently using FFT (cf. Section 2). It is then natural to treat the resulting system by standard iterative solvers, such as the Krylov subspace methods, instead of the series expansion technique. In Section 3, the potential of such approach is demonstrated by means of a numerical study comparing the performance of the original scheme and the conjugate- and biconjugate-gradient methods for two-dimensional scalar electrostatics.

2. Methodology

In this section, we briefly summarize the essential steps of the trigonometric collocation-based solution to the unit cell problem by adapting the original exposition of Vainikko [2] to the setting of electrical conduction in periodic composites. In what follows, a, $\boldsymbol{a}$ and $\boldsymbol{A}$ denote scalar, vector and second-order tensor quantities with Greek subscripts used when referring to the corresponding components, e.g. $A_{\alpha\beta}$. Matrices are denoted by a serif font (e.g. A) and a multi-index notation is employed, in which $\mathbb{R}^{\boldsymbol{N}}$ with $\boldsymbol{N} = (N_1,\dots,N_d)$ represents $\mathbb{R}^{N_1\times\cdots\times N_d}$ and $\mathsf{A}^{\boldsymbol{k}}$ stands for the $(k_1,\dots,k_d)$-th element of the matrix $\mathsf{A}\in\mathbb{R}^{\boldsymbol{N}}$.

2.1. Problem setting

We consider a composite material represented by a periodic unit cell $\mathcal{Y} = \prod_{\alpha=1}^{d}(-Y_\alpha, Y_\alpha) \subset \mathbb{R}^d$. In the context of linear electrostatics, the associated unit cell problem reads as

$$\nabla\times\boldsymbol{e}(\boldsymbol{x}) = \boldsymbol{0}, \quad \nabla\cdot\boldsymbol{j}(\boldsymbol{x}) = \boldsymbol{0}, \quad \boldsymbol{j}(\boldsymbol{x}) = \boldsymbol{L}(\boldsymbol{x})\cdot\boldsymbol{e}(\boldsymbol{x}), \quad \boldsymbol{x}\in\mathcal{Y} \tag{1}$$

where $\boldsymbol{e}$ is a $\mathcal{Y}$-periodic vectorial electric field, $\boldsymbol{j}$ denotes the corresponding vector of electric current and $\boldsymbol{L}$ is a second-order positive-definite tensor of electric conductivity. In addition, the field $\boldsymbol{e}$ is subject to a constraint

$$\boldsymbol{e}^0 = \frac{1}{|\mathcal{Y}|}\int_{\mathcal{Y}}\boldsymbol{e}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \tag{2}$$

where $\boldsymbol{e}^0$ denotes a prescribed macroscopic electric field and $|\mathcal{Y}|$ represents the d-dimensional measure of $\mathcal{Y}$.

Next, we introduce a homogeneous reference medium with constant conductivity $\boldsymbol{L}^0$, leading to a decomposition of the electric current field in the form

$$\boldsymbol{j}(\boldsymbol{x}) = \boldsymbol{L}^0\cdot\boldsymbol{e}(\boldsymbol{x}) + \delta\boldsymbol{L}(\boldsymbol{x})\cdot\boldsymbol{e}(\boldsymbol{x}), \quad \delta\boldsymbol{L}(\boldsymbol{x}) = \boldsymbol{L}(\boldsymbol{x}) - \boldsymbol{L}^0. \tag{3}$$

The original problem (1)–(2) is then equivalent to the periodic Lippmann–Schwinger integral equation, formally written as

$$\boldsymbol{e}(\boldsymbol{x}) + \int_{\mathcal{Y}}\boldsymbol{\Gamma}^0(\boldsymbol{x}-\boldsymbol{y})\cdot(\delta\boldsymbol{L}(\boldsymbol{y})\cdot\boldsymbol{e}(\boldsymbol{y}))\,\mathrm{d}\boldsymbol{y} = \boldsymbol{e}^0, \quad \boldsymbol{x}\in\mathcal{Y}, \tag{4}$$

where the $\boldsymbol{\Gamma}^0$ operator is derived from the Green's function of the problem (1)–(2) with $\boldsymbol{L}(\boldsymbol{x}) = \boldsymbol{L}^0$ and $\boldsymbol{e}^0 = \boldsymbol{0}$. Making use of the convolution theorem, Eq. (4) attains a local form in the Fourier space:

$$\hat{\boldsymbol{e}}(\boldsymbol{k}) = \begin{cases} |\mathcal{Y}|^{\frac{1}{2}}\boldsymbol{e}^0, & \boldsymbol{k} = \boldsymbol{0}, \\ -\widehat{\boldsymbol{\Gamma}}^0(\boldsymbol{k})\cdot\widehat{(\delta\boldsymbol{L}\cdot\boldsymbol{e})}(\boldsymbol{k}), & \boldsymbol{k}\in\mathbb{Z}^d\setminus\{\boldsymbol{0}\}, \end{cases} \tag{5}$$

where $\hat{f}(\boldsymbol{k})$ denotes the Fourier coefficient of $f(\boldsymbol{x})$ for the $\boldsymbol{k}$th frequency given by

$$\hat{f}(\boldsymbol{k}) = \int_{\mathcal{Y}} f(\boldsymbol{x})\varphi_{-\boldsymbol{k}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \quad \varphi_{\boldsymbol{k}}(\boldsymbol{x}) = |\mathcal{Y}|^{-\frac{1}{2}}\exp\left(\mathrm{i}\pi\sum_{\alpha=1}^{d}\frac{x_\alpha k_\alpha}{Y_\alpha}\right), \tag{6}$$

"i" is the imaginary unit and

$$\widehat{\boldsymbol{\Gamma}}^0(\boldsymbol{k}) = \begin{cases} \boldsymbol{0}, & \boldsymbol{k} = \boldsymbol{0}, \\ \frac{\boldsymbol{k}\otimes\boldsymbol{k}}{\boldsymbol{k}\boldsymbol{L}^0\cdot\boldsymbol{k}}, & \boldsymbol{k}\in\mathbb{Z}^d\setminus\{\boldsymbol{0}\}, \end{cases} \tag{7}$$

Here, we refer to [4,8] for additional details.

2.2. Discretization via trigonometric collocation

Numerical solution of the Lippmann–Schwinger equation is based on a discretization of a unit cell $\mathcal{Y}$ into a regular periodic grid with $N_1 \times \cdots \times N_d$ nodal points and grid spacings $\boldsymbol{h} = (2Y_1/N_1, \ldots, 2Y_d/N_d)$. The searched field $\boldsymbol{e}$ in (4) is approximated by a trigonometric polynomial $\boldsymbol{e}^{\boldsymbol{N}}$ in the form (cf. [2])

$$\boldsymbol{e}(\boldsymbol{x}) \approx \boldsymbol{e}^{\boldsymbol{N}}(\boldsymbol{x}) = \sum_{\boldsymbol{k} \in \mathbb{Z}^{\boldsymbol{N}}} \hat{\boldsymbol{e}}(\boldsymbol{k}) \varphi_{\boldsymbol{k}}(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{Y}, \tag{8}$$

where $\boldsymbol{N} = (N_1, \ldots, N_d)$, $\hat{\boldsymbol{e}}$ designates the Fourier coefficients defined in (6) and

$$\mathbb{Z}^{\boldsymbol{N}} = \left\{ \boldsymbol{k} \in \mathbb{Z}^d : -\frac{N_\alpha}{2} < k_\alpha \leqslant \frac{N_\alpha}{2}, \alpha = 1, \ldots, d \right\}. \tag{9}$$

We recall, e.g. from [2], that the αth component of the trigonometric polynomial expansion $e^{\boldsymbol{N}}_\alpha$ admits two equivalent finite-dimensional representations. The first one is based on a matrix $\hat{\mathsf{e}}_\alpha \in \mathbb{C}^N$ of the Fourier coefficients of the αth component and equation (8) with $\hat{e}_\alpha(\boldsymbol{k}) = \hat{\mathsf{e}}^{\boldsymbol{k}}_\alpha$. Second, the data can be entirely determined by interpolation of nodal values

$$e^{\boldsymbol{N}}_\alpha(\boldsymbol{x}) = \sum_{\boldsymbol{k} \in \mathbb{Z}^{\boldsymbol{N}}} \mathsf{e}^{\boldsymbol{k}}_\alpha \varphi^{\boldsymbol{N}}_{\boldsymbol{k}}(\boldsymbol{x}), \quad \alpha = 1, \ldots, d \tag{10}$$

where $\mathsf{e}_\alpha \in \mathbb{R}^{\boldsymbol{N}}$ is a matrix storing electric field values at grid points, $\mathsf{e}^{\boldsymbol{k}}_\alpha = e^{\boldsymbol{N}}_\alpha(\boldsymbol{x}^{\boldsymbol{k}})$ is the corresponding value at the $\boldsymbol{k}$th node with coordinates $\boldsymbol{x}^{\boldsymbol{k}} = (k_1 h_1, \ldots, k_d h_d)$ and basis functions

$$\varphi^{\boldsymbol{N}}_{\boldsymbol{k}}(\boldsymbol{x}) = |\boldsymbol{N}|^{-1} \sum_{\boldsymbol{m} \in \mathbb{Z}^N} \exp\left\{ \mathrm{i}\pi \sum_{\alpha=1}^{d} m_\alpha \left(\frac{x_\alpha}{Y_\alpha} - \frac{2k_\alpha}{N_\alpha} \right) \right\} \tag{11}$$

satisfy the Dirac delta property $\varphi^{\boldsymbol{N}}_{\boldsymbol{k}}(\boldsymbol{x}^{\boldsymbol{m}}) = \delta_{\boldsymbol{m}\boldsymbol{k}}$ with $|\boldsymbol{N}| = \prod_{\alpha=1}^{d} N_\alpha$. Both representations can be directly related to each other by

$$\hat{\mathsf{e}}_\alpha = \mathsf{F}\mathsf{e}_\alpha, \quad \mathsf{e}_\alpha = \mathsf{F}^{-1}\hat{\mathsf{e}}_\alpha, \tag{12}$$

where the Vandermonde matrices $\mathsf{F} \in \mathbb{C}^{N \times N}$ and $\mathsf{F}^{-1} \in \mathbb{C}^{N \times N}$

$$\mathsf{F}^{\boldsymbol{k}\boldsymbol{m}} = |\mathcal{Y}|^{-\frac{1}{2}} \exp\left(-\sum_{\alpha=1}^{d} 2\pi \mathrm{i} \frac{k_\alpha m_\alpha}{N_\alpha} \right), \tag{13}$$

$$(\mathsf{F}^{-1})^{\boldsymbol{k}\boldsymbol{m}} = |\mathcal{Y}|^{\frac{1}{2}} |N|^{-1} \exp\left(\sum_{\alpha=1}^{d} 2\pi \mathrm{i} \frac{k_\alpha m_\alpha}{N_\alpha} \right), \tag{14}$$

implement the forward and inverse Fourier transform, respectively, e.g. [9, Section 4.6].

The trigonometric collocation method is based on the projection of the Lippmann–Schwinger equation (4) to the space of the trigonometric polynomials of the form $\{\sum_{\boldsymbol{k} \in \mathbb{Z}^{\boldsymbol{N}}} c_{\boldsymbol{k}} \varphi_{\boldsymbol{k}}, c_{\boldsymbol{k}} \in \mathbb{C}\}$ (cf. [7,2]). In view of Eq. (10), this is equivalent to the collocation at grid points, with the action of $\boldsymbol{\Gamma}^0$ operator evaluated from the Fourier space expression (5) converted to the nodal representation by $(12)_2$. The resulting system of collocation equations reads

$$(\mathsf{I} + \mathsf{B})\mathsf{e} = \mathsf{e}^0, \tag{15}$$

where $\mathsf{e} \in \mathbb{R}^{d \times \boldsymbol{N}}$ and $\mathsf{e}^0 \in \mathbb{R}^{d \times \boldsymbol{N}}$ store the corresponding components of the solution and of the macroscopic field, respectively. Furthermore, I is the $d \times d \times \boldsymbol{N} \times \boldsymbol{N}$ unit matrix and the non-symmetric matrix B can be expressed, for the two-dimensional setting, in the partitioned format as

$$\mathsf{B} = \begin{bmatrix} \mathsf{F}^{-1} & 0 \\ 0 & \mathsf{F}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\Gamma}^0_{11} & \widehat{\Gamma}^0_{12} \\ \widehat{\Gamma}^0_{21} & \widehat{\Gamma}^0_{22} \end{bmatrix} \begin{bmatrix} \mathsf{F} & 0 \\ 0 & \mathsf{F} \end{bmatrix} \begin{bmatrix} \delta\mathsf{L}_{11} & \delta\mathsf{L}_{12} \\ \delta\mathsf{L}_{21} & \delta\mathsf{L}_{22} \end{bmatrix}, \tag{16}$$

with an obvious generalization to an arbitrary dimension. Here, $\widehat{\Gamma}^0_{\alpha\beta} \in \mathbb{R}^{N \times \boldsymbol{N}}$ and $\delta\mathsf{L}_{\alpha\beta} \in \mathbb{R}^{\boldsymbol{N} \times \boldsymbol{N}}$ are diagonal matrices storing the corresponding grid values, for which it holds

$$\left(\widehat{\Gamma}^0_{\alpha\beta}\right)^{\boldsymbol{k}\boldsymbol{k}} = \widehat{\Gamma}^0_{\alpha\beta}(\boldsymbol{k}), \quad \delta\mathsf{L}^{\boldsymbol{k}\boldsymbol{k}}_{\alpha\beta} = \delta L_{\alpha\beta}(\boldsymbol{x}^{\boldsymbol{k}}), \quad \alpha, \beta = 1, \ldots, d \quad \text{and} \quad \boldsymbol{k} \in \mathbb{Z}^{\boldsymbol{N}}. \tag{17}$$

2.3. Iterative solution of collocation equations

It follows from Eq. (16) that the cost of the multiplication by B or by B^{T} is driven by the forward and inverse Fourier transforms, which can be performed in $O(|\boldsymbol{N}|\log|\boldsymbol{N}|)$ operations by FFT techniques. This makes the resulting system (15) ideally suited for iterative solvers.

In particular, the original Fast Fourier Transform-based Homogenization (FFTH) scheme formulated by Moulinec and Suquet in [1] is based on the Neumann expansion of the matrix inverse $(\mathsf{I} + \mathsf{B})^{-1}$, so as to yield the mth iterate in the form

$$\mathsf{e}^{(m)} = \sum_{j=0}^{m} (-\mathsf{B})^j \mathsf{e}^0. \tag{18}$$

Convergence of the series (18) was comprehensively studied in [4,8], where it was shown that the optimal rate of convergence is achieved for

$$\boldsymbol{L}^0 = \frac{\lambda_{\min} + \lambda_{\max}}{2} \boldsymbol{I}, \tag{19}$$

with $\lambda_{\min}$ and $\lambda_{\max}$ denoting the minimum and maximum eigenvalues of $\boldsymbol{L}(\boldsymbol{x})$ on $\mathcal{Y}$ and $\boldsymbol{I}$ being the identity tensor.

Here, we propose to solve the non-symmetric system (15) by well-established Krylov subspace methods, in particular, exploiting the classical Conjugate Gradient (CG) method [10] and the biconjugate gradient (BiCG) algorithm [11]. Even though that CG algorithm is generally applicable to symmetric and positive-definite systems only, its convergence in the one-dimensional setting has been proven by Vondřejc [12, Section 6.2]. A successful application of CG method to a generalized Eshelby inhomogeneity problem has also been recently reported by Novák [13] and Kanaun [14].

3. Results

To assess the performance of the conjugate gradient algorithms, we consider a model problem of the transverse electric conduction in a square array of identical circular particles with 50% volume fraction. A uniform macroscopic field $\boldsymbol{e}^0 = (1,0)$ is imposed on the corresponding single-particle unit cell, discretized by $\boldsymbol{N} = (255,255)$ nodes[1] and the phases are considered to be isotropic with the conductivities set to $\boldsymbol{L} = \boldsymbol{I}$ for the matrix phase and to $\boldsymbol{L} = \varrho\boldsymbol{I}$ for the particle.

The conductivity of the homogeneous reference medium is parameterized as

$$\boldsymbol{L}^0(\omega) = (1 - \omega + \varrho\omega)\boldsymbol{I}, \tag{20}$$

where $\omega = 0.5$ corresponds to the optimal convergence of FFTH algorithm (19). All conjugate gradient-related results have been obtained using the implementations according to [16] and referred to as Algorithm 6.18 (CG method) and Algorithm 7.3 (BiCG scheme). Two termination criteria are considered. The first one is defined for the mth iteration as [15]

$$\left(\eta_{\mathrm{e}}^{(m)}\right)^2 = \frac{\sum_{\boldsymbol{k}\in\mathbb{Z}^N} \left(\boldsymbol{k}\cdot\hat{\mathsf{j}}^{\boldsymbol{k}(m)}\right)^2}{\|\hat{\mathsf{j}}^{\boldsymbol{0}(m)}\|_2^2} \leqslant \varepsilon^2, \tag{21}$$

and provides the test of the equilibrium condition $(1)_2$ in the Fourier space. An alternative expression, related to the standard residual norm for iterative solvers, has been proposed by Vinogradov and Milton in [8] and admits the form

$$\eta_{\mathrm{r}}^{(m)} = \frac{\|\mathsf{L}^0(\mathsf{e}^{(m+1)} - \mathsf{e}^{(m)})\|_2}{\|\mathsf{e}^0\|_2} \leqslant \varepsilon, \tag{22}$$

with the additional L^0 term ensuring the proportionality to (21) at convergence. From the numerical point of view, the latter criterion is more efficient than the equilibrium variant, which requires additional operations per iteration. From the theoretical point of view, its usage is justified only when supported by a convergence result for the iterative algorithm. In the opposite case, the equilibrium norm appears to be more appropriate, in order to avoid spurious non-physical solutions.

3.1. Choice of reference medium and norm

Since no results for the optimal choice of the reference medium are known for (Bi)CG-based solvers, we first estimate their sensitivity to this aspect numerically. The results appear in Fig. 1(a), plotting the relative number of iterations for CG and BiCG solvers against the conductivity of the reference medium parameterized by ω, recall Eq. (20).

As expected, both CG and BiCG solvers achieve a significant improvement over FFTH method in terms of the number of iterations, ranging from 50% for a mildly-contrasted composite down to 2% for $\varrho = 10^4$. Moreover, contrary to all other available methods, the number of iterations is almost independent of the choice of the reference medium. We also observe, in agreement with results in [12, Section 6.2] for the one-dimensional setting, that CG and BiCG algorithms generate identical sequences of iterates; the minor differences visible for $\omega > 1$ or $\varrho = 10^4$ can be therefore attributed to accumulation of round-off errors. These conclusions hold for both equilibrium- and residual-based norms, which appear to be roughly proportional for the considered range of the phase contrasts (cf. Fig. 1(b)). Therefore, the residual criterion (22) will mostly be used in what follows.

In Fig. 2, we supplement the comparison by considering the total CPU time required to achieve a convergence. The data indicate that the cost of one iteration is governed by the matrix–vector multiplication, recall Eq. (16): the overhead of CG scheme is about 10% with respect to FFTH method, while the application of BiCG algorithm, which involves B and $\mathsf{B}^{\mathbf{T}}$ products per iteration [11], is about twice as demanding. As a result, CG algorithm significantly reduces the overall computational

[1] Note that the odd number of discretization points is used to eliminate artificial high-frequency oscillations of the solution in the Fourier space, as reported in [15, Section 2.4].

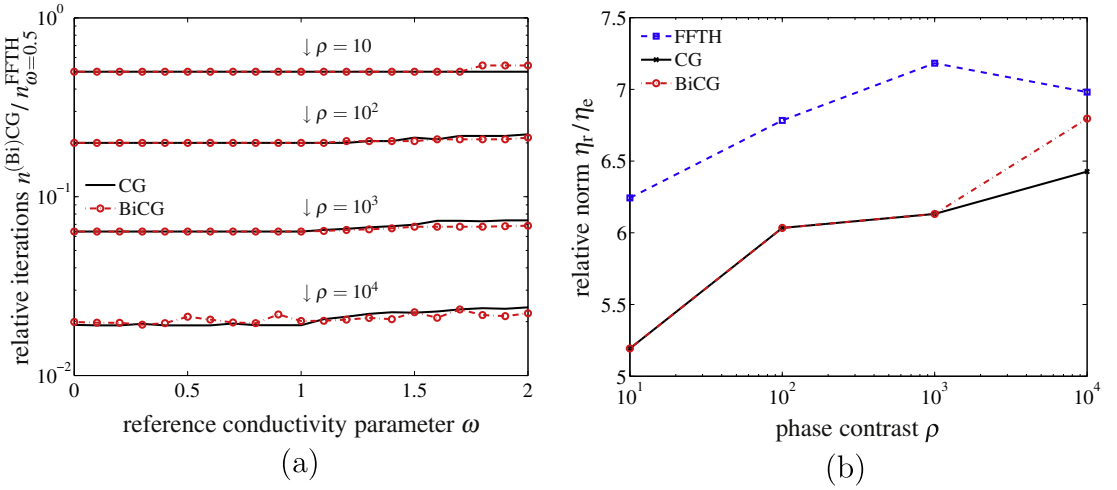


Fig. 1. (a) Relative number of iterations as a function of the reference medium parameter ω and (b) ratio between residual- and equilibrium-based norms at convergence for η_r termination condition with tolerance $\varepsilon = 10^{-4}$.

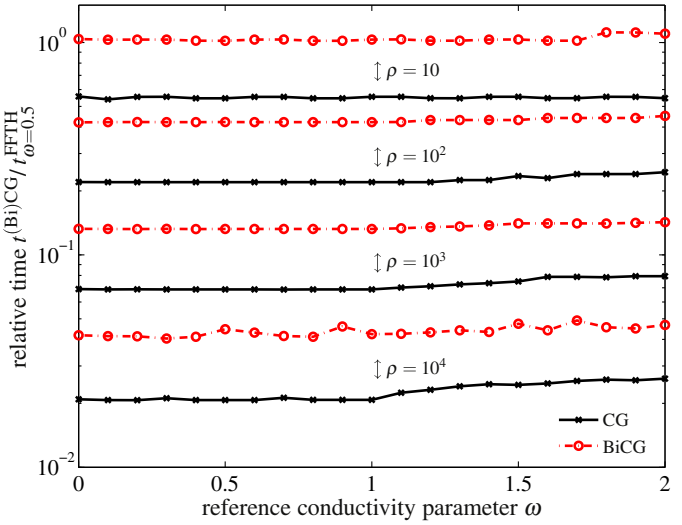


Fig. 2. Relative CPU time t of CG and BiCG solvers plotted against the conductivity parameter ω for η_r-based termination condition with tolerance $\varepsilon = 10^{-4}$.

time in the whole range of contrasts, whereas a similar effect has been reported for the candidate schemes only for $\varrho \geqslant 10^3$ (cf. [6]).

3.2. Influence of phase contrast

As confirmed by all previous works, the phase contrast ϱ is the critical parameter influencing the convergence of FFT-based iterative solvers. In Fig. 3, we compare the scaling of the total number of iterations with respect to phase contrast for CG and FFTH methods, respectively. The results clearly show that the number of iterations grows as $\sqrt{\varrho}$ instead of the linear increase for FFTH method. This follows from error bounds

$$\eta_\mathrm{r}^{(m)} \leqslant \gamma^m \eta_\mathrm{r}^{(0)}, \quad \gamma^{\mathrm{FFTH}} = \frac{\varrho - 1}{\varrho + 1}, \quad \gamma^{\mathrm{CG}} = \frac{\sqrt{\varrho} - 1}{\sqrt{\varrho} + 1}. \tag{23}$$

The first estimate was proven in [4], whereas the second expression is a direct consequence of the condition number of matrix B being proportional to ϱ and a well-known result for the conjugate gradient method, e.g. [16, Section 6.11.3]. The CG-based method, however, failed to converge for the infinite contrast limit. Such behavior is equivalent to the Eyre-Milton

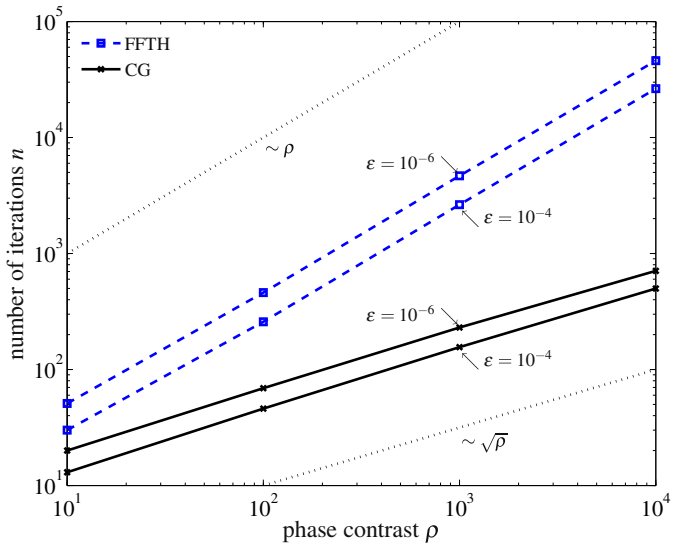


Fig. 3. Total number of iterations n plotted against phase contrast ϱ; the reference medium corresponds to for $\omega = 0.5$ and tolerance ε is related to η_r norm.

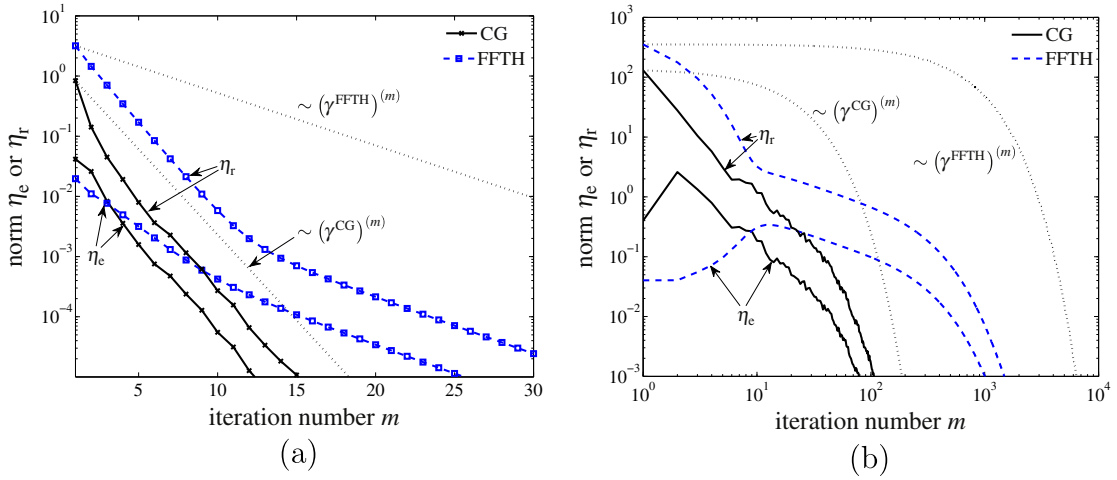


Fig. 4. Convergence progress of CG and FFTH methods for (a) $\varrho = 10^1$ and (b) $\varrho = 10^3$ as quantified by η_e and η_r norms; reference medium corresponds to $\omega = 0.5$ and the dot-and-dahsed curves indicate the convergence rates (23).

scheme [4]. It is, however, inferior to the Augmented Lagrangian algorithm, for which the convergence rate improves with increasing ρ and the method converges even as $\rho \to \infty$. Nonetheless, such results are obtained for optimal, but not always straightforward, choice of the parameters [5].

3.3. Convergence progress

The final illustration of the CG-based algorithm is provided by Fig. 4, displaying a detailed convergence behavior for both low- and high-contrast cases. The results in Fig. 4(a) correspond well with estimates (23) for both residual and equilibrium-based norms. Influence of a higher phase contrast is visible from Fig. 4(b), plotted in the full logarithmic scale. For FFTH algorithm, two regimes can be clearly distinguished. In the first few iterations, the residual error rapidly decreases, but the iterates tend to deviate from equilibrium. Then, both residuals are simultaneously reduced. For CG scheme, the increase of the equilibrium residual appears only in the first iteration and then the method rapidly converges to the correct solution. However, its convergence curve is irregular and the algorithm repeatedly stagnates in two consecutive iterations. Further analysis of this phenomenon remains a subject of future work.

4. Conclusions

In this short note, we have presented a conjugate gradient-based acceleration of the FFT-based homogenization solver originally proposed by Moulinec and Suquet [1] and illustrated its performance on a problem of electric conduction in a periodic two-phase composite with isotropic phases. On the basis of obtained results, we conjecture that:

(i) the non-symmetric system of linear equations (15), arising from discretization by the trigonometric collocation method [2], can be solved using the standard conjugate gradient algorithm,
(ii) the convergence rate of the method is proportional to the square root of the phase contrast,
(iii) the methods fails to converge in the infinite contrast limit,
(iv) contrary to available improvements of the original FFT-solver [4,5], the cost of one iteration remains comparable to the basic scheme and the method is insensitive to the choice of auxiliary reference medium.

The presented computational experiments provide the first step towards further improvements of the method, including a rigorous analysis of its convergence properties, acceleration by multi-grid solvers and preconditioning and the extension to non-linear problems.

Acknowledgements

The authors thank Milan Jirásek (Czech Technical University in Prague) and Christopher Quince (University of Glasgow) for helpful comments on the manuscript. This research was supported by the Czech Science Foundation, through projects No. GAČR 103/09/1748, No. GAČR 103/09/P490 and No. GAČR 201/09/1544, and by the Grant Agency of the Czech Technical University in Prague through project No. SGS10/124/OHK1/2T/11.

References

[1] H. Moulinec, P. Suquet, A fast numerical method for computing the linear and nonlinear mechanical properties of composites, Comptes rendus de l'Académie des sciences. Série II, Mécanique, physique, chimie, astronomie 318 (11) (1994) 1417–1423.
[2] G. Vainikko, Fast solvers of the Lippmann–Schwinger equation, in: R.P. Gilbert, J. Kajiwara, Y.S. Xu (Eds.), Direct and Inverse Problems of Mathematical Physics, International Society for Analysis Applications and Computation, vol. 5, Kluwer Academic Publishers., Dordrecht, The Netherlands, 2000, pp. 423–440.
[3] L. Greengard, J. Lee, Electrostatics and heat conduction in high contrast composite materials, Journal of Computational Physics 211 (1) (2006) 64–76.
[4] D.J. Eyre, G.W. Milton, A fast numerical scheme for computing the response of composites using grid refinement, The European Physical Journal Applied Physics 6 (1) (1999) 41–47.
[5] J.C. Michel, H. Moulinec, P. Suquet, A computational method based on augmented Lagrangians and fast Fourier transforms for composites with high contrast, CMES-Computer Modeling in Engineering and Sciences 1 (2) (2000) 79–88.
[6] H. Moulinec, P. Suquet, Comparison of FFT-based methods for computing the response of composites with highly contrasted mechanical properties, Physica B: Condensed Matter 338 (1–4) (2003) 58–60.
[7] J. Saranen, G. Vainikko, Trigonometric collocation methods with product integration for boundary integral equations on closed curves, SIAM Journal on Numerical Analysis 33 (4) (1996) 1577–1596.
[8] V. Vinogradov, G.W. Milton, An accelerated FFT algorithm for thermoelastic and non-linear composites, International Journal for Numerical Methods in Engineering 76 (11) (2008) 1678–1695.
[9] G. Golub, C.F. Van Loan, Matrix Computations, The Johns Hopkins University Press, Baltimore and London, 1996.
[10] M.R. Hestenes, E. Stiefel, Methods of conjugate gradients for solving linear systems, Journal of Research of the National Bureau of Standards 49 (6) (1952) 409–463.
[11] R. Fletcher, Conjugate gradient methods for indefinite systems, in: G. Watson (Ed.), Numerical Analysis, Proceedings of the Dundee Conference on Numerical Analysis, 1975, Lecture Notes in Mathematics, vol. 506, Springer-Verlag, New York, 1976, pp. 73–89.
[12] J. Vondřejc, Analysis of heterogeneous materials using efficient meshless algorithms: one-dimensional study, Master's thesis, Czech Technical University in Prague (2009). <http://mech.fsv.cvut.cz/~vondrejc/download/ING.pdf>.
[13] J. Novák, Calculation of elastic stresses and strains inside a medium with multiple isolated inclusions, in: M. Papadrakakis, B. Topping (Eds.), Proceedings of the Sixth International Conference on Engineering Computational Technology, Stirlingshire, UK, 2008, pp. 16, paper 127. doi:10.4203/ccp.89.127.
[14] S. Kanaun, Fast calculation of elastic fields in a homogeneous medium with isolated heterogeneous inclusions, International Journal of Multiscale Computational Engineering 7 (4) (2009) 263–276.
[15] H. Moulinec, P. Suquet, A numerical method for computing the overall response of nonlinear composites with complex microstructure, Computer Methods in Applied Mechanics and Engineering 157 (1–2) (1998) 69–94.
[16] Y. Saad, Iterative Methods for Sparse Linear Systems, second ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2003.

Part III
Paper 2

Authors:
Jaroslav Vondřejc, Jan Zeman, and Ivo Marek

Title:
Analysis of a Fast Fourier Transform Based Method for Modeling of Heterogeneous Materials

Conference:
8th International Conference on Large-Scale Scientific Computations, LSSC 2011; Sozopol; 6th June 2011–10th June 2011

Source:
Lecture Notes in Computer Science

Volume:
7116

Year:
2012

Pages:
515–522

DOI:
10.1007/978–3–642–29843–1_ 58

Analysis of a Fast Fourier Transform Based Method for Modeling of Heterogeneous Materials

Jaroslav Vondřejc[1], Jan Zeman[1], and Ivo Marek[2]

[1] Czech Technical University in Prague,
Faculty of Civil Engineering, Department of Mechanics
vondrejc@gmail.com,
[2] Czech Technical University in Prague,
Faculty of Civil Engineering, Department of Mathematics

Abstract. The focus of this paper is on the analysis of the Conjugate Gradient method applied to a non-symmetric system of linear equations, arising from a Fast Fourier Transform-based homogenization method due to Moulinec and Suquet [1]. Convergence of the method is proven by exploiting a certain projection operator reflecting physics of the underlying problem. These results are supported by a numerical example, demonstrating significant improvement of the Conjugate Gradient-based scheme over the original Moulinec-Suquet algorithm.

Keywords: Homogenization, Fast Fourier Transform, Conjugate Gradients

1 Introduction

The last decade has witnessed a rapid development in advanced experimental techniques and modeling tools for microstructural characterization, typically provided in the form of pixel- or voxel-based geometry. Such data now allow for the design of bottom-up predictive models of the overall behavior for a wide range of engineering materials. Of course, such step necessitates the development of specialized algorithms, capable of handling large-scale voxel-based data in an efficient manner. In the engineering community, perhaps the most successful solver meeting these criteria was proposed by Moulinec and Suquet in [1]. The algorithm is based on the Neumann series expansion of the inverse of an operator arising in the associated Lippmann-Schwinger equation and exploits the Fast Fourier Transform to evaluate the action of the operator efficiently for voxel-based data. In our recent work [2], we have offered a new approach to the Moulinec-Suquet scheme, by exploiting the trigonometric collocation method due to Saranen and Vainikko [3]. Here, the Lippman-Schwinger equation is projected to a space of trigonometric polynomials to yield a non-symmetric system of linear equations, see Section 2 below. Quite surprisingly, numerical experiments revealed that the system can be efficiently solved using the standard Conjugate

Gradient algorithm. The analysis of this phenomenon, as presented in Section 3, is at the heart of this contribution. The obtained results are further supported by a numerical example in Section 4 and summarized in Section 5.

The following notation is used throughout the paper. Symbols a, $\boldsymbol{a}$ and $\boldsymbol{A}$ denote scalar, vector and second-order tensor quantities, respectively, with Greek subscripts used when referring to the corresponding components, e.g. $A_{\alpha\beta}$. The outer product of two vectors is denoted as $\boldsymbol{a}\otimes\boldsymbol{a}$, whereas $\boldsymbol{a}\cdot\boldsymbol{b}$ or $\boldsymbol{A}\cdot\boldsymbol{b}$ represents the single contraction between vectors (or tensors). A multi-index notation is employed, in which $\mathbb{R}^{\boldsymbol{N}}$ with $\boldsymbol{N}=(N_1,\dots,N_d)$ represents $\mathbb{R}^{N_1\times\cdots\times N_d}$ and $|\boldsymbol{N}|$ abbreviates $\prod_{\alpha=1}^{d} N_\alpha$. Block matrices are denoted by capital letters typeset in a bold serif font, e.g. $\mathsf{\mathbf{A}}\in\mathbb{R}^{d\times d\times\boldsymbol{N}\times\boldsymbol{N}}$, and the superscript and subscript indexes are used to refer to the components, such that $\mathsf{\mathbf{A}}=[\mathsf{A}_{\alpha\beta}^{\boldsymbol{k}\boldsymbol{m}}]_{\alpha,\beta=1,\dots,d}^{\boldsymbol{k},\boldsymbol{m}\in\bar{\mathbb{Z}}^{\boldsymbol{N}}}$ with

$$\bar{\mathbb{Z}}^{\boldsymbol{N}}=\left\{\boldsymbol{k}\in\mathbb{Z}^d:-\frac{N_\alpha}{2}<k_\alpha\leqslant\frac{N_\alpha}{2},\alpha=1,\dots,d\right\}.$$

Sub-matrices of $\mathsf{\mathbf{A}}$ are denoted as

$$\mathsf{\mathbf{A}}_{\alpha\beta}=\left[\mathsf{A}_{\alpha\beta}^{\boldsymbol{k}\boldsymbol{m}}\right]^{\boldsymbol{k},\boldsymbol{m}\in\bar{\mathbb{Z}}^{\boldsymbol{N}}}\in\mathbb{R}^{\boldsymbol{N}\times\boldsymbol{N}},\quad \mathsf{\mathbf{A}}^{\boldsymbol{k}\boldsymbol{m}}=\left[\mathsf{A}_{\alpha\beta}^{\boldsymbol{k}\boldsymbol{m}}\right]_{\alpha,\beta=1,\dots,d}\in\mathbb{R}^{d\times d}$$

for $\alpha,\beta=1,\dots,d$ and $\boldsymbol{k},\boldsymbol{m}\in\bar{\mathbb{Z}}^{\boldsymbol{N}}$. Analogously, the block vectors are denoted by lower case letters, e.g. $\mathsf{\mathbf{e}}\in\mathbb{R}^{d\times\boldsymbol{N}}$ and the matrix-by-vector multiplication is defined as

$$[\mathsf{A}\mathsf{e}]_\alpha^{\boldsymbol{k}}=\sum_{\beta=1}^{d}\sum_{\boldsymbol{m}\in\bar{\mathbb{Z}}^{\boldsymbol{N}}}\mathsf{A}_{\alpha\beta}^{\boldsymbol{k}\boldsymbol{m}}\mathsf{e}_\beta^{\boldsymbol{m}}\in\mathbb{R}^{d\times\boldsymbol{N}},\tag{1}$$

with $\alpha=1,\dots,d$ and $\boldsymbol{k}\in\bar{\mathbb{Z}}^{\boldsymbol{N}}$.

2 Problem setting

Consider a composite material represented by a periodic unit cell

$$\mathcal{Y}=\prod_{\alpha=1}^{d}(-Y_\alpha,Y_\alpha)\subset\mathbb{R}^d.$$

In the context of linear electrostatics, the associated unit cell problem reads as

$$\boldsymbol{\nabla}\times\boldsymbol{e}(\boldsymbol{x})=\boldsymbol{0},\qquad\boldsymbol{\nabla}\cdot\boldsymbol{j}(\boldsymbol{x})=\boldsymbol{0},\qquad\boldsymbol{j}(\boldsymbol{x})=\boldsymbol{L}(\boldsymbol{x})\cdot\boldsymbol{e}(\boldsymbol{x}),\qquad\boldsymbol{x}\in\mathcal{Y}\tag{2}$$

where $\boldsymbol{e}$ is a $\mathcal{Y}$-periodic vectorial electric field, $\boldsymbol{j}$ denotes the corresponding vector of electric current and $\boldsymbol{L}$ is a second-order positive-definite tensor of electric conductivity. In addition, the field $\boldsymbol{e}$ is subject to a constraint

$$\langle\boldsymbol{e}(\boldsymbol{x})\rangle:=\frac{1}{|\mathcal{Y}|}\int_{\mathcal{Y}}\boldsymbol{e}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}=\boldsymbol{e}^0,\tag{3}$$

where $|\mathcal{Y}|$ denotes the d-dimensional measure of $\mathcal{Y}$ and $\boldsymbol{e}^0 \neq \boldsymbol{0}$ a prescribed macroscopic electric field.

The original problem (2)–(3) is then equivalent to the periodic Lippmann-Schwinger integral equation, formally written as

$$\boldsymbol{e}(\boldsymbol{x}) + \int_{\mathcal{Y}} \boldsymbol{\Gamma}(\boldsymbol{x}-\boldsymbol{y};\boldsymbol{L}^0)\cdot\Big(\boldsymbol{L}(\boldsymbol{y})-\boldsymbol{L}^0\Big)\cdot\boldsymbol{e}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = \boldsymbol{e}^0, \quad \boldsymbol{x}\in\mathcal{Y}, \tag{4}$$

where $\boldsymbol{L}^0 \in \mathbb{R}^{d\times d}$ denotes a homogeneous reference medium. The operator $\boldsymbol{\Gamma}(\boldsymbol{x},\boldsymbol{L}^0)$ is derived from the Green's function of the problem (2)–(3) with $\boldsymbol{L}(\boldsymbol{x}) = \boldsymbol{L}^0$ and can be simply expressed in the Fourier space

$$\hat{\boldsymbol{\Gamma}}(\boldsymbol{k};\boldsymbol{L}^0) = \begin{cases} \boldsymbol{0} & \boldsymbol{k}=\boldsymbol{0} \\ \dfrac{\boldsymbol{\xi}\otimes\boldsymbol{\xi}}{\boldsymbol{\xi}\cdot\boldsymbol{L}^0\cdot\boldsymbol{\xi}} & \boldsymbol{\xi}(\boldsymbol{k}) = \left(\frac{k_\alpha}{Y_\alpha}\right)_{\alpha=1}^{d}; \boldsymbol{k}\in\mathbb{Z}^d\backslash\boldsymbol{0}. \end{cases} \tag{5}$$

Operator $\widehat{f} = \widehat{f}(\boldsymbol{k})$ stands for the Fourier coefficient of $f(\boldsymbol{x})$ for the $\boldsymbol{k}$-th frequency given by

$$\widehat{f}(\boldsymbol{k}) = \int_{\mathcal{Y}} f(\boldsymbol{x})\varphi_{-\boldsymbol{k}}(\boldsymbol{x})\,\mathrm{d}x,\ \ \varphi_{\boldsymbol{k}}(\boldsymbol{x}) = |\mathcal{Y}|^{-\frac{1}{2}}\exp\left(\mathrm{i}\pi\sum_{\alpha=1}^{d}\frac{x_\alpha k_\alpha}{Y_\alpha}\right), \tag{6}$$

"i" is the imaginary unit ($\mathrm{i}^2 = -1$). We refer to [2,4] for additional details. Note that the linear electrostatics serves here as a model problem; the framework can be directly extended to e.g. elasticity [5], (visco-)plasticity [6] or to multiferroics [7].

2.1 Discretization via trigonometric collocation

The numerical solution of the Lippmann-Schwinger equation is based on a discretization of a unit cell $\mathcal{Y}$ into a regular periodic grid with $N_1\times\cdots\times N_d$ nodal points and grid spacings $\boldsymbol{h} = (2Y_1/N_1,\ldots,2Y_d/N_d)$. The searched field $\boldsymbol{e}$ in (4) is approximated by a trigonometric polynomial $\boldsymbol{e}^{\boldsymbol{N}}$ in the form (cf. [3, Chapter 10])

$$\boldsymbol{e}(\boldsymbol{x}) \approx \boldsymbol{e}^{\boldsymbol{N}}(\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\bar{\mathbb{Z}}^{\boldsymbol{N}}} \hat{\mathbf{e}}^{\boldsymbol{k}}\varphi_{\boldsymbol{k}}(\boldsymbol{x}),\ \ \boldsymbol{x}\in\mathcal{Y}, \tag{7}$$

where $\hat{\mathbf{e}}^{\boldsymbol{k}} = (\hat{\mathrm{e}}^{\boldsymbol{k}}_\alpha)_{\alpha=1,\ldots,d}$ designates the Fourier coefficients defined in (6). Notice that the trigonometrical polynomials are uniquely determined by a regular grid data, which makes them well-suited to problems with pixel- or voxel-based computations.

The trigonometric collocation method is based on the projection of the Lippmann-Schwinger equation (4) onto the space of the trigonometric polynomials

$$\mathcal{T}^{\boldsymbol{N}} = \Big\{\sum_{\boldsymbol{k}\in\bar{\mathbb{Z}}^{\boldsymbol{N}}} c_{\boldsymbol{k}}\varphi_{\boldsymbol{k}}, c_{\boldsymbol{k}}\in\mathbb{C}\Big\}, \tag{8}$$

leading to a to linear system in the form, cf. [2]

$$(\mathbf{I}+\mathbf{B})\mathbf{e}=\mathbf{e^0},\ \ \mathbf{B}=\mathbf{F}^{-1}\hat{\mathbf{\Gamma}}\mathbf{F}(\mathbf{L}-\mathbf{L^0}), \tag{9}$$

where $\mathbf{e}=\left(\mathsf{e}_\alpha^{\boldsymbol{k}}\right)_{\alpha=1,\dots,d}^{\boldsymbol{k}\in\bar{\mathbb{Z}}^{\boldsymbol{N}}}\in\mathbb{R}^{d\times\boldsymbol{N}}$ is the unknown vector, $\mathbf{I}=\left[\delta_{\alpha\beta}\delta_{\boldsymbol{km}}\right]_{\alpha,\beta=1,\dots,d}^{\boldsymbol{km}\in\bar{\mathbb{Z}}^{\boldsymbol{N}}}\in\mathbb{R}^{d\times d\times\boldsymbol{N}\times\boldsymbol{N}}$ is the identity matrix, expressed as the product of the Kronecker delta functions $\delta_{\alpha\beta}$ and $\delta_{\boldsymbol{km}}$, and $\mathbf{e^0}=(e_\alpha^0)_{\alpha=1,\dots,d}^{\boldsymbol{k}\in\bar{\mathbb{Z}}^{\boldsymbol{N}}}\in\mathbb{R}^{d\times\boldsymbol{N}}$.

All the matrices in (9) exhibit a block-diagonal structure. In particular,

$$\hat{\mathbf{\Gamma}}=\left[\delta_{\boldsymbol{km}}\hat{\mathsf{\Gamma}}_{\alpha\beta}^{\boldsymbol{km}}\right]_{\alpha,\beta=1,\dots,d}^{\boldsymbol{k},\boldsymbol{m}\in\bar{\mathbb{Z}}^{\boldsymbol{N}}},\ \mathbf{L}=\left[\delta_{\boldsymbol{km}}\mathsf{L}_{\alpha\beta}^{\boldsymbol{km}}\right]_{\alpha,\beta=1,\dots,d}^{\boldsymbol{k},\boldsymbol{m}\in\bar{\mathbb{Z}}^{\boldsymbol{N}}},\ \mathbf{L^0}=\left[\delta_{\boldsymbol{km}}\mathsf{L}^0{}_{\alpha\beta}\right]_{\alpha,\beta=1,\dots,d}^{\boldsymbol{k},\boldsymbol{m}\in\bar{\mathbb{Z}}^{\boldsymbol{N}}},$$

with $\hat{\mathsf{\Gamma}}_{\alpha\beta}^{\boldsymbol{kk}}=\hat{\mathit{\Gamma}}_{\alpha\beta}(\boldsymbol{k};\boldsymbol{L}^0)$, $\mathsf{L}_{\alpha\beta}^{\boldsymbol{kk}}=L_{\alpha\beta}(\boldsymbol{k})$ and $(\mathsf{L}^{\mathbf{0}})_{\alpha\beta}=L^0_{\alpha\beta}$. The matrix $\mathbf{F}$ implements the Discrete Fourier Transform and is defined as

$$\mathbf{F}=\left[\delta_{\alpha\beta}\mathsf{F}^{\boldsymbol{km}}\right]_{\alpha,\beta=1,\dots,d}^{\boldsymbol{k},\boldsymbol{m}\in\bar{\mathbb{Z}}^{\boldsymbol{N}}},\quad \mathsf{F}^{\boldsymbol{km}}=\frac{|\mathcal{Y}|^{\frac{1}{2}}}{\prod_{\alpha=1}^d N_\alpha}\exp\left(-\sum_{\alpha=1}^d 2\pi\mathrm{i}\frac{k_\alpha m_\alpha}{N_\alpha}\right), \tag{10}$$

with $\mathbf{F}^{-1}$ representing the inverse transform.

It follows from Eq. (1) that the cost of multiplication by $\mathbf{B}$ is dominated by the action of $\mathbf{F}$ and $\mathbf{F}^{-1}$, which can be performed in $O(|\boldsymbol{N}|\log|\boldsymbol{N}|)$ operations by the Fast Fourier Transform techniques. This makes the system (9) well-suited for applying some iterative solution technique. In particular, the original Fast Fourier Transform-based Homogenization scheme formulated by Moulinec and Suquet in [1] is based on the Neumann expansion of the matrix inverse $(\mathbf{I}+\mathbf{B})^{-1}$, so as to yield the m-th iterate in the form

$$\mathbf{e}^{(m)}=\sum_{j=0}^{m}(-\mathbf{B})^j\,\mathbf{e^0}. \tag{11}$$

As indicated earlier, our numerical experiments [2] suggest that the system can be efficiently solved using the Conjugate Gradient method, despite the non-symmetry of $\mathbf{B}$ evident from (9). This observation is studied in more detail in the next Section.

3 Solution by the Conjugate Gradient method

We start our analysis with recasting the system (9) into a more convenient form, by employing a certain operator and the associated sub-space introduced later. Note that for simplicity, the reference conductivity is taken as $\mathbf{L^0}=\lambda\mathbf{I}$.

Definition 1. *Given $\lambda>0$, we define operator $\mathbf{P}_{\mathcal{E}}=\lambda\mathbf{F}^{-1}\hat{\mathbf{\Gamma}}\mathbf{F}$ and associated sub-space as*

$$\mathcal{E}=\left\{\mathbf{P}_{\mathcal{E}}\mathbf{x}\text{ for }\mathbf{x}\in\mathbb{R}^{d\times\boldsymbol{N}}\right\}\subset\mathbb{R}^{d\times\boldsymbol{N}}.$$

Lemma 1. *The operator $\mathbf{P}_{\mathcal{E}}$ is an orthogonal projection.*

Proof. First, we will prove that $\mathbf{P}_{\mathcal{E}}$ is projection, i.e. $\mathbf{P}_{\mathcal{E}}^2 = \mathbf{P}_{\mathcal{E}}$. Since $\mathbf{F}$ is a unitary matrix, it is easy to see that

$$\mathbf{P}_{\mathcal{E}}^2 = (\lambda \mathbf{F}^{-1}\hat{\mathbf{\Gamma}}\mathbf{F})(\lambda \mathbf{F}^{-1}\hat{\mathbf{\Gamma}}\mathbf{F}) = \mathbf{F}^{-1}(\lambda\hat{\mathbf{\Gamma}})^2\mathbf{F}. \tag{12}$$

Hence, in view of the block-diagonal character of $\hat{\mathbf{\Gamma}}$, it it sufficient to prove the projection property of sub-matrices $(\lambda\hat{\mathbf{\Gamma}})^{\boldsymbol{kk}}$ only. This follows using a simple algebra, recall Eq. (5):

$$(\lambda\hat{\mathbf{\Gamma}})^{\boldsymbol{kk}}(\lambda\hat{\mathbf{\Gamma}})^{\boldsymbol{kk}} = \frac{\boldsymbol{\xi}(\boldsymbol{k})\otimes\boldsymbol{\xi}(\boldsymbol{k})}{\boldsymbol{\xi}(\boldsymbol{k})\cdot\boldsymbol{\xi}(\boldsymbol{k})}\cdot\frac{\boldsymbol{\xi}(\boldsymbol{k})\otimes\boldsymbol{\xi}(\boldsymbol{k})}{\boldsymbol{\xi}(\boldsymbol{k})\cdot\boldsymbol{\xi}(\boldsymbol{k})} = \frac{\boldsymbol{\xi}(\boldsymbol{k})\otimes\boldsymbol{\xi}(\boldsymbol{k})}{\boldsymbol{\xi}(\boldsymbol{k})\cdot\boldsymbol{\xi}(\boldsymbol{k})} = (\lambda\hat{\mathbf{\Gamma}})^{\boldsymbol{kk}}.$$

The orthogonality of $\mathbf{P}_{\mathcal{E}}$ now follows from

$$\mathbf{P}_{\mathcal{E}}^* = \left(\lambda\mathbf{F}^{-1}\hat{\mathbf{\Gamma}}\mathbf{F}\right)^* = \lambda\mathbf{F}^*\hat{\mathbf{\Gamma}}^*\left(\mathbf{F}^{-1}\right)^* = \lambda\mathbf{F}^{-1}\hat{\mathbf{\Gamma}}\mathbf{F} = \mathbf{P}_{\mathcal{E}},$$

according to a well-known result of linear algebra, e.g. Proposition 1.8 in [8]. □

Remark 1. It follows from the previous results that the subspace $\mathcal{E}$ collects the non-zero coefficients of trigonometric polynomials $\mathcal{T}^{\boldsymbol{N}}$ with zero rotation, which represent admissible solutions to the unit cell problem defined by (2). Note that the orthogonal space $\mathcal{E}^{\perp}$ contains the trigonometric representation of constant fields, cf. [4, Section 12.7].

Lemma 2. *The solution* $\mathbf{e}$ *to the linear system* (9) *admits the decomposition* $\mathbf{e} = \mathbf{e}^{\mathbf{0}} + \mathbf{e}_{\mathcal{E}}$*, with* $\mathbf{e}_{\mathcal{E}} \in \mathcal{E}$ *satisfying*

$$\mathbf{P}_{\mathcal{E}}\mathbf{L}\mathbf{e}_{\mathcal{E}} + \mathbf{P}_{\mathcal{E}}\mathbf{L}\mathbf{e}^{\mathbf{0}} = \mathbf{0}. \tag{13}$$

Proof. As $\mathbf{e} \in \mathbb{R}^{d\times\boldsymbol{N}}$, Lemma 1 ensures that it can be decomposed into two orthogonal parts $\mathbf{e}_{\mathcal{E}} = \mathbf{P}_{\mathcal{E}}\mathbf{e}$ and $\mathbf{e}_{\mathcal{E}^{\perp}} = (\mathbf{I} - \mathbf{P}_{\mathcal{E}})\mathbf{e}$. Substituting this expression into (9), and using the identity $\mathbf{B} = \lambda\mathbf{F}^{-1}\hat{\mathbf{\Gamma}}\mathbf{F}\left(\frac{\mathbf{L}}{\lambda} - \mathbf{I}\right)$, we arrive at

$$\frac{1}{\lambda}\mathbf{P}_{\mathcal{E}}\mathbf{L}\mathbf{e}_{\mathcal{E}} + \mathbf{e}_{\mathcal{E}^{\perp}} + \frac{1}{\lambda}\mathbf{P}_{\mathcal{E}}\mathbf{L}\mathbf{e}_{\mathcal{E}^{\perp}} = \mathbf{e}^{\mathbf{0}}. \tag{14}$$

Since $\mathbf{e}^0 \in \mathcal{E}^{\perp}$, we have $\mathbf{e}_{\mathcal{E}^{\perp}} = \mathbf{e}^{\mathbf{0}}$ and the proof is complete. □

With these auxiliary results in hand, we are in the position to present our main result.

Proposition 1. *The non-symmetric system of linear equations* (9) *is solvable by the Conjugate Gradient method for an initial vector* $\mathbf{e}_{(0)} = \mathbf{e}^0 + \tilde{\mathbf{e}}$ *with* $\tilde{\mathbf{e}} \in \mathcal{E}$*. Moreover, the sequence of iterates is independent of the parameter* λ*.*

Proof (outline). It follows from Lemma 2 that the solution to (9) admits yet another, optimization-based, characterization in the form

$$\mathbf{e} = \mathbf{e}^0 + \arg\min_{\bar{\mathbf{e}}\in\mathcal{E}}\left[\frac{1}{2}\left(\mathbf{L}\bar{\mathbf{e}},\bar{\mathbf{e}}\right)_{\mathbb{R}^{d\times\boldsymbol{N}}} + \left(\mathbf{L}\mathbf{e}^0,\bar{\mathbf{e}}\right)_{\mathbb{R}^{d\times\boldsymbol{N}}}\right]. \tag{15}$$

The residual corresponding to the initial vector $\mathbf{e}_{(0)}$ equals to

$$\mathbf{r}_{(0)} = \mathbf{e}^0 - (\mathbf{I} + \mathbf{B})\left(\mathbf{e}^0 + \tilde{\mathbf{e}}\right) = -\frac{1}{\lambda}\mathbf{P}_{\mathcal{E}}\mathbf{L}\mathbf{e}^0 - \frac{1}{\lambda}\mathbf{P}_{\mathcal{E}}\mathbf{L}\tilde{\mathbf{e}} \in \mathcal{E}.$$

It can be verified that the subspace $\mathcal{E}$ is $\mathbf{B}$-invariant, thus $(\mathbf{I}+\mathbf{B})\mathcal{E} \subset \mathcal{E}$. Therefore, the Krylov subspace

$$\mathscr{K}_m(\mathbf{I} + \mathbf{B}, \mathbf{r}_{(0)}) = \operatorname{span}\left\{\boldsymbol{r}_{(0)}, (\mathbf{I} + \mathbf{B})\mathbf{r}_{(0)}, \ldots, (\mathbf{I} + \mathbf{B})^m\mathbf{r}_{(0)}\right\} \subset \mathcal{E}$$

for arbitrary $m \in \mathbb{N}$. This implies that the residual $\mathbf{r}_{(m)}$ and the Conjugate Gradient search direction $\mathbf{p}_{(m)}$ at the m-th iteration satisfy $\mathbf{r}_{(m)} \in \mathcal{E}$ and $\mathbf{p}_{(m)} \in \mathcal{E}$. Since $\mathbf{B}$ is symmetric and positive-definite on $\mathcal{E}$, the convergence of CG algorithm now follows from standard arguments, e.g. Theorem 6.6 in [8]. Observe that different choices of λ generate identical Krylov subspaces, thus the sequence of iterates is independent of λ. □

Remark 2. Note that it is possible to show, using direct calculations based on the projection properties of $\mathbf{P}_{\mathcal{E}}$, that the Biconjugate Gradient algorithm produces exactly the same sequence of vectors as the Conjugate Gradient method, see [9].

4 Numerical example

To support our theoretical results, we consider a three-dimensional model problem of electric conduction in a cubic periodic unit cell $\mathcal{Y} = \prod_{\alpha=1}^{3}(-\frac{1}{2}, \frac{1}{2})$, representing a two-phase medium with spherical inclusions of 25% volume fraction. The conductivity parameters are defined as

$$\boldsymbol{L}(\boldsymbol{x}) = \begin{cases} \rho\boldsymbol{I}, & \|\boldsymbol{x}\|_2 < (\frac{3}{16\pi})^{\frac{1}{3}} \\ \begin{pmatrix} 1 & 0.2 & 0.2 \\ 0.2 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{pmatrix}, & \text{otherwise} \end{cases}$$

where $\rho > 0$ denotes the contrast of phase conductivities. We consider the macroscopic field $\boldsymbol{e}^0 = [1, 0, 0]$ and discretize the unit cell with $\boldsymbol{N} = [n, n, n]$ nodes[3]. The conductivity of the homogeneous reference medium $\boldsymbol{L}^0 \in \mathbb{R}^{d\times d}$ is parametrized as

$$\boldsymbol{L}^0 = \lambda\boldsymbol{I}, \qquad \lambda = 1 - \omega + \rho\omega, \tag{16}$$

where $\omega \approx 0.5$ delivers the optimal convergence of the original Moulinec-Suquet Fast-Fourier Transform-based Homogenization (FFTH) algorithm [1].

We first investigate the sensitivity of Conjugate Gradient (CG) algorithm to the choice of reference medium. The results appear in Fig. 1(a), plotting the

[3] In particular, n was taken consequently as $16, 32, 64, 128$ and 160 leading up to $3 \cdot 160^3 \doteq 12.2 \times 10^6$ unknowns

relative number of iterations for CG against the conductivity of the reference medium parametrized by ω, recall Eq. (16). As expected, CG solver achieve a significant improvement over FFTH method as it requires about 40% iterations of FFTH for a mildly-contrasted composite down to 4% for $\varrho = 10^3$. The minor differences visible especially for $\rho = 10^3$ can be therefore attributed to accumulation of round-off errors. These observations fully confirm our theoretical results presented earlier in Section 3.

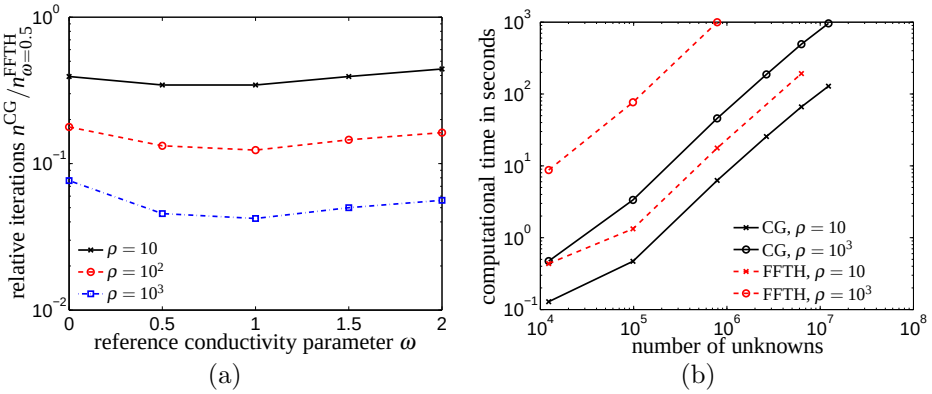


Fig. 1. (a) Relative number of iterations as a function of the reference medium parameter ω and (b) computational time as a function of the number of unknowns.

In Fig. 1(b), we present the total computational time[4] as a function of the number of degrees of freedom and the phase ratio ρ. The results confirm that the computational times scales linearly with the increasing number of degrees of freedom for both schemes for a fixed ρ [2]. The ratio of the computational time for CG and FFTH algorithms remains almost constant, which indicates that the cost of a single iteration of CG and FFTH method is comparable.

In addition, the memory requirements of both schemes are also comparable. This aspect represents the major advantage of the short-recurrence CG-based scheme over alternative schemes for non-symmetric systems, such as GMRES. Finally note that finer discretizations can be treated by a straightforward parallel implementation.

5 Conclusions

In this work, we have proven the convergence of Conjugate Gradient method for a non-symmetric system of linear equations arising from periodic unit cell ho-

[4] The problem was solved with a MATLAB® in-house code on a machine Intel® Core™2 Duo 3 GHz CPU, 3.28 GB computing memory with Debian linux 5.0 operating system.

mogenization problem and confirmed it by numerical experiment. The important conclusions to be pointed out are as follows:

1. The success of the Conjugate Gradient method follows from the projection properties of operator $\mathbf{P}_{\mathcal{E}}$ introduced in Definition 1, which reflect the structure of the underlying physical problem.
2. Contrary to all available extensions of the FFTH scheme, the performance of the Conjugate Gradient-based method is independent of the choice of reference medium. This offers an important starting point for further improvements of the method.

Apart from the already mentioned parallelization, performance of the scheme can further be improved by a suitable preconditioning procedure. This topic is currently under investigation.

Acknowledgments This work was supported by the Czech Science Foundation, through projects No. GAČR 103/09/1748, No. GAČR 103/09/P490, No. GAČR 201/09/1544, and by the Grant Agency of the Czech Technical University in Prague through project No. SGS10/124/OHK1/2T/11.

References

1. H. Moulinec, P. Suquet, A fast numerical method for computing the linear and nonlinear mechanical properties of composites, Comptes rendus de l'Académie des sciences. Série II, Mécanique, physique, chimie, astronomie 318 (11) (1994) 1417–1423.
2. J. Zeman, J. Vondřejc, J. Novák, I. Marek, Accelerating a FFT-based solver for numerical homogenization of periodic media by conjugate gradients, Journal of Computational Physics 229 (21) (2010) 8065–8071. `arXiv:1004.1122`.
3. J. Saranen, G. Vainikko, Periodic Integral and Pseudodifferential Equations with Numerical Approximation, Springer Monographs in Mathematics, Springer-Verlag, Berlin, Heidelberg, 2002.
4. G. W. Milton, The Theory of Composites, Vol. 6 of Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, UK, 2002.
5. V. Šmilauer, Z. Bittnar, Microstructure-based micromechanical prediction of elastic properties in hydrating cement paste, Cement and Concrete Research 36 (9) (2006) 1708–1718.
6. A. Prakash, R. A. Lebensohn, Simulation of micromechanical behavior of polycrystals: finite elements versus fast Fourier transforms, Modelling and Simulation in Materials Science and Engineering 17 (6) (2009) 064010+.
7. R. Brenner, J. Bravo-Castillero, Response of multiferroic composites inferred from a fast-Fourier-transform-based numerical scheme, Smart Materials and Structures 19 (11) (2010) 115004+.
8. Y. Saad, Iterative Methods for Sparse Linear Systems, second edition with corrections Edition, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2003.
9. J. Vondřejc, Analysis of heterogeneous materials using efficient meshless algorithms: One-dimensional study, Diploma thesis, Czech Technical University in Prague, URL: `http://mech.fsv.cvut.cz/~vondrejc/download/ING.pdf` (2009).

Part IV
Paper 3

Authors:
Jiři Němeček, Vlastimil Králík, and Jaroslav Vondřejc

Title:
Micromechanical analysis of heterogeneous structural materials

Source:
Cement and Concrete Composites

DOI:
10.1016/j.cemconcomp.2012.06.015

Published:
In Press, Available online 4th July 2012

Cement & Concrete Composites xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Cement & Concrete Composites

journal homepage: www.elsevier.com/locate/cemconcomp

Micromechanical analysis of heterogeneous structural materials

Jiří Němeček *, Vlastimil Králík, Jaroslav Vondřejc

Czech Technical University in Prague, Department of Mechanics, Faculty of Civil Engineering, Thákurova 7, 166 29 Prague 6, Czech Republic

ARTICLE INFO

Article history:
Received 31 January 2012
Received in revised form 25 June 2012
Accepted 26 June 2012
Available online xxxx

Keywords:
Micromechanics
Nanoindentation
Heterogeneous materials
Grid indentation
Deconvolution
Homogenization
FFT

ABSTRACT

This paper shows an efficient methodology based on micromechanical framework and grid nanoindentation for the assessment of effective elastic properties on several types of microscopically heterogeneous structural materials. Such task is a prerequisite for successful nano- and micro-structural material characterization, development and optimization. The grid nanoindentation and statistical deconvolution methods previously described in the literature e.g. for cementitious materials [1,2], alkali activated materials [3] or high-performance concretes [4] have been employed. In this paper we demonstrate their utilization also for other types of structural composites with crystalline nature and we validate the results by using enhanced numerical method based on fast Fourier transform (FFT). The direct procedure of using grid nanoindentation data in the FFT method simplifies the evaluation of effective composite properties and leads to the assemblage of the full stiffness matrix compared to simple analytical approaches.

The paper deals namely with cement paste, gypsum and aluminum alloy. Nanoindentation is used for the determination of phase properties in grid points at the scale below one micrometer. Statistical approach and deconvolution methods are applied to assess intrinsic phase properties. Elastic properties obtained by nanoindentation are homogenized in the frame of the representative volume element (RVE) by means of analytical and numerical FFT-based schemes. Good correlation of the results from all methods was found for the tested materials due to the close-to-isotropic nature of the composites in the RVE having dimensions ~100–200 μm. Results were also verified against macroscopic experimental results. The proposed and validated numerical approach can be successively used for the material modeling in finite element software or for optimization of materials with inhomogeneous microstructures.

1. Introduction

Structural composites such as concrete, gypsum, metals and others are often characterized by a heterogeneous microstructure at different length scales (nm to m). Traditionally, their mechanical properties are assessed from macroscopic tests on samples with cm to m dimensions that can only describe overall (averaged) properties like overall Young's modulus or strength. Nowadays, nanoindentation [5] can be successfully applied to access the nanometer scale and to assess individual phase properties like C–S–H gels, Portlandite or clinker. However, the properties extracted from nanoindentation are measured for small material volumes (nm to μm). The large gap between the scales can be crossed by using multiscale models and micromechanical framework which uses the concept of the representative volume element (RVE) [6] defined for each material level. Homogenization of individual contributions of the RVE microstructural components is provided by multiple micromechanical approaches that search for effective properties by solving matrix-inclusion problems. There is a variety of analytical methods and estimates (Voigt, Reuss or Hashin–Strikmann bounds, Mori–Tanaka method, self-consistent scheme and others [6]) that usually need to assess phase properties and their volume fractions prior to the analysis. Such assessment is not straightforward in the case of structural composites whose microstructure develops in space and time during their lifetime. Therefore, statistical estimates obtained from grid nanoindentation need to be employed. The grid nanoindentation and statistical deconvolution methods have been described and used e.g. by Ulm and coworkers [1,2] for cement based materials, Němeček et al. [3] for alkali activated materials or Sorelli et al. [4] for high performance concrete.

In the case of numerical methods (e.g. finite elements or FFT based methods), homogenization can be much easier due to the direct use of grid point mechanical data as will be demonstrated later in the paper.

2. Methods

In this paper, we first deal with the evaluation of nanoindentation data received from large statistical sets (hundreds of indents) on the scale of several hundreds of micrometers which is a scale

* Corresponding author. Tel.: +420 224 354 309.
E-mail addresses: jiri.nemecek@fsv.cvut.cz (J. Němeček), vlastimil.kralik@fsv.cvut.cz (V. Králík), jaroslav.vondrejc@fsv.cvut.cz (J. Vondřejc).

0958-9465/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cemconcomp.2012.06.015

that includes all material phases within RVE in a sufficient content. Since the microstructure of the composites is very complex in this scale and the determination of pure individual micromechanically distinct phases is not straightforward, we assess the individual properties by using grid indentation technique [2] with subsequent statistical deconvolution method [2–4]. Mathematically, the deconvolution is an ill-posed problem that can be regularized by a prior setting of the number of mechanically different phases that are determined. Therefore, we link this number with the number of chemically different phases or groups of mechanically similar constituents as described in Section 6. We also adapt the originally proposed deconvolution method [2] by using different minimizing criteria and modified Monte Carlo simulations as described in Němeček et al. [3]. Such methodology gives us mean phase properties together with the estimation of their volume fractions based on the experimental dataset from the whole grid.

After setting the RVE size and receiving phase properties, effective elastic properties are determined by both analytical and numerical homogenization schemes. The comparison of the methods is provided by comparing the differences between the output stiffness matrices. As mentioned earlier, the application of the numerical scheme does not require the knowledge of intrinsic phase properties and the direct use of grid data is utilized.

3. Tested materials and test setup

3.1. Cement paste

Selected heterogeneous structural materials were chosen for this study. At first, cement paste samples were prepared from Portland cement CEM-I 42,5 R (locality Mokrá, CZ) with water to cement weight ratio equal to 0.5 [7]. Samples were stored in water for two years. Therefore, high degree of hydration (over 90%) can be anticipated in the samples. The microstructure of cement paste in the tested volume includes several chemical phases known from cement chemistry, namely low- and high- density calcium–silica hydrates (LD and HD C–S–H), calcium hydroxide $Ca(OH)_2$, residual clinker, porosity and some other minor phases. The cement paste microstructure is shown in Fig. 1a. Very light areas in Fig. 1a can be attributed to the residual clinker, light grey areas are rich of $Ca(OH)_2$, dark grey zone belongs to C–S–H gels and black color represents very low density regions or capillary porosity. Note, that C–S–H gel and $Ca(OH)_2$ zones are spatially intermixed in small volumes (<<10 μm) and the resolution of SEM–BSE images does not allow for a direct separation of these phases from the image. The majority of the material volume mostly consists of poorly crystalline or amorphous phases (C–S–H) and partly of crystalline phases ($Ca(OH)_2$). Portlandite crystals are known for their anisotropy. Since their size and volume is not large in the sample and they can be mixed with C–S–H, all phases will be supposed to be mechanically isotropic for simplification in the analysis.

Cement paste includes also wide distribution of pores. Majority of pores lies in the nanometer range (<100 nm, as checked with He/Hg-porosimetry) and, on the other hand, large capillary pores are present in the scale above the indentation level (i.e. >>1 μm). Therefore, the indentation depth was chosen so that the nanoporosity was included in the tested volume but the large capillary porosity was not. The depth range ~100–300 nm was suitable for the analysis.

Cement paste was indented by a grid consisting of $20 \times 20 = 400$ indents with 10 μm spacing which yields the RVE size ~200 μm. The indents were prescribed as load controlled (maximum force 2 mN, loading/unloading rate 12 mN/min, holding for 30 s). Examples of load-penetration diagrams for different constituents are shown in Fig. 2a. The final penetration depths vary for the phases depending on their stiffness.

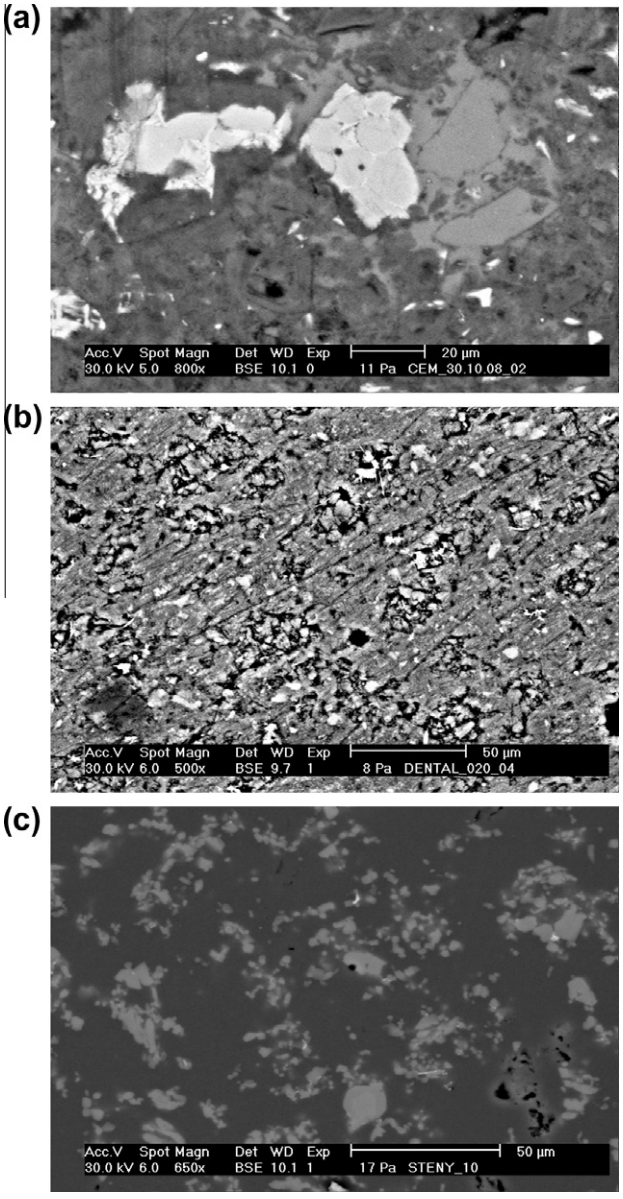


Fig. 1. Microstructures of (a) cement paste, (b) gypsum and (c) Al-alloy.

3.2. Gypsum

Secondly, dental gypsum (Interdent®) was chosen as a model representative for gypsum based materials. Samples were prepared with water to gypsum ratio 0.2 and matured in ambient conditions for two months. From the chemistry point of view, every gypsum binder is composed of three main components – calcium sulfate anhydrite ($CaSO_4$), calcium sulfate hemihydrate ($CaSO_4 \cdot ½H_2O$) in two modifications: α- or β-hemihydrate, and calcium sulfate dihydrate ($CaSO_4 \cdot 2H_2O$). The gypsum binder includes also some impurities and additives in the case of natural sources. The Interdent gypsum is a low-porosity purified α-hemihydrate used for dental purposes.

The hardened gypsum mass is a porous material with a relatively large internal surface consisting of interlocking crystals in the form of plates and needles (Singh and Middendorf [8]). Note

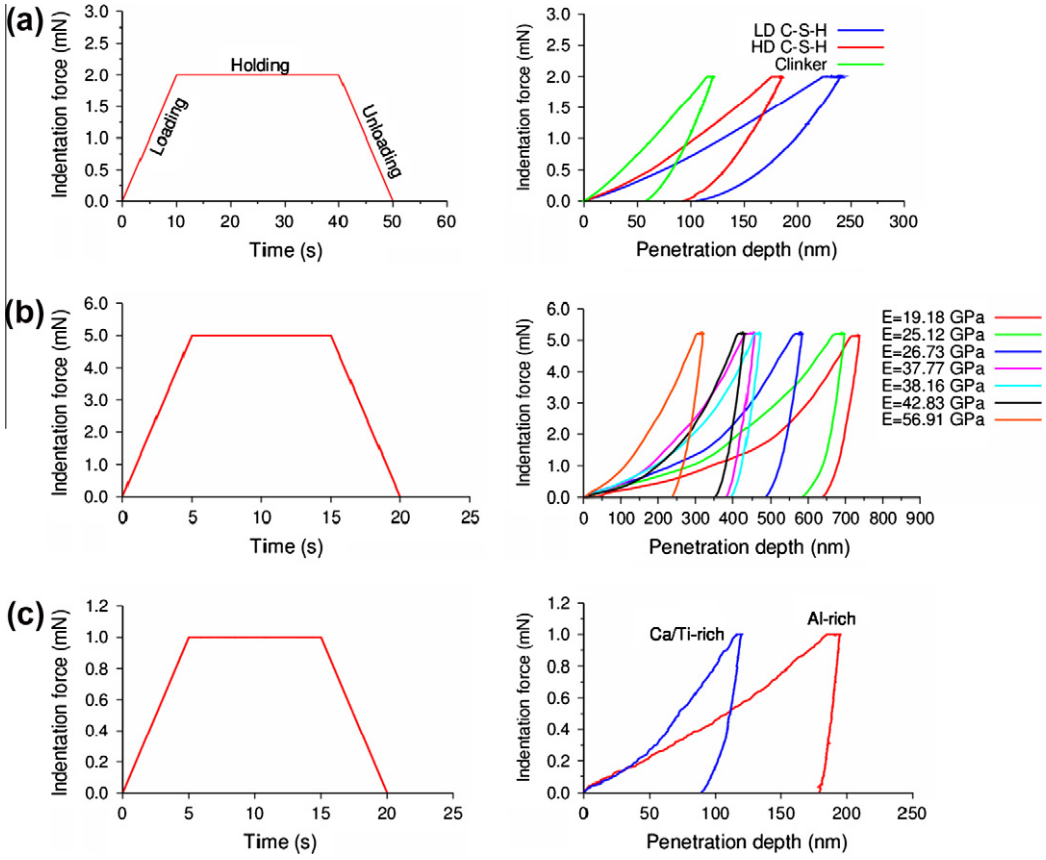


Fig. 2. Nanoindentation load–time and load–depth diagrams for (a) cement paste, (b) gypsum and (c) Al-alloy.

that in the case of β-hemihydrate hydration, the resulting sample porosity is typically very large (more than 50% for higher water to binder ratii) and crystals are interlocked very weakly. Therefore, ordinary gypsum systems used for building purposes which are based on β-hemihydrate are characterized with relatively low strengths (<10 MPa in compression). In contrast, hydration of our samples based on α-hemihydrate produced a dense matrix with total sample porosity just 19%. The majority of pores lay in the nano-range 0–300 nm (0–100 nm 7%, 100–300 nm 4%, 300–1000 nm 1%) and virtually no pores appeared between 1 and 100 μm (<0.5%). Due to the very low porosity, the strength of this material is much higher (>50 MPa in compression). The gypsum microstructure is depicted in Fig. 1b in which dark areas can be attributed to the porosity, very light areas belong to low hydrated $CaSO_4$ grains or carbonates. The majority of the sample volume in Fig. 1b composes of hydrated crystalline mass.

Two locations were tested on gypsum samples. Each place was covered by $15 \times 12 = 180$ indents with 15 μm spacing. Similar loading as in the case of cement was used (load controlled test to maximum force 5 mN). Typical loading diagrams are depicted in Fig. 2b. A bit wider range of final depths on indented phases (200–800 nm) was obtained due to larger differences in the phase stiffness. However, the majority of indents were performed to the mean final depths around 400–500 nm. The RVE size defined by the tested area is again ~200 μm in this case.

3.3. Aluminum alloy

For the sake of comparison with different kind of material, an aluminum alloy used for the production of lightweight aluminum foams Alporas® was also studied [9,10]. The material is produced from an aluminum intermixed with 1.5 wt.% of Ca and 1.6 wt.% TiH_2. Ca/Ti-rich discrete precipitates and diffuse Al_4Ca areas develop in the metal solid [11] that can be seen as lighter areas in Fig. 1c. Therefore, two distinct phases denoted as Al-rich and Ca/Ti-rich were separated in this study.

Nanoindentation was applied to the cell walls of the foam. Loading to maximum force 1 mN was used. Final depths arrived at ~100–200 nm. Typical differences between the loading diagrams of different phases obtained from nanoindentation are shown in Fig. 2c. Results from 200 indents (two locations 10×10 indents) with 10 μm spacing were evaluated. The RVE size related to the tested region is ~100 μm in this case.

4. Nanoindentation, sample preparation and evaluation of phase properties

As mentioned above, nanoindentation has been applied to receive elastic constants of individual material phases. Nanoindenter (Nanohardness tester, CSM Instruments) located in Prague's laboratory at the Czech Technical University was employed in our measurements. The apparatus was equipped with a diamond pyramidal Berkovich tip with the apex radius ~100 nm.

The already well-known principle of nanoindentation lies in bringing a very small tip (Berkovich in our case) to the surface of the material to make an imprint. Material constants are deduced from the measured load–displacement curves performed on flat surfaces.

For our measurements, the depth of penetration was kept around ~300 nm for cement paste, ~500 nm for gypsum and ~200 nm for aluminum in order to capture each material phase on one hand and to minimize phase interactions on the other hand. The depth of the affected volume under the indenter tip can be estimated as 3× the penetration depth [2], i.e. around 0.6^3–

1.5^3 μm^3 for the studied cases. Such size roughly corresponds to 1/10 of most of the grains or single phase areas which justifies the use of phase devonvolution [1,2]. The indentation volume contains also a part of nanoporosity that is naturally included in phase results.

All samples were mechanically polished prior to the testing in order to achieve smooth and flat surface with substantially smaller roughness compared to indentation depths. The surface roughness (evaluated as root-mean-square on the scanned area of 10×10 μm) was checked with AFM. It was found to be ~25 nm on cement paste, ~38 nm on gypsum and ~12 nm on Al-alloy. Therefore, the sample roughness was acceptable in relation to the awaited indentation depths.

The indentation loading history contained three segments: loading, holding and unloading periods. The holding period was included in order to minimize creep effects on the elastic unloading [7]. Elastic properties were evaluated for individual indents using analytical formulae derived by Oliver and Pharr [12], which account for an elasto-plastic contact of a conical indenter with an isotropic half-space. The reduced (combined) elastic modulus is then defined as:

$$E_r = \frac{1}{2\beta}\frac{\sqrt{\pi}}{\sqrt{A}}\frac{dP}{dh} \tag{1}$$

in which A is the projected contact area of the indenter at the peak load, β is geometrical constant (β = 1.034 for the used Berkovich tip) and dP/dh is a slope of the unloading branch evaluated at the peak. Elastic modulus E of the measured sample can be found using contact mechanics which accounts for the effect of non-rigid indenter as:

$$\frac{1}{E_r} = \frac{(1-\nu^2)}{E} + \frac{(1-\nu_i^2)}{E_i} \tag{2}$$

in which ν is the Poisson's ratio of the tested material, E_i a ν_i are known elastic modulus and Poisson's ratio of the indenter.

The solution of the contact problem for anisotropic materials can be found in [13,14]. In this work, all material phases were treated as elastically isotropic. Such simplification was adopted due to the following reasons. In cement paste, the degree of crystallinity is poor in the majority of specimen volume (e.g. in C–S–H gel). The content of crystalline $Ca(OH)_2$ phases is low and due to the limited space for the crystal growth the degree of crystallinity decreases.

On the other hand, gypsum is composed of a polycrystalline matter with locally anisotropic character. However, the response in grid nanoindentation is measured on differently oriented crystals and also on a combination of differently oriented crystals located under the indenter in the affected volume ~1.5^3 μm^3. The tested location can be viewed as a set of mechanically different phases that are physically averaged by an indenter. Apparent isotropic elasticity constants associated with the tested indentation volume can be derived in this case. Similarly, isotropic estimates were derived for the measured volume in case of Al-alloy disregarding the local anisotropy on a crystalline level.

The distinction of the chemically and/or mechanically different material phases is often not possible on the microlevel (<1 μm) even with the use of SEM–EDX images. In order to receive statistically relevant data from all material phases, we applied grid indentation over the tested RVE (Fig. 1). Large matrices containing hundreds of indents have been performed on tested samples (see Section 3). To assess individual phase properties, statistical deconvolution was employed [2,3]. In this method, experimental data are analyzed from the frequency plots. Mean elastic properties as well as phase volume fraction are estimated based on the best fit of the experimental data with a limited number of Gauss distributions (Fig. 3).

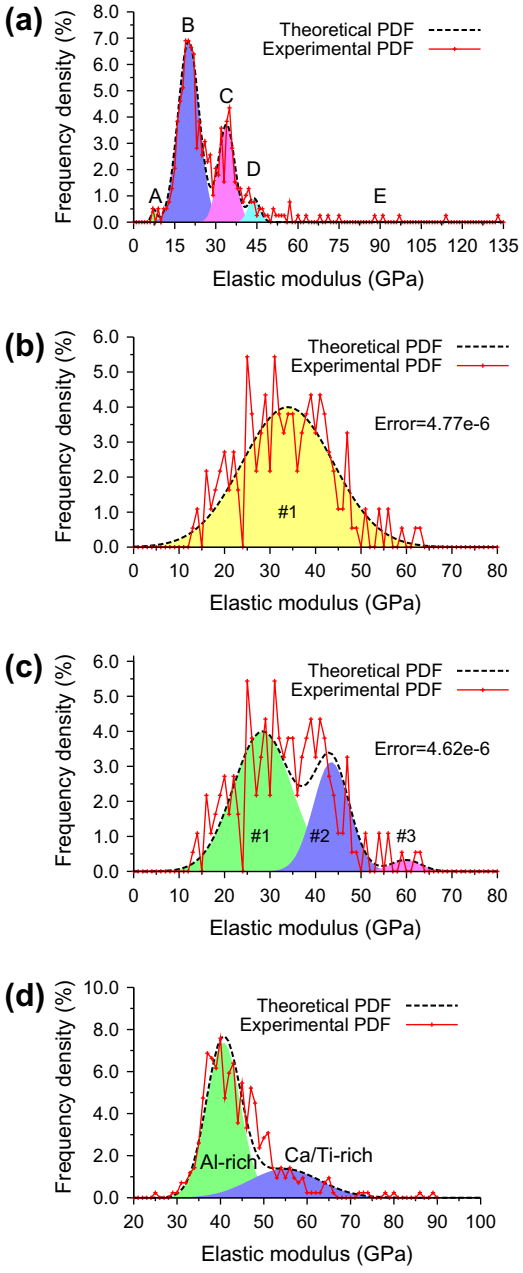


Fig. 3. Deconvolution of modulus of elasticity frequency plots into mechanical phases on (a) cement paste, (b) gypsum three phases fit (c) gypsum single phase fit and (d) Al-alloy.

5. Micromechanical homogenization

5.1. Analytical and numerical schemes

In general, homogenization methods search for effective material properties. The previously mentioned concept of RVE which includes all microstructural inhomogeneities that should be substantially smaller than the RVE size is utilized. The homogenization problem can be solved either by analytical methods or by numerical approximations. Analytical schemes often rely on simplified assumptions concerning inclusion geometry, boundary conditions or isotropy. More complex results can be obtained from numerical methods that are based on finite element solution or fast Fourier transformation (Moulinec and Suquet [17]), for instance.

Please cite this article in press as: Němeček J et al. Micromechanical analysis of heterogeneous structural materials. Cement Concrete Comp (2012), http://dx.doi.org/10.1016/j.cemconcomp.2012.06.015

The classical analytical solution based on constant stress/strain fields in individual microscale components for an ellipsoidal inclusion embedded in an infinite body was derived by Eshelby [15]. Effective elastic properties are then obtained through averaging over the local contributions. Various estimates considering different geometrical constraints or special choices of the reference medium such as the Mori–Tanaka method or the self-consistent scheme [6,16] can be used. For the case of a composite material with prevailing matrix and spherical inclusions the Mori–Tanaka method [16] was previously found to be a simple but powerful tool to estimate effective composite properties also for structural materials [2] and, therefore, it was used in this work. In the Mori–Tanaka method, the homogenized isotropic bulk and shear moduli of an r-phase composite are assessed as follows:

$$k_{\text{hom}} = \frac{\sum_r f_r k_r \left(1 + \alpha_0 \left(\frac{k_r}{k_0} - 1\right)\right)^{-1}}{\sum_r f_r \left(1 + \alpha_0 \left(\frac{k_r}{k_0} - 1\right)\right)^{-1}} \tag{3}$$

$$\mu_{\text{hom}} = \frac{\sum_r f_r \mu_r \left(1 + \beta_0 \left(\frac{\mu_r}{\mu_0} - 1\right)\right)^{-1}}{\sum_r f_r \left(1 + \beta_0 \left(\frac{\mu_r}{\mu_0} - 1\right)\right)^{-1}} \tag{4}$$

$$\alpha_0 = \frac{3k_0}{3k_0 + 4\mu_0}, \quad \beta_0 = \frac{6k_0 + 12\mu_0}{15k_0 + 20\mu_0} \tag{5}$$

where the subscript 0 corresponds to the reference medium and r corresponds to a particulate inclusion. Thus, k_0 and μ_0 are the bulk and shear moduli of the reference medium, while k_r and μ_r refer to the inclusion phases. Further, bulk and shear moduli can be recomputed to engineering values of elastic modulus and Poisson's ratio as:

$$E = \frac{9k\mu}{3k + \mu}, \quad \nu = \frac{3k - 2\mu}{6k + 2\mu} \tag{6}$$

Materials with no preference of matrix phase (e.g. polycrystalline metals) are usually modeled with the self-consistent scheme [6]. It is an implicit scheme, similar to the Mori–Tanaka method, in which the reference medium points back to the homogenized medium itself.

Local strain and stress fields in a RVE can also be found by numerical methods like finite element method or a method based on fast Fourier transformation (FFT). The later one was proven to be a reliable and computationally inexpensive method which only utilizes mechanical data in the regular grid (i.e. equidistant discretization points). Such a concept perfectly matches with the concept of the grid nanoindentation. Therefore, the FFT method was chosen for our purposes. The numerical scheme used here solves the problem of finding the effective elasticity tensor with a periodically repeating RVE by using discretization of an integral Lippmann–Schwinger equation:

$$\varepsilon(\mathbf{x}) = \varepsilon^0 - \int_\Omega \Gamma^0(\mathbf{x} - \mathbf{y}) : (L(\mathbf{y}) - L^0) : \varepsilon(\mathbf{y}) d\mathbf{y} \tag{7}$$

in which ε and L stand for the local strain and stiffness tensor, respectively, and ε^0 is the homogenized strain defined as a spatial average over RVE domain Ω as

$$\varepsilon^0 = \langle \varepsilon \rangle = \frac{1}{\Omega} \int_\Omega \varepsilon(\mathbf{x}) d\mathbf{x} \tag{8}$$

Γ^0 is the periodic Green operator associated with the reference elasticity tensor L^0 which is a parameter of the method [17,18]. The problem is further discretized using trigonometric collocation method [19,20] which leads to the assemblage of a nonsymmetrical linear system of equations:

$$[I + F^{-1} \hat{\Gamma} F (L - L^0)] e = e^0 \tag{9}$$

where the vector e stores a strain field at discretization points and e^0 the macroscopic strain, L and L^0 stores the material coefficients at discretization points and reference elasticity tensor respectively, I denotes the identity matrix, $\hat{\Gamma}$ stores the values corresponding to the integral kernel in the Fourier space, and F (F^{-1}) stores the (inverse) discrete Fourier transform matrices that can be provided by fast Fourier transform algorithm. The possibility to solve the nonsymmetric linear system by the conjugate gradient method (CG) is proposed by Zeman et al. in [21], where also the particular expression of individual matrices can be found for the problem of electric conductivity or heat transfer. The linear system (Eq. (9)) depends only on stiffness coefficients at grid points that can be obtained using nanoindentation and thus the homogenized (effective) tensor (further denoted as L_{eff}^{FFT}) can be calculated.

In practice, the homogenization procedure includes several steps:

(1) Definition of a periodic unit cell (PUC) with discretization points corresponding to indents' locations (a regular grid).
(2) Assessment of Young's moduli E and Poisson's ratii ν with the help of nanoindentation in grid points (Oliver and Pharr method [12] was used for the extraction of Young's moduli from load–displacement indentation curves).
(3) Assemblage of local elastic stiffness tensors in grid points (plane strain assumption used) which in Mandel's notation reads:

$$L = \frac{E}{(1+\nu)(1-2\nu)} \begin{bmatrix} 1-\nu & \nu & 0 \\ \nu & 1-\nu & 0 \\ 0 & 0 & 1-2\nu \end{bmatrix} \tag{10}$$

(4) Calculation of local strain (from a linear system, Eq. (9), using CG algorithm [21]) and stress fields ($\sigma = L{:}e$) in grid points when applying homogeneous macroscopic strain (unit loads e^0) to the PUC domain.
(5) Calculation of an average stress in the PUC by integration over its volume

$$\langle \sigma \rangle = \frac{1}{\Omega} \int_\Omega \sigma d\mathbf{x} \tag{11}$$

(6) Calculation of the homogenized elasticity tensor for PUC from average stress and prescribed macroscopic strain

$$L_{eff}^{\text{FFT}} : e^0 = \langle \sigma \rangle \tag{12}$$

The resulting homogenized stiffness matrix for PUC must be symmetric, positive definite, but generally anisotropic. The resulting anisotropy of the matrix depends on the topology of inclusions in PUC regardless of the fact that the individual points are treated as locally isotropic. Note also, that the FFT homogenization takes no assumptions on the morphology of the phases as in the case of analytical schemes. It works only with the stiffness coefficients distributed within the PUC and its accuracy depends only on the density of the grid points.

5.2. Comparison of analytical and numerical schemes

The simple analytical methods used in this work (Mori–Tanaka, self-consistent) operate with the assumption of isotropic effective properties. Such assumption is usually acceptable for disordered structural materials. In this case, the isotropic stiffness matrix and plane strain conditions takes the form (in Mandel's notation):

Please cite this article in press as: Němeček J et al. Micromechanical analysis of heterogeneous structural materials. Cement Concrete Comp (2012), http://dx.doi.org/10.1016/j.cemconcomp.2012.06.015

$$L^A_{eff} = \frac{E_{eff}}{(1+\nu_{eff})(1-2\nu_{eff})}\begin{bmatrix} 1-\nu_{eff} & \nu & 0 \\ \nu & 1-\nu_{eff} & 0 \\ 0 & 0 & 1-2\nu_{eff} \end{bmatrix} = \begin{bmatrix} k+\frac{4}{3}\mu & k-\frac{2}{3}\mu & 0 \\ k-\frac{2}{3}\mu & k+\frac{4}{3}\mu & 0 \\ 0 & 0 & 2\mu \end{bmatrix} \quad (13)$$

in which E_{eff} and ν_{eff} are analytically computed effective Young's modulus and Poisson's ratio, respectively. Alternatively, effective bulk and shear moduli k and μ can be used for the calculation. The difference between this stiffness matrix and that received from FFT homogenization can be expressed using a stiffness error norm:

$$\delta = \sqrt{\frac{\left(L^{FFT}_{eff} - L^A_{eff}\right)::\left(L^{FFT}_{eff} - L^A_{eff}\right)}{\left(L^{FFT}_{eff}::L^{FFT}_{eff}\right)}} \quad (14)$$

in which L^{FFT}_{eff} is the (anisotropic) effective stiffness matrix computed by the FFT method.

To assess the degree of anisotropy of the L^{FFT}_{eff} matrix, one can use different measures. Here, we define the degree of anisotropy as:

$$\delta_{ISO} = inf_{\,L_{ISO}} \sqrt{\frac{\left(L^{FFT}_{eff} - L_{ISO}\right)::\left(L^{FFT}_{eff} - L_{ISO}\right)}{\left(L^{FFT}_{eff}::L_{ISO}\right)}} \quad (15)$$

where the infimum is taken over all isotropic positive definite matrices. We simply calculate the upper estimate $\delta^{FFT}_{ISO} \geqslant \delta_{ISO}$:

$$\delta^{FFT}_{ISO} = \sqrt{\frac{\left(L^{FFT}_{eff} - L^{FFT}_{ISO}\right)::\left(L^{FFT}_{eff} - L^{FFT}_{ISO}\right)}{\left(L^{FFT}_{eff}::L^{FFT}_{ISO}\right)}} \quad (16)$$

by a particular choice of an isotropic matrix:

$$L^{FFT}_{ISO} = \begin{bmatrix} k_{ISO}+\frac{4}{3}\mu_{ISO} & k_{ISO}-\frac{2}{3}\mu_{ISO} & 0 \\ k_{ISO}-\frac{2}{3}\mu_{ISO} & k_{ISO}+\frac{4}{3}\mu_{ISO} & 0 \\ 0 & 0 & 2\mu_{ISO} \end{bmatrix} \quad (17)$$

with

$$\mu_{ISO} = \frac{L^{FFT}_{eff,33}}{2}, \quad k_{ISO} = \frac{L^{FFT}_{eff,11} + L^{FFT}_{eff,22}}{2} - \frac{4}{3}\mu_{ISO}.$$

6. Results and discussion

The resulting frequency plot of elastic moduli measured on cement paste merged from all positions (400 indents) was deconvoluted into five mechanical phases (that correspond to chemical ones) as specified in Table 1. Note, that the values in Table 1 (and similarly in Tables 2 and 3) were found as the best fit in the minimization problem solved by the deconvolution algorithm. The bin size was set to 1 GPa in the construction of probability density functions (PDFs). Ulm et al. [1] suggested the use of cumulative density function (CDF) in the deconvolution rather than PDF. Using CDF does not require the choice of a bin size. On the other hand, using PDF is more physically intuitive and in the case of large dataset leads to similar results.

The deconvoluted phases on cement paste correspond to the peaks shown in Fig. 3a. They are denoted as *A* = low stiffness phase, *B* = low density C–S–H, C = high density C–S–H, D = $Ca(OH)_2$, *E* = clinker. In this case, the notation of mechanically distinct phases matches well with the cement chemistry. Note, that the stiffest microstructural component, the clinker, is not captured well by nanoindentation since the stiffness contrast with respect to other components is too high [2,7]. However, the content of residual clinker is very low in the case of matured paste and it does not significantly influence the rest of the results. Nevertheless, the proper value of elastic modulus for homogenization was taken from ex situ measurements of clinker [7,22].

Table 1
Data received from statistical deconvolution and homogenized values on cement paste.

Deconvoluted phase	*E* (GPa)	Poisson's ratio	Volume fraction
Low stiffness phase (A)	7.45	0.2	0.011
Low density C–S–H (B)	20.09	0.2	0.632
High density C–S–H (C)	33.93	0.2	0.263
Portlandite (D)	43.88	0.3	0.046
Clinker (E)	121.0[a]	0.3	0.048
Homogenization			
C–S–H level (B + C) by M–T	23.36	0.2	
C–S–H level (B + C) by SCS	23.41	0.2	
Cement paste level (B + C) + A + D + E by M–T	25.39	0.207	1.0
Cement paste level (B + C) + A + D + E by SCS	25.44	0.208	1.0

M–T stands for the Mori–Tanaka scheme; SCS stands for the self-consistent scheme.
[a] *Note*: Clinker value was adjusted to 121 GPa according to [7].

Table 2
Data received from statistical deconvolution to the three phases and homogenized values on gypsum.

Deconvoluted phase	*E* (GPa)	Poisson's ratio	Volume fraction
#1	28.36	0.32	0.663
#2	43.46	0.32	0.310
#3	59.89	0.32	0.027
Homogenization method			
M–T	32.96	0.32	1.0
SCS	33.02	0.32	1.0

Note: M–T stands for the Mori–Tanaka scheme; SCS stands for the self-consistent scheme.

Table 3
Data received from statistical deconvolution and homogenized values on Al-alloy.

Deconvoluted phase	*E* (GPa)	Poisson's ratio	Volume fraction
Al-rich zone	61.88	0.35	0.64
Ca/Ti-rich zone	87.40	0.35	0.36
Homogenization method			
M–T	70.09	0.35	1.0
SCS	70.15	0.35	1.0

Note: M–T stands for the Mori–Tanaka scheme, SCS stands for the self-consistent scheme.

Two-step homogenization was used in the case of cement paste. Firstly, homogenized properties for the C–S–H level were obtained from low- and high-density C–S–H phases (RVE ~1 μm). Upper level homogenization for RVE (~200 μm) was performed in the second step in which homogenized C–S–H properties were considered together with the rest of the phases (i.e. low stiffness phase, Portlandite and clinker). Results for cement paste are summarized in Table 1. Very similar estimates have been obtained with the Mori–Tanaka and the self-consistent schemes.

Nanoindentation data received on gypsum samples (two locations with 180 indents each) revealed the polycrystalline nature of the composite with an anisotropic character. Since the gypsum crystals are dispersed in the sample volume in a random manner, surface measurements by nanoindentation show high scatter. As mentioned earlier, apparent isotropic moduli associated with the indentation volume ~1.5^3 μm^3 were assessed. The scatter in received results (Fig. 3b) can be treated as a set of mechanically different responses from different crystal orientations. As such, we

can either use deconvolution to separate mechanically significant groups of these orientations (further denoted as phases) or compute apparent elastic moduli of isotropic solid from all responses in an ensemble (i.e. compute average value from all results). Both approaches have been tested.

The physical motivation for identifying the mechanical phases lies in the fact that gypsum crystallizes in the monoclinic system which is characterized with three significant crystallographic orientations. Therefore, derivation of the three significant peaks by deconvolution of frequency plot was tested (Fig. 3b). Numerical results from this deconvolution are summarized in Table 2.

On the other hand, further simplification based on the assessment of only a single apparent isotropic phase is possible. Then, only one Gaussian distribution is assumed in the calculation. Such fit is depicted in Fig. 3c. The mean value derived from the histogram (E = 33.90 GPa) can be interpreted as an effective gypsum Young's modulus valid for the RVE (~100 μm) which includes also intrinsic nanoporosity.

The difference between the two solutions in terms of an error computed as a sum of squared differences between the experimental and theoretical curves in the deconvolution analysis [3] is very small (~3%). Thus, both fits are almost equally good as indicated in Fig. 3b and c. Also, the comparison of the resulting effective Young's moduli computed by the self-consistent scheme or the Mori–Tanaka method in the case of the three phase medium (Table 2) with an apparent Young's modulus in the case of a single phase (E = 33.90 GPa) shows small differences (~2.7%).

Two mechanically distinct phases were found by the statistical deconvolution (from 200 indents) on Al-alloy sample (Fig. 3c and Table 3). According to the SEM–EDX studies, the dominant phase was denoted as Al-rich zone, whereas the lower stiffness phase as Ca/Ti-rich area. The bin size in the frequency plots was set again to 1 GPa in both cases of gypsum and Al-alloy.

Based on the nanoindentation data analytical homogenization were employed for the assessment of effective RVE elastic properties at first (Tables 1–3). Very similar results have been produced by the Mori–Tanaka method or the self-consistent scheme.

At second, the comparison of stiffness matrices (Eq. (13)) derived from analytical results (Mori–Tanaka scheme was considered for cement pastes and Al-alloy; self-consistent scheme for gypsum), and those from FFT homogenization was performed. Results are specified in the following equations 18–20. The stiffness values are given in GPa. Respective error norms are computed in Eq. (21).

$$\text{cement}: L^A_{eff} = \begin{bmatrix} 28.44 & 7.43 & 0 \\ 7.43 & 28.44 & 0 \\ 0 & 0 & 21.02 \end{bmatrix}$$

$$L^{FFT}_{eff} = \begin{bmatrix} 26.177 & 6.778 & 0.068 \\ 6.778 & 26.224 & 0.014 \\ 0.068 & 0.014 & 19.818 \end{bmatrix} \quad (18)$$

$$\text{Gypsum : 3 phase fit}: L^A_{eff} = \begin{bmatrix} 47.25 & 22.24 & 0 \\ 22.24 & 47.25 & 0 \\ 0 & 0 & 25.02 \end{bmatrix}$$

$$\text{1phase fit}: L^A_{eff} = \begin{bmatrix} 48.51 & 22.84 & 0 \\ 22.84 & 48.51 & 0 \\ 0 & 0 & 25.69 \end{bmatrix}$$

$$L^{FFT}_{eff} = \begin{bmatrix} 45.302 & 21.185 & 0.101 \\ 21.185 & 45.497 & -0.008 \\ 0.101 & -0.008 & 24.396 \end{bmatrix} \quad (19)$$

$$\text{Al-alloy}: L^A_{eff} = \begin{bmatrix} 112.479 & 60.566 & 0 \\ 60.566 & 112.479 & 0 \\ 0 & 0 & 51.913 \end{bmatrix}$$

$$L^{FFT}_{eff} = \begin{bmatrix} 117.130 & 62.741 & -0.163 \\ 62.741 & 117.106 & -0.143 \\ -0.163 & -0.143 & 54.313 \end{bmatrix} \quad (20)$$

$$\text{Errors}: {}^{\text{cement}}\delta = 0.08, {}^{\text{gypsum}}\delta = 0.07, {}^{\text{Al-alloy}}\delta = 0.04 \quad (21)$$

It is clear from the above equations that both simple analytical and advanced FFT-based method give comparable results in our case. The differences given by error norms for cement and gypsum (7–8%) are acceptable and show good agreement of the results received from different methods. The best agreement of the methods was reached on Al-alloy (error 4%) which can be attributed to the fact that both material phases (Al-rich, and Ca/Ti-rich zones) are even more homogeneously dispersed at microscale RVE compared to the phases that appear in cement paste or gypsum.

The upper bound of the degree of anisotropy for the FFT-based stiffness matrices was assessed by an index defined in Eq. (16) with the following results:

$${}^{\text{cement}}\delta_{ISO} = 0.0132,\ {}^{\text{gypsum}}\delta_{ISO} = 0.0043,\ {}^{\text{Al-alloy}}\delta_{ISO} = 0.0016 \quad (22)$$

Low values in Eq. (22) (0.1–1.3%) show the close-to-isotropic nature of the tested materials within the specified RVE. In other words, microstructural inhomogeneities are uniformly dispersed in the RVE and consequently it also justifies the usage of analytical methods producing isotropic effective (homogenized) properties.

It must be emphasized again that although both analytical and numerical methods give similar results, there is a clear advantage of the FFT method which works directly with the grid indentation data compared to analytical Mori–Tanaka method which needs the assessment of phase properties and volume fractions. Moreover, the full stiffness matrix including possible anisotropy is captured by using the FFT method.

Comparison with macroscopic experimental values of elastic moduli for the given materials also shows good agreement with model predictions. Hydrated compound of cement paste was studied e.g. by Němeček [7] (E = 26.4 ± 1.8 GPa), Constantinides and Ulm [23,24] (E = 22.8 ± 0.5 GPa) or Hughes and Trtik [25] (E = 26.5 GPa). The values correspond well with our results (E = 25.4 GPa).

Gypsum elastic properties were studied e.g. by Meille and Garboczi [26,27] who estimated the plane strain values of the Young's modulus (computed as an angular average from anisotropic crystal elastic moduli tensor) as ~45.7 GPa. Such value was also reported for zero crystal porosity by Sanahuja et al. [28]. If one takes into account an intercrystalline porosity 12% (i.e. the gypsum nanoporosity measured for our specific case; see Section 3.2) the Young's modulus drops down to ~34 GPa [28] which is in excellent agreement with our homogenized value (E = 32–33.90 GPa).

Homogenized Al-alloy properties (E = 70.1 GPa) agree very well with experimental values reported e.g. by Jeon et al. [29] or Ashby et al. [30] (E = 70 GPa).

7. Conclusions

Nanoindentation was successfully used for the assessment of elastic parameters of intrinsic material constituents at the scale below one micrometer and effective composite properties were evaluated with analytical Mori–Tanaka, self-consistent and FFT numerical schemes for three typical structural composites with heterogeneous microstructure. Based on the micromechanical approaches and proposed methodologies we can draw the following conclusions.

Please cite this article in press as: Němeček J et al. Micromechanical analysis of heterogeneous structural materials. Cement Concrete Comp (2012), http://dx.doi.org/10.1016/j.cemconcomp.2012.06.015

(1) It has been shown that the use of grid indentation gives access to both phase properties as well as volume fractions in the case of testing highly heterogeneous microstructures of cement paste, gypsum and Al-alloy.
(2) Effective elastic properties of their microstructural RVEs (100–200 μm) were successfully determined with analytical Mori–Tanaka or self-consistent schemes. However, such approach assumes isotropic nature of the composite with spherical inclusions and several assumptions concerning mainly the number of mechanically different phases and bin size need to be made in the deconvolution algorithm. Therefore, an additional knowledge about the composite microstructure and its microstructural composition is necessary in this case.
(3) Further, numerical FFT-based method was used for the assessment of effective elastic composite properties. The direct use of grid indentation data is employed in this method. The method provides effective stiffness matrix and can capture also possible anisotropy.
(4) The performance of both analytical and numerical approaches was in good agreement for the tested materials mainly due to the close-to-isotropic nature in their RVEs.
(5) Comparison with macroscopic experimental data also shows good correlation of measured effective values and the predicted ones.
(6) The proposed numerical procedure for the estimation of effective elastic properties can be further applied also to other nano- or micro-heterogeneous structural composites in order to assess their anisotropic stiffness matrices or to optimize their composition.

Acknowledgment

Support of the Czech Science Foundation (P105/12/0824 and P105/12/0331) and the Grant Agency of the Czech Technical University in Prague (SGS12/116/OHK1/2T/11) is gratefully acknowledged.

References

[1] Ulm F-J, Vandamme M, Bobko C, Ortega JA. Statistical indentation techniques for hydrated nanocomposites: concrete, bone, and shale. J Am Ceram Soc 2007;90(9):2677–92.
[2] Constantinides G, Chandran KR, Ulm F-J, Vliet KV. Grid indentation analysis of composite microstructure and mechanics: principles and validation. Mater Sci Eng: A 2006;430(1–2):189–202.
[3] Němeček J, Šmilauer V, Kopecký L. Nanoindentation characteristics of alkali-activated aluminosilicate materials. Cem Concr Compos 2011;33(2):163–70.
[4] Sorelli L, Constantinides G, Ulm F-J, Toutlemonde F. The nano-mechanical signature of ultra high performance concrete by statistical nanoindentation techniques. Cem Concr Res 2008;38(12):1447–56.
[5] Fischer-Cripps AC. Nanoindentation. New York: Springer Verlag; 2002.
[6] Zaoui A. Continuum micromechanics: survey. J Eng Mech 2002;128:808–16.
[7] Němeček J. Creep effects in nanoindentation of hydrated phases of cement pastes. Mater Characterization 2009;60(9):1028–34.
[8] Singh NB, Middendorf B. Calcium sulfate hemihydrate hydration leading to gypsum crystallization. Prog Crystal Growth Characterization Mater 2007;53(1):57–77.
[9] Miyoshi T, Itoh M, Akiyama S, Kitahara A. Aluminium foam "ALPORAS": the production process, properties and application. Mater Res Soc Sympos Proc 1998:521.
[10] Němeček J, Králík V, Vondřejc J, Němečková J. Identification of micromechanical properties on metal foams using nanoindentation. In: Proceedings of the thirteenth international conference on civil, structural and environmental engineering computiing [CD-ROM]. Edinburgh: Civil-Comp Press; 2011. p. 1–12 [ISBN 978-1-905088-46-1].
[11] Simone AE, Gibson LJ. Aluminum foams produced by liquid-state processes. Acta Mater 1998;46(9):3109–23.
[12] Oliver WC, Pharr GM. An improved technique for determining hardness and elastic modulus using load and displacement sensing indentation experiments. J Mater Res 1992;7(6):1564–83.
[13] Swadener JG, Pharr GM. Indentation of elastically anisotropic half-spaces by cones and parabolae of revolution. Phil Mag A 2001;81(2):447–66.
[14] Vlassak JJ, Ciavarella M, Barber JR, Wang X. The indentation modulus of elastically anisotropic materials for indenters of arbitrary shape. J Mech Phys Solids 2003;51:1701–21.
[15] Eshelby JD. The determination of the elastic field of an ellipsoidal inclusion and related problems. Proc R Soc London A 1957;241:376–96.
[16] Mori T, Tanaka K. Average stress in the matrix and average elastic energy of materials with misfitting inclusions. Acta Metall 1973;21(5):571–4.
[17] Moulinec H, Suquet P. A fast numerical method for computing the linear and nonlinear mechanical properties of composites. Comptes rendus del'Académie des sciences. Série II, Mécanique, physique, chimie, astronomie 1994; 318(11): 1417–1423.
[18] Moulinec H, Suquet P. A numerical method for computing the overall response of nonlinear composites with complex microstructure. Comput Methods Appl Mech Eng 1998;157(1–2):69–94.
[19] Saranen J, Vainikko G. Periodic integral and pseudodifferential equations with numerical approximation. Berlin: Springer; 2002.
[20] Vainikko G. Fast solvers of the Lippmann–Schwinger equation. In: Gilbert RP, Kajiwara J, Xu YS, editors. Direct and inverse problems of mathematical physics, international society for analysis applications and computation, vol. 5. Dordrecht (The Netherlands): Kluwer Academic Publishers; 2000. p. 423–40.
[21] Zeman J, Vondřejc J, Novák J, Marek I. Accelerating a FFT-based solver for numerical homogenization of periodic media by conjugate gradients. J Comput Phys 2010;229(21):8065–71.
[22] Velez K et al. Determination of nanoindentation of elastic modulus and hardness of pure constituents of Portland cement clinker. Cem Concr Res 2001;31:555–61.
[23] Constantinides G, Ulm F-J. The nanogranular nature of C–S–H. J Mech Phys Solids 2007;55:64–90.
[24] Constantinides G, Ulm F-J. The effect of two types of C–S–H on the elasticity of cement-based materials: results from nanoindentation and micromechanical modeling. Cem Concr Res 2004;34(1):67–80.
[25] Hughes JJ, Trtik P. Micro-mechanical properties of cement paste measured by depth-sensing nanoindentation: a preliminary correlation of physical properties with phase type. Mater Characterization 2004;53:223–31.
[26] Meille S, Garboczi EJ. Linear elastic properties of 2D and 3D models of porous materials made from elongated objects. Modell Simul Mater Sci Eng 2001;9(5):371–90.
[27] Garboczi et al. Modeling and measuring the structure and properties of cement-based materials, National Institute of Standards and Technology; 1989–2011. <http://ciks.cbt.nist.gov/garbocz/monograph>.
[28] Sanahuja J, Dormieux L, Meille S, Hellmich C, Fritsch A. Micromechanical explanation of elasticity and strength of gypsum: from elongated anisotropic crystals to isotropic porous polycrystals. J Eng Mech 2010;136(2):239–53.
[29] Jeon I et al. Cell wall mechanical properties of closed-cell Al foam. Mech Mater 2009;41:60–73.
[30] Ashby MF, Evans AG, Fleck NA, Gibson LJ, Hutchinson JW, Wadley HNG. Metal Foams: A Design Guide. Butterworth-Heinemann; 2000.

Part V
Paper 4

Authors:

Jiři Němeček, Vlastimil Králík, and Jaroslav Vondřejc

Title:

A Two-scale micromechanical model for aluminium foam based on results from nanoindentation

A Two-scale micromechanical model for aluminium foam based on results from nanoindentation

Jiří Němeček[1], Vlastimil Králík[1], Jaroslav Vondřejc[1]

[1]Department of Mechanics, Faculty of Civil Engineering

Czech Technical University in Prague, Czech Republic

Abstract

The main aim of this paper is to develop and verify simple but effective model for assessing elastic properties of a porous aluminium foam system and to compare results received from experimental micromechanics with solutions given by simple analytical or more advanced numerical methods. The material is characterized by a closed pore system with very thin but microscopically inhomogeneous pore walls (~0.1 mm) and large air pores (~2.9 mm). Therefore, two material levels can be distinguished.

The lower level of the proposed model contains inhomogeneous solid matter of the foam cell walls produced from aluminium melt with admixtures. Elastic parameters as well as volume fractions of microstructural material phases at this level are assessed with nanoindentation and effective properties computed via analytical and numerical homogenization schemes. The effective Young's modulus of the cell walls was found close to 70 GPa irrespective to the used homogenization procedure.

The higher foam scale contains homogenized cell wall properties and a significant volume fraction of air voids (91.4%). Since analytical schemes fail to predict effective properties of this highly porous structure, numerical homogenization based on simple 2-D finite element model is utilized. The model geometry is based on foam optical images and the beam structure is produced using Voronoi tessellation. Effective foam Young's modulus was found to be 1.36-1.38 GPa which is in relation with ~1.45 GPa obtained from uniaxial compression experiments. The stiffness underestimation in the 2-D model is caused likely by the lack of the real 3-D confinement that can not be fully captured in the simplified model.

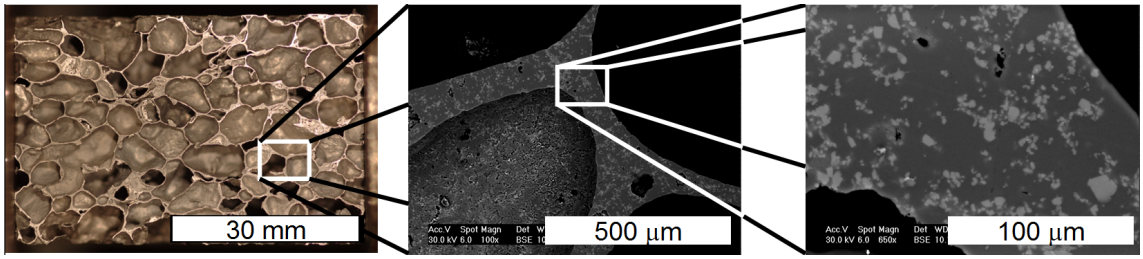
Keywords: aluminium foam, multi-scale model, nanoindentation, statistical deconvolution, elastic properties, image analysis, homogenization.

1 Introduction

Metal foams and especially lightweight aluminium foams belong to the group of up-to-date engineering materials with high potential to many applications. Metal foam is a highly porous hierarchical material with a cellular microstructure. Macroscopically, it can be characterized by attractive mechanical and physical properties such as high stiffness and strength in conjunction with very low weight, excellent impact energy absorption, high damping capacity and good sound absorption capability. The usual source material for the production of metal foams are aluminium and aluminium alloys because of low specific density (~2700 kg/m^3), low melting point (~660 °C), non-flammability, possibility of recycling and excellent corrosion resistance. Metal foams

are used in applications ranging from automotive and aerospace industries (e.g. bumpers, car body sills, motorcycle helmets) to building industry (e.g. sound proofing panels). Our aim was to characterize and to model a commercially available foam Alporas® produced by Shinko Wire Company, Ltd. .

Alporas is characterized with a hierarchical system of pores containing different cell morphologies (in shape and size) in dependence on the foam density and inhomogeneous material properties of the cell walls . A typical cross section of the foam can be seen in Figure 1 in which large pores (having typically 1-13 mm in diameter) with detailed view on thin walls (~100 μm thick) is shown.



(a) (b) (c)

Figure 1: (a) Overall view on a foam structure (further denoted as Level II); (b) ESEM image of a cell wall; (c) detailed ESEM image of a cell wall showing Al-rich (dark grey) and Ca/Ti-rich areas (light zones; denoted as Level I).

It follows from its hierarchical microstructure that the mechanical properties of metal foams are governed by two major factors:

(i) cell morphology (shape, size and distribution of cells) and

(ii) material properties of the cell walls .

Traditionally, mechanical properties of metal foams are obtained using conventional macroscopic testing techniques on large samples that can give overall (effective) properties, e.g. -. However, conventional measurements face significant obstacles in the form of very small dimensions of cell walls, low local bearing capacity, local yielding and bending of the cell walls. These problems can be overcome using micromechanical experimental methods in which the load–displacement curve is obtained in the sub-micrometer range. A few attempts have been carried out in the past, e.g. , .

In this paper, we focused on the prediction of overall foam elastic properties from microscopic measurements and on the model validation against experimental results. For accessing the cell wall properties we employed statistical nanoindentation and deconvolution technique for the phase separation ,. Compared to traditional macroscopic techniques nanoindentation can distinguish between individual inhomogeneous microstructural entities. The effective cell wall properties have been obtained through analytical and numerical up-scaling techniques . Finally, simple 2-D finite element model for the upper composite scale has been proposed and results validated by full-scale experiments.

2 Experimental part

2.1 Materials and sample preparation

Commercial aluminium foam Alporas® (Shinko Wire Company, Ltd) was used in this study. The manufacturing process of the Alporas is a batch casting process in which 1.5 wt.% of calcium is added to the aluminium molten at 680 °C. Calcium serves as a

thickening agent which increases viscosity and stabilizes the air bubbles. The alloy is poured into a casting mold and stirred with an admixture of 1.6 wt.% TiH_2 that is used as a blowing agent. Then, the foamed molten material is cooled down. A typical resulting internal structure of the aluminium foam is shown in Figure 1a.

Firstly, a large panel of Alporas (160 × 100 × 60 mm) was polished and scanned with a high resolution scanner. Acquired images were segmented to binary ones and further used in image analyses. Then, a smaller Alporas block was cut into thin slices (~5 mm) and embedded into epoxy resin to fill the pores. The surface was mechanically grinded and polished to reach minimum surface roughness suitable for nanoindentation. Very low roughness $R_q \approx 10$ nm was achieved on cell walls. The sample was investigated with electron microscopy (ESEM) and nanoindentation.

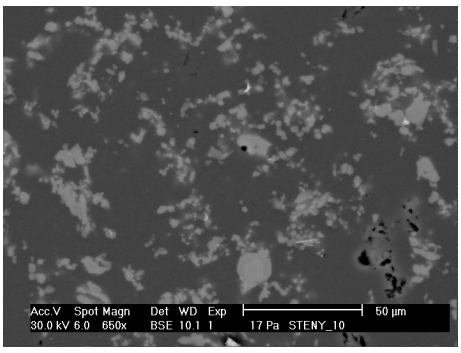
2.2 ESEM and microstructural analysis

The microstructure of cell walls was firstly studied in electron microscope (ESEM). It was found that a significant inhomogeneity of the microstrutural material phases exists on the level of tens of micrometers (Figure 1b,c). Two distinct phases, that exhibit different color in back-scattered electron (BSE) images, can be distinguished. The chemical composition of the two phases was checked with EDX element analysis in ESEM. It was found that the majority of the volume (dark zone in Figure 1c, 2a) consists of aluminium (~67 wt.%), oxygen (~32 wt.%) and further trace elements (Mg, Ti, Fe, Co, Ni, Cu, Si <2 wt.%). Lighter zones in Figure 2 consist of Al (~60 wt.%), O (~30 wt.%), Ca (~5 wt.%), Ti (~5 wt.%) and other elements (<1 wt.%). As expected, the majority of the volume (dark zone) is composed of aluminum and aluminium oxide

Al_2O_3 (further denoted as Al-rich area). Lighter zones contain significant amount of calcium and titanium (further denoted as Ca/Ti-rich area). The non-uniform distribution of these zones shows inhomogeneous mixing of the admixtures that are added during the production process.

2.3 Image analysis and porosity

In order to estimate the volume fractions of Al-rich and Ca/Ti-rich areas image analysis based on previously taken ESEM images was employed. Ten arbitrarily chosen areas on wall cross sections were explored. Images were segmented to two phases using a common threshold value of a grey level for all images (Figure 2). The Ca/Ti-rich area was estimated to cover 22±4% of the whole area.



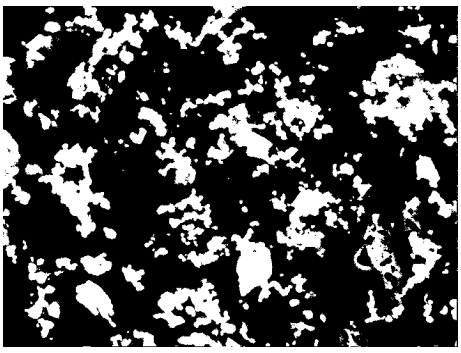
(a)

(b)

Figure 2: An example of (a) ESEM image of the cell wall and (b) processed image segmented to two phases (white=Ca/Ti-rich, black=Al-rich area).

The overall porosity of the sample was assessed by weighing of a large Alporas panel (knowing the sample dimensions and solid mass density 2700 $kg.m^{-3}$). The

porosity reached 91.4% which corresponds to e.g. ,. In other words, solid mass (i.e. cell walls) occupied only 8.6% of the total volume in the specimen.

Further, distribution of cell wall thicknesses and the distribution of the pore sizes were studied by means of pore contour detection in the Matlab environment. At first, the contours were generated for every pore in the image and areal characteristics (centroid, area, second moment of inertia) were computed (Figure 3). The wall thicknesses were calculated as the minimum distance between the neighboring contours. The distribution of the thicknesses is shown in Figure 4. It can be seen in Figure 4 that a significant peak occurs around ~60 μm which can be understood as a characteristic cell wall thickness.

Then, equivalent ellipses were constructed from contour areal characteristics under the condition that they have the same area and the same principal second moment of inertia. Such assumption led to the evaluation of two main half axes (a_i and b_i) for each equivalent ellipse. In order to characterize the shape of pores, an equivalent ellipse shape factor was defined as the ratio $e_i = \frac{a_i}{b_i}$. The distribution of the shape factor is depicted in Figure 5. It can be concluded that pores have typically a round shape with the shape factor lying mostly between 1 and 2. The peak with the highest occurrence in Figure 5 appears around e_i=1.15.

Due to the round shape of pores it makes sense to compute also an equivalent pore diameter using circular pore replacement. The distribution of equivalent circular pores is depicted in Figure 6. Wide distribution of pores with diameters 0-6 mm was found. The mean equivalent diameter was found to be 2.9±1.5 mm for the specific specimen.

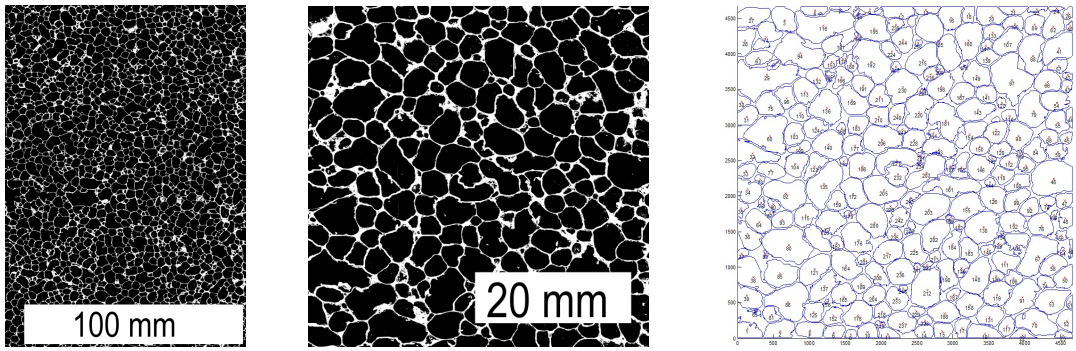


Figure 3: (a) Binary image of the polished foam panel. (b) Binary image of ~50×50 mm foam cut. (c) Cell contours in the cut (prepared in Matlab).

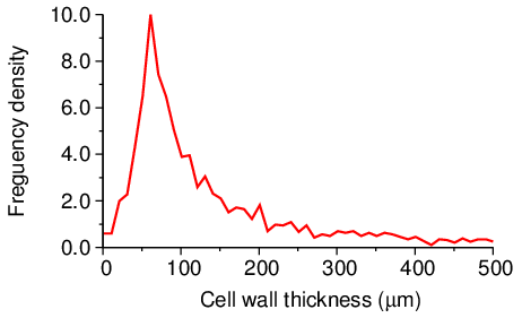


Figure 4: Distribution of cell wall thicknesses.

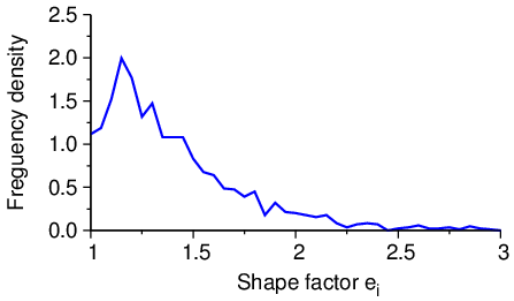


Figure 5: Distribution of equivalent ellipse shape factor.

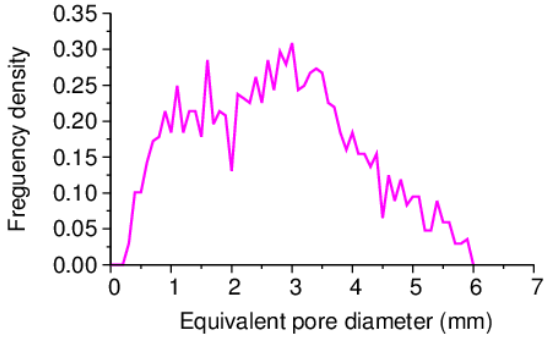


Figure 6: Distribution of equivalent circular pores.

2.4 Nanoindentation

Micromechanical properties of the cell walls were measured by means of nanoindentation. The tests were performed using a Hysitron Tribolab system® at the Czech Technical University in Prague. This system consists of in-situ SPM imaging which was used for scanning the sample surface. Three-sided pyramidal diamond tip (Berkovich type) was used for all measurements. Two distant locations were chosen on the sample to capture its heterogeneity. Each location was covered by a series of 10×10 indents with 10 μm spacing (Figure 7). It yields 200 indents in total which was considered to give sufficiently large statistical set of data. Standard load controlled test of an individual indent consisted of three segments: loading, holding at the peak and unloading. Loading and unloading of this trapezoidal loading function lasted for 5 seconds, the holding part lasted for 10 seconds. Maximum applied load was 1 mN. Maximum indentation depths were ranging between 100 and 300 nm depending on the stiffness of the indented phase. Elastic modulus was evaluated for individual indents using standard Oliver and Pharr methodology which accounts for elasto-plastic contact of a conical indenter with an isotropic half-space as

$$E_r = \frac{1}{2\beta}\frac{\sqrt{\pi}}{\sqrt{A}}\frac{dP}{dh} \qquad (1)$$

in which E_r is the reduced modulus measured in an experiment, A is the projected contact area of the indenter at the peak load, β is geometrical constant (β=1.034 for the

used Berkovich tip) and $\frac{dP}{dh}$ is a slope of the unloading branch evaluated at the peak. Elastic modulus E of the measured media can be found using contact mechanics which accounts for the effect of non-rigid indenter as

$$\frac{1}{E_r} = \frac{\left(1-v^2\right)}{E} + \frac{\left(1-v_i^2\right)}{E_i} \tag{2}$$

in which v is the Poisson's ratio of the tested material, E_i a v_i are known elastic modulus and Poisson's ratio of the indenter. In our case, v=0.35 was taken as an estimate for all indents.

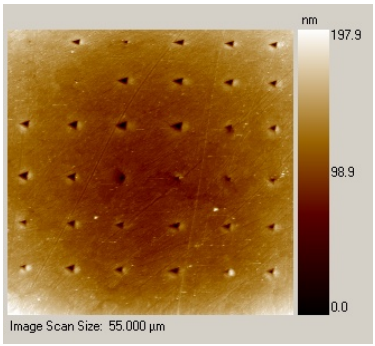


Figure 7: Part of the indentation matrix showing 6×6 indents with 10 µm spacing as scanned with Hysitron Tribolab.

Results of elastic moduli in the form of histograms have been further analyzed with the statistical deconvolution technique , . The technique seeks for parameters of individual phases included in overall results. It searches for n-Gauss distributions in an experimental Probability Density Function - PDF (Figure 9). Random seed and

minimizing criteria of the differences between the experimental and theoretical overall PDFs (particularly quadratic norm of the differences) are computed in the algorithm to find the best fit. Details on the deconvolution technique can be found in . Two-phase system (one dominant Al-rich phase and one minor Ca/Ti-rich phase) was assumed in our deconvolution.

3 Numerical part

3.1 Scale separation

In order to describe heterogeneous systems and their effective properties in a statistical sense, representative volume element (RVE) have been previously introduced . RVE statistically represents a higher structural level of the material and serves for evaluation of the effective (homogenized) properties within the defined volume. It includes all microstructural inhomogeneities that should be substantially smaller than the RVE size. The definition of the material scales can be defined through the scale separation inequality:

$$d \ll L \ll D \tag{3}$$

in which d is the characteristic size of the largest microstructural inhomogeneity, L is the RVE size and D is a characteristic structural length scale. Knowing the material and geometrical properties of the microstructural material phases a homogenization can be performed.

Nanoindentation is able to access intrinsic material properties of individual micro-scale phases provided the dimension of an indent (h) is small enough, i.e. $h<<d$. As a rule of thumb $h<d/10$ is usually used to access material properties of individual constituents without any dependence on the length scale.

As mentioned above, the metal foam material has a hierarchical microstructure. At least two levels need to be considered:

- **Level I** (the cell wall level) has a characteristic dimension defined by the mean midspan wall thickness L_I~60 μm. This level consists of prevailing aluminium matrix (Al-rich area) with embedded heterogeneities in the form of Ti/Ca-rich areas. Intrinsic elastic properties of the constituents were assessed by nanoindentation at this level. Individual indent size was prescribed to be considerably smaller (h≈100-300 nm) than a characteristic size of Ca/Ti inhomogeneities (~4 μm).
- **Level II** (the foam level) has a characteristic dimension of L_{II}~50 mm. At this level, large pores with an average equivalent diameter ~2.9 mm (assuming circular pores) occur in the total volume of 91.4%. At level II, cell walls are considered as homogeneous having the properties that come from the Level I homogenization.

3.2 Analytical homogenizations of Level I

The RVE with substantially smaller dimensions than the macroscale body allows imposing homogeneous boundary conditions over the RVE. Continuum micromechanics provides a framework, in which elastic properties of heterogeneous

microscale phases are homogenized to give overall effective properties on the upper scale . A significant group of analytical homogenization methods relies on the Eshelby's solution that is derived for ellipsoidal inclusions embedded in an infinite body. Then, uniform stress field appears in inclusions when macroscopic load is applied in infinity. Effective elastic properties are obtained through averaging over the local contributions.

From the material point of view, composite materials are usually characterized by a prevailing matrix phase, which serves as a reference medium in homogenization methods, reinforced with geometrically distinguishable inclusions. For example, the Mori-Tanaka method can be appropriate for these cases. In this method, the effective bulk k_{eff} and shear μ_{eff} moduli of the composite are computed as follows

$$k_{eff} = \frac{\sum_r f_r k_r (1+\alpha_0 (\frac{k_r}{k_0} - 1))^{-1}}{\sum_r f_r (1+\alpha_0 (\frac{k_r}{k_0} - 1))^{-1}}, \tag{4}$$

$$\mu_{eff} = \frac{\sum_r f_r \mu_r (1+\beta_0 (\frac{\mu_r}{\mu_0} - 1))^{-1}}{\sum_r f_r (1+\beta_0 (\frac{\mu_r}{\mu_0} - 1))^{-1}}, \tag{5}$$

$$\alpha_0 = \frac{3k_0}{3k_0 + 4\mu_0}, \beta_0 = \frac{6k_0 + 12\mu_0}{15k_0 + 20\mu_0} \tag{6}$$

where f_r is the volume fraction of the r^{th} phase, k_r its bulk modulus and the coefficients α_0 and β_0 describe bulk and shear properties of the 0^{th} phase, i.e. the reference medium , . The bulk and shear moduli can be directly linked with Young's modulus E and Poisson's ratio ν used in engineering computations as

$$E = \frac{9k\mu}{3k + \mu}, \tag{7}$$

$$\nu = \frac{3k - 2\mu}{6k + 2\mu}. \tag{8}$$

Polycrystalline metals, in which no preference of matrix phase exists, are usually modeled with the self-consistent scheme in which the reference medium refers back to the homogenized medium itself. Regardless the most suitable homogenization technique, which would be probably the Mori-Tanaka method in our case, we used multiple estimates assuming spherical inclusions. Namely, the Mori-Tanaka method, self-consistent scheme, Voigt and Reuss bounds (parallel or serial configuration of phases with perfect bonding). Results from nanoindentation have been used as input parameters to the methods.

3.3 Numerical homogenization of Level I based on FFT

In order to verify results from simple analytical schemes advanced homogenization method based on fast Fourier transformation (FFT) was used. The behavior of any heterogeneous materials consisting of periodically repeating RVE (occupying domain $\Omega = \prod_{i=1}^{d}(-Y_i, Y_i)$, where Y_i is the axial size and d denotes the space dimension) can be described with differential equations with periodic boundary conditions and prescribed macroscopic load as

$$\langle \boldsymbol{\varepsilon} \rangle := \frac{1}{|\Omega|}\int_{\Omega} \boldsymbol{\varepsilon}(\mathbf{x}) d\mathbf{x} = \boldsymbol{\varepsilon}^{0} \tag{9}$$

$$\boldsymbol{\sigma}(\mathbf{x}) = \mathrm{L}(\mathbf{x}) : \boldsymbol{\varepsilon}(\mathbf{x}) \qquad div\boldsymbol{\sigma}(\mathbf{x}) = \mathbf{0} \qquad \mathbf{x} \in \Omega \tag{10}$$

where $\boldsymbol{\sigma}$ denotes symmetric second order stress tensor, $\boldsymbol{\varepsilon}$ symmetric second order strain tensor and $\mathrm{L}(\mathbf{x})$ the fourth order tensor of elastic stiffness at individual locations $\mathbf{x}$. The effective (homogenized) material tensor $\mathrm{L}_{\mathrm{eff}}$ is such a tensor satisfying

$$\langle\boldsymbol{\sigma}\rangle = \mathrm{L}_{\mathrm{eff}}\langle\boldsymbol{\varepsilon}\rangle \tag{11}$$

for an arbitrary macroscopic strain $\boldsymbol{\varepsilon}^0$. Thus the problem of finding effective material tensor is composed of finding corresponding strain field $\boldsymbol{\varepsilon}$ and associated stress field $\boldsymbol{\sigma}$ for known elastic properties L and prescribed strain $\boldsymbol{\varepsilon}^0$ using differential Equation 10.

In addition to discretization of the weak formulation leading to classical finite element method, the problem can be solved by method based on fast Fourier transform, proposed by Moulinec and Suquet in , based on an integral (Lippmann–Schwinger) equation

$$\boldsymbol{\varepsilon}(\mathbf{x}) + \int_{\Omega}\Gamma^0(\mathbf{x}-\mathbf{y}) : (\mathrm{L}(\mathbf{y})-\mathrm{L}^0) : \boldsymbol{\varepsilon}(\mathbf{y})d\mathbf{y} = \boldsymbol{\varepsilon}^0 \tag{12}$$

where Γ^0 is the periodic Green's operator associated with the reference elasticity tensor L^0 which is a parameter of the method. The operator is expressed in the Fourier space as

$$\hat{\Gamma}^0_{ijkl}(\xi) = \frac{1}{4\mu|\xi|^2}(\delta_{ki}\xi_l\xi_j + \delta_{li}\xi_k\xi_j + \delta_{kj}\xi_l\xi_i + \delta_{lj}\xi_k\xi_i) - \frac{\lambda+\mu}{\mu(\lambda+2\mu)}\frac{\xi_i\xi_j\xi_k\xi_l}{|\xi|^4} \tag{13}$$

.

Numerical solution of Equation 13 is based on the discretization of a unit cell Ω into a regular periodic grid with $N_1 \times \cdots \times N_d$ nodal points and grid spacings $\mathbf{h} = (2\frac{Y_1}{N_1}, \ldots, 2\frac{Y_d}{N_d})$. The searched field $\boldsymbol{\varepsilon}$ is approximated by a trigonometric polynomial $\mathbf{e}_N$ in the form

$$\mathbf{e}(\mathbf{x}) \approx \mathbf{e}_N(\mathbf{x}) = \sum_{\mathbf{k} \in Z_N^d} \hat{\mathbf{e}}(\mathbf{k}) \varphi_{\mathbf{k}}(\mathbf{x}), \mathbf{x} \in \Omega, \tag{14}$$

where $\mathbf{N} = (N_1, \ldots, N_d)$, $\hat{\mathbf{e}}$ designates the Fourier coefficients and

$$Z_N^d = \left\{ \mathbf{k} \in Z^d : -\frac{N_\alpha}{2} < k_\alpha \leq \frac{N_\alpha}{2}, \alpha = 1, \ldots, d \right\}. \tag{15}$$

The discretization leads to a nonsymmetrical linear system of equations

$$[\mathbf{I} + \mathbf{F}^{-1} \hat{\Gamma} \mathbf{F} (\mathbf{L} - \mathbf{L}^0)] \mathbf{e} = \mathbf{e}^0 \tag{16}$$

where the vector $\mathbf{e}$ stores a strain field at discretization points and $\mathbf{e}^0$ the macroscopic strain, $\mathbf{L}$ and $\mathbf{L}^0$ stores the material coefficients at discretization points and reference elasticity tensor respectively, $\mathbf{I}$ denotes the identity matrix, $\hat{\Gamma}$ stores the values corresponding to the integral kernel in the Fourier space, and $\mathbf{F}$ ($\mathbf{F}^{-1}$) stores the

(inverse) discrete Fourier transform matrices that can be provided by fast Fourier transform algorithm. The possibility to solve the nonsymmetric linear system by the conjugate gradient method is proposed by Zeman et.al. in and justified in Vondřejc et al. , where also the particular expression of individual matrices can be found for the problem of electric conductivity or heat transfer. The linear system (Equation 16) depends only on stiffness coefficients at grid points that can be obtained using nanoindentation and thus the homogenized (effective) tensor $\mathbf{L}_{\mathrm{eff}}^{\mathrm{FFT}}$ can be calculated from Equation 11. The particular case of homogenization of elastic properties received from nanoindentation on a sample surface (half-space) also requires an assumption of plane strain conditions.

3.4 Comparison of analytical and numerical schemes

The comparison of analytical and FFT schemes includes an assessment of the stiffness matrix (here in Mandel's notation) for isotropic material assuming plane strain conditions (equally with the FFT scheme) as

$$\mathbf{L}_{\mathbf{eff}}^{\mathbf{A}} = \frac{E_{eff}}{(1+\nu_{eff})(1-2\nu_{eff})}\begin{bmatrix} 1-\nu_{eff} & \nu & 0 \\ \nu & 1-\nu_{eff} & 0 \\ 0 & 0 & 1-2\nu_{eff} \end{bmatrix}. \qquad (17)$$

The difference between the analytical results ($\mathbf{L}_{\mathbf{eff}}^{\mathbf{A}}$) and numerically computed stiffness matrix ($\mathbf{L}_{\mathbf{eff}}^{\mathbf{FFT}}$) can be expressed using a stiffness error norm as

$$\delta = \sqrt{\frac{\left(\mathbf{L}_{\mathrm{eff}}^{\mathrm{FFT}} - \mathbf{L}_{\mathrm{eff}}^{\mathrm{A}}\right) :: \left(\mathbf{L}_{\mathrm{eff}}^{\mathrm{FFT}} - \mathbf{L}_{\mathrm{eff}}^{\mathrm{A}}\right)}{\left(\mathbf{L}_{\mathrm{eff}}^{\mathrm{FFT}} :: \mathbf{L}_{\mathrm{eff}}^{\mathrm{FFT}}\right)}} \,. \tag{18}$$

4 Results and discussion

4.1 Nanoindentation

Results from nanoindentation clearly showed heterogeneity of the cell walls, i.e. the presence of mechanically different inclusions. An example of typical loading diagrams gained from nanoindentation at Al-rich area (dark zone in Figure 1b,c) and Ca/Ti-rich area (light zone in Figure 1b,c) are shown in Figure 8. Due to the load controlled nanoindentation test, the final penetration depth varied for differently stiff phases. An average maximum depth of penetration reached by the indenter was around ~180 nm. Higher values for more compliant Al-rich zone were reached (~190 nm) whereas the indentation depths to harder but less frequent Ca/Ti-rich areas were around 100 nm.

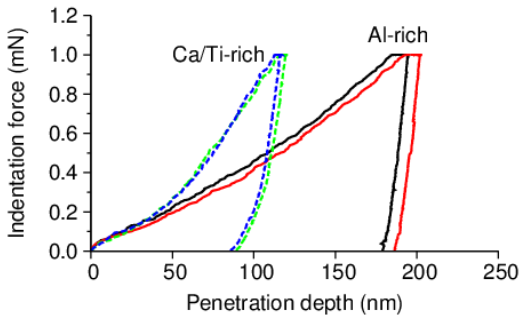
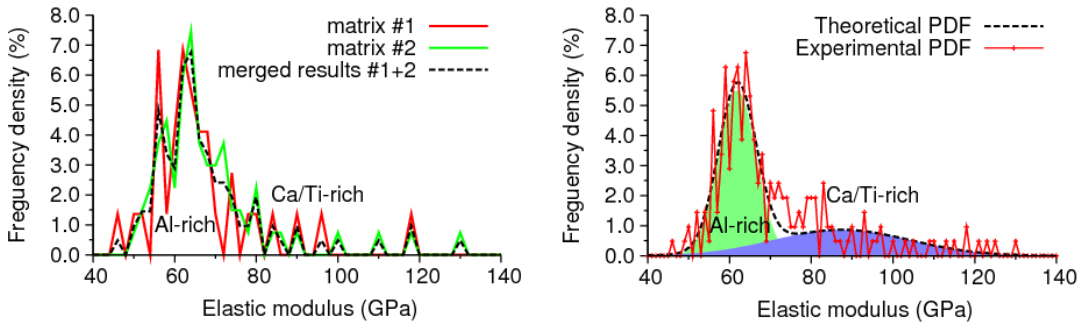


Figure 8: Typical loading diagrams for Al-rich and Ca/Ti-rich zones.

Elastic moduli were evaluated for each individual indent. Overall results are depicted in Figure 9a in which histogram of all elastic moduli from two different positions and results merged from both positions are shown. No significant differences between the positions were found. Therefore, merged results were further used in the deconvolution of phase elastic properties.

Two-phase system (one dominant Al-rich phase and one minor Ca/Ti-rich phase) was assumed in the deconvolution algorithm (Figure 9b). It can be seen in Figure 9b that a significant peak appears around 62 GPa. This value can be considered as a dominant characteristic of the prevailing phase (Al-rich). The rest of results can be attributed to the minor Ca/Ti-rich phase. Table 1 contains numerical results from the deconvolution with the estimated volume fractions of the phases.



(a) (b)

Figure 9: (a) Probability density functions of elastic moduli from two measured positions and (b) merged results with deconvoluted phases.

Table 1: Elastic moduli and volume fractions from deconvolution.

Phase	Mean (GPa)	St. dev. (GPa)	Volume fraction (-)
1 (Al-rich zone)	61.88	4.6	0.638
2 (Ca/Ti-rich zone)	87.40	16.7	0.362

The characteristic value for the first phase roughly corresponds to the elastic modulus of pure aluminium (70 GPa, ref.). The lower value obtained from nanoindentation suggests that probably some small-scale porosity or impurities (Ca) added to the molten are intrinsically included in the results of this mechanically dominant phase. The determined elastic modulus value of Al-rich zone is also in excellent agreement with the value 61.7 GPa measured by Jeon et al. on melted Al-1.5 wt.%Ca alloy.

4.2 Level I homogenization

It is clear from ESEM images (Figure 1 and 2) that the Ca/Ti-rich areas occupy much larger space of the solid compared to the initial batch volume fractions (Ca and TiH_2 content is less than 1 vol.%). Chemical reactions and precipitation during hardening form new compounds in the Al matrix. It follows from other studies that Ca/Ti-rich discrete precipitates and diffuse Al_4Ca areas develop in the metal solid. These areas are denoted as Ca/Ti-rich areas in this study. Based on the color in ESEM images the area is estimated as 22±4% by image analysis. Results from statistical nanoindentation (36.2%) suggest that a substantially larger part of the matrix is mechanically influenced by the Ca/Ti addition and a higher fraction of the volume belongs to this mechanically distinct phase.

The homogenized elastic modulus for the two considered microscale phases in the cell wall (i.e. Level I) is summarized in Table 2 for individual homogenization techniques. Very close bounds and insignificant differences in the elastic moduli estimated by the schemes were found.

Table 2: Values of the Level I effective Young's modulus computed by different homogenization schemes.

Scheme	Mori-Tanaka	Self-consist. scheme	Voigt bound	Reuss bound
E (GPa)	70.083	70.135	71.118	69.195

In the following considerations, we use the result received from the Mori-Tanaka scheme, i.e. we take the homogenized isotropic elastic constants (Young's modulus and Poisson's ratio) of the Level I as $E_{eff,I}$=70.083 GPa, $\nu_{eff,I}$=0.35. The stiffness matrices computed in Mandel's notation from analytical Mori-Tanaka results (using Equations 4-8) and from FFT homogenization are:

$$\mathbf{L}_{\mathbf{eff}}^{\mathbf{A}} = \begin{bmatrix} 112.479 & 60.566 & 0 \\ 60.566 & 112.479 & 0 \\ 0 & 0 & 51.913 \end{bmatrix} \text{(GPa)} \tag{19}$$

$$\mathbf{L}_{\mathbf{eff}}^{\mathbf{FFT}} = \begin{bmatrix} 117.1300 & 62.7413 & -0.1625 \\ 62.7413 & 117.1060 & -0.1430 \\ -0.1625 & -0.1430 & 54.3132 \end{bmatrix} \text{(GPa)}. \tag{20}$$

It is worth noting that the analytical form of the stiffness matrix $\mathbf{L}_{\mathbf{eff}}^{\mathbf{A}}$ as well as $\mathbf{L}_{\mathbf{eff}}^{\mathbf{FFT}}$ contains perfect symmetry by definition. Further, low values of off-axis components associated with shear strains in $\mathbf{L}_{\mathbf{eff}}^{\mathbf{FFT}}$ (that are zero in case of $\mathbf{L}_{\mathbf{eff}}^{\mathbf{A}}$) show close to isotropic nature of the material. In other words, microstructural inhomogeneities are uniformly dispersed in the RVE. Consequently, this finding also

justifies the usage of analytical methods producing isotropic effective (homogenized) properties. Also, the stiffness error evaluated by Equation 18 is δ=0.0393. The difference in schemes less than 4% shows a very good agreement of the methods for the studied case.

4.3 Level II homogenization

At this level, cell walls are considered as a homogeneous phase having the properties that come from the Level I homogenization. The solid phase is very sparse in the sample volume due to its porosity (91.4% of air). The walls create a matrix phase and the large air pores can be considered as inclusions in this homogenization.

Since analytical approaches are often used also for extreme cases of large stiffness contrast of phases or for large sample porosities (e.g. Šejnoha et al.), we firstly tried to estimate effective elastic properties with the same analytical schemes used at Level I. The result is summarized in Table 3. Voigt and Reuss bounds are quite distant in this case. Unfortunately, simple analytical schemes also fail to predict correctly the composite stiffness due to the extreme sample porosity. The Mori-Tanaka method approaches the arithmetic mean between the bounds, whereas the self-consistent scheme tends to reach the stiffness of the phase with higher occurrence (i.e. the air).

Table 3: Values of the Level II effective Young's modulus computed by different analytical homogenization schemes.

Scheme	Mori-Tanaka	Self-consist. scheme	Voigt bound	Reuss bound
E (GPa)	3.1510	0.0012	6.0200	0.0011

Most of analytical studies on the homogenization of foams are based on models with a regular periodic microstructure ,. Nevertheless, real foam microstructures are characterized with different sizes and shapes and sizes of pores rather than with periodic structures as shown in Section 2 of this paper. The solution can be to solve the problem of irregular microstructures by an analysis of a large representative volume element containing large enough number of pores. Such model can be solved in two or three dimensions.

Therefore, more appropriate (but still simple) two dimensional microstructure based FEM model was proposed. The model geometry was generated from high resolution optical image of polished foam cross-section (Figure 10a) converted to binary one. Square domain with 106×106 mm size (i.e. being much larger than average pore size ~2.9 mm) was extracted from the image. At this domain, pore centroids were detected, Delaunay triangulation applied and Voronoi cells created. Then, an equivalent 2D-beam structure was generated from Voronoi cell boundaries (Figure 10b). Based on several numerical studies performed for this purpose (but not shown here in details), it was found that the distribution of cross sectional areas and bending stiffness of individual beams do not play a significant role in the evaluation of the homogenized properties.
The overall stiffness is influenced mainly by the sum of the beam cross sectional areas and by the beam inclination to the load direction. The contribution of the beam bending

stiffness is diminished due to the very large beam length compared to its small cross sectional dimensions. Therefore, as an approximate but sufficient estimate, uniform cross-sectional area and uniform second moment of inertia were prescribed to all beams. The beam cross sectional area (A_{beam}) was computed from the total sample porosity (φ=0.086) and the total length of all beams (l_{total}) in the RVE with rectangular dimensions $a \times b$ as

$$A_{beam} = \frac{ab\varphi}{l_{total}}. \tag{21}$$

Taking into account 2-D case (i.e. unit thickness of the plane) beam height can be set as

$$h_{z,beam} = \frac{A_{beam}}{1} = A_{beam}. \tag{22}$$

Assuming rectangular shape of a cross section one can readily obtain the second moment of inertia as

$$I_{y,beam} = \frac{1}{12} 1 h_{z,beam}^3 = \frac{A_{beam}^3}{12}. \tag{23}$$

In the analysis, macroscopic strain $\boldsymbol{\varepsilon}^0$ is prescribed to the RVE and microscopic strains and stresses are solved. Volumetric averaging of microscopic stresses leads to the assessment of an average macroscopic stress and finally estimation of effective stiffness parameters. The key issue of the computation is the size of RVE and application of boundary conditions around the domain. Since the domain size is always smaller than an infinite body, any constraints can strongly influence the results. Application of the kinematic boundary conditions leads to the overestimation of effective stiffness and it can give an upper bound, whereas the static boundary conditions give a lower bound .

The best solution is usually provided by applying periodic boundary conditions to RVE which are, however, difficult to implement into commercial codes.

Nevertheless, the influence of the boundary conditions on microscopic strains and stresses in the domain decrease in distant points from the boundary. The size of our domain (106×106 mm) allowed us to solve the problem with kinematic boundary conditions. For homogenization, considerably smaller region (later found optimum 35-50 mm) in the central part was used. Microscopic strains and stresses were computed inside this smaller area which was assumed to be still sufficiently large to describe the material inhomogeneities and to serve as material RVE.

Kinematic constrains were applied on all domain sides. Free beams located around the boundary and not connected to any cell were deleted and supports put on the nodes located on the closest cell. Such arrangement of beams and supports prevented the structure from unreasonably large deformations of these free boundary beams. The whole domain (106×106 mm) was subjected to homogeneous macroscopic strain in one axial direction ($\boldsymbol{\varepsilon}^0 = \{1,0,0\}^T$) by imposing prescribed displacement to one domain side (Figure 10c). The test was performed using commercial Ansys FEM software and microscopic strains and stresses solved in the domain. Strains and stresses (structural forces for the case of beams, respectively) inside the smaller area (35-50 mm) were averaged and used for computation of the homogenized stiffness matrix (one column in the matrix, respectively). Assuming material isotropy, the first component (1,1) at the material stiffness matrix is given by:

$$L_{11} = E\frac{(1-\nu)}{(1+\nu)(1-2\nu)} \tag{24}$$

in which E is the Young's modulus and ν Poisson's ratio, respectively. Since the Poisson's ratio of the whole foam is close to zero (as confirmed by experimental measurements) the L_{11} member coincides with the Young's modulus E.

For the tension test in x-direction (Figure 10c), the homogenized Young's modulus was found to be RVE size dependent. Experimental investigations of the dependence of sample size on apparent elastic modulus and strength were conducted e.g. by Ashby et al. . They found, the modulus and strength become independent of size when the sample dimensions exceeded about seven cell diameters. This would imply minimum RVE size 20.3 mm for our typical cell size (2.9 mm). On the other hand, the RVE size should not exceed roughly 1/3 to 1/2 of the whole domain size not to be influenced by boundary conditions which implies maximum RVE width about 35 to 53 mm for our 106 mm wide domain. To find an optimum RVE size a numerical study was conducted for different RVE sizes in the range 20 to 90 mm (see Figure 11). An optimum RVE was confirmed to be between 35 and 50 mm for our specific domain. Results for smaller RVEs (<35 mm) are influenced by the beam inhomogeneity inside the RVE (in other words, such small RVE is not representative enough) whereas larger RVEs (>50 mm) are already influenced by the vicinity of boundary conditions. For optimum RVE sizes, the effective Young's modulus varied in the range $E_{eff,II}$=1.36-1.38 GPa.

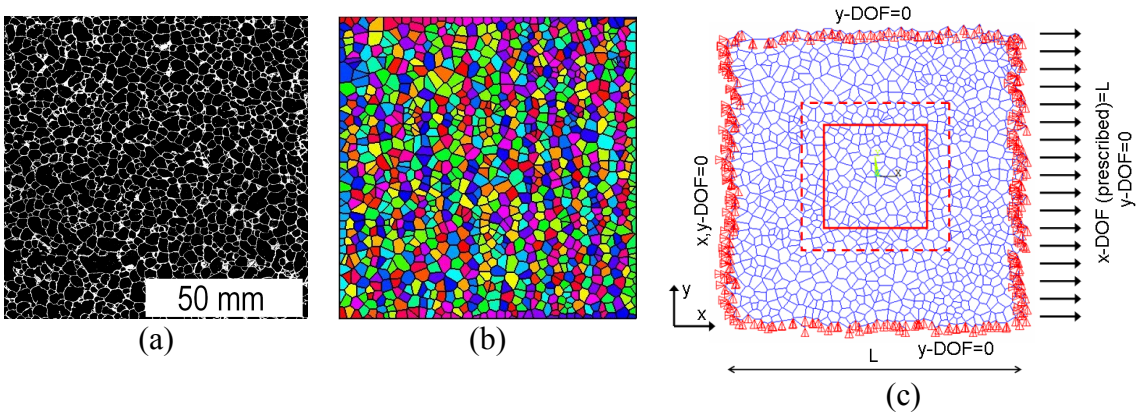


Figure 10: (a) Binary image of the foam (106×106 mm). (b) Voronoi tesselation. (c) 2-D beam model with boundary constraints (red squares indicate optimum RVE sizes from which homogenized properties have been obtained; solid line=35×35 mm, dashed line=50×50 mm).

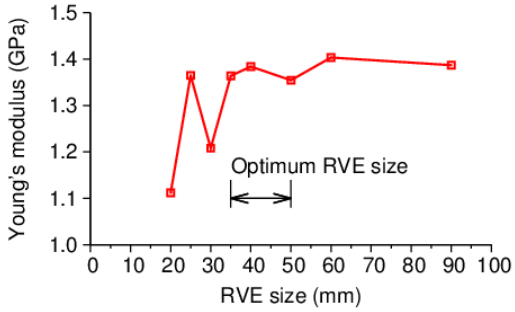


Figure 11: Dependence of effective Level II Young's modulus on RVE size.

The resulting homogenized Young's modulus is comparable with the range of experimental values (0.4−1 GPa) reported for Alporas® e.g. by Ashby et. al. . Experimental measurements in uniaxial compression performed on our samples (30×30×60/90 mm Alporas blocks) indicate E=1.45±0.15 GPa (see Section 4.4). The slightly lower stiffness obtained from two-dimensional model can be explained by the lack of additional confinement appearing in the three-dimensional case. The influence

of the RVE size can also play a role as described above. However, the obtained difference is small (~5%), probably also due to the almost zero foam Poisson's ratio. Anyway, results of the simplified 2-D model have to be treated as a relatively close but only the first estimate of the Level II material properties which should be refined e.g. by using more precise three-dimensional model.

4.4 Results from macroscopic measurements

Uniaxial compression tests on 30×30×60 and 30×30×90 mm Alporas blocks (Figure 12a) were performed in an electromechanical press to verify numerical results on the Level II. Specimens were loaded-unloaded by five to ten cycles at very low strains and than fully compressed up to ~5% longitudinal strain (Figure 12b,c). Longitudinal and transversal (engineering) strains were evaluated by means of digital image correlation from CCD camera images taken during the test . Negligible differences have been found between the slopes of loading/unloading cycles (Figure 12c) which justifies evaluation of elastic properties from this part of the loading diagram. Young's modulus was finally computed as the average slope from all relevant cycles (i.e. all cycles except the first and the last one that both can be influenced by non-linear effects). Young's modulus was determined as $E=1.45\pm0.15$ GPa on six foam samples. Poisson's ratio was found to be $\nu\approx0$ in the elastic regime.

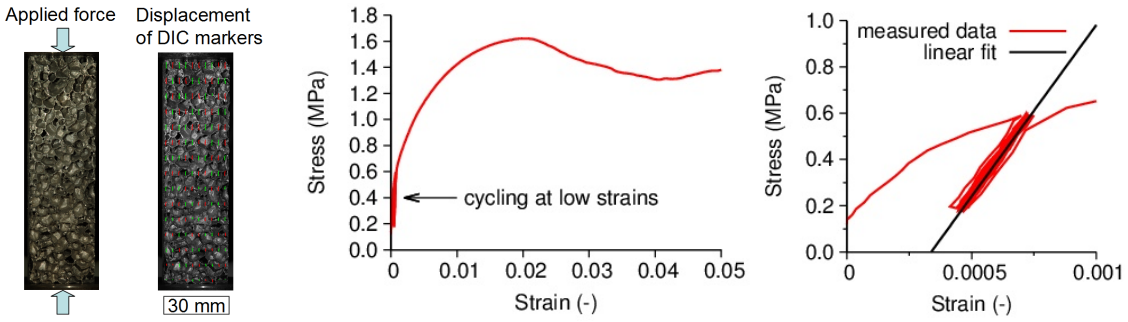


Figure 12: Compression test. (a) Foam sample and an image with digital image correlation markers, (b) stress strain diagram, (c) detail of loading-unloading cycles at low strains.

5 Conclusions

A simple but effective two-scale microstructure based model of closed-cell aluminium foam for the assessment of homogenized elastic properties was proposed in this paper. The first level characterized by thin cell walls (~60 μm) was successfully homogenized with several analytical continuum mechanics schemes. Two different material phases (Al-rich and Ca/Ti-rich) were detected at this lower scale by ESEM and statistical grid nanoindentation. Effective Young's modulus $E_{eff,I}$≈70 GPa was received regardless the used scheme. The value was also justified by numerical FFT-based homogenization with a very good agreement (error less than 4%).

The upper foam level (Level II) contained homogenized walls and large air pores. Here, analytical tools were applied without success. Very poor estimates were given by the Mori-Tanaka or self-consistent due to extremely high air content in the foam and large stiffness contrast. To better describe the real foam microstructure, a FEM model

was proposed for the numerical homogenization at the second level. The model geometry was generated from large optical scan of polished foam cross section converted to binary image. Delaunay triangulation and Voronoi tessellation have been applied and equivalent beam structure generated. The dependence of RVE size was solved in a large domain (106×106 mm) supported by kinematic boundary conditions. An optimum RVE size was found to be in the range 35-50 mm (i.e. 33-47% of the domain size) for which effective elastic properties were assessed ($E_{eff,II}$=1.36-1.38 GPa).

The model has proven to realistically describe macroscopic elastic properties of the foam. Nevertheless, two-dimensional approximation slightly underestimated the experimentally obtained stiffness (E≈1.45 GPa). It is likely due to the inability to capture additional confinement coming from three-dimensional material microstructure or due to the RVE size inaccuracy. Further development of the numerical model and generation of the model geometry from micro-CT data (i.e. extension to 3-D) are planned as future developments.

Acknowledgements

Financial support of the Czech Science Foundation is gratefully acknowledged (projects P105/12/0824 and P105/12/0331).

References

[1] Banhart J, Manufacture, characterisation and application of cellular metals and metal foams, Progress in Materials Science 46: 559-632, 2001.

[2] Miyoshi T, Itoh M, Akiyama S, Kitahara A, Aluminium foam "ALPORAS": The production process, properties and application" Materials Research Society Symp. Proc. 521, 1998.

[3] Hasan MA, Kim A, Lee H-J, Measuring the cell wall mechanical properties of Al-alloy foams using the nanoindentation method, Composite Structures 83: 180-188, 2008.

[4] Papadopoulos DP, Konstantinidis IC, Papanastasiou N, Skolianos S, Lefakis H and Tsipas DN, Mechanical properties of Al metal foams, Materials Letters 58 (21): 2574-2578, 2004.

[5] Jeon I et al., Cell wall mechanical properties of closed-cell Al foam, Mechanics of Materials, 41 (1): 60-73, 2009.

[6] Sugimura Y, Meyer J, He MY, Bart-Smith H, Grenstedt J, Evans AG, On the mechanical performance of closed cell Al alloy foams, Acta Mater. 45: 5245–5259, 1997.

[7] Idris MI, Vodenitcharova, Hoffman M, Mechanical behaviour and energy absorption of closed-cell aluminium foam panels in uniaxial compression, Materials Science and Engineering A 517, 37–45, 2009.

[8] Yongliang M, Guangchun Y, Hongjie L, Effect of cell shape anisotropy on the compressive behavior of closed-cellaluminum foams, Materials and Design 31, 1567–1569, 2010.

[9] De Giorgi M, Carofalo A, Dattoma V, Nobile R, Palano F, Aluminium foams structural modelling, Computers and Structures 88, 25–35, 2010.

[10] Constantinides G, Chandran FR, Ulm F-J, Vliet KV, Grid indentation analysis of composite microstructure and mechanics: Principles and validation, Materials Science and Engineering: A, 430 (1-2):189-202, 2006.

[11] Němeček J, Šmilauer V, Kopecký L, Nanoindentation characteristics of alkali-activated aluminosilicate materials, Cement and Concrete Composites 33 (2): 163-170, 2011.

[12] Zaoui A, Continuum Micromechanics: Survey, Journal of Engineering Mechanics 128, 2002.

[13] ISO 4287-1997, "Geometrical Product Specifications (GPS) - Surface texture: Profile method - Terms, definitions and surface texture parameters".

[14] Ashby MF et al., Metal Foams: A Design Guide, Elsevier, 2000.

[15] Oliver W, Pharr GM, An improved technique for determining hardness and elastic modulus using load and displacement sensing indentation experiments, J. Mater. Res. 7 (6): 1564–1583, 1992.

[16] Eshelby JD, The determination of the elastic field of an ellipsoidal inclusion and related problem, Proc. Roy. Soc. London A 241: 376–396 1957.

[17] Mori T, Tanaka K, Average stress in matrix and average elastic energy of materials with misfitting inclusions, Acta Metallurgica 21 (5): 571-574, 1973.

[18] Webelements on-line library, http://www.webelements.com/aluminium/physics.html

[19] Moulinec H, Suquet P, A fast numerical method for computing the linear and nonlinear mechanical properties of composites, Comptes rendus de l'Académie des sciences. Série II, Mécanique, physique, chimie, astronomie, 318(11): 1417-1423, 1994.

[20] Vainikko G, Fast solvers of the Lippmann-Schwinger equation, in R. P. Gilbert and J. Kajiwara and Y. S. Xu, editors, Direct and Inverse Problems of Mathematical Physics in International Society for Analysis, Applications and Computation, pages 423-440. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.

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

[22] Vondřejc J, Zeman J, Marek I, Analysis of a Fast Fourier Transform Based Method for Modeling of Heterogeneous Materials, Lecture Notes in Computer Science 7116: 512-522, 2012.

[23] Simone AE, Gibson LJ, Aluminum Foams Produced by Liquid-state Processes, Acta Mater 46 (9): 3109-3123, 1998.

[24] Šejnoha M, Šmilauer V, Němeček J, Kopecký L, Application of Micromechanics in Engineering Practice, Proceedings of the Eighth International Conference on Engineering Computational Technology, B.H.V. Topping, (Editor), Civil-Comp Press, Stirlingshire, Scotland, 2012.

[25] Gibson LJ, Ashby MF, Cellular Solids: Structure and Properties, Cambridge University Press, 1999.

[26] Hohe J, Becker W, A probabilistic approach to the numerical homogenization of irregular solid foams in the finite strain regime, International Journal of Solids and Structures 42, 3549–3569, 2005.

[27] Šmilauer V, Bittnar Z, Microstructure-based micromechanical prediction of elastic properties in hydrating cement paste, Cement and Concrete Research 36, 1708–1718, 2006.

[28] Ashby MF, Evans A, Fleck NA, Gibson LJ, Hutchinson JW, Wadley HN, Metal foams: a design guide, Butterworth-Heinemann, Oxford, UK, 2000.

[29] Jandejsek I, Jiroušek O, Vavřík D, Precise strain measurement in complex materials using Digital Volumetric Correlation and time lapse micro-CT data, Procedia Engineering 10, 1730–1735, 2011.

Figure captions

Figure 1: (a) Overall view on a foam structure (further denoted as Level II); (b) ESEM image of a cell wall; (c) detailed ESEM image of a cell wall showing Al-rich (dark grey) and Ca/Ti-rich areas (light zones; denoted as Level I).

Figure 2: An example of (a) ESEM image of the cell wall and (b) processed image segmented to two phases (white=Ca/Ti-rich, black=Al-rich area).

Figure 3: (a) Binary image of the polished foam panel. (b) Binary image of ~50×50 mm foam cut. (c) Cell contours in the cut (prepared in Matlab).

Figure 4: Distribution of cell wall thicknesses.

Figure 5: Distribution of equivalent ellipse shape factor.

Figure 6: Distribution of equivalent circular pores.

Figure 7: Part of the indentation matrix showing 6×6 indents with 10 μm spacing as scanned with Hysitron Tribolab.

Figure 8: Typical loading diagrams for Al-rich and Ca/Ti-rich zones.

Figure 9: (a) Probability density functions of elastic moduli from two measured positions and (b) merged results with deconvoluted phases.

Figure 10: (a) Binary image of the foam (106×106 mm). (b) Voronoi tesselation. (c) 2-D beam model with boundary constraints (red squares indicate optimum RVE sizes from which homogenized properties have been obtained; solid line=35×35 mm, dashed line=50×50 mm).

Figure 11: Dependence of effective Level II Young's modulus on RVE size.

Figure 12: Compression test. (a) Foam sample and an image with digital image correlation markers, (b) stress strain diagram, (c) detail of loading-unloading cycles at low strains.

Tables

Table 1: Elastic moduli and volume fractions from deconvolution.

Phase	Mean (GPa)	St. dev. (GPa)	Volume fraction (-)
1 (Al-rich zone)	61.88	4.6	0.638
2 (Ca/Ti-rich zone)	87.40	16.7	0.362

Table 2: Values of the Level I effective Young's modulus computed by different homogenization schemes.

Scheme	Mori-Tanaka	Self-consist. scheme	Voigt bound	Reuss bound
E (GPa)	70.083	70.135	71.118	69.195

Table 3: Values of the Level II effective Young's modulus computed by different analytical homogenization schemes.

Scheme	Mori-Tanaka	Self-consist. scheme	Voigt bound	Reuss bound
E (GPa)	3.1510	0.0012	6.0200	0.0011

Part VI
Paper 5

Authors:
Jaroslav Vondřejc, Jan Zeman, and Ivo Marek

Title:
FFT-based Finite element method for homogenization

FFT-based Finite element method for homogenization

J.Vondřejc, J.Zeman and I.Marek

January, 2013

Abstract

We present a mathematical theory to a FFT-based homogenization, the numerical method introduced by Moulinec and Suquet [14], The method is based on the Lippmann-Schwinger integral equation including the Green function. We show its equivalence to weak formulation in the sense the unique solution coincide. Then we provide an unifying concept of the discretization with trigonometric polynomials [19] that is applicable for both formulations. Moreover, we explain the solution of the resulting non-symmetric linear system by Conjugate gradients as proposed in [24]. Finally, the convergence of discrete solutions to continuous one is provided.

1 Introduction

A majority of computational homogenization techniques rely on the solution to the unite cell problem, which concerns the determination of local fields in a representative sample of a heterogeneous material under periodic boundary conditions. In order to show the approach, a scalar problem modeling heat transfer, diffusion, or electric conductivity (our choice), is considered and formulated as

$$\operatorname{curl} \boldsymbol{e}(\boldsymbol{x}) = \boldsymbol{0}, \quad \operatorname{div} \boldsymbol{j}(\boldsymbol{x}) = \boldsymbol{0}, \quad \boldsymbol{j}(\boldsymbol{x}) = \boldsymbol{A}(\boldsymbol{x})\boldsymbol{e}(\boldsymbol{x}),\ \boldsymbol{x} \in \mathcal{Y} \tag{1.1}$$

where $\mathcal{Y} = \prod_{\alpha=1}^{d}(-Y_\alpha, Y_\alpha) \subset \mathbb{R}^d$ denotes a periodic unit cell in d-dimensional space, $\boldsymbol{e} = \left(e_\alpha\right)_{\alpha=1}^{d} : \mathcal{Y} \to \mathbb{R}^d$ a vector valued electric field, $\boldsymbol{j} = \left(j_\alpha\right)_{\alpha=1}^{d} : \mathcal{Y} \to \mathbb{R}^d$ the corresponding vector of electric current, $\boldsymbol{A} = \left(A_{\alpha\beta}\right)_{\alpha,\beta=1}^{d} : \mathcal{Y} \to \mathbb{R}^{d\times d}$ a second-order uniformly elliptic bounded tensor of electric conductivity, and $\boldsymbol{Ae}$ express the tensor by vector field multiplication, i.e. $\boldsymbol{Ae} = \left(\sum_{\beta=1}^{d} A_{\alpha\beta}e_\beta\right)_{\alpha=1}^{d}$. The differential equation (1.1) is constrained to a periodic boundary condition with period $\boldsymbol{Y} \in \mathbb{R}^d$ and the prescribed average load coming from the macroscopic level average electric field $\boldsymbol{E}$, i.e.

$$\langle \boldsymbol{e} \rangle := \frac{1}{|\mathcal{Y}|_d} \int_{\mathcal{Y}} \boldsymbol{e}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \boldsymbol{E}, \tag{1.2}$$

where $\boldsymbol{E}$ denotes a prescribed macroscopic electric field and $|\mathcal{Y}|_d$ represents the d-dimensional measure of $\mathcal{Y}$.

The solution of initial problem, (1.1) and (1.2), can be very demanding especially for complicated microstructures and high contrasts in conductivity coefficients in combination with higher-dimensional problems. There are various works dealing with discretization of weak formulation, Def. 2.20, especially by Finite Element Method (FEM) with polynomials as the basis functions [1, 6, 7, 22, 25].

Another method based on the Lippmann-Schwinger equation uses discretization with trigonometric collocation method proposed by Vainikko in [19] and leading to the non-symmetric non-sparse linear system. The numerical solution based on Neumann series expansion was introduced by Moulinec and Suquet in [14]. Nowadays, the proposed algorithm is in extensive investigation and some improvements and modifications were proposed [15, 8, 20, 8, 20, 13, 2, 3, 23], however no rigorous theory has been published yet.

Under the assumption of positive definiteness of material coefficients $\boldsymbol{A}$ at periodic unit cell $\mathcal{Y}$, we show in Lemma 2.29 the connection between Lippmann-Schwinger integral equation, Def. (2.23), and weak formulation, Def. (2.20). The possibility to obtain the equivalent linear systems from both formulations has been already published in [21]. The discretization via trigonometric polynomials requires the values of material coefficients $\boldsymbol{A}$ at regular grid, thus it is fitted for data obtained as digital images from e.g. X-ray tomography [4] or nanoindentation [16, 17].

Since the multiplication with the matrix from the linear system can be efficiently provided using Fast Fourier Transform (FFT), the solution is appropriate for Krylov subspace methods. Zeman et.al. in [24] proposed the solution using Conjugate gradients in spite of the absence of positive definiteness and symmetry of linear system and showed the behavior of the algorithm, its equivalence with Biconjugate gradient method and especially its independence on reference conductivity $\boldsymbol{A}^0$, the parameter occurring in the Lippmann-Schwinger type integral formulation.

In Section 3.4, we show the convergence of discrete solution expressed as linear combination of trigonometric polynomials to the continuous solution and Section 3.6 validates the usage of Conjugate gradients for suitable initial approximation.

For the later use in this work, the following notation is introduced. The letter d denotes the dimension of the problem, assuming $d = 2, 3$; the Greek letters $\alpha, \beta, \gamma, \zeta, \theta$ are reserved to indices relating dimension, thus ranging $1, \dots, d$ (the range is for simplicity often omitted).

The set $\mathbb{N}_0$ represent natural numbers including zero. The sets $\mathbb{C}^d$ and $\mathbb{R}^d$ are spaces of complex and real vectors with canonical basis $\{\boldsymbol{\epsilon}_\alpha\}$ and are equipped with the Lebesgue measure $\mathrm{d}\boldsymbol{x}$. We denote by $|\Omega|_d$ the d-dimensional Lebesgue measure of a measurable set $\Omega \subset \mathbb{R}^d$. The norm $\|\cdot\|_2$ on $\mathbb{C}^d$ is induced by scalar product $\left(\boldsymbol{u}, \boldsymbol{v}\right)_{\mathbb{C}^d} = \sum_\alpha u_\alpha \overline{v_\alpha}$ for $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^d$.

The set $\mathbb{R}^{d\times d}_{\mathrm{spd}}$ denotes the space of symmetric positive definite matrices of size $d \times d$ with norm $\|\boldsymbol{C}\|_2 = \max_{\boldsymbol{x}\in\mathbb{R}^d, \|\boldsymbol{x}\|=1} \|\boldsymbol{C}\boldsymbol{x}\|_2$ that equals to a largest eigenvalue.

A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is $\boldsymbol{Y}$-periodic (with period $\boldsymbol{Y} \in \mathbb{R}^d$) if $f(\boldsymbol{x} + \boldsymbol{Y} \odot \boldsymbol{k}) = f(\boldsymbol{x})$ for arbitrary $\boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{k} \in \mathbb{Z}^d$, where operator $\odot$ denotes element-wise multiplication. The $\boldsymbol{Y}$-periodic functions are sufficient to define only on a periodic unit cell (PUC), set to $\mathcal{Y} := (-Y_\alpha, Y_\alpha)_{\alpha=1}^d \subset \mathbb{R}^d$. Two integrable functions which are almost everywhere equal are identified.

Finally, operator $\oplus^\perp$ denotes the direct sum of mutually orthogonal subspaces, e.g. $\mathbb{R}^d = \boldsymbol{\epsilon}_1 \oplus^\perp \boldsymbol{\epsilon}_2 \oplus^\perp \dots \oplus^\perp \boldsymbol{\epsilon}_d$.

2 Continuous formulation

This section describing the continuous solution to initial problem (1.1) and (1.2) is split into two parts. Sec. 2.1 summarize the well known facts about function spaces and their properties that are used in Sec. 2.2 defining a weak formulation and describing the Lippmann-Schwinger integral equation and showing their equivalence.

2.1 Definitions

This section provides some facts about periodic distribution, the periodic L^2_{per} space and its splitting into subspaces used thorough the work. The facts are well known and can be found in e.g. books of Saranen and Vainikko [18], Jikov, Kozlov, and Oleinik [9], or in other books dealing with function spaces and the equations of mathematical physics.

We begin with definitions of fundamental spaces and their properties.

Definition 2.1 ($\boldsymbol{Y}$-periodic L^p_{per} space)**.** *Let $p \in \mathbb{R}$ with $1 \le p \le \infty$. Define a $\boldsymbol{Y}$-periodic space of square integrable functions as*

$$L^p_{\mathrm{per}}(\mathcal{Y}) = \left\{ f \in L^p_{loc}(\mathbb{R}^d) : f \text{ is } \boldsymbol{Y}\text{-periodic a.e.} \right\}$$

where $L^p_{loc}(\mathcal{Y})$ *denotes*

$$\begin{cases} \{f : \mathbb{R}^d \to \mathbb{C}; f \text{ measurable} : \int_\Omega f^p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} < \infty \text{ with } \Omega \subset \mathbb{R}^d \text{ bounded}\}, & 1 \le p < \infty \\ \left\{f : \mathbb{R}^d \to \mathbb{C}; f \text{ measurable such that there exists } C \in \mathbb{R} \text{ with } |f| < C \text{ a.e.}\right\}, & p = \infty. \end{cases}$$

unifying the functions equaling one another almost everywhere.

The space $L^2_{\mathrm{per}}(\mathcal{Y})$ is a Hilbert space with a scalar product of functions f and g defined as

$$\big(f, g\big)_{L^2_{\mathrm{per}}(\mathcal{Y})} := |\mathcal{Y}|_d^{-1} \int_{\mathcal{Y}} f(\boldsymbol{x})\overline{g(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} \tag{2.1}$$

and induced norm

$$\|f\|_{L^2_{\mathrm{per}}(\mathcal{Y})} := \sqrt{\big(f, f\big)_{L^2_{\mathrm{per}}(\mathcal{Y})}} \tag{2.2}$$

where $|\mathcal{Y}|_d$ denotes the d-dimensional measure of the PUC $\mathcal{Y} \subset \mathbb{R}^d$. The space $L^p_{\mathrm{per}}(\mathcal{Y})$ for $1 \le p \le \infty$ is a Banach space with norm

$$\|f\|_{L^p_{\mathrm{per}}(\mathcal{Y})} = \begin{cases} \left(\int_{\mathcal{Y}} |f(\boldsymbol{x})|^p\,\mathrm{d}\boldsymbol{x}\right)^{\frac{1}{p}}, & 1 \le p < \infty \\ \inf\{C \in \mathbb{R} : |f(\boldsymbol{x})| < C \text{ for almost all } \boldsymbol{x}\}, & p = \infty. \end{cases} \tag{2.3}$$

The powerful method to analyze functions is the technique of Fourier Transform and/or, in the periodic case, the Fourier series working with trigonometric orthonormal basis of space $L^2_{\mathrm{per}}(\mathcal{Y})$ that is given by the set $\{\varphi_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\mathbb{Z}^d}$ with $\varphi_{\boldsymbol{k}}$ defined as

$$\varphi_{\boldsymbol{k}}(\boldsymbol{x}) = \exp\left(\mathrm{i}\pi \sum_\alpha \frac{k_\alpha x_\alpha}{Y_\alpha}\right), \quad \boldsymbol{k} \in \mathbb{Z}^d. \tag{2.4}$$

Definition 2.2 (Fourier representation of functions)**.** *We define the Fourier representation* $f_{\mathcal{F}} \in L^2_{\mathrm{per}}(\mathcal{Y})$ *of function* $f \in L^2_{\mathrm{per}}(\mathcal{Y})$ *as*

$$f_{\mathcal{F}}(\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \hat{f}(\boldsymbol{k})\varphi_{\boldsymbol{k}}(\boldsymbol{x})$$

where the numbers $\hat{f}(\boldsymbol{k}) \in \mathbb{C}, \boldsymbol{k} \in \mathbb{Z}^d$ *denotes the Fourier coefficients of function* f *defined as*

$$\hat{f}(\boldsymbol{k}) := \big(f, \varphi_{\boldsymbol{k}}\big)_{L^2_{\mathrm{per}}(\mathcal{Y})} = |\mathcal{Y}|_d^{-1} \int_{\mathcal{Y}} f(\boldsymbol{x})\varphi_{-\boldsymbol{k}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

where $\overline{\varphi_{\boldsymbol{k}}} = \varphi_{-\boldsymbol{k}}$.

It can be shown that Fourier representation $f_{\mathcal{F}}$ equals to the initial function f almost everywhere thus they can be identified. Moreover function $f \in C^1(\mathcal{Y})$, function with continuous partial derivatives, are represented by the Fourier series everywhere. Next, we can alternatively, due to Parseval's identity, express the scalar product of functions $f, g \in L^2_{\mathrm{per}}(\mathcal{Y})$ as

$$\big(f, g\big)_{L^2_{\mathrm{per}}(\mathcal{Y})} := \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \hat{f}(\boldsymbol{k})\overline{\hat{g}(\boldsymbol{k})} \tag{2.5}$$

and thus the norm as $\|f\|_{L^2_{\mathrm{per}}(\mathcal{Y})} = \left(\sum_{\boldsymbol{k}\in\mathbb{Z}^d} |\hat{f}(\boldsymbol{k})|^2\right)^{\frac{1}{2}}$.

In order to work with derivatives of L^2_{per} functions, we have to weaken the notion of derivative using distributions, the generalized functions. First, it is necessary to define the periodic test functions being the linear function space and a proper convergence on it. Next we follow with the Fourier representation of a derivative.

Lemma 2.3 (Fourier coefficients for a derivative)**.** *For a function $\boldsymbol{f} \in C^1_{\mathrm{per}}(\mathcal{Y})$ we can express the Fourier coefficient for a partial derivative as*

$$\widehat{\frac{\partial f}{\partial x_\alpha}}(\boldsymbol{k}) = \mathrm{i}\pi\xi_\alpha(\boldsymbol{k})\hat{f}(\boldsymbol{k})$$

where $\boldsymbol{k} \in \mathbb{Z}^d$ and $\xi_\alpha(\boldsymbol{k}) = \frac{k_\alpha}{Y_\alpha}$.

Definition 2.4 (Test functions)**.** *For $d \in \mathbb{N}$ and $\boldsymbol{Y} \in \mathbb{R}^d$ regarding a dimension and a period, we define spaces $\mathcal{D}(\mathbb{R}^d)$ and $\mathcal{D}_{\mathrm{per}}(\mathbb{R}^d)$ of test functions and $\boldsymbol{Y}$-periodic test functions on $\mathbb{R}^d$ as*

$$\mathcal{D}(\mathbb{R}^d) = \{\psi \in C^\infty(\mathbb{R}^d) : \operatorname{supp}\psi \text{ is compact}\}$$
$$\mathcal{D}_{\mathrm{per}}(\mathbb{R}^d) = \{\psi \in C^\infty(\mathbb{R}^d) : \mathcal{T}^{\boldsymbol{k}}_{\boldsymbol{Y}}\psi(\boldsymbol{x}) = \psi(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathbb{R}^d \forall \boldsymbol{k} \in \mathbb{Z}^d\}$$

where $C^\infty(\mathbb{R}^d)$ denotes the set of infinitely differentiable scalar functions on $\mathbb{R}^d$, operator $\operatorname{supp}\psi$ denotes support of a function defined as a set $\operatorname{supp}\psi = \overline{\{\boldsymbol{x} \in \mathbb{R}^d : \psi(\boldsymbol{x}) \neq 0\}}$, and operator $\mathcal{T}^{\boldsymbol{k}}_{\boldsymbol{Y}}\psi$ denotes the translation of function ψ, i.e.

$$\left(\mathcal{T}^{\boldsymbol{k}}_{\boldsymbol{Y}}\psi\right)(\boldsymbol{x}) = \psi\left(\boldsymbol{x} + \sum_{\alpha=1}^{d} k_\alpha Y_\alpha \boldsymbol{\epsilon}_\alpha\right)$$

with $\boldsymbol{\epsilon}_\alpha = (\delta_{\alpha\beta})_{\beta=1}^d$ for $\alpha = 1, \ldots, d$ being the vectors of the canonical basis of $\mathbb{R}^d$ and $\delta_{\alpha\beta}$ denoting the Kronecker delta.

Definition 2.5 (Convergence in $\mathcal{D}(\mathbb{R}^d)$)**.** *Having the sequence of test functions, i.e. $\{\psi_n\}_{n=1}^\infty, \psi_n \in \mathcal{D}(\mathbb{R}^d), n \in \mathbb{N}_0$, we say that the sequence $\{\psi_n\}_{n=1}^\infty$ converges to φ in $\mathcal{D}(\mathbb{R}^d)$, and write $\psi_n \to \psi$ in $\mathcal{D}(\mathbb{R})$, if the following properties are valid*

- *there exists a compact subset Ω of $\mathbb{R}^d$ such that $\operatorname{supp}\psi_n \subset \Omega$ for all $n \in \mathbb{N}$*
- *for the arbitrary $k_1, \ldots, k_d \in \mathbb{N}_0$, the uniform convergence holds*

$$\frac{\partial^k \psi_n(\boldsymbol{x})}{\partial x_1^{k_1}\partial x_2^{k_2}\ldots\partial x_d^{k_d}} \to \frac{\partial^k \psi(\boldsymbol{x})}{\partial x_1^{k_1}\partial x_2^{k_2}\ldots\partial x_d^{k_d}}, \quad \text{for } n \to \infty$$

with $k = \sum_\alpha^d k_\alpha$.

Definition 2.6 (distribution)**.** *A linear functional $T : \mathcal{D}(\mathbb{R}^d) \to \mathbb{R}$, with $T(\cdot)$ denoted as $\langle T, \cdot\rangle$, is a distribution on $\mathbb{R}^d$ if it is continuous with respect to the convergence of $\mathcal{D}(\mathbb{R}^d)$, i.e.*

$$\psi_n \to \psi \text{ in } \mathcal{D}(\mathbb{R}) \quad \Rightarrow \quad \langle T, \psi_n\rangle \to \langle T, \psi\rangle \tag{2.6}$$

The space of distributions is denoted by $\mathcal{D}'(\mathbb{R}^n)$.

Definition 2.7 ($\boldsymbol{Y}$-periodic distribution)**.** *Distribution $T \in \mathcal{D}'(\mathbb{R}^n)$ is $\boldsymbol{Y}$-periodic distribution, if*

$$\langle T, \mathcal{T}^Y_n\psi\rangle = \langle T, \psi\rangle, \quad \forall\psi \in \mathcal{D}(\mathbb{R}^d)$$

where $\mathcal{T}^{\boldsymbol{k}}_{\boldsymbol{Y}}\psi$ denotes the translation of function ψ described in Def. (2.4). The space of Y-periodic distributions, i.e. $\{T \in \mathcal{D}'(\mathbb{R}) : T \text{is } \boldsymbol{Y}\text{-periodic}\}$, is denoted as $\mathcal{D}'_{\mathrm{per}}(\mathbb{R}^d)$.

In order to define the derivative of integrable function without any requirement for regularity, the notions of the derivative of distribution and convergence of distribution are introduced.

Definition 2.8 (derivative of periodic distribution)**.** *For a periodic distribution $f \in \mathcal{D}'_{\mathrm{per}}(\mathbb{R}^d)$ we define its partial derivative as distribution $\frac{\partial f}{\partial x_\alpha} \in \mathcal{D}'_{\mathrm{per}}(\mathbb{R}^d)$ satisfying for $\alpha = 1, \ldots, d$*

$$\langle \frac{\partial f}{\partial x_\alpha}, \psi\rangle = -\langle f, \frac{\partial \psi}{\partial x_\alpha}\rangle, \quad \forall\psi \in \mathcal{D}_{\mathrm{per}}(\mathbb{R}^d). \tag{2.7}$$

Definition 2.9 (convergence of periodic distribution)**.** *We say that a periodic distributions $f_n \in \mathcal{D}'_{\mathrm{per}}(\mathbb{R}^d)$ converge to $f \in \mathcal{D}'_{\mathrm{per}}(\mathbb{R}^d)$, and write*

$$\lim_{n\to\infty} f_n = f,$$

if

$$\lim_{n\to\infty} \langle f_n, \psi\rangle = \langle f, \psi\rangle, \quad \forall \psi \in \mathcal{D}_{\mathrm{per}}(\mathbb{R}^d). \tag{2.8}$$

After a definition of Fourier coefficients of periodic distribution we follows with a Lemma about the representation of periodic distribution by Fourier series.

Definition 2.10 (Fourier coefficients of periodic distribution)**.** *We define the Fourier coefficients $\hat{f}(\boldsymbol{k}) \in \mathbb{C}^d$ for $\boldsymbol{k} \in \mathbb{Z}^d$ of a periodic distribution $f \in \mathcal{D}'_{\mathrm{per}}(\mathbb{R}^d)$ as*

$$\hat{f}(\boldsymbol{k}) = \langle f, \varphi_{\boldsymbol{k}}\rangle \tag{2.9}$$

as $\varphi_{\boldsymbol{k}} \in \mathcal{D}_{\mathrm{per}}(\mathbb{R}^d)$.

Lemma 2.11 (Fourier series of distributions)**.** *Assume $f \in \mathcal{D}'_{\mathrm{per}}(\mathbb{R}^d)$ and $\psi \in \mathcal{D}_{\mathrm{per}}(\mathbb{R}^d)$. Then there holds*

(a) there exists $p \in \mathbb{N}$ and $c_r \in \mathbb{R}$ such that $|\hat{f}(\boldsymbol{k})| \le c_p \|\boldsymbol{k}\|_2^p, \forall \boldsymbol{k} \in \mathbb{Z}^d \setminus \mathbf{0}$

(b) there exists $c_r \in \mathbb{R}$ such that for arbitrary $r > 0$ we have $|\hat{\psi}(\boldsymbol{k})| \ne c_r \|\boldsymbol{k}\|_2^r, \forall \boldsymbol{k} \in \mathbb{Z}^d \setminus \mathbf{0}$

(c) $\langle f, \psi\rangle = \sum_{\boldsymbol{k}\in\mathbb{Z}^n} \hat{f}(\boldsymbol{k})\hat{\psi}(-\boldsymbol{k})$

(d) $\lim_{\min N_1,\dots,N_d\to\infty} \sum_{\boldsymbol{k}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \hat{f}(\boldsymbol{k})\varphi_{\boldsymbol{k}}(\boldsymbol{x}) \to f(\boldsymbol{x})$ in $\mathcal{D}'_{\mathrm{per}}(\mathbb{R}^d)$

where $\underline{\mathbb{Z}}^d_{\boldsymbol{N}} = \left\{\boldsymbol{k} \in \mathbb{Z}^d : -\frac{N_\alpha}{2} \le k_\alpha < \frac{N_\alpha}{2}, \alpha = 1, \dots, d\right\}$.

The Lemma and the proof that can be founded in e.g. [18] says in (a) and (b) that the higher the regularity of a function/distribution the higher the dacay of Fourier coefficients in infinity giving the sense of expression (c). The property (c) is important especially with the definition of derivative of distribution, cf. Def. (2.8). The property (d) legitimize the expression of the distribution $f \in \mathcal{D}'_{\mathrm{per}}(\mathbb{R}^d)$ as a Fourier series.

Next notes are dedicated to Sobolev spaces.

Definition 2.12 (Derivative with multiindex notation)**.** *Let $m \in \mathbb{N}_0$, $\boldsymbol{l} \in \mathbb{N}_0^d$, and $u(\boldsymbol{x}) \in L^2_{\mathrm{per}}(\mathcal{Y})$, we define*

$$D^{\boldsymbol{l}} u(\boldsymbol{x}) = \frac{\partial^{\|\boldsymbol{l}\|_1} u(\boldsymbol{x})}{\prod_{\alpha=1}^d \partial^{l_\alpha} x_\alpha}$$
$$D^m u(\boldsymbol{x}) = \{D^{\boldsymbol{l}} u : \|\boldsymbol{l}\|_1 = m\}$$

with a partial derivative in the sense of distributions.

Definition 2.13 (Sobolev space)**.** *Let $1 \le p \le \infty$ and $\mu \in \mathbb{N}$. Sobolev space $W^{\mu,p}_{\mathrm{per}}(\mathcal{Y})$ of $\boldsymbol{Y}$-periodic functions (distributions) is defined as*

$$W^{\mu,p}_{\mathrm{per}}(\mathcal{Y}) = \left\{f \in L^p_{\mathrm{per}}(\mathcal{Y}); D^{\boldsymbol{l}} u \in L^p_{\mathrm{per}} \text{ for } \boldsymbol{l} \in \mathbb{N}_0^d : \|\boldsymbol{l}\|_1 \le \mu\right\}.$$

For a special case $p = 2$, we define

$$H^{\mu}_{\mathrm{per}}(\mathcal{Y}) := W^{\mu,2}_{\mathrm{per}}(\mathcal{Y}).$$

We define a norm of Sobolev space $\|\cdot\|_{W^{\mu,p}_{\mathrm{per}}(\mathcal{Y})}$ as

$$\|u\|_{W^{\mu,p}_{\mathrm{per}}(\mathcal{Y})} = \begin{cases} \left(\sum_{\boldsymbol{l}\in\mathbb{N}_0^d:\|\boldsymbol{l}\|_1<\mu} \|D^{\boldsymbol{l}} u\|^p_{L^p_{\mathrm{per}}(\mathcal{Y})}\right)^{\frac{1}{p}}, & \text{for } 1 \le p < \infty \\ \max_{\boldsymbol{l}\in\mathbb{N}_0^d:\|\boldsymbol{l}\|_1<\mu} \|D^{\boldsymbol{l}} u\|_{L^\infty_{\mathrm{per}}(\mathcal{Y})}, & \text{for } p = \infty. \end{cases}$$

where norm $\|\cdot\|_{L^p_{\mathrm{per}}(\mathcal{Y})}$ is defined in Eq. (2.3).

Remark 2.14. *For a Sobolev space $H^\mu_{\mathrm{per}}(\mathcal{Y})$ we also use equivalent norm*

$$\|u\|_{H^\mu_{\mathrm{per}}(\mathcal{Y})} = \left(\sum_{\boldsymbol{k}\in\mathbb{Z}^d} \|\underline{\xi(\boldsymbol{k})}\|_2^{2\mu} |\hat{u}(\boldsymbol{k})|^2 \right)^{\frac{1}{2}}.$$

where $\hat{u}(\boldsymbol{k})$ denotes Fourier coefficients of function u and

$$\underline{\boldsymbol{\xi}(\boldsymbol{k})} = \begin{cases} 1, & \text{for } \boldsymbol{k} = \boldsymbol{0}, \\ \left(\frac{k_\alpha}{2Y_\alpha}\right)_{\alpha=1}^d, & \text{for } \boldsymbol{0} \neq \boldsymbol{k} \in \mathbb{Z}^d. \end{cases}$$

Remark 2.15 (vector valued functions)**.** *Let $\mu \in \mathbb{N}$ and $1 \leq p \leq \infty$. We define a space of vector valued functions as*

$$L^p_{\mathrm{per}}(\mathcal{Y};\mathbb{C}^d) = \{\boldsymbol{f} : \mathcal{Y} \to \mathbb{C}^d | \boldsymbol{f} = (f_\alpha)_{\alpha=1}^d, f_\alpha \in L^p_{\mathrm{per}}(\mathcal{Y}) \text{ for } \alpha = 1,\dots,d\}$$
$$W^{\mu,p}_{\mathrm{per}}(\mathcal{Y};\mathbb{C}^d) = \{\boldsymbol{f} : \mathcal{Y} \to \mathbb{C}^d | \boldsymbol{f} = (f_\alpha)_{\alpha=1}^d, f_\alpha \in W^{\mu,p}_{\mathrm{per}}(\mathcal{Y}) \text{ for } \alpha = 1,\dots,d\}$$

with norms

$$\|\boldsymbol{f}\|_{L^p_{\mathrm{per}}(\mathcal{Y};\mathbb{C}^d)} = \|(\|\boldsymbol{f}(\boldsymbol{x})\|_p)\|_{L^p_{\mathrm{per}}(\mathcal{Y})} \qquad \|\boldsymbol{f}\|_{W^{\mu,p}_{\mathrm{per}}(\mathcal{Y};\mathbb{C}^d)} = \|(\|\boldsymbol{f}(\boldsymbol{x})\|_p)\|_{W^{\mu,p}_{\mathrm{per}}(\mathcal{Y})}$$

and scalar product

$$\big(\boldsymbol{f},\boldsymbol{g}\big)_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{C}^d)} = \sum_\alpha \big(f_\alpha, g_\alpha\big)_{L^2_{\mathrm{per}}(\mathcal{Y})}$$

where $\|\boldsymbol{v}\|_p := (\sum_\alpha v_\alpha^p)^{\frac{1}{p}}$ denotes the norm on space $\mathbb{R}^d$. The above space naturally inherits the properties stated in previous parts.

The Fourier representation $\boldsymbol{f}_{\mathcal{F}} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{C}^d)$ of the vector valued function $\boldsymbol{f} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{C}^d)$ can be defined as a Fourier coefficients of its components, i.e.

$$\boldsymbol{f}_{\mathcal{F}}(\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \hat{\boldsymbol{f}}(\boldsymbol{k})\varphi_{\boldsymbol{k}}(\boldsymbol{x}) \qquad \hat{\boldsymbol{f}}(\boldsymbol{k}) = \Big(\big(f_\alpha, \varphi_{\boldsymbol{k}}\big)_{L^2_{\mathrm{per}}(\mathcal{Y})}\Big)_\alpha^d.$$

The matrix valued functions or tensor, i.e. $L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{C}^{d\times d})$ and $W^{\mu,p}_{\mathrm{per}}(\mathcal{Y};\mathbb{C}^{d\times d})$, are defined analogically with norms

$$\|\boldsymbol{f}\|_{L^p_{\mathrm{per}}(\mathcal{Y};\mathbb{C}^d)} = \| \left(\|f_{\alpha\beta}\|_{L^2_{\mathrm{per}}(\mathcal{Y})}\right)_{\alpha,\beta=1}^d \|_p \qquad \|\boldsymbol{f}\|_{W^{\mu,p}_{\mathrm{per}}(\mathcal{Y};\mathbb{C}^d)} = \| \left(\|f_{\alpha\beta}\|_{W^{\mu,p}_{\mathrm{per}}(\mathcal{Y})}\right)_{\alpha,\beta=1}^d \|_p$$

where $\|\cdot\|_p$ denotes a matrix norm on $\mathbb{R}^{d\times d}$ defined as

$$\|\boldsymbol{A}\|_p = \sup_{\boldsymbol{x}\in\mathbb{R}^d, \boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p}.$$

Now, we will define the differential operators divergence and curl on the space L^2_{per}, i.e. in the sense of distributions.

Definition 2.16 (divergence and curl)**.** *Let $\boldsymbol{f} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ be vector valued function, then we define divergence and curl in distributional sense as*

$$\operatorname{div}\boldsymbol{f} := (\operatorname{div}\boldsymbol{f},\psi) = -\sum_{\alpha=1}^d \langle f_\alpha, \frac{\partial\psi}{\partial x_\alpha}\rangle, \quad \forall\psi\in\mathcal{D}_{\mathrm{per}}(\mathbb{R}^d),$$
$$\operatorname{curl}\boldsymbol{f} := (\operatorname{curl}\boldsymbol{f},\psi) = -\left(\langle f_\alpha, \frac{\partial\psi}{\partial x_\beta}\rangle - \langle f_\beta, \frac{\partial\psi}{\partial x_\alpha}\rangle\right)_{\alpha,\beta=1,\dots,d}, \quad \forall\psi\in\mathcal{D}_{\mathrm{per}}(\mathbb{R}^d).$$

Remark 2.17 (Helmholtz decomposition)**.** *Next, we introduce Helmholtz decomposition* $L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) = \mathscr{U} \oplus^\perp \mathscr{E} \oplus^\perp \mathscr{J}$ *to the spaces of constant, curl-free with zero mean, and divergence free with zero mean fields*

$$\mathscr{U} = \{\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) : \boldsymbol{v}(\boldsymbol{x}) = \text{const.}\}, \tag{2.10a}$$

$$\mathscr{E} = \{\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) : \operatorname{curl}\boldsymbol{v} = \boldsymbol{0}, \langle\boldsymbol{v}\rangle = \boldsymbol{0}\}, \tag{2.10b}$$

$$\mathscr{J} = \{\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) : \operatorname{div}\boldsymbol{v} = 0, \langle\boldsymbol{v}\rangle = \boldsymbol{0}\}, \tag{2.10c}$$

where $\langle\boldsymbol{v}\rangle := \frac{1}{|\mathcal{Y}|_d}\int_{\mathcal{Y}} \boldsymbol{v}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \in \mathbb{R}^d$ *denotes the mean value of function* $\boldsymbol{v}$ *over periodic unit cell* $\mathcal{Y}$*. Since the space* $\mathscr{U}$ *consists of constant functions, we identify the space* $\mathscr{U}$ *with* $\mathbb{R}^d$*; this validates the operations such as* $\boldsymbol{E}+\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ *for* $\boldsymbol{E} \in \mathbb{R}^d$*,* $\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ *and* $\boldsymbol{C}\boldsymbol{J} \in \mathbb{R}^d$ *for* $\boldsymbol{C} \in \mathbb{R}^{d\times d}$ *and* $\boldsymbol{J} \in \mathscr{U}$*.*

Lemma 2.18 (Existence of potential)**.** *Vector valued function* $\boldsymbol{e} \in \mathscr{E}$ *has potential, i.e. there exist scalar function* $u \in H^1_{\mathrm{per}}(\mathcal{Y}) := \{u \in L^2_{\mathrm{per}}(\mathcal{Y}); |\hat{u}(0)|^2 + \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \|\boldsymbol{\xi}(\boldsymbol{k})\|_2^2 |\hat{u}(\boldsymbol{k})|^2 < \infty\}$*. Moreover, space* $\mathscr{E}$ *can be alternatively expressed as* $\mathscr{E} = \{\nabla u; u \in H^1_{\mathrm{per}}(\mathcal{Y})\}$*.*

Proof. Vector valued function $\boldsymbol{e} \in \mathscr{E} \subset L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ can be expressed using Fourier series

$$\boldsymbol{e} = \sum_{\boldsymbol{k}\in\mathbb{Z}^d\setminus\{0\}} \hat{f}_\alpha(\boldsymbol{k})\varphi_{\boldsymbol{k}}$$

noting $\langle\boldsymbol{e}\rangle = \boldsymbol{0}$. Then we define a scalar function $u \in L^2_{\mathrm{per}}(\mathcal{Y})$ as

$$u(\boldsymbol{x}) := \sum_{\boldsymbol{k}\in\mathbb{Z}^d\setminus\{0\}} \frac{\hat{e}_1(\boldsymbol{k})}{\mathrm{i}\pi\xi_1(\boldsymbol{k})}\varphi_{\boldsymbol{k}}(\boldsymbol{x})$$

and we show that it is a potential for $\boldsymbol{e}$.

Next, we show that $\frac{\partial u}{x_\alpha} = e_\alpha$ in $L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ by testing with function $\frac{\partial\psi}{x_1}$, i.e.

$$\begin{aligned}\Big\langle \frac{\partial u}{\partial x_\alpha} - e_\alpha, \frac{\partial\psi}{x_1}\Big\rangle &= \Big\langle \frac{\partial u}{\partial x_\alpha}, \frac{\partial\psi}{x_1}\Big\rangle - \Big\langle e_1, \frac{\partial\psi}{x_\alpha}\Big\rangle \\ &= \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \left(\mathrm{i}\pi\xi_\alpha(\boldsymbol{k})\frac{\hat{e}_1(\boldsymbol{k})}{\mathrm{i}\pi\xi_1(\boldsymbol{k})}\mathrm{i}\pi\xi_1(\boldsymbol{k})\hat{\psi}(\boldsymbol{k}) - \hat{e}_1(\boldsymbol{k})\mathrm{i}\pi\xi_\alpha(\boldsymbol{k})\hat{\psi}(\boldsymbol{k})\right) = 0\end{aligned}$$

where the identity $\big\langle e_\alpha, \frac{\psi}{\partial x_1}\big\rangle = \big\langle e_1, \frac{\psi}{\partial x_\alpha}\big\rangle$ was used as the function $\boldsymbol{e}$ is curl free from definition.

The inclusion $\{\nabla u; u \in H^1_{\mathrm{per}}(\mathcal{Y})\} \subseteq \mathscr{E}$ follows from

$$\begin{aligned}\operatorname{curl}\nabla u := (\operatorname{curl}\nabla u, \psi) &= -\left(\Big\langle \frac{\partial u}{\partial x_\alpha}, \frac{\partial\psi}{\partial x_\beta}\Big\rangle - \Big\langle \frac{\partial u}{\partial x_\beta}, \frac{\partial\psi}{\partial x_\alpha}\Big\rangle\right)_{\alpha,\beta=1,\dots,d} \\ &= \left(\Big\langle u, \frac{\partial^2\psi}{\partial x_\alpha\partial x_\beta}\Big\rangle - \Big\langle u, \frac{\partial^2\psi}{\partial x_\alpha\partial x_\beta}\Big\rangle\right)_{\alpha,\beta=1,\dots,d} = \big(0\big)_{\alpha,\beta=1,\dots,d}\end{aligned}$$

for arbitrary $\psi \in \mathcal{D}_{\mathrm{per}}(\mathbb{R})$.

Finally, we show that u is from Sobolev space, i.e. $u \in H^1_{\mathrm{per}}(\mathcal{Y})$. It could be shown that functions $u_\alpha \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ for $\alpha = 2,\dots,d$ defined as

$$u_\alpha(\boldsymbol{x}) := \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \frac{\hat{e}_\alpha(\boldsymbol{k})}{\mathrm{i}\pi\xi_\alpha(\boldsymbol{k})}\varphi_{\boldsymbol{k}}(\boldsymbol{x})$$

are also potentials of function $\boldsymbol{e}$. Clearly we have $\nabla u_\alpha = \boldsymbol{e}$ for $\alpha = 2, \dots, d$ and thus $\nabla u_\alpha = \nabla u$ and consequently even $u_\alpha = u$. Then $u \in H^1_{\mathrm{per}}(\mathcal{Y})$ as

$$\begin{aligned}
\|u\|^2_{H^1_{\mathrm{per}}} &= \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \|\underline{\boldsymbol{\xi}(\boldsymbol{k})}\|_2^2 |\hat{u}(\boldsymbol{k})|^2 = \sum_{\boldsymbol{k}\in\mathbb{Z}^d\setminus\{0\}} \|\boldsymbol{\xi}(\boldsymbol{k})\|_2^2 |\hat{u}(\boldsymbol{k})|^2 \\
&\leq \sum_{\boldsymbol{k}\in\mathbb{Z}^d\setminus\{0\}} |\xi_1(\boldsymbol{k})|^2|\hat{u}(\boldsymbol{k})|^2 + |\xi_2(\boldsymbol{k})|^2|\hat{u}(\boldsymbol{k})|^2 + \ldots + |\xi_d(\boldsymbol{k})|^2|\hat{u}(\boldsymbol{k})|^2 \\
&= \sum_{\boldsymbol{k}\in\mathbb{Z}^d\setminus\{0\}} |\xi_1(\boldsymbol{k})|^2 \left|\frac{\hat{e}_1(\boldsymbol{k})}{\mathrm{i}\pi\xi_1(\boldsymbol{k})}\right|^2 + |\xi_2(\boldsymbol{k})|^2 \left|\frac{\hat{e}_2(\boldsymbol{k})}{\mathrm{i}\pi\xi_2(\boldsymbol{k})}\right|^2 + \ldots + |\xi_d(\boldsymbol{k})|^2 \left|\frac{\hat{e}_d(\boldsymbol{k})}{\mathrm{i}\pi\xi_d(\boldsymbol{k})}\right|^2 \\
&\leq \|\boldsymbol{e}\|^2_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)}.
\end{aligned}$$

□

2.2 Integral and weak formulation

In this section, we define weak formulation Def. 2.20 and the formulation based on Lippmann-Schwinger equation Def. 2.23. In Lemma 2.29 we show the equivalence of both formulations. It is based on projection operator that is derived from Green function occurring in the Lippmann-Schwinger equation, see Lem. 2.28.

Notation 2.19. *Here and in the sequel, $\boldsymbol{A} \in L^\infty_{\mathrm{per}}(\mathcal{Y}, \mathbb{R}^{d\times d}_{\mathrm{spd}})$ denotes symmetric[1] and uniformly elliptic material coefficients of electric conductivity. It means that there exists positive constant $c_A > 0$ such that for almost all $\boldsymbol{x} \in \mathcal{Y}$ and all nonzero $\boldsymbol{u} \in \mathbb{R}^d$, inequality*

$$c_A\|u\|_2^2 \leq \big(\boldsymbol{A}(\boldsymbol{x})\boldsymbol{u}, \boldsymbol{u}\big)_{\mathbb{R}^d}$$

holds. For the next use we define constant of upper bound $C_A := \|\boldsymbol{A}\|_{L^\infty_{\mathrm{per}}}$.

Moreover, function $\boldsymbol{e} \in \mathscr{E}$ denotes perturbation of electric field, and $\boldsymbol{E} \in \mathscr{U}$ its macroscopic counterparts. Then their summation $(\boldsymbol{E} + \boldsymbol{e}) \in \mathscr{U} \oplus^\perp \mathscr{E}$ represents microscopic field.

Weak formulation in the next definition is derived with the multiplication of differential equation (1.1) by a test function $\boldsymbol{v} \in \mathcal{D}_{\mathrm{per}}(\mathcal{Y})$, the application of the Green theorem and some algebraic emendation.

Definition 2.20 (Weak formulation)**.** *Let $\boldsymbol{A} \in L^\infty_{\mathrm{per}}(\mathcal{Y}, \mathbb{R}^{d\times d})$ be symmetric positive definite material coefficients. Then we define bilinear form $B[\cdot,\cdot] : L^2_{\mathrm{per}}(\mathcal{Y}, \mathbb{R}^d) \times L^2_{\mathrm{per}}(\mathcal{Y}, \mathbb{R}^d) \to \mathbb{R}$ as*

$$B[\boldsymbol{e}, \boldsymbol{v}] := \big(\boldsymbol{A}\boldsymbol{e}, \boldsymbol{v}\big)_{L^2_{\mathrm{per}}(\mathcal{Y},\mathbb{R}^d)} \tag{2.11}$$

and the linear functional $f[\cdot] : L^2_{\mathrm{per}}(\mathcal{Y}, \mathbb{R}^d) \to \mathbb{R}$ as

$$f[\boldsymbol{v}] := -\big(\boldsymbol{A}\boldsymbol{E}, \boldsymbol{v}\big)_{L^2_{\mathrm{per}}(\mathcal{Y},\mathbb{R}^d)}. \tag{2.12}$$

where $\boldsymbol{E} \in \mathscr{U}$ is a macroscopic load. Function $\tilde{\boldsymbol{e}} \in \mathscr{E}$ is the weak solution of the differential equation (1.1) with periodic boundary condition and constraint $\langle \boldsymbol{e} \rangle = \boldsymbol{E}$ if

$$B[\tilde{\boldsymbol{e}}, \boldsymbol{v}] = f[\boldsymbol{v}], \quad \forall \boldsymbol{v} \in \mathscr{E}. \tag{2.13}$$

We say that $\boldsymbol{e} = \boldsymbol{E} + \tilde{\boldsymbol{e}}$ is microscopic field for macroscopic load $\boldsymbol{E} \in \mathscr{U}$.

Remark 2.21 (Existence of the unique solution)**.** *Since the bilinear form from Def. (2.20) is positive definite $c_A\|\boldsymbol{v}\|_{L^2_{\mathrm{per}}} \leq B[\boldsymbol{v}, \boldsymbol{v}]$ and bounded $B[\boldsymbol{u}, \boldsymbol{v}] \leq C_A\|\boldsymbol{u}\|_{L^2_{\mathrm{per}}}\|\boldsymbol{v}\|_{L^2_{\mathrm{per}}}$ for all $\boldsymbol{u}, \boldsymbol{v} \in \mathscr{E}$ and the linear functional f is bounded $\|f\| \leq C_A\|\boldsymbol{e}_0\|_{L^2_{\mathrm{per}}}$, the well known Lax-Milgram theorem provide the existence and uniqueness of the solution.*

[1] For almost all $\boldsymbol{x} \in \mathcal{Y}$, equality $\boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{A}(\boldsymbol{x})^T$ holds.

Remark 2.22 (The solution of weak formulation as a minimizer)**.** *The bilinear form is symmetric, i.e.* $B[\boldsymbol{u},\boldsymbol{v}] = B[\boldsymbol{v},\boldsymbol{u}]$ *for all* $\boldsymbol{u},\boldsymbol{v} \in L^2_{\mathrm{per}}$*, as* $\boldsymbol{A}$ *is symmetric, thus we can express the solution as a minimizer of quadratic functional*

$$\boldsymbol{e} = \boldsymbol{E} + \tilde{\boldsymbol{e}} = \boldsymbol{E} + \operatorname{argmin}_{\boldsymbol{v}\in\mathscr{E}} \frac{1}{2} B[\boldsymbol{v},\boldsymbol{v}] - f[\boldsymbol{v}]$$

In order to formulate Lippmann-Schwinger equation, we introduce a homogeneous reference medium with constant conductivity $\boldsymbol{A}^0 \in \mathbb{R}^{d\times d}$ that is symmetric and positive definite. Then we can decompose the electric current from Eq. (1.1) into the form

$$\operatorname{div}\Big[\boldsymbol{A}^0\boldsymbol{e}(\boldsymbol{x})\Big] = -\operatorname{div}\Big[\big(\boldsymbol{A}(\boldsymbol{x}) - \boldsymbol{A}^0\big)\boldsymbol{e}(\boldsymbol{x})\Big] \tag{2.14}$$

that is appropriate for the use of Fourier Transform technique for a homogeneous problem with conductivity $\boldsymbol{A}^0$; the right-hand side of Eq. (2.14) then represents the sources, the divergence of so called polarization field. The original problem (1.1)–(1.2) is then equivalent to the periodic Lippmann-Schwinger equation with a derivative of Green function as an integral kernel.

Definition 2.23 (Integral formulation)**.** *Having the same assumption as in Def.* (2.20) *and in addition let* $\boldsymbol{A}^0 \in \mathbb{R}^{d\times d}_{\mathrm{spd}}$ *be symmetric positive definite matrix. We say that* $\boldsymbol{e} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ *is the solution of Lippmann-Schwinger equation if it satisfies*

$$\boldsymbol{e}(\boldsymbol{x}) + \int_{\mathcal{Y}} \boldsymbol{\Gamma}(\boldsymbol{x}-\boldsymbol{y};\boldsymbol{A}^0)(\boldsymbol{A}(\boldsymbol{y}) - \boldsymbol{A}^0)\boldsymbol{e}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = \boldsymbol{E}, \quad \boldsymbol{x}\in\mathcal{Y} \tag{2.15}$$

where the convolution integral $\boldsymbol{f} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) \to \int_{\mathcal{Y}} \boldsymbol{\Gamma}(\boldsymbol{x}-\boldsymbol{y};\boldsymbol{A}^0)\boldsymbol{f}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ *is defined with the help of Fourier series as*

$$\int_{\mathcal{Y}} \boldsymbol{\Gamma}(\boldsymbol{x}-\boldsymbol{y};\boldsymbol{A}^0)\boldsymbol{f}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} := \sum_{\boldsymbol{k}\in\mathbb{Z}^d\setminus\{\boldsymbol{0}\}} \frac{\boldsymbol{\xi}(\boldsymbol{k})\otimes\boldsymbol{\xi}(\boldsymbol{k})}{\big(\boldsymbol{A}^0\boldsymbol{\xi}(\boldsymbol{k}),\boldsymbol{\xi}(\boldsymbol{k})\big)_{\mathbb{R}^d}} \hat{\boldsymbol{f}}(\boldsymbol{k})\varphi_{\boldsymbol{k}}(\boldsymbol{x}) \tag{2.16a}$$

$$= \sum_{\boldsymbol{k}\in\mathbb{Z}^d\setminus\{\boldsymbol{0}\}} \left(\sum_{\beta=1}^{d} \frac{\frac{k_\alpha k_\beta}{Y_\alpha Y_\beta}}{\sum_{\mu,\nu=1}^{d} A^0_{\mu\nu}\frac{k_\mu k_\nu}{Y_\mu Y_\nu}} \hat{f}_\beta(\boldsymbol{k})\varphi_{\boldsymbol{k}}(\boldsymbol{x}) \right)_{\alpha=1,\dots,d} \tag{2.16b}$$

where $\boldsymbol{\xi}(\boldsymbol{k}) \in \mathbb{R}^d$ *is a vector with components* $\xi_\alpha(\boldsymbol{k}) = \frac{k_\alpha}{Y_\alpha}$*, binary operator* $\otimes$ *denotes tensor product and* $\big(\boldsymbol{A}^0\boldsymbol{\xi}(\boldsymbol{k}),\boldsymbol{\xi}(\boldsymbol{k})\big)_{\mathbb{R}^d}$ *is a quadratic form in* $\mathbb{R}^d$ *and* $\hat{\boldsymbol{f}}(\boldsymbol{k})$ *for* $\boldsymbol{k}\in\mathbb{Z}^d$ *are the Fourier coefficients.*

Remark 2.24. *The definition of convolution integral in Eq.* (2.16) *is correct, it really maps the function to the space* $L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ *as*

$$\Big\| \int_{\mathcal{Y}} \boldsymbol{\Gamma}(\boldsymbol{x}-\boldsymbol{y};\boldsymbol{A}^0)\boldsymbol{f}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \Big\|_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)} \le \frac{1}{c_A}\|\boldsymbol{f}\|_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)}$$

In order to show the equivalence of both formulations in Def. (2.20) and (2.23), we introduce an operator $\mathcal{G}[\cdot]$ based on convolution integral of Lippmann-Schwinger equation. In the following lemma, we show that it is a projection on space $\mathscr{E}$, the space of curl-free fields with zero mean.

Definition 2.25. *We define an operator* $\mathcal{G}[\cdot] : L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) \to L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ *as*

$$\mathcal{G}[\boldsymbol{f}](\boldsymbol{x}) := \int_{\mathcal{Y}} \boldsymbol{\Gamma}(\boldsymbol{x}-\boldsymbol{y};\boldsymbol{A}^0)\boldsymbol{A}^0\boldsymbol{f}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}$$

where the convolution integral is defined in Def. (2.23)*.*

Remark 2.26. *The operator $\mathcal{G}$ can be explicitly expressed as*

$$\mathcal{G}[\boldsymbol{f}](\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^d\setminus\{\boldsymbol{0}\}} \sum_{\beta,\gamma} \frac{\xi_\alpha(\boldsymbol{k})\xi_\beta(\boldsymbol{k})}{\big(\boldsymbol{A}^0\boldsymbol{\xi}(\boldsymbol{k}),\boldsymbol{\xi}(\boldsymbol{k})\big)_{\mathbb{R}^d}} A^0_{\beta\gamma}\widehat{f}_\gamma(\boldsymbol{k})\varphi_{\boldsymbol{k}}(\boldsymbol{x}) \tag{2.17}$$

where $\xi_\alpha(\boldsymbol{k}) = \frac{k_\alpha}{Y_\alpha} \in \mathbb{R}$ for $\alpha = 1,\dots,d$ as in Def. (2.23).

Remark 2.27 (The adjoint of $\mathcal{G}$)**.** *For the next use we also discuss the adjoint operator $\mathcal{G}^*$ to $\mathcal{G}$, i.e. the operator satisfying $\big(\mathcal{G}\boldsymbol{v},\boldsymbol{w}\big) = \big(\boldsymbol{v},\mathcal{G}^*\boldsymbol{w}\big)$ for all $\boldsymbol{v},\boldsymbol{w}\in L^2_{\mathrm{per}}$. With the help of previous Remark 2.26 we can easily express the adjoint operator $\mathcal{G}^*$ as*

$$\mathcal{G}^*[\boldsymbol{f}](\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^d\setminus\{\boldsymbol{0}\}} \sum_{\beta,\gamma} A^0_{\alpha\beta} \frac{\xi_\beta(\boldsymbol{k})\xi_\gamma(\boldsymbol{k})}{\big(\boldsymbol{A}^0\boldsymbol{\xi}(\boldsymbol{k}),\boldsymbol{\xi}(\boldsymbol{k})\big)_{\mathbb{R}^d}} \widehat{f}_\gamma(\boldsymbol{k})\varphi_{\boldsymbol{k}}(\boldsymbol{x}) \tag{2.18}$$

leading to a property $\boldsymbol{A}^0\mathcal{G} = \mathcal{G}^*\boldsymbol{A}^0$.

The following lemma provides a projection on space $\mathscr{E}$ and it is based on the well known results about projections of matrices in Fourier space $\frac{\boldsymbol{\xi}(\boldsymbol{k})\otimes\boldsymbol{\xi}(\boldsymbol{k})}{\big(\boldsymbol{\xi}(\boldsymbol{k}),\boldsymbol{\xi}(\boldsymbol{k})\big)_{\mathbb{R}^d}}$ that are used in homogenization theory, see e.g. [12]. However, the operator $\mathcal{G}$, including parameter $\boldsymbol{A}^0$ and acting on L^2_{per}, has not been presented yet, to the best of our knowledge.

Lemma 2.28 (Projection on curl-free fields)**.** *Operator $\mathcal{G}[\cdot] : L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) \to L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ defined in Def. (2.25) is a projection on the space $\mathscr{E}$.*

Proof. First, we note that Def. (2.25) is correct, i.e. it maps the functions into the space $L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ as $\|\mathcal{G}[\boldsymbol{f}]\|_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)} \le \frac{C_A}{c_A}\|\boldsymbol{f}\|_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)}$. The Fourier coefficients of the real functions satisfies the following symmetry $\hat{\boldsymbol{f}}(\boldsymbol{k}) = \operatorname{conj}\hat{\boldsymbol{f}}(-\boldsymbol{k})$ for all $\boldsymbol{k}\in\mathbb{Z}^d$. The images of operator $\mathcal{G}$ provides such symmetry

$$\begin{aligned}
\operatorname{conj}\big(\widehat{\mathcal{G}\boldsymbol{f}}(-\boldsymbol{k})\big) &= \operatorname{conj}\left(\sum_{\beta,\gamma=1}^{d} \frac{\xi_\alpha\xi_\beta}{\big(\boldsymbol{A}^0\boldsymbol{\xi},\boldsymbol{\xi}\big)_{\mathbb{R}^d}} A^0_{\beta\gamma} f_\gamma(-\boldsymbol{k})\right)_{\alpha=1}^{d} \\
&= \left(\sum_{\beta,\gamma=1}^{d} \frac{\xi_\alpha\xi_\beta}{\big(\boldsymbol{A}^0\boldsymbol{\xi},\boldsymbol{\xi}\big)_{\mathbb{R}^d}} A^0_{\beta\gamma} \operatorname{conj}\big(f_\gamma(-\boldsymbol{k})\big)\right)_{\alpha=1}^{d} \\
&= \left(\sum_{\beta,\gamma=1}^{d} \frac{\xi_\alpha\xi_\beta}{\big(\boldsymbol{A}^0\boldsymbol{\xi},\boldsymbol{\xi}\big)_{\mathbb{R}^d}} A^0_{\beta\gamma} f_\gamma(\boldsymbol{k})\right)_{\alpha=1}^{d} = \widehat{\mathcal{G}\boldsymbol{f}}(\boldsymbol{k})
\end{aligned}$$

We show that operator $\mathcal{G}$ is the projection by showing $\mathcal{G}^2 = \mathcal{G}$, hence

$$\begin{aligned}
\mathcal{G}^2[\boldsymbol{f}](\boldsymbol{x}) = \mathcal{G}\big[\mathcal{G}[\boldsymbol{f}]\big](\boldsymbol{x}) &= \mathcal{G}\left[\sum_{\boldsymbol{k}\in\mathbb{Z}\setminus\{\boldsymbol{0}\}} \frac{\boldsymbol{\xi}(\boldsymbol{k})\otimes\boldsymbol{\xi}(\boldsymbol{k})}{\big(\boldsymbol{A}^0\boldsymbol{\xi}(\boldsymbol{k}),\boldsymbol{\xi}(\boldsymbol{k})\big)_{\mathbb{R}^d}} \boldsymbol{A}^0\widehat{\boldsymbol{f}}(\boldsymbol{k})\varphi_{\boldsymbol{k}}(\boldsymbol{x})\right] \\
&= \sum_{\boldsymbol{k}\in\mathbb{Z}\setminus\{\boldsymbol{0}\}} \frac{\boldsymbol{\xi}(\boldsymbol{k})\otimes\boldsymbol{\xi}(\boldsymbol{k})}{\big(\boldsymbol{A}^0\boldsymbol{\xi}(\boldsymbol{k}),\boldsymbol{\xi}(\boldsymbol{k})\big)_{\mathbb{R}^d}} \boldsymbol{A}^0 \frac{\boldsymbol{\xi}(\boldsymbol{k})\otimes\boldsymbol{\xi}(\boldsymbol{k})}{\big(\boldsymbol{A}^0\boldsymbol{\xi}(\boldsymbol{k}),\boldsymbol{\xi}(\boldsymbol{k})\big)_{\mathbb{R}^d}} \boldsymbol{A}^0\widehat{\boldsymbol{f}}(\boldsymbol{k})\varphi_{\boldsymbol{k}}(\boldsymbol{x}) \\
&= \sum_{\boldsymbol{k}\in\mathbb{Z}\setminus\{\boldsymbol{0}\}} \frac{\boldsymbol{\xi}(\boldsymbol{k})\otimes\boldsymbol{\xi}(\boldsymbol{k})}{\big(\boldsymbol{A}^0\boldsymbol{\xi}(\boldsymbol{k}),\boldsymbol{\xi}(\boldsymbol{k})\big)_{\mathbb{R}^d}} \boldsymbol{A}^0\widehat{\boldsymbol{f}}(\boldsymbol{k})\varphi_{\boldsymbol{k}}(\boldsymbol{x}) = \mathcal{G}[\boldsymbol{f}](\boldsymbol{x})
\end{aligned}$$

Next, we prove that it maps into the space $\mathscr{E}$ by showing $\operatorname{curl}\mathcal{G}[\boldsymbol{f}] = 0$ and $\big\langle\mathcal{G}[\boldsymbol{f}]\big\rangle = 0$ for all $\boldsymbol{f}\in L^2_{\mathrm{per}}(\mathcal{Y},\mathbb{R}^d)$. The later case is trivial as we sum in Eq. (2.17) over the set $\mathbb{Z}^d\setminus\{\boldsymbol{0}\}$, i.e. the constant

term is omitted. Hence, the first case follows for arbitrary test function $\psi \in \mathcal{D}_{\text{per}}(\mathbb{R}^d)$ as

$$\operatorname{curl} \mathcal{G}[\boldsymbol{f}] := (\operatorname{curl} \mathcal{G}[\boldsymbol{f}], \psi) = -\left(\langle (\mathcal{G}[\boldsymbol{f}])_\alpha, \frac{\partial \psi}{\partial x_\beta} \rangle - \langle (\mathcal{G}[\boldsymbol{f}])_\beta, \frac{\partial \psi}{\partial x_\alpha} \rangle \right)_{\alpha,\beta=1}^{d} \tag{2.19a}$$

$$= -\left(\sum_{\boldsymbol{k}\in\mathbb{Z}} \widehat{(\mathcal{G}[\boldsymbol{f}])}_\alpha(\boldsymbol{k}) \widehat{\frac{\partial \psi}{\partial x_\beta}}(-\boldsymbol{k}) - \sum_{\boldsymbol{k}\in\mathbb{Z}} \widehat{(\mathcal{G}[\boldsymbol{f}])}_\beta(\boldsymbol{k}) \widehat{\frac{\partial \psi}{\partial x_\alpha}}(-\boldsymbol{k}) \right)_{\alpha,\beta=1}^{d} \tag{2.19b}$$

$$= \left(\sum_{\boldsymbol{k}\in\mathbb{Z}\setminus\{\boldsymbol{0}\}} \sum_{\zeta,\theta} \frac{\xi_\alpha \xi_\zeta A^0_{\zeta\theta} \widehat{f}_\theta}{\left(\boldsymbol{A}^0 \boldsymbol{\xi}(\boldsymbol{k}), \boldsymbol{\xi}(\boldsymbol{k})\right)_{\mathbb{R}^d}} \mathrm{i}\pi \xi_\beta \widehat{\psi} - \sum_{\boldsymbol{k}\in\mathbb{Z}\setminus\{\boldsymbol{0}\}} \sum_{\zeta,\theta} \frac{\xi_\beta \xi_\zeta A^0_{\zeta\theta} \widehat{f}_\theta}{\left(\boldsymbol{A}^0 \boldsymbol{\xi}(\boldsymbol{k}), \boldsymbol{\xi}(\boldsymbol{k})\right)_{\mathbb{R}^d}} \mathrm{i}\pi \xi_\alpha \widehat{\psi} \right)_{\alpha,\beta=1}^{d} \tag{2.19c}$$

$$= (0)_{\alpha,\beta=1}^{d} \tag{2.19d}$$

where Eq. (2.19a) is just from Def. (2.16), Eq. (2.19b) comes from a property (c) in Lemma 2.11, and Eq. (2.19c) follows from the definition of operator $\mathcal{G}$ and Lemma 2.3.

We show that the projection maps on space $\mathscr{E}$ by showing $\mathcal{G}[\boldsymbol{f}] = \boldsymbol{f}$ for all $\boldsymbol{f} \in \mathscr{E}$. Using again property (c) in Lemma 2.11, we can write for an arbitrary test function $\psi \in \mathcal{D}_{\text{per}}(\mathbb{R}^d)$ and arbitrary α

$$\left(f_\alpha - \mathcal{G}[\boldsymbol{f}]_\alpha, \psi\right)_{L^2_{\text{per}}(\mathcal{Y})} = \sum_{\boldsymbol{k}\in\mathbb{Z}\setminus\{\boldsymbol{0}\}} \left(\hat{f}_\alpha(\boldsymbol{k}) - \sum_{\beta,\gamma} \frac{\xi_\alpha(\boldsymbol{k}) \xi_\beta(\boldsymbol{k})}{\left(\boldsymbol{A}^0 \boldsymbol{\xi}(\boldsymbol{k}), \boldsymbol{\xi}(\boldsymbol{k})\right)_{\mathbb{R}^d}} A^0_{\beta\gamma} \widehat{f}_\gamma(\boldsymbol{k}) \right) \widehat{\psi}(\boldsymbol{k})$$

Using Lemma 2.18 giving a potential $u \in H^1_{\text{per}}(\mathcal{Y})$ such that $\nabla u = \boldsymbol{f}$, we can continue as

$$= \sum_{\boldsymbol{k}\in\mathbb{Z}\setminus\{\boldsymbol{0}\}} \left(\widehat{\frac{\partial u}{\partial x_\alpha}}(\boldsymbol{k}) - \sum_{\beta,\gamma} \frac{\xi_\alpha(\boldsymbol{k}) \xi_\beta(\boldsymbol{k})}{\left(\boldsymbol{A}^0 \boldsymbol{\xi}(\boldsymbol{k}), \boldsymbol{\xi}(\boldsymbol{k})\right)_{\mathbb{R}^d}} A^0_{\beta\gamma} \widehat{\frac{\partial u}{\partial x_\gamma}}(\boldsymbol{k}) \right) \widehat{\psi}(\boldsymbol{k})$$

$$= \sum_{\boldsymbol{k}\in\mathbb{Z}\setminus\{\boldsymbol{0}\}} \left(\mathrm{i}\pi \xi_\alpha(\boldsymbol{k}) \hat{u}(\boldsymbol{k}) - \sum_{\beta,\gamma} \frac{\xi_\alpha(\boldsymbol{k}) \xi_\beta(\boldsymbol{k})}{\left(\boldsymbol{A}^0 \boldsymbol{\xi}(\boldsymbol{k}), \boldsymbol{\xi}(\boldsymbol{k})\right)_{\mathbb{R}^d}} A^0_{\beta\gamma} \mathrm{i}\pi \xi_\gamma(\boldsymbol{k}) \widehat{u}(\boldsymbol{k}) \right) \widehat{\psi}(\boldsymbol{k}) = 0$$

where we have used Lemma 2.3 about the Fourier coefficients of a derivative. □

Using previously proven Lemma, we show the connection between both formulations. We notice that the solutions of both formulations differ by a constant; the unique solution of weak formulation $\tilde{\boldsymbol{e}} \in \mathscr{E}$ has zero mean while the solution of Lippmann-Schwinger equation $\boldsymbol{e}_{\text{LS}}$, if exists, has the mean value equal to macroscopic load $\langle \boldsymbol{e}_{\text{LS}} \rangle = \boldsymbol{E}$.

Theorem 2.29. *Weak formulation, Def.* (2.20)*, and Lippmann-Schwinger equation, Def.* (2.23)*, are equivalent in the sense the unique solution coincide.*

Proof. The unique microscopic field $\boldsymbol{e} = \boldsymbol{E} + \tilde{\boldsymbol{e}}$ with $\tilde{\boldsymbol{e}} \in \mathscr{E}$ and $\boldsymbol{E} \in \mathscr{U}$ coming from a weak formulation satisfies

$$\left(\boldsymbol{A}\tilde{\boldsymbol{e}}, \boldsymbol{w}\right)_{L^2_{\text{per}}(\mathcal{Y};\mathbb{R}^d)} = -\left(\boldsymbol{A}\boldsymbol{E}, \boldsymbol{w}\right)_{L^2_{\text{per}}(\mathcal{Y};\mathbb{R}^d)}, \quad \forall \boldsymbol{w} \in \mathscr{E}.$$

Using the property of operator $\mathcal{G}$, i.e. $\mathcal{G}[\boldsymbol{w}] = \boldsymbol{w}$ for all $\boldsymbol{w} \in \mathscr{E}$, and the identity $\boldsymbol{A}^0 (\boldsymbol{A}^0)^{-1}$ noting that $\boldsymbol{A}^0$ is positive definite and symmetric, we can rewrite the weak formulation into

$$\left(\boldsymbol{A}^0 (\boldsymbol{A}^0)^{-1} \boldsymbol{A}(\boldsymbol{E} + \tilde{\boldsymbol{e}}), \mathcal{G}\boldsymbol{w}\right)_{L^2_{\text{per}}(\mathcal{Y};\mathbb{R}^d)} = 0, \quad \forall \boldsymbol{w} \in L^2_{\text{per}}(\mathcal{Y};\mathbb{R}^d),$$
$$\left((\boldsymbol{A}^0)^{-1} \boldsymbol{A}(\boldsymbol{E} + \tilde{\boldsymbol{e}}), \boldsymbol{A}^0 \mathcal{G}\boldsymbol{w}\right)_{L^2_{\text{per}}(\mathcal{Y};\mathbb{R}^d)} = 0, \quad \forall \boldsymbol{w} \in L^2_{\text{per}}(\mathcal{Y};\mathbb{R}^d),$$

where we have enlarged the space of test functions. Next, using the properties of adjoint operator $\mathcal{G}^*$ stated in Remark 2.27 and the fact that $\boldsymbol{A}^0(\cdot) : \boldsymbol{w} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) \to \boldsymbol{A}^0\boldsymbol{w} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ is isomorphism, we write

$$\begin{aligned}\big((\boldsymbol{A}^0)^{-1}\boldsymbol{A}(\boldsymbol{E}+\tilde{\boldsymbol{e}}),\mathcal{G}^*\boldsymbol{A}^0\boldsymbol{w}\big)_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)} = 0, \quad \forall \boldsymbol{w} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)\\ \big(\mathcal{G}(\boldsymbol{A}^0)^{-1}\boldsymbol{A}\boldsymbol{e},\boldsymbol{v}\big)_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)} = 0, \quad \forall \boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)\end{aligned}$$

where $\boldsymbol{v} = \boldsymbol{A}^0\boldsymbol{w}$.

Adding and substracting the terms $\big(\tilde{\boldsymbol{e}},\boldsymbol{v}\big)_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)}$ and $\big(\boldsymbol{E},\boldsymbol{v}\big)_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)}$ and using a property $\mathcal{G}\boldsymbol{E} = 0$ heads to

$$\begin{aligned}\big(\tilde{\boldsymbol{e}},\boldsymbol{v}\big)_{L^2_{\mathrm{per}}} + \big(\boldsymbol{E},\boldsymbol{v}\big)_{L^2_{\mathrm{per}}} + \big(\mathcal{G}(\boldsymbol{A}^0)^{-1}\boldsymbol{A}\boldsymbol{e},\boldsymbol{v}\big)_{L^2_{\mathrm{per}}} - \big(\tilde{\boldsymbol{e}},\boldsymbol{v}\big)_{L^2_{\mathrm{per}}} = \big(\boldsymbol{E},\boldsymbol{v}\big)_{L^2_{\mathrm{per}}}, \quad \forall \boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)\\ \big(\boldsymbol{e}+\mathcal{G}(\boldsymbol{A}^0)^{-1}(\boldsymbol{A}-\boldsymbol{A}^0)\boldsymbol{e},\boldsymbol{v}\big)_{L^2(\mathcal{Y};\mathbb{R}^d)} = \big(\boldsymbol{E},\boldsymbol{v}\big)_{L^2(\mathcal{Y};\mathbb{R}^d)}, \quad \forall \boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)\end{aligned}$$

Since it is tested for all test function $\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$, we abandon the scalar product to equation holding almost everywhere

$$\begin{aligned}\boldsymbol{e}(\boldsymbol{x}) + \mathcal{G}[(\boldsymbol{A}^0)^{-1}(\boldsymbol{A}-\boldsymbol{A}^0)\boldsymbol{e}](\boldsymbol{x}) = \boldsymbol{E}, \quad \text{for a.a. } \boldsymbol{x} \in \mathcal{Y}\\ \boldsymbol{e}(\boldsymbol{x}) + \int_{\mathcal{Y}} \boldsymbol{\Gamma}(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{A}(\boldsymbol{y})-\boldsymbol{A}^0)\boldsymbol{e}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = \boldsymbol{E}, \quad \text{for a.a. } \boldsymbol{x} \in \mathcal{Y}\end{aligned}$$

The functions from $L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ equal to each other if they are the same almost everywhere, thus we can choose an appropriate representative in order to satisfy the equation everywhere and the first implication is done.

The contrary case can be done by multiplication with a test function $\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$, by integration over $\mathcal{Y}$, by split of the solution with projection $\boldsymbol{e} = \mathcal{G}\boldsymbol{e} + (\boldsymbol{I}-\mathcal{G})\boldsymbol{e}$, and by the same arguments about the projections $\mathcal{G}$ and its adjoint $\mathcal{G}^*$ as in the first part of the proof. □

3 Discrete solutions

This section is dedicated to the discretization of weak formulation in Def. 2.20. The trigonometric polynomials are used as the basis functions, see Sec. 3.1, and their approximation properties are stated in Sec. 3.2.

Next, in Sec. 3.3, we define Galerkin approximation (GA) and Galerkin approximation with numerical integration (GAwNI) and we show the fully discrete formulation of GAwNI that is appropriate for the solution by Conjugate gradients. In Sec. 3.4, we provide convergence of discrete solutions to the solution of weak formulation, then in Sec. 3.5 the convergence is shown for rough material coefficients by their regularization. Finally, Sec. 3.6 provides information about the numerical solution of a corresponding linear system by Conjugate gradients.

In the sequel, vector $\boldsymbol{N} \in \mathbb{N}^d$ is reserved for a number of discretization points, then scalar $|\boldsymbol{N}|_\Pi := \prod_\alpha N_\alpha$ denotes the number of degrees of freedom. If N_α is odd for all α we talk about the odd number of discretization points. We use a simplification in order to make the theory of real valued trigonometric polynomials uncomplicated — in the sequel, only the odd number of discretization points is considered, Sec. 3.1.

A multi-index notation is employed, in which $\mathbb{R}^{\boldsymbol{N}}$ represents $\mathbb{R}^{N_1\times\cdots\times N_d}$. Set $\mathbb{R}^{d\times\boldsymbol{N}}$ denotes the space of vectors v with components $\mathsf{v}^{\boldsymbol{n}}_\alpha$ and $\mathbb{R}^{d\times d\times\boldsymbol{N}\times\boldsymbol{N}}$ the space of matrices A with components $\mathsf{A}^{\boldsymbol{nm}}_{\alpha\beta}$ for α,β and $\boldsymbol{n},\boldsymbol{m} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}$. Next, vectors $\mathsf{v}^{\boldsymbol{n}} \in \mathbb{R}^d$ for $\boldsymbol{n} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}$ and $\mathsf{v}_\alpha \in \mathbb{R}^{\boldsymbol{N}}$ for α represent subvectors of v with components $\mathsf{v}^{\boldsymbol{n}}_\alpha$. Analogically, submatrices $\mathsf{A}^{\boldsymbol{nm}} \in \mathbb{R}^{d\times d}$ and $\mathsf{A}_{\alpha\beta} \in \mathbb{R}^{\boldsymbol{N}\times\boldsymbol{N}}$ can be defined. A scalar product on set $\mathbb{R}^{d\times\boldsymbol{N}}$ is defined as $\big(\mathsf{u},\mathsf{v}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}} := \sum_\alpha \sum_{\boldsymbol{n}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \mathsf{u}^{\boldsymbol{n}}_\alpha \mathsf{v}^{\boldsymbol{n}}_\alpha$ and matrix A by vector v multiplication as $(\mathsf{A}\mathsf{v})^{\boldsymbol{n}}_\alpha := \sum_\beta \sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \mathsf{A}^{\boldsymbol{nm}}_{\alpha\beta}\mathsf{v}^{\boldsymbol{m}}_\beta$. Matrix A is symmetric positive definite if relation

$\mathsf{A}^{\boldsymbol{mn}}_{\alpha\beta} = \mathsf{A}^{\boldsymbol{nm}}_{\beta\alpha}$ holds for all components and inequality $\left(\mathbf{A}\mathbf{v}, \mathbf{v}\right)_{\mathbb{R}^{d\times \boldsymbol{N}}} > 0$ applies for arbitrary $\mathbf{v} \in \mathbb{R}^{d\times \boldsymbol{N}}$. We use the serif font for vectors $\mathbf{v} \in \mathbb{R}^{d\times \boldsymbol{N}}$ and matrices $\mathbf{A} \in \mathbb{R}^{d\times d\times \boldsymbol{N}\times \boldsymbol{N}}$ to distinguish from vectors $\boldsymbol{E} \in \mathbb{R}^d$ and matrices $\boldsymbol{A}(\boldsymbol{x}) \in \mathbb{R}^{d\times d}$ and from vector valued functions $\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y}; \mathbb{R}^d)$. In order to differentiate vectors and matrices for different number of discretization points $\boldsymbol{N}$, we often write them with subscript $\boldsymbol{N}$, i.e. $\mathbf{v}_{\boldsymbol{N}}$ and $\mathbf{A}_{\boldsymbol{N}}$.

3.1 Finite dimensional space of trigonometric polynomials

In this section we define a finite dimensional space of trigonometric polynomials and provide its properties.

Definition 3.1. *For odd $\boldsymbol{N}$, we define the space of trigonometric polynomials as*

$$\mathscr{T}_{\boldsymbol{N}} = \Big\{ \sum_{\boldsymbol{k}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \hat{\mathsf{c}}^{\boldsymbol{k}} \varphi_{\boldsymbol{k}}(\boldsymbol{x}); \hat{\mathsf{c}}^{\boldsymbol{k}} \in \mathbb{C}, \hat{\mathsf{c}}^{\boldsymbol{k}} = \overline{(\hat{\mathsf{c}}^{-\boldsymbol{k}})} \text{ for } \boldsymbol{k} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}} \Big\} \tag{3.1}$$

where the index set $\underline{\mathbb{Z}}^d_{\boldsymbol{N}}$, already defined in Lemma 2.11, is expressed as

$$\underline{\mathbb{Z}}^d_{\boldsymbol{N}} = \left\{ \boldsymbol{k} \in \mathbb{Z}^d : -\frac{N_\alpha}{2} \le k_\alpha < \frac{N_\alpha}{2}, \right\} \tag{3.2}$$

and $\varphi_{\boldsymbol{k}}$ are basis functions defined already in Eq. (2.4), i.e. $\varphi_{\boldsymbol{k}}(\boldsymbol{x}) = \exp\left(\mathrm{i}\pi \sum_{\alpha=1}^d \frac{k_\alpha x_\alpha}{Y_\alpha}\right)$. The scalar valued functions from the space $\mathscr{T}_{\boldsymbol{N}}$ are denoted with a subscript $\boldsymbol{N}$, e.g. $f_{\boldsymbol{N}} \in \mathscr{T}_{\boldsymbol{N}}$. Analogically, we define the vector valued trigonometric polynomials as

$$\mathscr{T}^d_{\boldsymbol{N}} = \big\{ \boldsymbol{f} \in L^2_{\mathrm{per}}(\mathcal{Y}; \mathbb{R}^d) : f_\alpha \in \mathscr{T}_{\boldsymbol{N}} \text{ for } \alpha = 1, \dots, d \big\},$$

Remark 3.2. *The condition $\hat{\mathsf{c}}^{\boldsymbol{k}} = \overline{(\hat{\mathsf{c}}^{-\boldsymbol{k}})}$ for $\boldsymbol{k} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}$ in previous definition satisfies that trigonometric polynomials are real valued. The fact that $\boldsymbol{N}$ is odd satisfies that the boundary frequencies $\frac{N_\alpha}{2}$ do not occur at all. This assures that the index set $\underline{\mathbb{Z}}^d_{\boldsymbol{N}}$ is symmetric around zero, thus all frequencies $\boldsymbol{k} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}$ has the opposite counterpart $-\boldsymbol{k}$ as well as the Fourier coefficients.*

Now, we show two possible representations of trigonometric polynomials and the connection between them. They can be expressed through their Fourier coefficients and through their function values at nodal points.

Remark 3.3 (Representation of $\boldsymbol{v}_{\boldsymbol{N}} \in \mathscr{T}^d_{\boldsymbol{N}}$ through its Fourier coefficients)**.** *The function of trigonometric polynomials, $v_{\boldsymbol{N}} \in \mathscr{T}_{\boldsymbol{N}}$, can be expressed as*

$$v_{\boldsymbol{N}}(\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \hat{v}_{\boldsymbol{N}}(\boldsymbol{k}) \varphi_{\boldsymbol{k}}(\boldsymbol{x}),$$

where $\hat{v}_{\boldsymbol{N}}(\boldsymbol{k}) \in \mathbb{C}^d$ are Fourier coefficients defined in Def. (2.2).

Definition 3.4 (nodal points)**.** *We define the nodal points of a periodic unit cell $\mathcal{Y}$ as points $\boldsymbol{x}^{\boldsymbol{k}}_{\boldsymbol{N}} \in \mathcal{Y}$ for $\boldsymbol{k} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}$ explicitly expressed as*

$$\boldsymbol{x}^{\boldsymbol{k}}_{\boldsymbol{N}} := \sum_\alpha \frac{2Y_\alpha k_\alpha}{N_\alpha} \boldsymbol{\epsilon}_\alpha = \left(\frac{2Y_1 k_1}{N_1}, \frac{2Y_2 k_2}{N_2}, \dots, \frac{2Y_d k_d}{N_d} \right).$$

Thus each vector $\boldsymbol{k} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}$ points out one point $\boldsymbol{x}^{\boldsymbol{k}}_{\boldsymbol{N}} \in \mathcal{Y}$ from the regular grid $\{\boldsymbol{x}^{\boldsymbol{k}}_{\boldsymbol{N}} \in \mathcal{Y}; \boldsymbol{k} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}\}$.

Remark 3.5 (Representation of $\boldsymbol{v}_{\boldsymbol{N}} \in \mathscr{T}^d_{\boldsymbol{N}}$ through its nodal values)**.** *The trigonometric polynomials can be represented through its nodal values as*

$$\boldsymbol{v}_{\boldsymbol{N}}(\boldsymbol{x}) = \sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \boldsymbol{v}_{\boldsymbol{N}}(\boldsymbol{x}^{\boldsymbol{m}}_{\boldsymbol{N}}) \varphi_{\boldsymbol{N},\boldsymbol{m}}(\boldsymbol{x})$$

where the functions $\varphi_{\boldsymbol{N},\boldsymbol{k}}(\boldsymbol{x}) \in \mathscr{T}_{\boldsymbol{N}}^d$ *state for*

$$\varphi_{\boldsymbol{N},\boldsymbol{m}}(\boldsymbol{x}) = \frac{1}{|\boldsymbol{N}|_\Pi} \sum_{\boldsymbol{k}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} \varphi_{\boldsymbol{k}}(\boldsymbol{x}) \omega_{\boldsymbol{N}}^{-\boldsymbol{m}\boldsymbol{k}} = \frac{1}{|\boldsymbol{N}|_\Pi} \sum_{\boldsymbol{k}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} \exp\left\{ \mathrm{i}\pi \sum_{\alpha=1}^d k_\alpha \left(\frac{x_\alpha}{Y_\alpha} - \frac{2m_\alpha}{N_\alpha} \right) \right\} \tag{3.3}$$

The are called shape basis functions and satisfy the Dirac delta property at nodal points

$$\varphi_{\boldsymbol{N},\boldsymbol{m}}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{k}}) = \delta_{\boldsymbol{k}\boldsymbol{m}}, \tag{3.4}$$

i.e. it is such trigonometric polynomial having the value one in a particular nodal point of regular grid and value zero in the rest of the nodal points.

Remark 3.6 (Connection of represention)**.** *Let* $v_{\boldsymbol{N}}(\boldsymbol{x}) \in \mathscr{T}_{\boldsymbol{N}}$*. Then Fourier coefficients* $\hat{v}_{\boldsymbol{N}}(\boldsymbol{k})$*,* $\boldsymbol{k} \in \underline{\mathbb{Z}}_{\boldsymbol{N}}^d$ *can be obtained with the Discrete Fourier Transform from the values at nodal points* $\left(v_{\boldsymbol{N}}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{m}})\right)^{\boldsymbol{m}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d}$*. Hence*

$$\begin{aligned} \hat{\boldsymbol{v}}(\boldsymbol{k}) &= \frac{1}{|\boldsymbol{N}|_\Pi} \sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} \boldsymbol{v}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{m}}) \omega_{\boldsymbol{N}}^{-\boldsymbol{k}\boldsymbol{m}}, \quad \boldsymbol{k} \in \underline{\mathbb{Z}}_{\boldsymbol{N}}^d \\ \boldsymbol{v}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{m}}) &= \sum_{\boldsymbol{k}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} \hat{\boldsymbol{v}}(\boldsymbol{k}) \omega_{\boldsymbol{N}}^{\boldsymbol{k}\boldsymbol{m}}, \quad \boldsymbol{m} \in \underline{\mathbb{Z}}_{\boldsymbol{N}}^d \end{aligned}$$

where $\omega_{\boldsymbol{N}}^{\boldsymbol{k}\boldsymbol{m}} = \exp\left(2\pi\mathrm{i} \sum_\alpha \frac{k_\alpha m_\alpha}{N_\alpha}\right)$ *are the coefficients of DFT with the property*

$$\sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} \omega_{\boldsymbol{N}}^{\boldsymbol{k}\boldsymbol{m}} \overline{\omega_{\boldsymbol{N}}^{\boldsymbol{m}\boldsymbol{n}}} = \delta_{\boldsymbol{k}\boldsymbol{n}} |\boldsymbol{N}|_\Pi \tag{3.5}$$

Remark 3.7 (Orthogonality of $\varphi_{\boldsymbol{N},\boldsymbol{k}}$)**.** *For shape basis functions, the following holds*

$$\left(\varphi_{\boldsymbol{N},\boldsymbol{k}}, \varphi_{\boldsymbol{N},\boldsymbol{l}}\right) = \frac{\delta_{\boldsymbol{k}\boldsymbol{l}}}{|\boldsymbol{N}|_\Pi}$$

Proof.

$$\begin{aligned} \left(\varphi_{\boldsymbol{N},\boldsymbol{k}}, \varphi_{\boldsymbol{N},\boldsymbol{l}}\right)_{L^2_{\mathrm{per}}} &= \frac{1}{|\boldsymbol{N}|_\Pi} \sum_{\boldsymbol{m}\in\mathbb{Z}_{\boldsymbol{N}}^d} \omega_{\boldsymbol{N}}^{\boldsymbol{m}\boldsymbol{k}} \left(\varphi_{\boldsymbol{m}}, \varphi_{\boldsymbol{N},\boldsymbol{l}}\right)_{L^2_{\mathrm{per}}} \\ &= \left(\frac{1}{|\boldsymbol{N}|_\Pi}\right)^2 \sum_{\boldsymbol{m}\in\mathbb{Z}_{\boldsymbol{N}}^d} \sum_{\boldsymbol{n}\in\mathbb{Z}_{\boldsymbol{N}}^d} \omega_{\boldsymbol{N}}^{\boldsymbol{m}\boldsymbol{k}} \overline{\omega_{\boldsymbol{N}}^{\boldsymbol{n}\boldsymbol{l}}} \left(\varphi_{\boldsymbol{m}}, \varphi_{\boldsymbol{n}}\right)_{L^2_{\mathrm{per}}} \\ &= \left(\frac{1}{|\boldsymbol{N}|_\Pi}\right)^2 \sum_{\boldsymbol{m}\in\mathbb{Z}_{\boldsymbol{N}}^d} \sum_{\boldsymbol{n}\in\mathbb{Z}_{\boldsymbol{N}}^d} \omega_{\boldsymbol{N}}^{\boldsymbol{m}\boldsymbol{k}} \overline{\omega_{\boldsymbol{N}}^{\boldsymbol{n}\boldsymbol{l}}} \delta_{\boldsymbol{m}\boldsymbol{n}} \\ &= \left(\frac{1}{|\boldsymbol{N}|_\Pi}\right)^2 \sum_{\boldsymbol{m}\in\mathbb{Z}_{\boldsymbol{N}}^d} \omega_{\boldsymbol{N}}^{\boldsymbol{m}\boldsymbol{k}} \omega_{\boldsymbol{N}}^{-\boldsymbol{m}\boldsymbol{l}} = \frac{\delta_{\boldsymbol{k}\boldsymbol{l}}}{|\boldsymbol{N}|_\Pi} \end{aligned}$$

□

Definition 3.8. *Operator* $\mathcal{I}_{\boldsymbol{N}} : \mathscr{T}_{\boldsymbol{N}}^d \to \mathbb{R}^{d\times\boldsymbol{N}}$ *stocks the values of the trigonometric polynomials at the nodal points to a vector* $\mathcal{I}_{\boldsymbol{N}}[\boldsymbol{v}_{\boldsymbol{N}}] = \left(\boldsymbol{v}_{\boldsymbol{N},\alpha}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{n}})\right)_{\alpha=1,\ldots,d}^{\boldsymbol{n}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d}$*.*

Lemma 3.9. *Operator* $\mathcal{I}_{\boldsymbol{N}}$ *from the previous definition is an isomorphism.*

Proof. The proof is the consequence of its Def. 3.8 and the Dirac delta property (3.4). □

3.2 Orthogonal and Interpolation Projection

This section provides lemmas about trigonometric approximations, i.e. about the estimates of orthogonal projection and interpolation projection. In [18, Section 8], this topic is well described for a one-dimensional setting and partially for a two-dimensional setting with equal number of discretization points in each direction. Thus we present here the generalization of the estimates for arbitrary dimension and for arbitrary number of discretization points. However, the proofs follows the same ideas as in [18, Section 8].

Definition 3.10 (Orthogonal projection)**.** *For odd $\boldsymbol{N} \in \mathbb{N}$, we define an operator $P_{\boldsymbol{N}}[\cdot] : L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) \to \mathscr{T}^d_N$ as*

$$P_{\boldsymbol{N}}[\boldsymbol{v}](\boldsymbol{x}) = \sum_{k \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \hat{\boldsymbol{v}}(\boldsymbol{k}) \varphi_{\boldsymbol{k}}(\boldsymbol{x})$$

where $\hat{\boldsymbol{v}}(\boldsymbol{k})$ denotes Fourier coefficients of $\boldsymbol{v}$.

Remark 3.11. *It can be easily shown that operator $P_{\boldsymbol{N}}[\cdot]$ is the orthogonal projection.*

The possibility of expression functions as a Fourier series is stated in the following lemma with a proof that is based on the density of the set of trigonometric polynomials $\{\varphi_{\boldsymbol{k}}\}^{\boldsymbol{k}\in\mathbb{Z}^d}$ in space L^2_{per}, e.g. [11].

Lemma 3.12 (Approximation by orthogonal projection)**.** *For a function $\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$, the following holds*

$$\lim_{\boldsymbol{N}\to\infty} \|\boldsymbol{v} - P_{\boldsymbol{N}}[\boldsymbol{v}]\|_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)} = 0$$

where $\boldsymbol{N} \to \infty$ means $\min_\alpha N_\alpha \to \infty$.

Lemma 3.13 (Approximation by orthogonal projection)**.** *For a function $\boldsymbol{v} \in H^\mu_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ with $\mu \in \mathbb{R}$ we have*

$$\|\boldsymbol{v} - P_{\boldsymbol{N}}[\boldsymbol{v}]\|_{H^\lambda_{\mathrm{per}}} \le \left(\min_{\alpha=1,\dots,d} \frac{N_\alpha}{2Y_\alpha}\right)^{(\lambda-\mu)} \|\boldsymbol{v}\|_{H^\mu_{\mathrm{per}}} \tag{3.6}$$

where $\lambda \le \mu$ and $\boldsymbol{N} \neq \boldsymbol{0}$.

Proof. The proof is a consequence of direct calculation.

$$\begin{aligned}
\|\boldsymbol{v} - P_{\boldsymbol{N}}[\boldsymbol{v}]\|^2_{H^\lambda_{\mathrm{per}}} &= \sum_{\mathbb{Z}^d\setminus\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \|\underline{\boldsymbol{\xi}(\boldsymbol{k})}\|_2^{2\lambda} \|\hat{\boldsymbol{v}}(\boldsymbol{k})\|_2^2 \\
&= \sum_{\mathbb{Z}^d\setminus\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \|\underline{\boldsymbol{\xi}(\boldsymbol{k})}\|_2^{2(\lambda-\mu)} \|\underline{\boldsymbol{\xi}(\boldsymbol{k})}\|_2^{2\mu} \|\hat{\boldsymbol{v}}(\boldsymbol{k})\|_2^2 \\
&\le \left(\min_{\alpha=1,\dots,d} \frac{N_\alpha}{2Y_\alpha}\right)^{2(\lambda-\mu)} \sum_{\mathbb{Z}^d\setminus\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \|\underline{\boldsymbol{\xi}(\boldsymbol{k})}\|_2^{2\mu} \|\hat{\boldsymbol{v}}(\boldsymbol{k})\|_2^2 \\
&\le \left(\min_{\alpha=1,\dots,d} \frac{N_\alpha}{2Y_\alpha}\right)^{2(\lambda-\mu)} \|\boldsymbol{v}\|^2_{H^\mu_{\mathrm{per}}}
\end{aligned}$$

□

Definition 3.14 (Interpolation projection)**.** *For $\boldsymbol{Y}$-periodic continuous functions $C_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ and odd $\boldsymbol{N}$, we define an operator $Q_{\boldsymbol{N}}[\cdot] : C_{\mathrm{per}} \to \mathscr{T}^d_{\boldsymbol{N}}$ with conditions*

$$Q_{\boldsymbol{N}}[\boldsymbol{v}] \in \mathscr{T}^d_{\boldsymbol{N}} \qquad\qquad Q_{\boldsymbol{N}}[\boldsymbol{v}](\boldsymbol{x}^{\boldsymbol{k}}_{\boldsymbol{N}}) = \boldsymbol{v}(\boldsymbol{x}^{\boldsymbol{k}}_{\boldsymbol{N}}) \quad \text{for } \boldsymbol{k} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}.$$

Alternatively, the interpolation operator can be defined as

$$Q_{\boldsymbol{N}}[\boldsymbol{v}] = \sum_{\boldsymbol{k}\in\mathbb{Z}^d_{\boldsymbol{N}}} \boldsymbol{v}(\boldsymbol{x}^{\boldsymbol{k}}_{\boldsymbol{N}})\varphi_{\boldsymbol{N},\boldsymbol{k}}.$$

This is based on two possibilities of trigonometric polynomial representation and their connection Rem. 3.6.

Remark 3.15. *The interpolation projection is well defined for $H^\mu_{\mathrm{per}}(\mathcal{Y})$ with $\mu > \frac{d}{2}$ as the space $H^\mu_{\mathrm{per}}(\mathcal{Y})$ is embedded into the space of continuous functions C_{per}.*

Remark 3.16. *The mapping $Q_{\boldsymbol{N}}$ from previous definition is clearly a projection, nevertheless not orthogonal anymore.*

Next, we state the lemma about interpolation operator in Fourier space; the lemma is used for the facts about approximation by interpolation projection.

Lemma 3.17 (Interpolation operator Q_N in Fourier space)**.** *For $\boldsymbol{v} \in H^\mu_{\mathrm{per}}(\mathcal{Y})$, $\mu > \frac{d}{2}$*

$$Q_N[\boldsymbol{v}](\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^d_{\boldsymbol{N}}} \left[\sum_{\boldsymbol{l}\in\mathbb{Z}^d} \hat{\boldsymbol{v}}(\boldsymbol{k}+\boldsymbol{l}\odot\boldsymbol{N})\right]\varphi_{\boldsymbol{k}}(\boldsymbol{x})$$

where $\odot$ denotes element-wise multiplication.

Proof. Using the property of operator $Q_{\boldsymbol{N}}[\cdot]$ to being a projection on $\mathscr{T}_{\boldsymbol{N}}$, we have $Q_{\boldsymbol{N}}[\varphi_{\boldsymbol{k}}] = \varphi_{\boldsymbol{k}}$ as $\varphi_{\boldsymbol{k}} \in \mathscr{T}_{\boldsymbol{N}}$. Moreover for all $\boldsymbol{k},\boldsymbol{m} \in \mathbb{Z}^d_{\boldsymbol{N}}$ and $\boldsymbol{l} \in \mathbb{Z}^d$, the basis functions $\varphi_{\boldsymbol{k}}(\boldsymbol{x})$ and $\varphi_{\boldsymbol{k}+\boldsymbol{l}\odot\boldsymbol{N}}(\boldsymbol{x})$ equals one another[2] at nodal points $\boldsymbol{x}^{\boldsymbol{m}}_{\boldsymbol{N}} = \sum_{\alpha=1}^d \frac{2Y_\alpha k_\alpha}{N_\alpha}\boldsymbol{\epsilon}_\alpha$ as

$$\begin{aligned}\varphi_{\boldsymbol{k}+\boldsymbol{l}\odot\boldsymbol{N}}(\boldsymbol{x}^{\boldsymbol{m}}_{\boldsymbol{N}}) &= \varphi_{\boldsymbol{k}}(\boldsymbol{x}^{\boldsymbol{m}}_{\boldsymbol{N}}) \cdot \exp\left(\mathrm{i}\pi \sum_{\alpha=1}^d \frac{l_\alpha N_\alpha 2Y_\alpha m_\alpha}{Y_\alpha N_\alpha}\right)\\ &= \varphi_{\boldsymbol{k}}(\boldsymbol{x}^{\boldsymbol{m}}_{\boldsymbol{N}}) \cdot \exp\left(2\mathrm{i}\pi \sum_{\alpha=1}^d l_\alpha m_\alpha\right) = \varphi_{\boldsymbol{k}}(\boldsymbol{x}^{\boldsymbol{m}}_{\boldsymbol{N}})\end{aligned}$$

and thus we even have

$$Q_{\boldsymbol{N}}[\varphi_{\boldsymbol{k}+\boldsymbol{l}\odot\boldsymbol{N}}] = Q_{\boldsymbol{N}}[\varphi_{\boldsymbol{k}}] = \varphi_{\boldsymbol{k}}, \quad (\boldsymbol{k} \in \mathbb{Z}^d_{\boldsymbol{N}}, \boldsymbol{l} \in \mathbb{Z}^d).$$

As the operator $Q_{\boldsymbol{N}}$ is obviously linear, we can finish the proof

$$\begin{aligned}Q_N[\boldsymbol{v}](\boldsymbol{x}) &= Q_N\left[\sum_{\boldsymbol{m}\in\mathbb{Z}^d} \hat{\boldsymbol{v}}(\boldsymbol{m})\varphi_{\boldsymbol{m}}(\boldsymbol{x})\right](\boldsymbol{x}) = Q_N\left[\sum_{\boldsymbol{k}\in\mathbb{Z}^d_{\boldsymbol{N}}}\sum_{\boldsymbol{l}\in\mathbb{Z}^d} \hat{\boldsymbol{v}}(\boldsymbol{k}+\boldsymbol{l}\odot\boldsymbol{N})\varphi_{\boldsymbol{k}+\boldsymbol{l}\odot\boldsymbol{N}}(\boldsymbol{x})\right](\boldsymbol{x})\\ &= \sum_{\boldsymbol{k}\in\mathbb{Z}^d_{\boldsymbol{N}}} \left[\sum_{\boldsymbol{l}\in\mathbb{Z}^d} \hat{\boldsymbol{v}}(\boldsymbol{k}+\boldsymbol{l}\odot\boldsymbol{N})\right]\varphi_{\boldsymbol{k}}(\boldsymbol{x})\end{aligned}$$

□

Lemma 3.18 (Approximation by interpolation projection)**.** *Let $v \in H^\mu_{\mathrm{per}}(\mathcal{Y})$ with $\mu > \frac{d}{2}$, then*

$$\|v - Q_{\boldsymbol{N}}[v]\|_{H^\lambda_{\mathrm{per}}(\mathcal{Y})} \le c_{\lambda,\mu}\left(\min_{\alpha=1,\dots,d}\frac{N_\alpha}{2Y_\alpha}\right)^{\lambda-\mu}\|v\|_{H^\mu_{\mathrm{per}}(\mathcal{Y})} \tag{3.7}$$

[2] operator $\odot$ denotes element-wise multiplication

for $0 \le \lambda \le \mu$ *and*

$$c_{\lambda,\mu} = \left(1 + d^\lambda \rho^{2\lambda} \sum_{\boldsymbol{l} \in \mathbb{N}_0^d \setminus \{\boldsymbol{0}\}} \|\boldsymbol{l}\|_2^{-2\mu}\right)^{\frac{1}{2}}$$
$$\rho = \frac{\max_{\alpha=1,\dots,d} \frac{N_\alpha}{2Y_\alpha}}{\min_{\alpha=1,\dots,d} \frac{N_\alpha}{2Y_\alpha}}.$$

Proof. The proof generalizes the proof of Theorems 8.2.1 and 8.5.3 in [18]. Using the identity

$$\|v - Q_{\boldsymbol{N}}[v]\|^2_{H^\lambda_{\mathrm{per}}(\mathcal{Y})} = \|v - P_{\boldsymbol{N}}[v]\|^2_{H^\lambda_{\mathrm{per}}(\mathcal{Y})} + \|P_{\boldsymbol{N}}[v] - Q_{\boldsymbol{N}}[v]\|^2_{H^\lambda_{\mathrm{per}}(\mathcal{Y})}$$

and Lem. 3.13 about approximation with orthogonal projection, we have to estimate the second term $\|P_{\boldsymbol{N}}[v] - Q_{\boldsymbol{N}}[v]\|^2_{H^\lambda_{\mathrm{per}}(\mathcal{Y})}$.

Using Lem. 3.17, we can express Fourier coefficients of undergoing term

$$\widehat{(P_{\boldsymbol{N}}[v] - Q_{\boldsymbol{N}}[v])}(\boldsymbol{k}) = \begin{cases} \sum_{\boldsymbol{l} \in \mathbb{Z}^d \setminus \{\boldsymbol{0}\}} \hat{v}(\boldsymbol{k} + \boldsymbol{l} \odot \boldsymbol{N}), & \text{for } \boldsymbol{k} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}} \\ 0, & \text{for } \boldsymbol{k} \in \mathbb{Z}^d \setminus \underline{\mathbb{Z}}^d_{\boldsymbol{N}}. \end{cases}$$

Then some algebra with Cauchy inequality yield

$$\begin{aligned}
\|P_{\boldsymbol{N}}[v] - Q_{\boldsymbol{N}}[v]\|^2_{H^\lambda_{\mathrm{per}}(\mathcal{Y})} &= \sum_{\boldsymbol{k} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \|\underline{\boldsymbol{\xi}(\boldsymbol{k})}\|_2^{2\lambda} \left| \sum_{\boldsymbol{l} \in \mathbb{Z}^d \setminus \{\boldsymbol{0}\}} \hat{v}(\boldsymbol{k} + \boldsymbol{l} \odot \boldsymbol{N}) \right|^2 \\
&\le \sum_{\boldsymbol{k} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \left(\sum_{\boldsymbol{l} \in \mathbb{Z}^d \setminus \{\boldsymbol{0}\}} \frac{\|\underline{\boldsymbol{\xi}(\boldsymbol{k})}\|_2^\lambda}{\|\boldsymbol{\xi}(\boldsymbol{k} + \boldsymbol{l} \odot \boldsymbol{N})\|_2^\mu} \cdot \|\boldsymbol{\xi}(\boldsymbol{k} + \boldsymbol{l} \odot \boldsymbol{N})\|_2^\mu |\hat{v}(\boldsymbol{k} + \boldsymbol{l} \odot \boldsymbol{N})| \right)^2 \\
&\le \sum_{\boldsymbol{k} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \left(\sum_{\boldsymbol{l} \in \mathbb{Z}^d \setminus \{\boldsymbol{0}\}} \frac{\|\underline{\boldsymbol{\xi}(\boldsymbol{k})}\|_2^{2\lambda}}{\|\boldsymbol{\xi}(\boldsymbol{k} + \boldsymbol{l} \odot \boldsymbol{N})\|_2^{2\mu}} \right) \left(\sum_{\boldsymbol{l} \in \mathbb{Z}^d \setminus \{\boldsymbol{0}\}} \|\boldsymbol{\xi}(\boldsymbol{k} + \boldsymbol{l} \odot \boldsymbol{N})\|_2^{2\mu} |\hat{v}(\boldsymbol{k} + \boldsymbol{l} \odot \boldsymbol{N})|^2 \right) \\
&\le \varepsilon^2_{\boldsymbol{N}} \|\boldsymbol{v}\|^2_{H^\mu_{\mathrm{per}}}(\mathcal{Y})
\end{aligned}$$

where

$$\begin{aligned}
\varepsilon^2_{\boldsymbol{N}} &= \max_{\boldsymbol{k} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \left(\sum_{\boldsymbol{l} \in \mathbb{Z}^d \setminus \{\boldsymbol{0}\}} \frac{\|\underline{\boldsymbol{\xi}(\boldsymbol{k})}\|_2^{2\lambda}}{\|\boldsymbol{\xi}(\boldsymbol{k} + \boldsymbol{l} \odot \boldsymbol{N})\|_2^{2\mu}} \right) \\
&\le \|\boldsymbol{\xi}(\boldsymbol{N}/2)\|_2^{2\lambda} \max_{\boldsymbol{k} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \left(\sum_{\boldsymbol{l} \in \mathbb{Z}^d \setminus \{\boldsymbol{0}\}} \frac{1}{\left[\sum_{\alpha=1}^d (\frac{k_\alpha + l_\alpha N_\alpha}{Y_\alpha})^2\right]^\mu} \right) \\
&= \|\boldsymbol{\xi}(\boldsymbol{N}/2)\|_2^{2\lambda} \max_{\boldsymbol{k} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \left(\sum_{\boldsymbol{l} \in \mathbb{Z}^d \setminus \{\boldsymbol{0}\}} \frac{1}{\left[\sum_{\alpha=1}^d (\frac{N_\alpha}{2Y_\alpha})^2 (\frac{k_\alpha}{2N_\alpha} + 2l_\alpha)^2\right]^\mu} \right) \\
&= \|\boldsymbol{\xi}(\boldsymbol{N}/2)\|_2^{2\lambda} \left(\sum_{\boldsymbol{l} \in \mathbb{Z}^d \setminus \{\boldsymbol{0}\}} \frac{1}{\left[\sum_{\alpha=1}^d (\frac{N_\alpha}{2Y_\alpha})^2 (1 + 2l_\alpha)^2\right]^\mu} \right) \\
&\le \|\boldsymbol{\xi}(\boldsymbol{N}/2)\|_2^{2\lambda} \left(\min_{\alpha=1,\dots,d} \frac{N_\alpha}{2Y_\alpha} \right)^{-2\mu} \left(\sum_{\boldsymbol{l} \in \mathbb{Z}^d \setminus \{\boldsymbol{0}\}} \frac{1}{\left[\sum_{\alpha=1}^d (1 + 2l_\alpha)^2\right]^\mu} \right) \\
&\le d^\lambda \left(\max_{\alpha=1,\dots,d} \frac{N_\alpha}{2Y_\alpha} \right)^{2\lambda} \left(\min_{\alpha=1,\dots,d} \frac{N_\alpha}{2Y_\alpha} \right)^{-2\mu} \left(\sum_{\boldsymbol{l} \in \mathbb{N}_0^d \setminus \{\boldsymbol{0}\}} \frac{1}{\left[\sum_{\alpha=1}^d l_\alpha^2\right]^\mu} \right).
\end{aligned}$$

Hence

$$\|P_{\boldsymbol{N}}[v]-Q_{\boldsymbol{N}}[v]\|^2_{H^\lambda_{\mathrm{per}}(\mathcal{Y})} \leq d^\lambda \rho^{2\lambda} \sum_{\boldsymbol{l}\in\mathbb{N}_0^d\setminus\{\boldsymbol{0}\}} \frac{1}{\|\boldsymbol{l}\|_2^{2\mu}} \left(\min_{\alpha=1,\dots,d} \frac{N_\alpha}{2Y_\alpha}\right)^{2(\lambda-\mu)} \|\boldsymbol{v}\|^2_{H^\mu_{\mathrm{per}}}(\mathcal{Y})$$

and together with Eq. (3.6), estimate in Eq. (3.7) follows. □

3.3 Galerkin approximation with numerical integration

This section presents Galerkin approximation (GA) and Galerkin approximation with numerical integration (GAwNI) that leads to a linear system equivalent with Moulinec-Suquet algorithm, cf. [21].

Definition 3.19 (Finite dimensional space)**.** *We define the finite dimensional space used for discretization $\mathscr{E}_{\boldsymbol{N}} = \mathscr{T}^d_{\boldsymbol{N}} \bigcap \mathscr{E}$, the space of curl-free trigonometric polynomials with zero mean.*

There is no doubt that the space $\mathscr{E}_{\boldsymbol{N}}$ is empty. Since we have projection $\mathcal{G}$ from Def. 2.25, the Helmholtz decomposition for the space of trigonometric polynomials can be written down

$$\mathscr{T}^d_{\boldsymbol{N}} = \mathscr{U} \oplus^\perp \mathscr{E}_{\boldsymbol{N}} \oplus^\perp \mathscr{J}_{\boldsymbol{N}}. \tag{3.8}$$

Definition 3.20 (Galerkin approximation)**.** *Let $\mathscr{E}_{\boldsymbol{N}}$ be a space from Def. 3.19. The function $\tilde{\boldsymbol{e}}_{\boldsymbol{N}} \in \mathscr{E}_{\boldsymbol{N}}$ is called Galerkin approximation if it satisfies*

$$B[\tilde{\boldsymbol{e}}_{\boldsymbol{N}}, \boldsymbol{v}_{\boldsymbol{N}}] = f[\boldsymbol{v}_{\boldsymbol{N}}], \quad \forall \boldsymbol{v}_{\boldsymbol{N}} \in \mathscr{E}_{\boldsymbol{N}}$$

Remark 3.21 (Existence of the unique solution)**.** *The same arguments (positive definiteness and boundedness) with Lax-Milgram lemma as in Remark 2.21 provides the existence of the unique solution.*

Lemma 3.22 (Convergence of Galerkin approximation)**.** *The solution of Galerkin approximation stated in Def. 3.20 converges to the solution of weak formulation Def. 2.20.*

Proof. Using Cea's lemma with constants c_A and C_A from 2.20, we can write

$$\begin{aligned}\|\boldsymbol{e}-\boldsymbol{e}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}} &\leq \frac{C_A}{c_A} \inf_{\boldsymbol{v}_{\boldsymbol{N}}\in\mathscr{E}_{\boldsymbol{N}}} \|\boldsymbol{e}-\boldsymbol{v}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}} \\ &\leq \frac{C_A}{c_A} \|\boldsymbol{e}-P_{\boldsymbol{N}}[\boldsymbol{e}]\|_{L^2_{\mathrm{per}}}.\end{aligned}$$

The limit passage with Lemma 3.12 provides the converge result. □

Remark 3.23. *The better result than in previous lemma can be obtained with Lemma 3.13 for higher regularity of conductivity coefficients and consequently the higher regularity of the solution.*

Definition 3.24 (Galerkin approximation with numerical integration)**.** *Let $\boldsymbol{A} \in L^2_{\mathrm{per}}(\mathcal{Y}; \mathbb{R}^{d\times d}_{\mathrm{spd}}) \bigcap C_{\mathrm{per}}$ be material coefficients that are additionally continuous. Then the function $\tilde{\boldsymbol{e}}_{\boldsymbol{N}} \in \mathscr{E}_{\boldsymbol{N}}$ is the solution of GAwNI if it satisfies*

$$B_{\boldsymbol{N}}[\tilde{\boldsymbol{e}}_{\boldsymbol{N}}, \boldsymbol{v}_{\boldsymbol{N}}] = f_{\boldsymbol{N}}[\boldsymbol{v}_{\boldsymbol{N}}], \quad \forall \boldsymbol{v}_{\boldsymbol{N}} \in \mathscr{E}_{\boldsymbol{N}} \tag{3.9}$$

where

$$B_{\boldsymbol{N}}[\boldsymbol{e}_{\boldsymbol{N}}, \boldsymbol{v}_{\boldsymbol{N}}] := \big(Q_{\boldsymbol{N}}[\boldsymbol{A}\boldsymbol{e}_{\boldsymbol{N}}], \boldsymbol{v}_{\boldsymbol{N}}\big)_{L^2_{\mathrm{per}}(\mathcal{Y},\mathbb{R}^d)} \tag{3.10a}$$

$$f_{\boldsymbol{N}}[\boldsymbol{v}_{\boldsymbol{N}}] := \big(Q_{\boldsymbol{N}}[\boldsymbol{A}\boldsymbol{E}], \boldsymbol{v}_{\boldsymbol{N}}\big)_{L^2_{\mathrm{per}}(\mathcal{Y},\mathbb{R}^d)}. \tag{3.10b}$$

Remark 3.25. *The regularity requirement of conductivity coefficients $\boldsymbol{A}$ in previous lemma satisfies that the function $\boldsymbol{A}\boldsymbol{e}_{\boldsymbol{N}}$ is continuous. Hence, the interpolation operator $Q_{\boldsymbol{N}}[\cdot]$ is well defined.*

Next, we define the fully discrete space, the space of function values at regular grid.

Definition 3.26 (Curl-free vectors with zero mean)**.** *We define the space of curl-free vectors with zero mean as* $\mathbb{E}_{\boldsymbol{N}} := \mathcal{I}_{\boldsymbol{N}}[\mathscr{E}_{\boldsymbol{N}}] \subset \mathbb{R}^{d\times \boldsymbol{N}}$ *where operator* $\mathcal{I}_{\boldsymbol{N}}$ *is defined in Def. 3.8.*

With the help of isomorphism $\mathcal{I}_{\boldsymbol{N}}$, space $\mathbb{R}^{d\times \boldsymbol{N}}$ can be split with Helmholtz decomposition into $\mathbb{R}^{d\times \boldsymbol{N}} = \mathbb{U}_{\boldsymbol{N}} \oplus^{\perp} \mathbb{E}_{\boldsymbol{N}} \oplus^{\perp} \mathbb{J}_{\boldsymbol{N}}$, cf. (3.8). The following definition provides the fully discrete formulation of GAwNI and the next lemma shows its correctness.

Definition 3.27 (Fully discrete representation of Galerkin approximation with numerical integration)**.** *Let* $\boldsymbol{A} \in C_{\mathrm{per}}$ *be the conductivity coefficients and*

$$\mathbf{A}_{\boldsymbol{N}} = \left[\delta_{\boldsymbol{km}} A_{\alpha\beta}(\boldsymbol{x}^{\boldsymbol{k}}_{\boldsymbol{N}})\right]^{\boldsymbol{km}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}}_{\alpha,\beta=1,\dots,d} \in \mathbb{R}^{d\times d\times \boldsymbol{N}\times \boldsymbol{N}} \tag{3.11}$$

be the matrix composed of its values at regular grid. Analogically let $\boldsymbol{E}$ *be the prescribed macroscopic electric field and*

$$\mathbf{E}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}[\boldsymbol{E}] \in \mathbb{U}_{\boldsymbol{N}} \tag{3.12}$$

its fully discrete counterpart. Than we define the matrix representation of Galerkin approximation with numerical integration as the following problem:
find $\tilde{\mathbf{e}}_{\boldsymbol{N}} \in \mathbb{E}_{\boldsymbol{N}}$ *such that*

$$\left(\mathbf{A}_{\boldsymbol{N}}\tilde{\mathbf{e}}_{\boldsymbol{N}}, \mathbf{v}_{\boldsymbol{N}}\right)_{\mathbb{R}^{d\times \boldsymbol{N}}} = -\left(\mathbf{A}_{\boldsymbol{N}}\mathbf{E}_{\boldsymbol{N}}, \mathbf{v}_{\boldsymbol{N}}\right)_{\mathbb{R}^{d\times \boldsymbol{N}}}, \quad \forall \mathbf{v}_{\boldsymbol{N}} \in \mathscr{E}_{\boldsymbol{N}}.$$

Lemma 3.28 (Fully discrete representation of Galerkin approximation with numerical integration)**.** *Let* $\boldsymbol{A} \in C_{\mathrm{per}}$ *be the conductivity coefficients. Then the solution of Galerkin approximation with numerical integration* $\tilde{\boldsymbol{e}}_{\boldsymbol{N}} \in \mathscr{E}_{\boldsymbol{N}}$ *from Def. 3.24 is equivalent to the solution of fully discrete representation of Galerkin approximation with numerical integration* $\tilde{\mathbf{e}}_{\boldsymbol{N}} \in \mathbb{E}_{\boldsymbol{N}}$ *defined in Def. 3.27 as the solutions are related with isomorphism*

$$\tilde{\mathbf{e}}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}[\tilde{\boldsymbol{e}}_{\boldsymbol{N}}].$$

Proof. We substitute a function $\boldsymbol{v}_{\boldsymbol{N}}$ expressed as the Fourier series

$$\boldsymbol{v}_{\boldsymbol{N}}(\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \boldsymbol{v}_{\boldsymbol{N}}(\boldsymbol{x}^{\boldsymbol{k}}_{\boldsymbol{N}})\varphi_{\boldsymbol{N},\boldsymbol{k}}(\boldsymbol{x})$$

into the bilinear form to obtain

$$B_{\boldsymbol{N}}[\boldsymbol{e}_{\boldsymbol{N}}, \boldsymbol{v}_{\boldsymbol{N}}] = \sum_{\boldsymbol{k}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \boldsymbol{v}_{\boldsymbol{N}}(\boldsymbol{x}^{\boldsymbol{k}}_{\boldsymbol{N}}) \cdot \left(Q_{\boldsymbol{N}}[\boldsymbol{A}\boldsymbol{e}_{\boldsymbol{N}}], \varphi_{\boldsymbol{N},\boldsymbol{k}}\right)_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)} \tag{3.13a}$$

$$= \sum_{\boldsymbol{k}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \mathbf{v}^{\boldsymbol{k}} \cdot \left(Q_{\boldsymbol{N}}[\boldsymbol{A}\boldsymbol{e}_{\boldsymbol{N}}], \varphi_{\boldsymbol{N},\boldsymbol{k}}\right)_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)} \tag{3.13b}$$

where with a term $\left(Q_{\boldsymbol{N}}[\boldsymbol{A}\boldsymbol{e}_{\boldsymbol{N}}], \varphi_{\boldsymbol{N},\boldsymbol{k}}\right)_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)}$ we understand $\left(\left(Q_{\boldsymbol{N}}[\boldsymbol{A}\boldsymbol{e}_{\boldsymbol{N}}]_\alpha, \varphi_{\boldsymbol{N},\boldsymbol{k}}\right)_{L^2_{\mathrm{per}}(\mathcal{Y})}\right)^d_{\alpha=1}$, operator '$\cdot$' thus denotes the scalar product on $\mathbb{R}^d$, and vector $\mathbf{v}^{\boldsymbol{k}} \in \mathbb{R}^d$ for $\boldsymbol{k} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}$ is a subvector of vector $\mathbf{v}_{\boldsymbol{N}} := \mathcal{I}_{\boldsymbol{N}}[\boldsymbol{v}_{\boldsymbol{N}}] \in \mathbb{R}^{d\times \boldsymbol{N}}$.

Next, we analogically express the term with interpolation projection $Q_{\boldsymbol{N}}[\boldsymbol{A}\boldsymbol{e}_{\boldsymbol{N}}]$ through its nodal points

$$Q_{\boldsymbol{N}}[\boldsymbol{A}\boldsymbol{e}_{\boldsymbol{N}}](\boldsymbol{x}) = \sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}} \boldsymbol{A}(\boldsymbol{x}^{\boldsymbol{m}}_{\boldsymbol{N}})\boldsymbol{e}_{\boldsymbol{N}}(\boldsymbol{x}^{\boldsymbol{m}}_{\boldsymbol{N}})\varphi_{\boldsymbol{N},\boldsymbol{m}}(\boldsymbol{x}). \tag{3.14}$$

It can be observed that nodal values $\boldsymbol{A}(\boldsymbol{x}^{\boldsymbol{m}}_{\boldsymbol{N}})\boldsymbol{e}_{\boldsymbol{N}}(\boldsymbol{x}^{\boldsymbol{m}}_{\boldsymbol{N}})$ for $\boldsymbol{m} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}$ can be expressed with matrix

$$\mathbf{A}_{\boldsymbol{N}} = \left[\delta_{\boldsymbol{km}} A_{\alpha\beta}(\boldsymbol{x}^{\boldsymbol{k}}_{\boldsymbol{N}})\right]^{\boldsymbol{km}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}}_{\alpha,\beta=1,\dots,d} \in \mathbb{R}^{d\times d\times \boldsymbol{N}\times \boldsymbol{N}}$$

and vector $\mathbf{e}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}[e_{\boldsymbol{N}}]$ as

$$\boldsymbol{A}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{m}})\boldsymbol{e}_{\boldsymbol{N}}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{m}}) = (\mathbf{A}_{\boldsymbol{N}}\mathbf{e}_{\boldsymbol{N}})^{\boldsymbol{m}} \in \mathbb{R}^d. \tag{3.15}$$

Substitution of Eq. (3.14) with (3.15) into bilinear form (3.13), we deduce

$$B_{\boldsymbol{N}}[\boldsymbol{e}_{\boldsymbol{N}}, \boldsymbol{v}_{\boldsymbol{N}}] = \sum_{\boldsymbol{k}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} \sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} (\mathbf{A}_{\boldsymbol{N}}\mathbf{e}_{\boldsymbol{N}})^{\boldsymbol{m}} \cdot \mathbf{v}_{\boldsymbol{N}}^{\boldsymbol{k}} \big(\varphi_{\boldsymbol{N},\boldsymbol{m}}, \varphi_{\boldsymbol{N},\boldsymbol{k}}\big)_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)} \tag{3.16a}$$

$$= \sum_{\boldsymbol{k}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} \sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} (\mathbf{A}_{\boldsymbol{N}}\mathbf{e}_{\boldsymbol{N}})^{\boldsymbol{m}} \cdot \mathbf{v}_{\boldsymbol{N}}^{\boldsymbol{k}} \frac{\delta_{\boldsymbol{m}\boldsymbol{k}}}{|\boldsymbol{N}|_\Pi} \tag{3.16b}$$

$$= \sum_{\boldsymbol{k}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} (\mathbf{A}_{\boldsymbol{N}}\mathbf{e}_{\boldsymbol{N}})^{\boldsymbol{k}} \cdot \mathbf{v}_{\boldsymbol{N}}^{\boldsymbol{k}} \frac{1}{|\boldsymbol{N}|_\Pi} = \frac{\big(\mathbf{A}_{\boldsymbol{N}}\mathbf{e}_{\boldsymbol{N}}, \mathbf{v}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}}}{|\boldsymbol{N}|_\Pi}. \tag{3.16c}$$

where we have used the orthogonality property of functions $\varphi_{\boldsymbol{N},\boldsymbol{k}}$ stated in Remark 3.7.
The formula for a linear functional

$$f_{\boldsymbol{N}}[\boldsymbol{v}_{\boldsymbol{N}}] = \frac{\big(\mathbf{A}_{\boldsymbol{N}}\mathbf{E}_{\boldsymbol{N}}, \mathbf{v}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}}}{|\boldsymbol{N}|_\Pi}. \tag{3.17}$$

with $\mathbf{E}_{\boldsymbol{N}} := \mathcal{I}_{\boldsymbol{N}}[\boldsymbol{E}]$ can be deduced accordingly. Using Eq. (3.16) and (3.17), the bilinear form

$$B_{\boldsymbol{N}}[\boldsymbol{e}_{\boldsymbol{N}}, \boldsymbol{v}_{\boldsymbol{N}}] = f_{\boldsymbol{N}}[\boldsymbol{v}_{\boldsymbol{N}}], \quad \forall \boldsymbol{v}_{\boldsymbol{N}} \in \mathscr{E}_{\boldsymbol{N}}$$

can be transformed to

$$\frac{\big(\mathbf{A}_{\boldsymbol{N}}\mathbf{e}_{\boldsymbol{N}}, \mathbf{v}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}}}{|\boldsymbol{N}|_\Pi} = \frac{\big(\mathbf{A}_{\boldsymbol{N}}\mathbf{E}_{\boldsymbol{N}}, \mathbf{v}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}}}{|\boldsymbol{N}|_\Pi}, \quad \forall \mathbf{v}_{\boldsymbol{N}} \in \mathbb{E}_{\boldsymbol{N}}$$
$$\big(\mathbf{A}_{\boldsymbol{N}}\mathbf{e}_{\boldsymbol{N}}, \mathbf{v}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}} = \big(\mathbf{A}_{\boldsymbol{N}}\mathbf{E}_{\boldsymbol{N}}, \mathbf{v}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}}, \quad \forall \mathbf{v}_{\boldsymbol{N}} \in \mathbb{E}_{\boldsymbol{N}}$$

where $\mathbf{e}_{\boldsymbol{N}}$ and $\boldsymbol{e}_{\boldsymbol{N}}$ are related with isomorphism $\mathbf{e}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}[\boldsymbol{e}_{\boldsymbol{N}}]$. □

Lemma 3.29 (Existence of the unique solution of GAwNI)**.** *The approximated bilinear form* (3.10a) *is uniformly elliptic, i.e.*

$$\frac{1}{c_A}\|\boldsymbol{v}_{\boldsymbol{N}}\|^2_{L^2_{\mathrm{per}}} \le B_{\boldsymbol{N}}[\boldsymbol{v}_{\boldsymbol{N}}, \boldsymbol{v}_{\boldsymbol{N}}], \tag{3.18}$$

and bounded

$$B_{\boldsymbol{N}}[\boldsymbol{e}_{\boldsymbol{N}}, \boldsymbol{v}_{\boldsymbol{N}}] \le C_A\|\boldsymbol{e}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}}\|\boldsymbol{v}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}}. \tag{3.19}$$

The linear function is bounded with

$$f_{\boldsymbol{N}}[\boldsymbol{v}_{\boldsymbol{N}}] \le C_A\|\boldsymbol{e}_0\|_{L^2_{\mathrm{per}}}\|\boldsymbol{v}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}} \tag{3.20}$$

Moreover, there exists the unique solution of Galerkin approximation with numerical integration.

Proof. The formulas (3.18), (3.19), and (3.20) follow easily from expressions in Remark 3.28.

$$\frac{1}{c_A}\|\boldsymbol{v}_{\boldsymbol{N}}\|^2_{L^2_{\mathrm{per}}} = \frac{1}{c_A}\|\mathbf{v}\|_2^2 \le \big(\mathbf{A}\mathbf{v}, \mathbf{v}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}} = B_{\boldsymbol{N}}[\boldsymbol{v}_{\boldsymbol{N}}, \boldsymbol{v}_{\boldsymbol{N}}]$$
$$B_{\boldsymbol{N}}[\boldsymbol{e}_{\boldsymbol{N}}, \boldsymbol{v}_{\boldsymbol{N}}] = \big(\mathbf{A}\mathbf{e}, \mathbf{v}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}} \le \|\mathbf{A}\|_\infty\|\mathbf{e}\|_2\|\mathbf{v}\|_2 = C_A\|\boldsymbol{e}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}}\|\boldsymbol{v}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}}$$
$$f_{\boldsymbol{N}}[\boldsymbol{v}_{\boldsymbol{N}}] = \big(\mathbf{A}\mathbf{E}, \mathbf{v}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}} \le \|\mathbf{A}\|_\infty\|\mathbf{E}\|_2\|\mathbf{v}\|_2 = C_A\|\boldsymbol{E}\|_{L^2_{\mathrm{per}}}\|\boldsymbol{v}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}}$$

Now, the existence and uniqueness of the solution is provided using Lax-Milgram theorem. □

3.4 Convergence to continuous solution

This section provides convergence of discrete solutions obtained with GAwNI to the solution of weak formulation. The rate of converge is primarily based on the regularity of the solution; for completeness, the basic regularity results are also provided.

We start with a definition of difference quotients and Lemma 3.31 that characterizes the Sobolev space H^1_{per}.

Definition 3.30 (difference quotient)**.** *For a function $\boldsymbol{u} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$, the i-th difference quotient of size h is $D^h_i\boldsymbol{u}(\boldsymbol{x}) \in \mathbb{R}^d$ defined as*

$$\bar{D}^h_i\boldsymbol{u}(\boldsymbol{x}) := \frac{\boldsymbol{u}(\boldsymbol{x}+h\boldsymbol{\epsilon}_i)-\boldsymbol{u}(\boldsymbol{x})}{h}, \quad i=1,\dots,d$$

for $\boldsymbol{x} \in \mathbb{R}^d$, $h \in \mathbb{R}$ and $\boldsymbol{\epsilon}_\alpha = (\delta_{\alpha\beta})^d_{\beta=1}$ being a vector of canonical basis of space $\mathbb{R}^d$. Moreover we note

$$\bar{D}^h\boldsymbol{u}(\boldsymbol{x}) := \big(\bar{D}^h_\alpha\boldsymbol{u}(\boldsymbol{x})\big)^d_{\alpha=1}.$$

Lemma 3.31 (Difference quotient and Sobolev space)**.** *The following two statements hold:*

- *Assume $1<p<\infty$, $\boldsymbol{u} \in L^p_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$, and there exists constant C independent of h such that*

  $$\|\bar{D}^h\boldsymbol{u}\|_{L^p_{\mathrm{per}}} \le C.$$

  Then

  $$\boldsymbol{u} \in W^{1,p}_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$$

  with $\|D^1\boldsymbol{u}\|_{L^p_{\mathrm{per}}} \le C$.

- *Assume $1<p<\infty$ and $\boldsymbol{u} \in W^{1,p}_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$. Then*

  $$\|\bar{D}^h\boldsymbol{u}\|_{L^p_{\mathrm{per}}} \le \|D^1\boldsymbol{u}\|_{L^p_{\mathrm{per}}}$$

  for $h \in \mathbb{R}$.

Proof. For proof see Theorem 3 in Section 5.8.2 in [10]. □

Lemma 3.32 (Regularity result)**.** *Let $\boldsymbol{A} \in W^{1,\infty}_{\mathrm{per}}(\mathcal{Y},\mathbb{R}^{d\times d})$ and $\boldsymbol{e} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ be a solution of weak formulation, c.f. Def. 2.20. Then $\boldsymbol{e} \in H^1_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ and it satisfies*

$$\begin{aligned}
\|\boldsymbol{e}\|_{H^1_{\mathrm{per}}} &\le \frac{1}{c_A}\left(\|\boldsymbol{A}\boldsymbol{E}\|_{H^1_{\mathrm{per}}} + \|\boldsymbol{A}\|_{W^{1,\infty}_{\mathrm{per}}}\|\boldsymbol{e}\|_{L^2_{\mathrm{per}}}\right)\\
&\le \frac{1}{c_A}\|\boldsymbol{A}\|_{W^{1,\infty}_{\mathrm{per}}}\left(\|\boldsymbol{E}\|_{L^2_{\mathrm{per}}} + \|\boldsymbol{e}\|_{L^2_{\mathrm{per}}}\right)\\
&\le \frac{1}{c_A}\|\boldsymbol{A}\|_{W^{1,\infty}_{\mathrm{per}}}\|\boldsymbol{E}\|_{L^2_{\mathrm{per}}}\left(1+\frac{C_A}{c_A}\right).
\end{aligned}$$

Proof. The proof based on Lem. 3.31 can be found for a more general case in Section 6.3.1 in [10] as Theorem 1. In order to work with difference quotients, we use simple formulas

$$\big(\boldsymbol{v},\bar{D}^{-h}_k\boldsymbol{w}\big)_{L^2_{\mathrm{per}}} = -\big(\bar{D}^h_k\boldsymbol{v},\boldsymbol{w}\big)_{L^2_{\mathrm{per}}} \tag{3.21}$$

$$\bar{D}^h_k(\boldsymbol{v}\boldsymbol{w}) = \boldsymbol{v}^h\bar{D}^h_k\boldsymbol{w} + \boldsymbol{w}\bar{D}^h_k\boldsymbol{v} \tag{3.22}$$

where $\boldsymbol{v}^h(\boldsymbol{x}) = \boldsymbol{v}(\boldsymbol{x}+h\boldsymbol{\epsilon}_k)$. Then we start with weak formulation, c.f. Def. 2.20, i.e.

$$\big(\boldsymbol{A}\boldsymbol{e},\boldsymbol{v}\big)_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)} = -\big(\boldsymbol{f},\boldsymbol{v}\big)_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)}$$

where $\boldsymbol{f} := \boldsymbol{A}\boldsymbol{E} \in H^1_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$, with a special choice of test function $\boldsymbol{v} = -\bar{D}_k^{-h}\big(\bar{D}_k^h \boldsymbol{e}\big)$ leading to

$$\begin{aligned}
-\big(\boldsymbol{A}\boldsymbol{e}, -\bar{D}_k^{-h}\big(\bar{D}_k^h\boldsymbol{e}\big)\big)_{L^2_{\mathrm{per}}} &= \big(\boldsymbol{f}, \bar{D}_k^{-h}(\bar{D}_k^h\boldsymbol{e})\big)_{L^2_{\mathrm{per}}}\\
\big(\bar{D}_k^h\boldsymbol{A}\boldsymbol{e}, \bar{D}_k^h\boldsymbol{e}\big)_{L^2_{\mathrm{per}}} &= -\big(\bar{D}_k^h\boldsymbol{f}, \bar{D}_k^h\boldsymbol{e}\big)_{L^2_{\mathrm{per}}}\\
\big(\boldsymbol{A}^h(\boldsymbol{x})\bar{D}_k^h\boldsymbol{e}, \bar{D}_k^h\boldsymbol{e}\big)_{L^2_{\mathrm{per}}} &= -\big(\bar{D}_k^h\boldsymbol{f}, \bar{D}_k^h\boldsymbol{e}\big)_{L^2_{\mathrm{per}}} - \big(\bar{D}_k^h\boldsymbol{A}\boldsymbol{e}, \bar{D}_k^h\boldsymbol{e}\big)_{L^2_{\mathrm{per}}}.
\end{aligned}$$

The positive definiteness and boundedness of $\boldsymbol{A}$ implies

$$\begin{aligned}
c_A\|\bar{D}_k^h\boldsymbol{e}\|^2_{L^2_{\mathrm{per}}} &\le \|\bar{D}_k^h\boldsymbol{f}\|_{L^2_{\mathrm{per}}}\|\bar{D}_k^h\boldsymbol{e}\|_{L^2_{\mathrm{per}}} + \|\bar{D}_k^h\boldsymbol{A}\|_{L^\infty_{\mathrm{per}}}\|\boldsymbol{e}\|_{L^2_{\mathrm{per}}}\|\bar{D}_k^h\boldsymbol{e}\|_{L^2_{\mathrm{per}}}\\
\|\bar{D}_k^h\boldsymbol{e}\|_{L^2_{\mathrm{per}}} &\le \frac{1}{c_A}\|\bar{D}_k^h\boldsymbol{f}\|_{L^2_{\mathrm{per}}} + \|\bar{D}_k^h\boldsymbol{A}\|_{L^\infty_{\mathrm{per}}}\|\boldsymbol{e}\|_{L^2_{\mathrm{per}}}\\
&\le \frac{1}{c_A}\|\boldsymbol{f}\|_{H^1_{\mathrm{per}}} + \|\boldsymbol{A}\|_{W^{1,\infty}_{\mathrm{per}}}\|\boldsymbol{e}\|_{L^2_{\mathrm{per}}}\\
&\le \frac{1}{c_A}\|\boldsymbol{A}\|_{W^{1,\infty}_{\mathrm{per}}}\left(\|\boldsymbol{E}\|_{L^2_{\mathrm{per}}} + \|\boldsymbol{e}\|_{L^2_{\mathrm{per}}}\right).
\end{aligned}$$

The solution of weak formulation $\boldsymbol{e}$ is bounded in L^2_{per}, i.e.

$$c_A\|\boldsymbol{e}\|^2_{L^2_{\mathrm{per}}} < B[\boldsymbol{e},\boldsymbol{e}] = f[\boldsymbol{e}] < C_A\|\boldsymbol{E}\|_{L^2_{\mathrm{per}}}\|\boldsymbol{e}\|_{L^2_{\mathrm{per}}},$$

the difference quotient $\|\bar{D}_k^h L\|_{L^\infty_{\mathrm{per}}}$ is bounded due to Lem. 3.31 and thus the required Eq. (3.21) follows. □

Lemma 3.33 (Higher regularity result)**.** *Let $\boldsymbol{A} \in W^{m,\infty}_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^{d\times d})$ for $m \in \mathbb{N}$ and $\boldsymbol{e} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ be a solution of weak formulation, Def. 2.20. Then*

$$\boldsymbol{e} \in H^m_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) \tag{3.23}$$

and it satisfies

$$\|\boldsymbol{e}\|_{H^m_{\mathrm{per}}} \le C(\boldsymbol{E})\|\boldsymbol{A}\|_{W^{m,\infty}_{\mathrm{per}}}\|\boldsymbol{E}\|_{L^2_{\mathrm{per}}} \tag{3.24}$$

where $C(\boldsymbol{E})$ depends only on $\boldsymbol{E}$.

Proof. The proof based on Lem. 3.31 can be found for a more general case in Section 6.3.1 in [10] as Theorem 2.
We will establish (3.23) and estimate (3.24) by induction hypothesis.
Base case: The case $m = 1$ is proven in Lem. 3.32.
Induction hypothesis: Assume now (3.23) and estimate (3.24) be valid for $\boldsymbol{A} \in W^{m,\infty}_{\mathrm{per}}$.
Induction step: Suppose additionally $\boldsymbol{A} \in W^{(m+1),\infty}_{\mathrm{per}}$ to prove $\boldsymbol{e} \in H^{m+1}_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ with estimate

$$\|\boldsymbol{e}\|_{H^{m+1}_{\mathrm{per}}} < C\|\boldsymbol{A}\|_{W^{m+1,\infty}_{\mathrm{per}}}\|\boldsymbol{E}\|_{L^2_{\mathrm{per}}}.$$

Let $\boldsymbol{l}$ be any multiindex with $\|\boldsymbol{l}\|_1 = m$ and take any test function $\tilde{\boldsymbol{v}} \in \mathcal{D}_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ and put $\boldsymbol{v} := (-1)^{\|\boldsymbol{l}\|_1}D^{\boldsymbol{l}}\tilde{\boldsymbol{v}}$ into the formula of weak formulation

$$\begin{aligned}
B[\boldsymbol{e},\boldsymbol{v}] &= f[\boldsymbol{v}]\\
\big(\boldsymbol{A}\boldsymbol{e},\boldsymbol{v}\big)_{L^2_{\mathrm{per}}} &= \big(\boldsymbol{A}\boldsymbol{E},\boldsymbol{v}\big)_{L^2_{\mathrm{per}}}
\end{aligned}$$

Integration by parts leads to

$$B[\tilde{\boldsymbol{e}},\tilde{\boldsymbol{v}}] = \tilde{f}[\tilde{\boldsymbol{v}}] \tag{3.25}$$

where

$$\begin{aligned}\tilde{\boldsymbol{e}}(\boldsymbol{x}) &= D^{\boldsymbol{l}}\boldsymbol{e}(\boldsymbol{x}) \in L^2_{\mathrm{per}}\\ \tilde{f}[\tilde{\boldsymbol{v}}] &= \left(\tilde{f}, \tilde{\boldsymbol{v}}\right)_{L^2_{\mathrm{per}}}\\ \tilde{f}(\boldsymbol{x}) &= D^{\boldsymbol{l}}\boldsymbol{A}(\boldsymbol{x})\boldsymbol{E} - \sum_{\boldsymbol{k}:\boldsymbol{k}\le\boldsymbol{l},\boldsymbol{l}\neq\boldsymbol{k}} \binom{\boldsymbol{l}}{\boldsymbol{k}} D^{\boldsymbol{l}-\boldsymbol{k}}\boldsymbol{A}(\boldsymbol{x}) D^{\boldsymbol{k}}\boldsymbol{e}(\boldsymbol{x}).\end{aligned}$$

and $\boldsymbol{x} \in \mathbb{R}^d$. Since the test function $\tilde{\boldsymbol{v}}$ is arbitrary, $\tilde{\boldsymbol{e}}(\boldsymbol{x})$ is a weak solution of

$$B[\tilde{\boldsymbol{e}}, \tilde{\boldsymbol{v}}] = \tilde{f}[\tilde{\boldsymbol{v}}], \quad \forall \tilde{\boldsymbol{v}} \in L^2_{\mathrm{per}}(\mathcal{Y}; \mathbb{R}^d),$$

Next, we can observe that $\tilde{\boldsymbol{f}} \in H^1_{\mathrm{per}}$ as

$$\begin{aligned}\|\tilde{\boldsymbol{f}}\|_{H^1_{\mathrm{per}}} &\le \|D^{\boldsymbol{l}}\boldsymbol{A}\|_{W^{1,\infty}_{\mathrm{per}}} \|\boldsymbol{E}\|_{L^2_{\mathrm{per}}} + \sum_{\boldsymbol{k}:\boldsymbol{k}\le\boldsymbol{l},\boldsymbol{l}\neq\boldsymbol{k}} \binom{\boldsymbol{l}}{\boldsymbol{k}} \|D^{\boldsymbol{l}-\boldsymbol{k}}\boldsymbol{A}\|_{W^{1,\infty}_{\mathrm{per}}} \|D^{\boldsymbol{k}}\boldsymbol{e}\|_{H^1_{\mathrm{per}}}\\ &\le \|\boldsymbol{A}\|_{W^{m+1,\infty}_{\mathrm{per}}} \|\boldsymbol{E}\|_{L^2_{\mathrm{per}}} + C\|\boldsymbol{A}\|_{W^{m+1,\infty}_{\mathrm{per}}} \|\boldsymbol{e}\|_{H^m_{\mathrm{per}}}\\ &\le C\|\boldsymbol{A}\|_{W^{m+1,\infty}_{\mathrm{per}}} \left(\|\boldsymbol{E}\|_{L^2_{\mathrm{per}}} + \|\boldsymbol{e}\|_{H^m_{\mathrm{per}}}\right)\\ &\le C\|\boldsymbol{A}(\boldsymbol{x})\|_{W^{m+1,\infty}_{\mathrm{per}}} \|\boldsymbol{E}\|_{L^2_{\mathrm{per}}}\end{aligned}$$

Using Lem. 3.32 on problem stated in (3.25), we estimate

$$\begin{aligned}\|\tilde{\boldsymbol{e}}\|_{H^1_{\mathrm{per}}} &\le C\left(\|\tilde{\boldsymbol{f}}\|_{H^1_{\mathrm{per}}} + \|\boldsymbol{A}\|_{W^{1,\infty}_{\mathrm{per}}} \|\tilde{\boldsymbol{e}}\|_{L^2_{\mathrm{per}}}\right)\\ &\le C\left(\|\boldsymbol{A}\|_{W^{m+1,\infty}_{\mathrm{per}}} \|\boldsymbol{E}\|_{L^2_{\mathrm{per}}} + \|\boldsymbol{A}\|_{W^{1,\infty}_{\mathrm{per}}} \|\boldsymbol{E}\|_{L^2_{\mathrm{per}}}\right)\\ &\le C\|\boldsymbol{A}\|_{W^{m+1,\infty}_{\mathrm{per}}} \|\boldsymbol{E}\|_{L^2_{\mathrm{per}}}.\end{aligned}$$

This inequality holds for each multiindex $\boldsymbol{l}$ such that $\|\boldsymbol{l}\|_1 = m$ and $\tilde{\boldsymbol{e}}(\boldsymbol{x}) = D^{\boldsymbol{l}}\boldsymbol{e}(\boldsymbol{x})$, thus we conclude $\boldsymbol{e} \in H^{m+1}_{\mathrm{per}}$ with

$$\|\boldsymbol{e}\|_{H^{m+1}_{\mathrm{per}}} \le C\|\boldsymbol{A}\|_{W^{m+1,\infty}_{\mathrm{per}}} \|\boldsymbol{E}\|_{L^2_{\mathrm{per}}}.$$

□

Lemma 3.34 (Strang)**.** *Consider a family of discrete problems whose associated approximate bilinear forms are uniformly elliptic. Then there exists a constant C independent of the space $\mathcal{T}_{\boldsymbol{N}}$ such that*

$$\begin{aligned}\|\boldsymbol{e} - \boldsymbol{e}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}} \le \inf_{\boldsymbol{v}_{\boldsymbol{N}} \in \mathcal{E}_{\boldsymbol{N}}} \left\{ \left(\frac{C_A}{c_A} + 1\right) \|\boldsymbol{e} - \boldsymbol{v}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}} + \frac{1}{c_A} \sup_{\boldsymbol{w}_{\boldsymbol{N}} \in \mathcal{E}_{\boldsymbol{N}}} \frac{|B[\boldsymbol{v}_{\boldsymbol{N}}, \boldsymbol{w}_{\boldsymbol{N}}] - B_{\boldsymbol{N}}[\boldsymbol{v}_{\boldsymbol{N}}, \boldsymbol{w}_{\boldsymbol{N}}]|}{\|\boldsymbol{w}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}}} \right\}\\ + \frac{1}{c_A} \sup_{\boldsymbol{w}_{\boldsymbol{N}} \in \mathcal{E}_{\boldsymbol{N}}} \frac{|f[\boldsymbol{w}_{\boldsymbol{N}}] - f_{\boldsymbol{N}}[\boldsymbol{w}_{\boldsymbol{N}}]|}{\|\boldsymbol{w}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}}}\end{aligned}$$

where the constants are from Def. 2.20.

Proof. The proof can be found as Theorem 26.1 in [5]. □

Lemma 3.35 (Convergence of discrete solutions to continuous one)**.** *Let the conductivity coefficients be $\boldsymbol{A} \in W^{\mu,\infty}_{\mathrm{per}}(\mathcal{Y}; \mathbb{R}^{d\times d})$ with $\mu \in \mathbb{N}$ such that $W^{\mu,\infty}_{\mathrm{per}}$ is embedded into continuous functions. Then the sequence of solutions of Galerkin approximation with numerical integration, c.f. Def. 3.24, converge to the solution of weak formulation, Def. 2.20, in $\|\cdot\|_{L^2_{\mathrm{per}}}$ norm, i.e.*

$$\|\boldsymbol{e} - \boldsymbol{e}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}} \le C\left(\min_{\alpha} \frac{N_\alpha}{2Y_\alpha}\right)^{-\mu} \to 0 \quad \textit{for } \min_{\alpha} \frac{N_\alpha}{2Y_\alpha} \to \infty$$

where

$$C = \left(\frac{C_A}{c_A} + 1\right) \|\boldsymbol{e}\|_{H^\mu_{\mathrm{per}}} + \frac{c_{0,\mu}}{c_A}\|\boldsymbol{A}\|_{W^{\mu,\infty}_{\mathrm{per}}}\|\boldsymbol{e}\|_{H^\mu_{\mathrm{per}}} + \frac{c_{0,\mu}}{c_A}\|\boldsymbol{A}\|_{W^{\mu,\infty}_{\mathrm{per}}}\|\boldsymbol{E}\|_{L^2_{\mathrm{per}}}.$$

Proof. We use Lem. 3.34 due to Strang, results about regularization Lem. 3.32 and 3.33, and approximation by orthogonal projection Lem. 3.13 and by interpolation projection Lem. 3.18.

$$\begin{aligned}
\|\boldsymbol{e} - \boldsymbol{e}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}} &\le \inf_{\boldsymbol{v}_{\boldsymbol{N}} \in \mathscr{E}_{\boldsymbol{N}}} \left\{ \left(\frac{C_A}{c_A} + 1\right) \|\boldsymbol{e} - \boldsymbol{v}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}} + \frac{1}{c_A} \sup_{\boldsymbol{w}_{\boldsymbol{N}} \in \mathscr{E}_{\boldsymbol{N}}} \frac{|\left(\boldsymbol{A}\boldsymbol{v}_{\boldsymbol{N}}, \boldsymbol{w}_{\boldsymbol{N}}\right)_{L^2_{\mathrm{per}}} - \left(Q_{\boldsymbol{N}}[\boldsymbol{A}\boldsymbol{v}_{\boldsymbol{N}}], \boldsymbol{w}_{\boldsymbol{N}}\right)_{L^2_{\mathrm{per}}}|}{\|\boldsymbol{w}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}}} \right\} \\
&\quad + \frac{1}{c_A} \sup_{\boldsymbol{w}_{\boldsymbol{N}} \in \mathscr{E}_{\boldsymbol{N}}} \frac{|\left(\boldsymbol{A}\boldsymbol{E}, \boldsymbol{w}_{\boldsymbol{N}}\right)_{L^2_{\mathrm{per}}} - \left(Q_{\boldsymbol{N}}[\boldsymbol{A}\boldsymbol{E}], \boldsymbol{w}_{\boldsymbol{N}}\right)_{L^2_{\mathrm{per}}}|}{\|\boldsymbol{w}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}}} \\
&\le \inf_{\boldsymbol{v}_{\boldsymbol{N}} \in \mathscr{E}_{\boldsymbol{N}}} \left\{ \left(\frac{C_A}{c_A} + 1\right) \|\boldsymbol{e} - \boldsymbol{v}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}} + \frac{1}{c_A}\|\boldsymbol{A}\boldsymbol{v}_{\boldsymbol{N}} - Q_{\boldsymbol{N}}[\boldsymbol{A}\boldsymbol{v}_{\boldsymbol{N}}]\|_{L^2_{\mathrm{per}}} \right\} + \frac{1}{c_A}\|\boldsymbol{A}\boldsymbol{E} - Q_{\boldsymbol{N}}[\boldsymbol{A}\boldsymbol{E}]\|_{L^2_{\mathrm{per}}} \\
&\le \left(\frac{C_A}{c_A} + 1\right) \|\boldsymbol{e} - P_{\boldsymbol{N}}\boldsymbol{e}\|_{L^2_{\mathrm{per}}} + \frac{1}{c_A}\|\boldsymbol{A}P_{\boldsymbol{N}}\boldsymbol{e} - Q_{\boldsymbol{N}}[\boldsymbol{A}P_{\boldsymbol{N}}\boldsymbol{e}]\|_{L^2_{\mathrm{per}}} + \frac{1}{c_A}\|\boldsymbol{A} - Q_{\boldsymbol{N}}[\boldsymbol{A}]\|_{L^2_{\mathrm{per}}}\|\boldsymbol{E}\|_{L^\infty_{\mathrm{per}}} \\
&\le \left(\frac{C_A}{c_A} + 1\right) \left(\min_\alpha \frac{N_\alpha}{2Y_\alpha}\right)^{-\mu} \|\boldsymbol{e}\|_{H^\mu_{\mathrm{per}}} + \frac{c_{0,\mu}}{c_A} \left(\min_\alpha \frac{N_\alpha}{2Y_\alpha}\right)^{-\mu} \|\boldsymbol{A}\|_{W^{\mu,\infty}_{\mathrm{per}}}\|\boldsymbol{e}\|_{H^\mu_{\mathrm{per}}} \\
&\quad + \frac{c_{0,\mu}}{c_A} \left(\min_\alpha \frac{N_\alpha}{2Y_\alpha}\right)^{-\mu} \|\boldsymbol{A}\|_{W^{\mu,\infty}_{\mathrm{per}}}\|\boldsymbol{E}\|_{L^2_{\mathrm{per}}} \\
&\le \left(\min_\alpha \frac{N_\alpha}{2Y_\alpha}\right)^{-\mu} \left[\left(\frac{C_A}{c_A} + 1\right) \|\boldsymbol{e}\|_{H^\mu_{\mathrm{per}}} + \frac{c_{0,\mu}}{c_A}\|\boldsymbol{A}\|_{W^{\mu,\infty}_{\mathrm{per}}}\|\boldsymbol{e}\|_{H^\mu_{\mathrm{per}}} + \frac{c_{0,\mu}}{c_A}\|\boldsymbol{A}\|_{W^{\mu,\infty}_{\mathrm{per}}}\|\boldsymbol{E}\|_{L^2_{\mathrm{per}}} \right]
\end{aligned}$$

Now, the limit passage reveals the proof. □

3.5 Regularization of rough material coefficients

This section provide a method for calculation of discrete solutions if the material coefficients $\boldsymbol{A}$ are rough, do not possess sufficient regularity. The method is based on regularization of material coefficient. Although the convergence is provided, the convergence can be arbitrarily slow — we cannot expect any order of convergence.

Lemma 3.36. *Let $\boldsymbol{A} \in L^\infty_{\mathrm{per}}(\mathcal{Y}; \mathbb{R}^{d\times d})$ be a symmetric and uniformly elliptic material coefficients and $\boldsymbol{A}_h$ be their regularization such that*

$$\begin{aligned}
\boldsymbol{A}_h &\in L^\infty_{\mathrm{per}}(\mathcal{Y}; \mathbb{R}^{d\times d}) \\
\|\boldsymbol{A} - \boldsymbol{A}_h\|_{L^2_{\mathrm{per}}(\mathcal{Y}; \mathbb{R}^{d\times d})} &\to 0 \quad \textit{for } h \to 0
\end{aligned}$$

and be positive definite and bounded with constants $c_{A,h}$ and $C_{A,h}$, i.e.

$$c_{A,h} \le \frac{\sum_{\alpha,\beta=1}^d (A_h)_{\alpha\beta}(\boldsymbol{x})\xi_\alpha\xi_\beta}{\|\boldsymbol{\xi}\|^2} \le C_{A,h}, \quad \forall \boldsymbol{\xi} \in \mathbb{R}^d, \quad \textit{and for a.a. } \boldsymbol{x} \in \mathcal{Y}.$$

Suppose $\boldsymbol{e}, \boldsymbol{e}_h \in L^2_{\mathrm{per}}(\mathcal{Y}; \mathbb{R}^d)$ be solutions of weak formulation, c.f. Def. 2.20, for material coefficients $\boldsymbol{A}$ and $\boldsymbol{A}_h$ resp. Then

$$\lim_{h\to 0} \|\boldsymbol{e} - \boldsymbol{e}_h\|_{L^2_{\mathrm{per}}(\mathcal{Y}; \mathbb{R}^d)} = 0 \tag{3.26}$$

Proof. First, define a bilinear form and linear functional for regularized coefficients $\boldsymbol{A}_h$ as

$$B_h[\boldsymbol{e}_h, \boldsymbol{v}] = \left(\boldsymbol{A}_h\boldsymbol{e}_h, \boldsymbol{v}\right)_{L^2_{\mathrm{per}}(\mathcal{Y}; \mathbb{R}^d)} \qquad f_h[\boldsymbol{v}] = \left(\boldsymbol{A}_h\boldsymbol{E}, \boldsymbol{v}\right)_{L^2_{\mathrm{per}}(\mathcal{Y}; \mathbb{R}^d)}$$

and a weak formulation

$$B_h[\boldsymbol{e}_h, \boldsymbol{v}] = f_h[\boldsymbol{v}], \quad \forall \boldsymbol{v} \in \mathscr{E}$$

where $\boldsymbol{e}_h$ denotes its solution.

We show from positive definiteness and boundedness of $\boldsymbol{A}$ that solutions of regularized problems $\boldsymbol{e}_h$ are uniformly bounded

$$\|\boldsymbol{e}_h\|_{L^2}^2 \leq \frac{1}{c_{A,h}}\big(\boldsymbol{A}_h\boldsymbol{e}_h, \boldsymbol{e}_h\big) = -\frac{1}{c_{A,h}}\big(\boldsymbol{A}_h\boldsymbol{E}, \boldsymbol{e}_h\big) \leq \frac{C_{A,h}}{c_{A,h}}\|\boldsymbol{E}\|_{L^2}\|\boldsymbol{e}_h\|_{L^2}.$$

Thus using Banach selection principle for sequence $\{\boldsymbol{e}_h\}$ under a reflexive and separable space such as L^2_{per}, we conclude there exists some subsequence converging weakly to some function $\bar{\boldsymbol{e}}$, i.e.

$$\boldsymbol{e}_h \rightharpoonup \tilde{\boldsymbol{e}} \quad \Leftrightarrow \quad \lim_{h\to 0}\big(\boldsymbol{e}_h, \boldsymbol{v}\big)_{L^2_{\mathrm{per}}} = \big(\bar{\boldsymbol{e}}, \boldsymbol{v}\big)_{L^2_{\mathrm{per}}}, \quad \forall \boldsymbol{v} \in L^2_{\mathrm{per}}.$$

Next, we show that the function $\tilde{\boldsymbol{e}}$ satisfies the weak formulation with coefficients $\boldsymbol{A}$, hence for arbitrary $\boldsymbol{v} \in \mathscr{E}$

$$\big(\boldsymbol{A}_h\boldsymbol{e}_h, \boldsymbol{v}\big)_{L^2_{\mathrm{per}}} = \big(\boldsymbol{A}_h\boldsymbol{E}, \boldsymbol{v}\big)_{L^2_{\mathrm{per}}} \overset{h\to 0}{\longrightarrow} \big(\boldsymbol{A}\boldsymbol{E}, \boldsymbol{v}\big)_{L^2_{\mathrm{per}}} = \big(\boldsymbol{A}\bar{\boldsymbol{e}}, \boldsymbol{v}\big)_{L^2_{\mathrm{per}}}$$

and since $\big(\boldsymbol{A}_h\boldsymbol{e}_h, \boldsymbol{v}\big)_{L^2_{\mathrm{per}}} \overset{h\to 0}{\longrightarrow} \big(\boldsymbol{A}\tilde{\boldsymbol{e}}, \boldsymbol{v}\big)_{L^2_{\mathrm{per}}}$, we can conclude that $\tilde{\boldsymbol{e}} = \boldsymbol{e}$.
Next we show that the subsequence $\boldsymbol{e}_h$ converge even strongly

$$\begin{aligned}
c_{A,h}\|\boldsymbol{e} - \boldsymbol{e}_h\|_{L^2_{\mathrm{per}}}^2 &\leq \big(\boldsymbol{A}_h(\boldsymbol{e} - \boldsymbol{e}_h), \boldsymbol{e} - \boldsymbol{e}_h\big)_{L^2_{\mathrm{per}}} \\
&= \big(\boldsymbol{A}_h\boldsymbol{e}, \boldsymbol{e} - \boldsymbol{e}_h\big)_{L^2_{\mathrm{per}}} - \big(\boldsymbol{A}_h\boldsymbol{e}_h, \boldsymbol{e} - \boldsymbol{e}_h\big)_{L^2_{\mathrm{per}}} \pm \big(\boldsymbol{A}\boldsymbol{e}, \boldsymbol{e} - \boldsymbol{e}_h\big)_{L^2_{\mathrm{per}}} \\
&= \big(\boldsymbol{A}\boldsymbol{e}, \boldsymbol{e} - \boldsymbol{e}_h\big)_{L^2_{\mathrm{per}}} - \big(\boldsymbol{A}_h\boldsymbol{e}_h, \boldsymbol{e} - \boldsymbol{e}_h\big)_{L^2_{\mathrm{per}}} + \big(\boldsymbol{A}_h\boldsymbol{e}, \boldsymbol{e} - \boldsymbol{e}_h\big)_{L^2_{\mathrm{per}}} - \big(\boldsymbol{A}\boldsymbol{e}, \boldsymbol{e} - \boldsymbol{e}_h\big)_{L^2_{\mathrm{per}}} \\
&= -\big((\boldsymbol{A} - \boldsymbol{A}_h)\boldsymbol{E}, \boldsymbol{e} - \boldsymbol{e}_h\big)_{L^2_{\mathrm{per}}} + \big(\boldsymbol{A}_h\boldsymbol{e}, \boldsymbol{e} - \boldsymbol{e}_h\big)_{L^2_{\mathrm{per}}} - \big(\boldsymbol{A}\boldsymbol{e}, \boldsymbol{e} - \boldsymbol{e}_h\big)_{L^2_{\mathrm{per}}} \\
&\leq \|(\boldsymbol{A} - \boldsymbol{A}_h)\|_{L^\infty_{\mathrm{per}}}\big(\boldsymbol{E}, \boldsymbol{e} - \boldsymbol{e}_h\big)_{L^2_{\mathrm{per}}} + \|\boldsymbol{A}_h - \boldsymbol{A}\|_{L^\infty_{\mathrm{per}}}\big(\boldsymbol{e}, \boldsymbol{e} - \boldsymbol{e}_h\big)
\end{aligned}$$

concluding that there is a subsequence converging strongly $\boldsymbol{e}_h \to \boldsymbol{e}$.
Finally, since every convergent subsequence satisfies the same weak formulation with an unique solution, all limit points have to equal each other. Thus not only the subsequence but the whole sequence converge to the unique solution of weak formulation. □

Lemma 3.37. *Let the assumptions from previous Lem. 3.36 be satisfied. Moreover let $\boldsymbol{e}_{h,\boldsymbol{N}}$ be solution of Galerkin approximation with numerical integration for regularized conductivity coefficients $\boldsymbol{A}_h$, c.f. Def. 3.24. Then for arbitrary $\varepsilon > 0$ there exists $h \in \mathbb{R}$ and $\boldsymbol{N} \in \mathbb{N}^d$ (not the other way around) such that*

$$\|\boldsymbol{e} - \boldsymbol{e}_{h,\boldsymbol{N}}\|_{L^2_{\mathrm{per}}} \leq \varepsilon$$

Proof. From the previous Lem. 3.36, we can choose fixed $h > 0$ sufficiently small such that $\|\boldsymbol{e} - \boldsymbol{e}_h\| < \frac{\varepsilon}{2}$. Next, we can choose $\boldsymbol{N}$ such that

$$\min_{\alpha} \frac{N_\alpha}{2Y_\alpha} \geq \left(\frac{2C}{\varepsilon}\right)^{\frac{1}{\mu}}$$

with constant C from Lem. 3.35. Then Lem. 3.35 implies

$$\|\boldsymbol{e}_h - \boldsymbol{e}_{h,\boldsymbol{N}}\|_{L^2_{\mathrm{per}}} \leq \frac{\varepsilon}{2}$$

and the triangle inequality finishes the proof

$$\|\boldsymbol{e} - \boldsymbol{e}_{h,\boldsymbol{N}}\|_{L^2_{\mathrm{per}}} \leq \|\boldsymbol{e} - \boldsymbol{e}_h\|_{L^2_{\mathrm{per}}} + \|\boldsymbol{e}_h - \boldsymbol{e}_{h,\boldsymbol{N}}\|_{L^2_{\mathrm{per}}} \leq \varepsilon.$$

□

Remark 3.38 (Example of regularization of conductivity coefficients)**.** *The example of regularized data $\boldsymbol{A}_h$ satisfying conditions in Lem. 3.36 is a standard mollification, i.e.*

$$\boldsymbol{A}_h(\boldsymbol{x}) = \int_{\mathbb{R}^d} \eta_h(\boldsymbol{x}-\boldsymbol{y})\boldsymbol{A}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}$$

where $\eta_h(\boldsymbol{x}) = \frac{1}{h^d}\eta(\frac{\boldsymbol{x}}{h})$,

$$\eta(\boldsymbol{x}) = \begin{cases} C\exp\left(\frac{1}{|\boldsymbol{x}|^2-1}\right), & \text{for } |\boldsymbol{x}|<1,\\ 0, & \text{otherwise}\end{cases}$$

with constant C chosen so that $\int_{\mathbb{R}^d}\eta(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}=1$.

3.6 Solution of Linear systems using Conjugate gradients

In this section, we describe the solution of Galerkin approximation with numerical integration by Conjugate gradients. The next lemma shows that the solution by Conjugate gradients corresponds to minimization of quadratic functional over subspace $\mathbb{E}_{\boldsymbol{N}} \subset \mathbb{R}^{d\times\boldsymbol{N}}$. The further definition and lemma provide an projection on $\mathbb{E}_{\boldsymbol{N}}$. This satisfies Conjugate gradients are working in appropriate space. Moreover in [21], the relation of fully discrete formulation of GAwNI to Moulinec-Suquet algorithm [14] is provided.

Lemma 3.39. *The Galerkin approximation with numerical integration, defined in Def. 3.24, is solvable by Conjugate gradients.*

Proof. In Lemma 3.28, we have shown that Galerkin approximation with numerical integration can be characterized with the problem: find $\tilde{\mathbf{e}}_{\boldsymbol{N}} \in \mathbb{E}_{\boldsymbol{N}}$ such that

$$\left(\mathbf{A}_{\boldsymbol{N}}\tilde{\mathbf{e}}_{\boldsymbol{N}}, \mathbf{v}_{\boldsymbol{N}}\right)_{\mathbb{R}^{d\times\boldsymbol{N}}} = -\left(\mathbf{A}_{\boldsymbol{N}}\mathbf{E}_{\boldsymbol{N}}, \mathbf{v}_{\boldsymbol{N}}\right)_{\mathbb{R}^{d\times\boldsymbol{N}}}, \quad \forall \mathbf{v}_{\boldsymbol{N}} \in \mathbb{E}_{\boldsymbol{N}}. \tag{3.27}$$

where

$$\mathbf{A}_{\boldsymbol{N}} = \left[\delta_{\boldsymbol{km}} A_{\alpha\beta}(\boldsymbol{x}^{\boldsymbol{k}}_{\boldsymbol{N}})\right]^{\boldsymbol{k},\boldsymbol{m}\in\mathbb{Z}^d_{\boldsymbol{N}}}_{\alpha,\beta=1,\dots,d} \in \mathbb{R}^{d\times d\times\boldsymbol{N}\times\boldsymbol{N}} \qquad \mathbf{E}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}[\boldsymbol{E}] \in \mathbb{R}^{d\times\boldsymbol{N}}$$

being dependent on conductivity $\boldsymbol{A}$ and macroscopic electric field $\boldsymbol{E}$. Since $\boldsymbol{A}$ is symmetric and positive definite, the matrix $\mathbf{A}_{\boldsymbol{N}}$ is. Then the solution $\tilde{\mathbf{e}}_{\boldsymbol{N}}$ of Eq. (3.27) can be expressed as

$$\tilde{\mathbf{e}}_{\boldsymbol{N}} = \operatorname{argmin}_{\mathbf{v}_{\boldsymbol{N}}\in\mathbb{E}_{\boldsymbol{N}}} \frac{1}{2}\left(\mathbf{A}_{\boldsymbol{N}}\mathbf{v}_{\boldsymbol{N}}, \mathbf{v}_{\boldsymbol{N}}\right)_{\mathbb{R}^{d\times\boldsymbol{N}}} + \left(\mathbf{A}_{\boldsymbol{N}}\mathbf{E}_{\boldsymbol{N}}, \mathbf{v}_{\boldsymbol{N}}\right)_{\mathbb{R}^{d\times\boldsymbol{N}}} \tag{3.28}$$

and thus it is solvable by Conjugate gradients. □

However, the previous lemma does not show how to effectively minimize the quadratic functional, Eq. (3.28), on a subspace $\mathbb{E}_{\boldsymbol{N}} \subset \mathbb{R}^{d\times\boldsymbol{N}}$.

Definition 3.40. *We define a matrix composed of integral kernel in Fourier space, c.f. Def. 2.23 and especially Eq.* (2.16)*, as*

$$\hat{\mathbf{G}} = \left[\delta_{\boldsymbol{km}}\frac{\xi_\alpha(\boldsymbol{k})\xi_\beta(\boldsymbol{k})}{\left(\boldsymbol{\xi}(\boldsymbol{k}),\boldsymbol{\xi}(\boldsymbol{k})\right)_{\mathbb{R}^d}}\right]^{\boldsymbol{k},\boldsymbol{m}\in\mathbb{Z}^d_{\boldsymbol{N}}}_{\alpha,\beta=1,\dots,d} \in \mathbb{R}^{d\times d\times\boldsymbol{N}\times\boldsymbol{N}}$$

where $\boldsymbol{\xi}(\boldsymbol{k}) = \left(\frac{k_\alpha}{Y_\alpha}\right)_{\alpha=1,\dots,d} \in \mathbb{R}^d$*. Next, we define matrices composed of (inverse) Discrete Fourier Transform*

$$\mathbf{F} = \left[\frac{\delta_{\alpha\beta}\omega^{-\boldsymbol{km}}_{\boldsymbol{N}}}{|\boldsymbol{N}|_\Pi}\right]^{\boldsymbol{k},\boldsymbol{m}\in\mathbb{Z}^d_{\boldsymbol{N}}}_{\alpha,\beta=1,\dots,d} \in \mathbb{C}^{d\times d\times\boldsymbol{N}\times\boldsymbol{N}} \qquad \mathbf{F}^{-1} = \left[\delta_{\alpha\beta}\omega^{\boldsymbol{km}}_{\boldsymbol{N}}\right]^{\boldsymbol{k},\boldsymbol{m}\in\mathbb{Z}^d_{\boldsymbol{N}}}_{\alpha,\beta=1,\dots,d} \in \mathbb{C}^{d\times d\times\boldsymbol{N}\times\boldsymbol{N}}.$$

where

$$\omega_{\boldsymbol{N}}^{\boldsymbol{km}} = \exp\left(2\pi\mathrm{i}\sum_{\alpha=1}^{d}\frac{k_\alpha m_\alpha}{N_\alpha}\right), \quad \boldsymbol{m},\boldsymbol{k} \in \underline{\mathbb{Z}}_{\boldsymbol{N}}^d.$$

Finally, we define matrix

$$\mathbf{G} = \mathbf{F}^{-1}\hat{\mathbf{G}}\mathbf{F} \in \mathbb{C}^{d\times d\times \boldsymbol{N}\times \boldsymbol{N}} \tag{3.29}$$

Remark 3.41. *Submatrix $\mathbf{F}_{\alpha\alpha}$ is exactly the matrix of Discrete Fourier Transform. Thus the linear operator $\mathbf{F} : \mathbf{v} \in \mathbb{R}^{d\times \boldsymbol{N}} \to \mathbf{Fv} \in \mathbb{C}^{d\times \boldsymbol{N}}$ can be treated as the d-dimensional fast Fourier Transform routine applied on each submatrix $\mathbf{v}_\alpha$ for $\alpha = 1,\ldots,d$.*

Lemma 3.42. *The matrix $\mathbf{G}$ from Def. 3.40 defining a linear operator $\mathbf{G} : \mathbb{R}^{d\times \boldsymbol{N}} \to \mathbb{C}^{d\times \boldsymbol{N}}$ is a projection on $\mathbb{E}_{\boldsymbol{N}}$.*

Proof. The proof is a consequence of the fact that matrix $\mathbf{G}$ is defined via continuous projection $\mathcal{G}$ Def. 2.25 and that operator $\mathcal{I}_{\boldsymbol{N}}$ from Def. 3.8 is isomorphism. □

Acknowledgments This work was supported by the Czech Science Foundation, through projects No. GAČR 103/09/1748, No. GAČR 103/09/P490, No. GAČR P105/12/0331 and by the Grant Agency of the Czech Technical University in Prague through project No. SGS10/124/OHK1/2T/11 and No. SGS 12/027/OHK1/1T/11.

References

[1] N. S. Bakhvalov and A. V. Knyazev, *Efficient computation of averaged characteristics of composites of a periodic structure of essentially different materials*, Soviet mathematics - Doklady **42** (1991), no. 1, 57–62.

[2] S. Brisard and L. Dormieux, *FFT-based methods for the mechanics of composites: A general variational framework*, Computational Materials Science **49** (2010), no. 3, 663–671.

[3] ———, *Combining Galerkin approximation techniques with the principle of Hashin and Shtrikman to derive a new FFT-based numerical method for the homogenization of composites*, Computer Methods in Applied Mechanics and Engineering (2012).

[4] J. Y. Buffière, P. Cloetens, W. Ludwig, E. Maire, and L. Salvo, *In situ X-ray tomography studies of microstructural evolution combined with 3D modeling*, MRS Bulletin **33** (2008), no. 6, 611–619.

[5] P.G. Ciarlet, *Handbook of numerical analysis: Finite element methods*, vol. II, Elsevier Science Publishers B.V. (North-Holland), 1991.

[6] J. Dvořák, *Optimization of composite materials*, Master's thesis, The Charles University in Prague, June 1993.

[7] ———, *A reliable numerical method for computing homogenized coefficients*, Tech. report, available at CiteSeerX `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.45.1190`, 1995.

[8] 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** (1999), no. 1, 41–47.

[9] V.V. Jikov, S.M. Kozlov, and O.A. Oleinik, *Homogenization of differential operators and integral functionals*, Springer-Verlag, 1994.

[10] L.C.Evans, *Partial differential equations*, vol. 19, American Mathematical Society, 2000.

[11] J. Lukeš and J. Malý, *Measure and integral*, Matfyzpress Prague, 1995.

[12] G. W. Milton and R. V. Kohn, *Variational bounds on the effective moduli of anisotropic composites*, Journal of Mechanics and Physics of Solids **36** (1988), no. 6, 597–629.

[13] V. Monchiet and G. Bonnet, *A polarization-based FFT iterative scheme for computing the effective properties of elastic composites with arbitrary contrast*, International Journal for Numerical Methods in Engineering **89** (2012), no. 11, 1419–1436.

[14] H. Moulinec and P. Suquet, *A fast numerical method for computing the linear and nonlinear mechanical properties of composites*, Comptes rendus de l'Académie des sciences. Série II, Mécanique, physique, chimie, astronomie **318** (1994), no. 11, 1417–1423.

[15] ———, *A numerical method for computing the overall response of nonlinear composites with complex microstructure*, Computer Methods in Applied Mechanics and Engineering **157** (1997), no. 1–2, 69–94.

[16] J. Němeček, V. Králík, and J. Vondřejc, *Micromechanical analysis of heterogeneous structural materials*, Cement and Concrete Composites (2012).

[17] J. Němeček, V. Králík, and J. Vondřejc, *A two-scale micromechanical model for aluminium foam based on results from nanoindentation*, (2012), Submitted.

[18] J. Saranen and G. Vainikko, *Periodic integral and pseudodifferential equations with numerical approximation*, Springer Monographs Mathematics, 2000.

[19] G. Vainikko, *Fast solvers of the Lippmann-Schwinger equation*, Direct and Inverse Problems of Mathematical Physics (R. P. Gilbert, J. Kajiwara, and Y. S. Xu, eds.), International Society for Analysis, Applications and Computation, vol. 5, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 423–440.

[20] V. Vinogradov and G. W. Milton, *An accelerated FFT algorithm for thermoelastic and non-linear composites*, International Journal for Numerical Methods in Engineering **76** (2008), no. 11, 1678–1695.

[21] J. Vondřejc, J. Zeman, and I. Marek, *Analysis of a Fast Fourier Transform based method for modeling of heterogeneous materials*, Lecture Notes in Computer Science **7116** (2012), 512–522.

[22] Z. Wieçkowski, *Dual finite element methods in mechanics of composite materials*, Journal of Theoretical and Applied Mechanics **2** (1995), no. 33, 233–252.

[23] J. Yvonnet, *A fast method for solving microstructural problems defined by digital images: a space lippmann–schwinger scheme*, International Journal for Numerical Methods in Engineering (2012).

[24] 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** (2010), no. 21, 8065–8071.

[25] J. Zeman and M. Šejnoha, *Numerical evaluation of effective elastic properties of graphite fiber tow impregnated by polymer matrix*, Journal of the Mechanics and Physics of Solids **49** (2001), no. 1, 69–90.

Part VII
Paper 6

Authors:

Jaroslav Vondřejc, Jan Zeman, and Ivo Marek

Title:

Arbitrary precise guaranteed bounds of homogenized material coefficients by FFT-based finite element method

ARBITRARY PRECISE GUARANTEED BOUNDS OF HOMOGENIZED MATERIAL COEFFICIENTS BY FFT-BASED FINITE ELEMENT METHOD *

JAROSLAV VONDŘEJC†, JAN ZEMAN‡§, AND IVO MAREK¶

Abstract. In this work, the guaranteed bounds of homogenized material coefficients are calculated with an arbitrary precision; primal and dual variational formulations are evaluated with approximate microscopic (local) solutions to produce the upper-lower bounds. Contrary to Dvořák [5, 6] and Wieçkowski [32] employing the approach with local solutions from h and p-version of Finite Element Method (FEM), we utilize the FFT-based FEM using trigonometric polynomials as the basis functions [31]. The connection between the primal and the dual formulations is investigated in discretized form; the solution of the dual formulation can be avoided, however, the theory substantially differs for even and odd number of discretization points. Numerical examples confirm the theoretical results about the rates of convergence and about the upper-lower bounds of the homogenized matrix.

Key words. FFT, homogenization, guaranteed bounds, Finite Element Method

AMS subject classifications. 65N15, 74Q20

1. Introduction. Guaranteed bounds of homogenized (effective) material properties, with some confidence in its accuracy, is essential for a reliable design [24]; a special attention is addressed to the upper-lower bounds for linear periodic elliptic problems, a scalar one or elasticity. The majority of bounds — e.g. Voight and Reuss bounds [29, 25], Hashin and Shtrikman bounds [9, 10, 11] — are derived for arbitrary geometry of material phases, only the knowledge of volume fractions is assumed. However, these a priori estimates are rather wide especially for materials with highly oscillating coefficients.

Besides, assuming material properties are well known, there are a posteriori estimates producing reliable and guaranteed homogenized bounds with an arbitrary precision independently introduced in [6, 32] for a scalar problem and elasticity resp., see Sec. 2.2 for a summary.

This approach is based on the primal and the dual variational formulations. Approximated microscopic fields, satisfying linear second order elliptic partial differential equations with periodic boundary conditions and prescribed macroscopic (averaged) values, are calculated and their a posteriori evaluation in both formulations produces the guaranteed bounds of homogenized properties (matrix).

Contrary to already mentioned works [6, 32] incorporating classical p- and h-version resp. of Finite Element Method (FEM) for computation of local fields, we rather employ the FFT-based FEM originated from [19] and theoretically described in [31], see Sec. 3.3 for its overview. Noting, the method — a widely used in engineering problems, recently in [22, 23, 14, 28, 12, 15, 21] — has obtained miscellaneous variants

*This work was supported by the Czech Science Foundation through project No. GAČR P105/12/0331 and by the Grant Agency of the Czech Technical University in Prague through project No. SGS 12/027/OHK1/1T/11.

†Department of Mechanics, Faculty of Civil Engineering, Czech Technical University in Prague, Thákurova 7, 166 29 Prague 6, Czech Republic (vondrejc@gmail.com).

‡Department of Mechanics, Faculty of Civil Engineering, Czech Technical University in Prague, Thákurova 7, 166 29 Prague 6, Czech Republic and Centre of Excellence IT4Innovations, VŠB-TU Ostrava, 17. listopadu 15/2172 708 33 Ostrava-Poruba, Czech Republic (zemanj@cml.fsv.cvut.cz).

§Department of Mathematics, Faculty of Civil Engineering, Czech Technical University in Prague, Thákurova 7, 166 29 Prague 6, Czech Republic (marekivo@mat.fsv.cvut.cz).

and modifications [8, 16, 2, 3, 18].

In this work, for simplicity, the scalar problem of electric conduction is considered, however, an extension to elasticity is feasible. Numerical homogenization of a periodic media with its upper and lower bounds, according to [5, 6], is summarized in Section 2.

Section 3.3 deals with the FFT-based FEM based on the discretization with trigonometric polynomials. In our previous works [30, 31], only the discretization with the odd number of points in each direction is considered; the problem with even number of discretization points was identified and partially solved in [20, section 2.4.2]. We describe the theory for the even number of discretization points in a way to satisfy the conformity of the method; finite dimensional spaces are subspaces of the trial space — it is the natural requirement for guaranteed bounds of the homogenized matrix.

In section 3.4, the connection between discretized primal-dual formulations is studied. It provides a computational simplification: the solution of the dual formulation, necessary for the lower bound, can be avoided; moreover, for the odd number of discretization points, it can be fully omitted.

The upper-lower bounds of the homogenized matrix are calculated in Section 3.5. Although, their exact values, dependent on arbitrary non-regular material coefficients, are able to exactly evaluate only for some special cases. Hence, we propose their approximations; moreover, a careful computation still guarantees the upper-lower bounds structure.

Numerical examples in last Section 4 validate the theory and demonstrate the upper-lower bounds structure of homogenized matrices.

2. Preliminaries to homogenization and to guaranteed bounds. In this section, we summarize the well established theory of homogenization of periodic media and add the part about upper and lower bounds of the homogenized material coefficients. In the sequel, letter d denotes the dimension of the model problem, assuming $d = 2, 3$; Greek letters α, β are reserved to indices relating dimension, thus ranging $1, \dots, d$ — further, it is for simplicity omitted.

Sets $\mathbb{C}^d$ and $\mathbb{R}^d$ are the spaces of complex and real vectors with canonical basis $\{\boldsymbol{\epsilon}_\alpha\}$ and are equipped with Lebesgue measure $\mathrm{d}\boldsymbol{x}$. We denote by $|\Omega|_d$ the d-dimensional Lebesgue measure of a measurable set $\Omega \subset \mathbb{R}^d$. Standard Euklidean norm $\|\cdot\|_2$ on $\mathbb{C}^d$ is induced by scalar product $\big(\boldsymbol{u}, \boldsymbol{v}\big)_{\mathbb{C}^d} = \sum_\alpha u_\alpha \overline{v_\alpha}$ for $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^d$.

Set $\mathbb{R}^{d\times d}_{\mathrm{spd}}$ denotes the space of symmetric positive definite matrices of size $d \times d$ with norm $\|\boldsymbol{A}\|_2 = \max_{\boldsymbol{x}\in\mathbb{R}^d, \|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{x}\|_2$ that equals to the largest eigenvalue.

Function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is $\boldsymbol{Y}$-periodic (with period $\boldsymbol{Y} \in \mathbb{R}^d$) if $f(\boldsymbol{x} + \boldsymbol{Y} \odot \boldsymbol{k}) = f(\boldsymbol{x})$ for arbitrary $\boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{k} \in \mathbb{Z}^d$, where operator $\odot$ denotes the element-wise multiplication. Then, $\boldsymbol{Y}$-periodic functions suffice to define only on a periodic unit cell (PUC), the set defined as $\mathcal{Y} := (-Y_\alpha, Y_\alpha)_{\alpha=1}^d \subset \mathbb{R}^d$. We will identify two integrable functions which are equal almost everywhere. The mean value of function $\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y}; \mathbb{R}^d)$ over periodic unit cell $\mathcal{Y}$ is denoted as $\langle \boldsymbol{v} \rangle := \frac{1}{|\mathcal{Y}|_d} \int_{\mathcal{Y}} \boldsymbol{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \in \mathbb{R}^{\mathrm{d}}$.

We define space $C_{\mathrm{per}}(\mathcal{Y}; \mathbb{X})$ of continuous $\boldsymbol{Y}$-periodic functions $\mathbb{R}^d \mapsto \mathbb{X}$, where $\mathbb{X}$ is some finite dimensional vector space, e.g. $\mathbb{C}, \mathbb{R}, \mathbb{C}^d, \mathbb{R}^d$. Vector valued functions, for $\mathbb{X} = \mathbb{C}^d$ or $\mathbb{X} = \mathbb{R}^d$, are denoted with small bold letters $\boldsymbol{v}$ and have components v_α.

Spaces $L^2_{\mathrm{per}}(\mathcal{Y}; \mathbb{X})$ or $L^\infty_{\mathrm{per}}(\mathcal{Y}; \mathbb{R}^{d\times d}_{\mathrm{spd}})$ are composed of functions $\boldsymbol{v} : \mathbb{R}^d \mapsto \mathbb{X}$ or $\boldsymbol{A} : \mathbb{R}^d \mapsto \mathbb{R}^{d\times d}_{\mathrm{spd}}$ having $\boldsymbol{Y}$-periodic, measurable components v_α or $A_{\alpha\beta}$ and having finite norm, i.e. $\|\boldsymbol{v}\|_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{X})} < \infty$ or $\|\boldsymbol{A}\|_{L^\infty_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^{d\times d}_{\mathrm{spd}})} < \infty$. The first norm is deduced from scalar product $\big(\boldsymbol{u}, \boldsymbol{v}\big)_{L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{X})} = \frac{1}{|\mathcal{Y}|_d} \int_{\mathcal{Y}} \big(\boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}(\boldsymbol{x})\big)_{\mathbb{X}} \mathrm{d}\boldsymbol{x}$ while the second norm

is defined as $\|\boldsymbol{A}\|_{L^\infty_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^{d\times d}_{\mathrm{spd}})} = \operatorname{ess\,sup}_{\boldsymbol{x}\in\mathcal{Y}} \|\boldsymbol{A}(\boldsymbol{x})\|_2$. If there is no ambiguity, both the norms and the scalar products are denoted with subscript L^2_{per} or L^∞_{per} rather than $L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{X})$ or $L^\infty_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^{d\times d}_{\mathrm{spd}})$.

Next, we define the spaces of Helmholtz decomposition of $L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) = \mathscr{U} \oplus^\perp \mathscr{E} \oplus^\perp \mathscr{J}$, i.e. the spaces of constant, curl-free with zero mean, and divergence free with zero mean fields

$$\mathscr{U} = \{\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) : \boldsymbol{v}(\boldsymbol{x}) = \mathrm{const}\}, \tag{2.1a}$$

$$\mathscr{E} = \{\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) : \nabla \times \boldsymbol{v} = \boldsymbol{0}, \langle \boldsymbol{v}\rangle = \boldsymbol{0}\}, \tag{2.1b}$$

$$\mathscr{J} = \{\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) : \nabla \cdot \boldsymbol{v} = 0, \langle \boldsymbol{v}\rangle = \boldsymbol{0}\}, \tag{2.1c}$$

where differential operator $\nabla = (\frac{\partial}{\partial x_\alpha})_{\alpha=1}^d$ is meant in the distributional sense. For dimension $d \neq 3$, the curl-free condition in (2.1b) means $(\nabla \times \boldsymbol{v})_{\alpha\beta} := \frac{\partial v_\alpha}{\partial x_\beta} - \frac{\partial v_\beta}{\partial x_\alpha} = 0$. Since space $\mathscr{U}$ consists of constant functions, we identify spaces $\mathscr{U}$ and $\mathbb{R}^d$; it validates the operation such as $\boldsymbol{E} + \boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ for $\boldsymbol{E} \in \mathbb{R}^d$, $\boldsymbol{v} \in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ and $\boldsymbol{A}\boldsymbol{J} \in \mathbb{R}^d$ for $\boldsymbol{A} \in \mathbb{R}^{d\times d}_{\mathrm{spd}}$ and $\boldsymbol{J} \in \mathscr{U}$.

2.1. Homogenization in primal and dual formulations. The theory is demonstrated for a scalar problem modeling: diffusion, stationary heat transfer, or electric conductivity — our choice. The well established theory of periodic homogenization for linear elliptic partial differential equations can be found in [1, 13, 4]. Although, we focus only on the numerical solution of microscopic fields and the consequence evaluation of the homogenized matrix and its bounds.

Notation 2.1. *Here and in the sequel, $\boldsymbol{A} \in L^\infty_{\mathrm{per}}(\mathcal{Y},\mathbb{R}^{d\times d}_{\mathrm{spd}})$ denotes symmetric and uniformly elliptic*[1] *material coefficients of electric conductivity, $\boldsymbol{e} \in \mathscr{E}$ and $\boldsymbol{\jmath} \in \mathscr{J}$ perturbation of electric field and electric current resp., and $\boldsymbol{E}, \boldsymbol{J} \in \mathscr{U}$ their macroscopic counterparts. Then their summation $(\boldsymbol{E}+\boldsymbol{e})$ and $(\boldsymbol{J}+\boldsymbol{\jmath})$ represent microscopic fields.*

Remark 2.2. *The variables from previous notations, being additionally sufficiently smooth, satisfy the differential equations*

$$\boldsymbol{J} + \boldsymbol{\jmath} = \boldsymbol{A}(\boldsymbol{E} + \boldsymbol{e}) \qquad \nabla \cdot \boldsymbol{\jmath} = 0 \qquad \nabla \times \boldsymbol{e} = 0 \tag{2.2}$$

that clarify the definition of subspaces $\mathscr{U}$,$\mathscr{E}$, and $\mathscr{J}$; the addition of periodic boundary conditions and prescription of macroscopic loads $\boldsymbol{E}, \boldsymbol{J} \in \mathbb{R}^d$ sets the homogenization problem.

Definition 2.3 (Homogenization problem). *The primal and the dual homogenization problem states: find homogenized matrices $\boldsymbol{A}_{\mathrm{eff}}, \boldsymbol{A}_{\mathrm{eff,D}} \in \mathbb{R}^{d\times d}$ such that for arbitrary fixed macroscopic quantities $\boldsymbol{E}, \boldsymbol{J} \in \mathbb{R}^d$, the following relations hold*

$$\big(\boldsymbol{A}_{\mathrm{eff}}\boldsymbol{E}, \boldsymbol{E}\big)_{\mathbb{R}^d} = \inf_{\boldsymbol{e}\in\mathscr{E}} \big(\boldsymbol{A}(\boldsymbol{E}+\boldsymbol{e}), \boldsymbol{E}+\boldsymbol{e}\big)_{L^2_{\mathrm{per}}} \tag{2.3a}$$

$$\big(\boldsymbol{A}^{-1}_{\mathrm{eff,D}}\boldsymbol{J}, \boldsymbol{J}\big)_{\mathbb{R}^d} = \inf_{\boldsymbol{\jmath}\in\mathscr{J}} \big(\boldsymbol{A}^{-1}(\boldsymbol{J}+\boldsymbol{\jmath}), \boldsymbol{J}+\boldsymbol{\jmath}\big)_{L^2_{\mathrm{per}}} \tag{2.3b}$$

Remark 2.4. *The homogenized matrices are symmetric positive definite and they equal to one another $\boldsymbol{A}_{\mathrm{eff}} = \boldsymbol{A}_{\mathrm{eff,D}}$ — it follows from the perturbation duality*

[1] There exists positive constant $c_A > 0$ such that inequality $c_A\|u\|_2^2 \le (\boldsymbol{A}(\boldsymbol{x})u, u)_{\mathbb{R}^d}$ holds for almost all $x \in \mathcal{Y}$ and all nonzero $u \in \mathbb{R}^d$.

theorem, see e.g. [7, 27, 6] and compare it to Lem. 3.24 and 3.25 for a discrete setting. The symmetry comes from upcoming Eq. (2.6) *and symmetry of* $\boldsymbol{A}$*. The positive definiteness is the consequence of the uniform ellipticity of* $\boldsymbol{A}$ *and simultaneously the consequence of Voight* $\big(\boldsymbol{A}_{\rm eff}\boldsymbol{E},\boldsymbol{E}\big)_{\mathbb{R}^d} \le \big(\boldsymbol{A}\boldsymbol{E},\boldsymbol{E}\big)_{L^2_{\rm per}}$ *and Reuss* $\big(\boldsymbol{A}^{-1}_{\rm eff,D}\boldsymbol{J},\boldsymbol{J}\big)_{\mathbb{R}^d} \le \big(\boldsymbol{A}^{-1}\boldsymbol{J},\boldsymbol{J}\big)_{L^2_{\rm per}}$ *bounds according to [29, 25]; the bounds come from* (2.3) *for* $\boldsymbol{e}=\boldsymbol{\jmath}\equiv 0$.

The minimizers of the variational formulations (2.3a) and (2.3b) can also be found as a solution of weak formulations with existence and uniqueness provided by Lax-Milgram lemma. Thanks to linearity, the minimizers can be found only for unitary macroscopic loads.

Definition 2.5 (auxiliary problems). *We say that* $\tilde{\boldsymbol{e}}^{(\alpha)}\in\mathscr{E}$ *and* $\tilde{\boldsymbol{\jmath}}^{(\alpha)}\in\mathscr{J}$ *are unitary minimizers if*

$$\big(\boldsymbol{A}\tilde{\boldsymbol{e}}^{(\alpha)},\boldsymbol{v}\big)_{L^2_{\rm per}} = -\big(\boldsymbol{A}\boldsymbol{\epsilon}_\alpha,\boldsymbol{v}\big)_{L^2_{\rm per}},\quad \forall \boldsymbol{v}\in\mathscr{E}, \tag{2.4}$$

$$\big(\boldsymbol{A}^{-1}\tilde{\boldsymbol{\jmath}}^{(\alpha)},\boldsymbol{v}\big)_{L^2_{\rm per}} = -\big(\boldsymbol{A}^{-1}\boldsymbol{\epsilon}_\alpha,\boldsymbol{v}\big)_{L^2_{\rm per}},\quad \forall \boldsymbol{v}\in\mathscr{J}. \tag{2.5}$$

Next, we define unitary microscopic fields $\boldsymbol{e}^{(\alpha)}:=\boldsymbol{\epsilon}_\alpha+\tilde{\boldsymbol{e}}^{(\alpha)}$ *and* $\boldsymbol{\jmath}^{(\alpha)}:=\boldsymbol{\epsilon}_\alpha+\tilde{\boldsymbol{\jmath}}^{(\alpha)}$.

Remark 2.6 (Consequences of the linearity). *Minimizers* $\tilde{\boldsymbol{e}}^{(\boldsymbol{E})}\in\mathscr{E}$ *and* $\tilde{\boldsymbol{\jmath}}^{(J)}\in\mathscr{J}$ *of the homogenization problems for macroscopic fields* $\boldsymbol{E},\boldsymbol{J}\in\mathbb{R}^d$ *can be obtain, due to linear structure, from unitary minimizers*

$$\tilde{\boldsymbol{e}}^{(\boldsymbol{E})}=\sum_\alpha E_\alpha\tilde{\boldsymbol{e}}^{(\alpha)},\qquad\qquad \tilde{\boldsymbol{\jmath}}^{(J)}=\sum_\alpha J_\alpha\tilde{\boldsymbol{\jmath}}^{(\alpha)}.$$

Alike, the components of homogenized material coefficients states

$$(\boldsymbol{A}_{\rm eff})_{\alpha\beta}=\big(\boldsymbol{A}\boldsymbol{e}^{(\beta)},\boldsymbol{e}^{(\alpha)}\big)_{L^2_{\rm per}},\qquad (\boldsymbol{A}^{-1}_{\rm eff})_{\alpha\beta}=\big(\boldsymbol{A}^{-1}\boldsymbol{\jmath}^{(\beta)},\boldsymbol{\jmath}^{(\alpha)}\big)_{L^2_{\rm per}}. \tag{2.6}$$

Remark 2.7. *The dual unitary microscopic fields* $\boldsymbol{\jmath}^{(\beta)}$ *can be expressed as a linear combination of primal ones* $\boldsymbol{e}^{(\alpha)}$*, hence* $\boldsymbol{\jmath}^{(\beta)}=\boldsymbol{A}\sum_{\alpha=1}^d E_\alpha\boldsymbol{e}^{(\alpha)}$ *where* $\boldsymbol{E}=\boldsymbol{A}^{-1}_{\rm eff}\boldsymbol{\epsilon}_\beta$*. Similarly to Rem. 2.4, it comes from the perturbation duality theorem, see e.g. [7, 27, 6] and compare it to Lem. 3.24 and 3.25 for a discrete setting.*

2.2. Upper-lower bounds of homogenized matrix. The upper-lower bounds obtained from a posteriori error estimates were introduced by Dvořák [5, 6] for a scalar problem and later independently by Wieçkowski [32] for linear elasticity; this section provides a summary of results in [6]. In what follows, relation $\boldsymbol{C}\preceq\boldsymbol{D}$ between matrices $\boldsymbol{C},\boldsymbol{D}\in\mathbb{R}^{d\times d}_{\rm spd}$ stands for ordering in the sense of quadratic forms; it is equivalent to $\boldsymbol{E}\cdot\boldsymbol{C}\boldsymbol{E}\le\boldsymbol{E}\cdot\boldsymbol{D}\boldsymbol{E}$ for all $\boldsymbol{E}\in\mathbb{R}^d$. In this section, we will work with some conforming approximations of unitary minimizers $\tilde{\boldsymbol{e}}^{(\alpha)}$ and $\tilde{\boldsymbol{\jmath}}^{(\alpha)}$, namely with $\tilde{\boldsymbol{e}}^{(\alpha)}_{\boldsymbol{N}}=(\boldsymbol{e}^{(\alpha)}_{\boldsymbol{N}}-\boldsymbol{\epsilon}_\alpha)\in\mathscr{E}$ and $\tilde{\boldsymbol{\jmath}}^{(\alpha)}_{\boldsymbol{N}}=(\boldsymbol{\jmath}^{(\alpha)}_{\boldsymbol{N}}-\boldsymbol{\epsilon}_\alpha)\in\mathscr{J}$; parameter $\boldsymbol{N}$ represent inverse of discretization size of FEM or the number of discretization points in our case of FFT-based FEM, for detail see 3.3.

Definition 2.8. *We say that matrices* $\overline{\boldsymbol{A}}_{{\rm eff},\boldsymbol{N}},\underline{\boldsymbol{A}}_{{\rm eff},\boldsymbol{N}}\in\mathbb{R}^{d\times d}$ *defined as*

$$(\overline{\boldsymbol{A}}_{{\rm eff},\boldsymbol{N}})_{\alpha\beta}=\big(\boldsymbol{A}\boldsymbol{e}^{(\beta)}_{\boldsymbol{N}},\boldsymbol{e}^{(\alpha)}_{\boldsymbol{N}}\big)_{L^2_{\rm per}},\qquad (\underline{\boldsymbol{A}}^{-1}_{{\rm eff},\boldsymbol{N}})_{\alpha\beta}=\big(\boldsymbol{A}^{-1}\boldsymbol{\jmath}^{(\beta)}_{\boldsymbol{N}},\boldsymbol{\jmath}^{(\alpha)}_{\boldsymbol{N}}\big)_{L^2_{\rm per}}$$

are upper and lower bounds of homogenized properties $\boldsymbol{A}_{\rm eff}$.

Correctness of the definition, the fact that they are truly the upper-lower bounds, is stated in Lem. 2.10 that is based on a following statement.

LEMMA 2.9. *Let* $\boldsymbol{C}, \boldsymbol{D} \in \mathbb{R}^{d\times d}_{\text{spd}}$ *be such that* $\boldsymbol{C} \preceq \boldsymbol{D}$. *Then* $\boldsymbol{D}^{-1} \preceq \boldsymbol{C}^{-1}$.

Proof. From assumption $\boldsymbol{C} \preceq \boldsymbol{D}$, we claim that $\boldsymbol{R}^T\boldsymbol{C}\boldsymbol{R} \preceq \boldsymbol{R}^T\boldsymbol{D}\boldsymbol{R}$ for any regular matrix $\boldsymbol{R} \in \mathbb{R}^{d\times d}$; it comes from $(\boldsymbol{R}\boldsymbol{E})\cdot\boldsymbol{C}(\boldsymbol{R}\boldsymbol{E}) \leq (\boldsymbol{R}\boldsymbol{E})\cdot\boldsymbol{D}(\boldsymbol{R}\boldsymbol{E})$ holding for arbitrary $\boldsymbol{E} \in \mathbb{R}^d$ and the invertibility of matrix $\boldsymbol{R}$, namely the property $\boldsymbol{R}(\mathbb{R}^d) = \mathbb{R}^d$.

Since $\boldsymbol{D}$ is positive definite, matrix $\boldsymbol{D}^{-\frac{1}{2}}$ exists. Consequently, the multiplication of inequality $\boldsymbol{C} \preceq \boldsymbol{D}$ by matrix $\boldsymbol{D}^{-\frac{1}{2}}$ produces inequality $\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{C}\boldsymbol{D}^{-\frac{1}{2}} \preceq \boldsymbol{I}$. Matrix $\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{C}\boldsymbol{D}^{-\frac{1}{2}}$ possesses all eigenvalues real (from symmetry) and smaller or equal to 1. Thus its inverse $\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{C}^{-1}\boldsymbol{D}^{\frac{1}{2}}$ have eigenvalues larger or equal to 1. The next multiplication of inequality $\boldsymbol{I} \preceq \boldsymbol{D}^{\frac{1}{2}}\boldsymbol{C}^{-1}\boldsymbol{D}^{\frac{1}{2}}$ by matrix $\boldsymbol{D}^{-\frac{1}{2}}$ reveals the proof. □

LEMMA 2.10. *The matrices from previous definition 2.8 are symmetric positive definite and satisfy the upper-lower bound structure*

$$\boldsymbol{A}_{\text{eff}} \preceq \overline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}}, \qquad \boldsymbol{A}^{-1}_{\text{eff}} \preceq \underline{\boldsymbol{A}}^{-1}_{\text{eff},\boldsymbol{N}}, \qquad \underline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}} \preceq \boldsymbol{A}_{\text{eff}} \preceq \overline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}}. \tag{2.7}$$

Proof. The first two inequalities follow from the minimization problems (2.3) that are evaluated approximately, for a particular choice of microscopic fields $\tilde{\boldsymbol{e}}^{(\alpha)}_{\boldsymbol{N}}$, $\tilde{\boldsymbol{j}}^{(\alpha)}_{\boldsymbol{N}}$ rather than minimizers $\tilde{\boldsymbol{e}}^{(\alpha)}$, $\tilde{\boldsymbol{e}}^{(\alpha)}$. The last inequality is a consequence of previous Lem. 2.9.

The symmetry of the homogenized matrices comes from their definition 2.8 and the symmetry of material coefficients $\boldsymbol{A}$. The positive definiteness is a consequence of first two inequalities in (2.7) with the positive definiteness of homogenized matrix $\boldsymbol{A}_{\text{eff}} \in \mathbb{R}^{d\times d}_{\text{spd}}$. □

DEFINITION 2.11 (Approximate homogenized matrix with guaranteed error). *The mean of the upper-lower bounds from Def. 2.8,* $\boldsymbol{A}_{\text{eff},\boldsymbol{N}} = \frac{1}{2}(\overline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}} + \underline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}})$, *is called approximate homogenized matrix with guaranteed error* $\boldsymbol{D}_{\boldsymbol{N}} = \frac{1}{2}(\overline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}} - \underline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}})$.

LEMMA 2.12 (Element-wise upper-lower bounds). *The upper-lower bounds from Def. 2.8 imply element-wise bounds, explicitly for diagonal components*

$$\underline{A}_{\text{eff},\boldsymbol{N},\alpha\alpha} \leq A_{\text{eff},\alpha\alpha} \leq \overline{A}_{\text{eff},\boldsymbol{N},\alpha\alpha}, \tag{2.8}$$

and for non-diagonal components, i.e. for $\alpha \neq \beta$

$$\underline{A}_{\text{eff},\boldsymbol{N},\alpha\beta} - D_{\text{eff},\boldsymbol{N},\alpha\alpha} - D_{\text{eff},\boldsymbol{N},\beta\beta} \leq A_{\text{eff},\alpha\beta} \leq \overline{A}_{\text{eff},\boldsymbol{N},\alpha\beta} + D_{\text{eff},\boldsymbol{N},\alpha\alpha} + D_{\text{eff},\boldsymbol{N},\beta\beta}.$$

Proof. The proof for the diagonal terms (2.8) comes from the inequality (2.7) tested with $\boldsymbol{\epsilon}_\alpha$. The estimates for the non-diagonal terms come from equality

$$2A_{\text{eff},\alpha\beta} = (\boldsymbol{\epsilon}_\alpha + \boldsymbol{\epsilon}_\beta)\cdot\boldsymbol{A}_{\text{eff}}(\boldsymbol{\epsilon}_\alpha + \boldsymbol{\epsilon}_\beta) - A_{\text{eff},\alpha\alpha} - A_{\text{eff},\beta\beta},$$

the first inequality in (2.7) tested with $\boldsymbol{\epsilon}_\alpha + \boldsymbol{\epsilon}_\beta$, i.e. $(\boldsymbol{\epsilon}_\alpha + \boldsymbol{\epsilon}_\beta)\cdot\boldsymbol{A}_{\text{eff}}(\boldsymbol{\epsilon}_\alpha + \boldsymbol{\epsilon}_\beta) \leq (\boldsymbol{\epsilon}_\alpha + \boldsymbol{\epsilon}_\beta)\cdot\overline{\boldsymbol{A}}_{\text{eff},\boldsymbol{N}}(\boldsymbol{\epsilon}_\alpha + \boldsymbol{\epsilon}_\beta)$, and the inequalities for diagonal components (2.8). An analogy yields the lower bound. □

LEMMA 2.13 (Properties of trace operator). *Let* $\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D} \in \mathbb{R}^{d\times d}_{\text{spd}}$. *Then*

$$\operatorname{tr}(\boldsymbol{B}+\boldsymbol{C}) = \operatorname{tr}\boldsymbol{B} + \operatorname{tr}\boldsymbol{C}, \qquad \operatorname{tr}(\boldsymbol{B}\boldsymbol{C}\boldsymbol{D}) \leq (\operatorname{tr}\boldsymbol{B})(\operatorname{tr}\boldsymbol{C})(\operatorname{tr}\boldsymbol{D}).$$

Moreover, let $\boldsymbol{C} \preceq \boldsymbol{D}$, *then* $0 \leq \operatorname{tr}(\boldsymbol{D}-\boldsymbol{C}) \leq (\operatorname{tr}\boldsymbol{D})^2(\boldsymbol{C}^{-1} - \boldsymbol{D}^{-1})$.

Proof. The proof can be found in [6, Lemmas 4.3 and 4.4]. ∎

LEMMA 2.14 (The rate of convergence of homogenized properties). *The trace of guaranteed error $\boldsymbol{D}_{\boldsymbol{N}}$, from Def. 2.11, satisfies following inequality*

$$\operatorname{tr}\boldsymbol{D}_{\boldsymbol{N}} \leq \|\boldsymbol{A}\|_{L^\infty_{\mathrm{per}}} \sum_\alpha \|\tilde{\boldsymbol{e}}^{(\alpha)} - \tilde{\boldsymbol{e}}^{(\alpha)}_{\boldsymbol{N}}\|^2_{L^2_{\mathrm{per}}} + (\operatorname{tr}\boldsymbol{A}_{\mathrm{eff}})^2 \|\boldsymbol{A}^{-1}\|_{L^\infty_{\mathrm{per}}} \sum_\alpha \|\tilde{\boldsymbol{j}}^{(\alpha)} - \tilde{\boldsymbol{j}}^{(\alpha)}_{\boldsymbol{N}}\|^2_{L^2_{\mathrm{per}}}.$$

Proof. First, we prove two statements

$$0 \leq \overline{A}_{\mathrm{eff},\boldsymbol{N},\alpha\alpha} - A_{\mathrm{eff},\alpha\alpha} \leq \|\boldsymbol{A}\|_{L^\infty} \|\tilde{\boldsymbol{e}}^{(\alpha)} - \tilde{\boldsymbol{e}}^{(\alpha)}_{\boldsymbol{N}}\|^2_{L^2_{\mathrm{per}}}, \tag{2.9a}$$

$$0 \leq \underline{A}^{-1}_{\mathrm{eff},\boldsymbol{N},\alpha\alpha} - A^{-1}_{\mathrm{eff},\alpha\alpha} \leq \|\boldsymbol{A}^{-1}\|_{L^\infty} \|\tilde{\boldsymbol{j}}^{(\alpha)} - \tilde{\boldsymbol{j}}^{(\alpha)}_{\boldsymbol{N}}\|^2_{L^2_{\mathrm{per}}}. \tag{2.9b}$$

Both inequalities follow from Lem. 2.10 and Hölder inequality; we demonstrate the calculation for the first one

$$\begin{aligned} 0 \leq \overline{A}_{\mathrm{eff},\boldsymbol{N},\alpha\alpha} - A_{\mathrm{eff},\alpha\alpha} &= |\big(\boldsymbol{A}\boldsymbol{e}^{(\alpha)}, \boldsymbol{e}^{(\alpha)}\big)_{L^2_{\mathrm{per}}} - \big(\boldsymbol{A}\boldsymbol{e}^{(\alpha)}_{\boldsymbol{N}}, \boldsymbol{e}^{(\alpha)}_{\boldsymbol{N}}\big)_{L^2_{\mathrm{per}}}| \\ &\leq |\big(\boldsymbol{A}(\tilde{\boldsymbol{e}}^{(\alpha)} - \tilde{\boldsymbol{e}}^{(\alpha)}_{\boldsymbol{N}}), (\tilde{\boldsymbol{e}}^{(\alpha)} - \tilde{\boldsymbol{e}}^{(\alpha)}_{\boldsymbol{N}})\big)_{L^2_{\mathrm{per}}}| \\ &\leq \|\boldsymbol{A}\|_{L^\infty} \|\tilde{\boldsymbol{e}}^{(\alpha)} - \tilde{\boldsymbol{e}}^{(\alpha)}_{\boldsymbol{N}}\|^2_{L^2_{\mathrm{per}}}. \end{aligned}$$

Using the properties of the trace operator stated in Lem. 2.13 and previously proven statements (2.9), we finish the proof with direct calculation

$$\begin{aligned} \operatorname{tr}\boldsymbol{D}_{\boldsymbol{N}} &= \operatorname{tr}(\overline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}} - \boldsymbol{A}_{\mathrm{eff}}) + \operatorname{tr}(\boldsymbol{A}_{\mathrm{eff}} - \underline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}) \\ &\leq \operatorname{tr}(\overline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}} - \boldsymbol{A}_{\mathrm{eff}}) + (\operatorname{tr}\boldsymbol{A}_{\mathrm{eff}})^2 \operatorname{tr}(\underline{\boldsymbol{A}}^{-1}_{\mathrm{eff},\boldsymbol{N}} - A^{-1}_{\mathrm{eff},\boldsymbol{N}}) \\ &\leq \|\boldsymbol{A}\|_{L^\infty_{\mathrm{per}}} \sum_\alpha \|\tilde{\boldsymbol{e}}^{(\alpha)} - \tilde{\boldsymbol{e}}^{(\alpha)}_{\boldsymbol{N}}\|^2_{L^2_{\mathrm{per}}} + (\operatorname{tr}\boldsymbol{A}_{\mathrm{eff}})^2 \|\boldsymbol{A}^{-1}\|_{L^\infty_{\mathrm{per}}} \sum_\alpha \|\tilde{\boldsymbol{j}}^{(\alpha)} - \tilde{\boldsymbol{j}}^{(\alpha)}_{\boldsymbol{N}}\|^2_{L^2_{\mathrm{per}}}. \end{aligned}$$

∎

REMARK 2.15. *The trace of $\boldsymbol{D}_{\boldsymbol{N}}$ will converges to zero for $\min_\alpha N_\alpha \to \infty$, if approximate minimizers $\tilde{\boldsymbol{e}}^\alpha_{\boldsymbol{N}}$, $\tilde{\boldsymbol{j}}^\alpha_{\boldsymbol{N}}$ converge to minimizers $\tilde{\boldsymbol{e}}^\alpha$, $\tilde{\boldsymbol{e}}^\alpha$ for $\min_\alpha N_\alpha \to \infty$. Moreover, the trace operator is a norm on the set of symmetric positive definite matrices showing the convergence of $D_{\boldsymbol{N}}$ to zero for any matrix norm.*

3. Guaranteed bounds using FFT-based FEM. This section is the core of the work. It provides the theory for the arbitrary precise guaranteed bounds of the homogenized matrix calculated with local fields provided by the FFT-based FEM. We start in Sec. 3.1 with the definition of finite dimensional spaces, the spaces of trigonometric polynomials. Then we follow with fully discrete spaces that are the analogue to Helmholtz decomposition spaces. The FFT-based FEM is described in Sec. 3.3; additionally to [31], the theory for the non-odd number of discretization points is provided. Sec. 3.4 is dedicated to the connection between the primal and the dual formulations in the fully discrete setting. Finally, the calculation of the upper-lower bounds is explained in Sec. 3.5.

In the sequel, vector $\boldsymbol{N} \in \mathbb{N}^d$ is reserved for a number of discretization points, then scalar $|\boldsymbol{N}|_\Pi := \prod_\alpha N_\alpha$ denotes the number of degrees of freedom. If N_α is odd (even) for all α we talk about the odd (even) number of discretization points, otherwise about the non-odd ones. A reduced and a full index sets state for

$$\mathbb{Z}^d_{\boldsymbol{N}} = \left\{\boldsymbol{k} \in \mathbb{Z}^d : |k_\alpha| < \frac{N_\alpha}{2}\right\}, \qquad \underline{\mathbb{Z}}^d_{\boldsymbol{N}} = \left\{\boldsymbol{k} \in \mathbb{Z}^d : -\frac{N_\alpha}{2} \leq k_\alpha < \frac{N_\alpha}{2}\right\}. \tag{3.1}$$

A multi-index notation is employed, in which $\mathbb{R}^{\boldsymbol{N}}$ represents $\mathbb{R}^{N_1\times\dots\times N_d}$. Set $\mathbb{R}^{d\times\boldsymbol{N}}$ represents the space of vectors v with components $\mathsf{v}_\alpha^{\boldsymbol{n}}$ and $\mathbb{R}^{d\times d\times\boldsymbol{N}\times\boldsymbol{N}}$ the space of matrices A with components $\mathsf{A}_{\alpha\beta}^{\boldsymbol{nm}}$ for α,β and $\boldsymbol{n},\boldsymbol{m}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}$. Next, vectors $\mathsf{v}^{\boldsymbol{n}}\in\mathbb{R}^d$ for $\boldsymbol{n}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}$ and $\mathsf{v}_\alpha\in\mathbb{R}^{\boldsymbol{N}}$ for α represent subvectors of v with components $\mathsf{v}_\alpha^{\boldsymbol{n}}$. Analogically, submatrices $\mathsf{A}^{\boldsymbol{nm}}\in\mathbb{R}^{d\times d}$ and $\mathsf{A}_{\alpha\beta}\in\mathbb{R}^{\boldsymbol{N}\times\boldsymbol{N}}$ can be defined. A scalar product on set $\mathbb{R}^{d\times\boldsymbol{N}}$ is defined as $\big(\mathsf{u},\mathsf{v}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}}:=\sum_\alpha\sum_{\boldsymbol{n}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}}\mathsf{u}_\alpha^{\boldsymbol{n}}\mathsf{v}_\alpha^{\boldsymbol{n}}$ and matrix A by vector v multiplication as $(\mathsf{Av})_\alpha^{\boldsymbol{n}}:=\sum_\beta\sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}}\mathsf{A}_{\alpha\beta}^{\boldsymbol{nm}}\mathsf{v}_\beta^{\boldsymbol{m}}$. Matrix A is symmetric positive definite if relation $\mathsf{A}_{\alpha\beta}^{\boldsymbol{mn}}=\mathsf{A}_{\beta\alpha}^{\boldsymbol{nm}}$ holds for all components and inequality $\big(\mathsf{Av},\mathsf{v}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}}>0$ applies for arbitrary $\mathsf{v}\in\mathbb{R}^{d\times\boldsymbol{N}}$. We use the serif font for vectors $\mathsf{v}\in\mathbb{R}^{d\times\boldsymbol{N}}$ and matrices $\mathsf{A}\in\mathbb{R}^{d\times d\times\boldsymbol{N}\times\boldsymbol{N}}$ to distinguish from vectors $\boldsymbol{E}\in\mathbb{R}^d$ and matrices $\boldsymbol{A}_{\mathrm{eff}}\in\mathbb{R}^{d\times d}$ and from vector valued functions $\boldsymbol{v}\in L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$. In order to differentiate vectors and matrices for different number of discretization points $\boldsymbol{N}$, we write them with subscript, i.e. $\mathsf{v}_{\boldsymbol{N}}$ and $\mathsf{A}_{\boldsymbol{N}}$. Finally, operator $\oplus^\perp$ denotes the direct sum of mutually orthogonal subspaces, e.g. $\mathbb{R}^d=\boldsymbol{\epsilon}_1\oplus^\perp\boldsymbol{\epsilon}_2\oplus^\perp\dots\oplus^\perp\boldsymbol{\epsilon}_d$.

3.1. Trigonometric polynomials and their properties. In this section, we introduce the finite dimensional space of trigonometric polynomials and their properties. Definitions and lemmas adopted from [26] are amended for the non-odd number of discretization points $\boldsymbol{N}$ in order to satisfy the conformity of discretization — the requirement for the guaranteed upper-lower bounds.

NOTATION 3.1 (DFT). *For $\boldsymbol{N}\in\mathbb{N}^d$ we define, up to constant, unitary matrices $\mathsf{F}_{\boldsymbol{N}},\mathsf{F}_{\boldsymbol{N}}^{-1}\in\mathbb{C}^{d\times d\times\boldsymbol{N}\times\boldsymbol{N}}$ of the Discrete Fourier transform (DFT) and its inverse (iDFT) as*

$$\mathsf{F}_{\boldsymbol{N}}=\frac{1}{|\boldsymbol{N}|_\Pi}\big(\delta_{\alpha\beta}\omega_{\boldsymbol{N}}^{-\boldsymbol{mn}}\big)_{\alpha,\beta=1,\dots,d}^{\boldsymbol{m},\boldsymbol{n}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}}\qquad\qquad\mathsf{F}_{\boldsymbol{N}}^{-1}=\big(\delta_{\alpha\beta}\omega_{\boldsymbol{N}}^{\boldsymbol{mn}}\big)_{\alpha,\beta=1,\dots,d}^{\boldsymbol{m},\boldsymbol{n}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}}.$$

where $\delta_{\alpha\beta}$ is Kronecker delta and $\omega_{\boldsymbol{N}}^{\boldsymbol{mn}}=\exp\left(2\pi\mathrm{i}\sum_\alpha\frac{m_\alpha n_\alpha}{N_\alpha}\right)$ with $\boldsymbol{m},\boldsymbol{n}\in\mathbb{Z}^d$ are their components.

DEFINITION 3.2 (nodal points, basis functions). *Let $\boldsymbol{N}\in\mathbb{N}^d$. We define nodal points of periodic unit cell $\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{n}}=\sum_\alpha\frac{2Y_\alpha n_\alpha}{N_\alpha}\boldsymbol{\epsilon}_\alpha$ and Fourier and shape basis functions*

$$\varphi_{\boldsymbol{n}}(\boldsymbol{x})=\exp\left(\pi\mathrm{i}\sum_\alpha\frac{n_\alpha x_\alpha}{Y_\alpha}\right),\quad\varphi_{\boldsymbol{N},\boldsymbol{n}}(\boldsymbol{x})=\frac{1}{|\boldsymbol{N}|_\Pi}\sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}}\omega_{\boldsymbol{N}}^{-\boldsymbol{mn}}\varphi_{\boldsymbol{m}}(\boldsymbol{x}),\quad\boldsymbol{n}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}.$$

LEMMA 3.3 (Properties of $\varphi_{\boldsymbol{m}}$ and $\varphi_{\boldsymbol{N},\boldsymbol{m}}$). *Let $\boldsymbol{m},\boldsymbol{n}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}$, then*

$$\big(\varphi_{\boldsymbol{m}},\varphi_{\boldsymbol{n}}\big)_{L^2_{\mathrm{per}}}=\delta_{\boldsymbol{mn}}\qquad\qquad\varphi_{\boldsymbol{n}}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{m}})=\omega_{\boldsymbol{N}}^{\boldsymbol{mn}}\tag{3.2a}$$

$$\varphi_{\boldsymbol{N},\boldsymbol{m}}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{n}})=\delta_{\boldsymbol{mn}}\qquad\qquad\big(\varphi_{\boldsymbol{N},\boldsymbol{m}},\varphi_{\boldsymbol{N},\boldsymbol{n}}\big)_{L^2_{\mathrm{per}}}=\frac{\delta_{\boldsymbol{mn}}}{|\boldsymbol{N}|_\Pi}\tag{3.2b}$$

Proof. The proof, which can be also found in [31], is the consequence of direct calculation; the proof of last equality comes from orthogonality of vectors $(\omega_{\boldsymbol{N}}^{\boldsymbol{mn}})^{\boldsymbol{m}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}}$ for $\boldsymbol{n}\in\underline{\mathbb{Z}}^d_{\boldsymbol{N}}$ in $\mathbb{C}^{\boldsymbol{N}}$. □

DEFINITION 3.4 (Trigonometric polynomials). *For $\boldsymbol{N} \in \mathbb{N}^d$, we define the spaces of trigonometric polynomials $\mathscr{T}_{\boldsymbol{N}}$, $\tilde{\mathscr{T}}_{\boldsymbol{N}}$ and their vector valued versions $\mathscr{T}_{\boldsymbol{N}}^d$,$\tilde{\mathscr{T}}_{\boldsymbol{N}}^d$ as*

$$\mathscr{T}_{\boldsymbol{N}} = \Big\{ \sum_{\boldsymbol{n}\in\mathbb{Z}_{\boldsymbol{N}}^d} \hat{\mathsf{v}}^{\boldsymbol{n}} \varphi_{\boldsymbol{n}} : \hat{\mathsf{v}}^{\boldsymbol{n}} \in \mathbb{C}, \hat{\mathsf{v}}^{\boldsymbol{n}} = \overline{(\hat{\mathsf{v}}^{-\boldsymbol{n}})} \Big\}, \qquad \mathscr{T}_{\boldsymbol{N}}^d = \big\{ \boldsymbol{v} : v_\alpha \in \mathscr{T}_{\boldsymbol{N}} \big\}.$$

$$\tilde{\mathscr{T}}_{\boldsymbol{N}} = \Big\{ \sum_{\boldsymbol{n}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} \mathsf{v}^{\boldsymbol{n}} \varphi_{\boldsymbol{N},\boldsymbol{n}} : \mathsf{v}^{\boldsymbol{n}} \in \mathbb{R} \Big\}, \qquad \tilde{\mathscr{T}}_{\boldsymbol{N}}^d = \big\{ \boldsymbol{v} : v_\alpha \in \tilde{\mathscr{T}}_{\boldsymbol{N}} \big\}.$$

DEFINITION 3.5 (Interpolation projection). *Interpolation operator $\mathcal{Q}_{\boldsymbol{N}} : C_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) \to L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{C}^d)$ is defined as*

$$\mathcal{Q}_{\boldsymbol{N}}[f] = \sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} f(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{m}}) \varphi_{\boldsymbol{N},\boldsymbol{m}}.$$

LEMMA 3.6. *Interpolation operator $\mathcal{Q}_{\boldsymbol{N}}$ is a projection and its image is $\tilde{\mathscr{T}}_{\boldsymbol{N}}^d$.*

Proof. It comes from the definition of operator $\mathcal{Q}_{\boldsymbol{N}}$ in Def. 3.5, space $\tilde{\mathscr{T}}_{\boldsymbol{N}}^d$ in Def. 3.4, and second property in Lem. (3.2b). ∎

DEFINITION 3.7. *Operator $\mathcal{I}_{\boldsymbol{N}} : \tilde{\mathscr{T}}_{\boldsymbol{N}}^d \to \mathbb{R}^{d\times\boldsymbol{N}}$ stocks the values of the trigonometric polynomials at the nodal points to a vector $\mathcal{I}_{\boldsymbol{N}}[\boldsymbol{v}_{\boldsymbol{N}}] = \big(\boldsymbol{v}_{\boldsymbol{N},\alpha}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{n}})\big)_{\alpha=1,\dots,d}^{\boldsymbol{n}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d}$.*

LEMMA 3.8. *Operator $\mathcal{I}_{\boldsymbol{N}}$ from the previous definition is an isomorphism.*

Proof. The proof is the consequence of Def. 3.7 and of the second property in Eq. 3.2a. ∎

REMARK 3.9 (Connection of representation). *Trigonometric polynomial $\boldsymbol{v}_{\boldsymbol{N}} \in \tilde{\mathscr{T}}_{\boldsymbol{N}}^d$ can be uniquely expressed using both the Fourier coefficients and the function values at the nodal points*

$$\boldsymbol{v}_{\boldsymbol{N}} = \sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} \boldsymbol{v}_{\boldsymbol{N}}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{m}}) \varphi_{\boldsymbol{N},\boldsymbol{m}} = \sum_{\boldsymbol{n}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} \hat{\boldsymbol{v}}_{\boldsymbol{N}}(\boldsymbol{n}) \varphi_{\boldsymbol{n}}. \tag{3.3}$$

with a connection through the DFT as $\hat{\mathbf{v}}_{\boldsymbol{N}} = \mathsf{F}_{\boldsymbol{N}} \mathbf{v}_{\boldsymbol{N}}$, where $\mathbf{v}_{\boldsymbol{N}} := \mathcal{I}_{\boldsymbol{N}}[\boldsymbol{v}_{\boldsymbol{N}}]$ and the vector of Fourier coefficients $\hat{\mathbf{v}}_{\boldsymbol{N}}$ has components $(\hat{\mathbf{v}}_{\boldsymbol{N}})_\alpha^{\boldsymbol{m}} := \hat{v}_{\boldsymbol{N},\alpha}(\boldsymbol{m})$. Thus, space $\tilde{\mathscr{T}}_{\boldsymbol{N}}^d$ can be alternatively characterized with the Fourier coefficients as

$$\tilde{\mathscr{T}}_{\boldsymbol{N}}^d = \{ \sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} \hat{\mathbf{v}}_{\boldsymbol{N}}^{\boldsymbol{m}} \varphi_{\boldsymbol{m}} : \hat{\mathbf{v}}_{\boldsymbol{N}} \in \mathsf{F}_{\boldsymbol{N}}(\mathbb{R}^{d\times\boldsymbol{N}}) \}.$$

REMARK 3.10. *The trigonometric polynomials are real valued if the Fourier coefficients obey conjugate symmetry $\hat{\boldsymbol{v}}(\boldsymbol{n}) = \overline{\hat{\boldsymbol{v}}(-\boldsymbol{n})}$ for all $\boldsymbol{n} \in \mathbb{Z}^d$. In Def. 3.4, it is valid only for trigonometric polynomials $\mathscr{T}_{\boldsymbol{N}} \subset L^2_{\mathrm{per}}(\mathcal{Y})$ and $\mathscr{T}_{\boldsymbol{N}}^d \subset L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$.*

A peculiar situation occurs for space $\tilde{\mathscr{T}}_{\boldsymbol{N}}^d$. If $\boldsymbol{N}$ is odd, both spaces coincide $\mathscr{T}_{\boldsymbol{N}}^d = \tilde{\mathscr{T}}_{\boldsymbol{N}}^d$ as the index sets do $\mathbb{Z}_{\boldsymbol{N}}^d = \underline{\mathbb{Z}}_{\boldsymbol{N}}^d$; generally, the inclusion $\mathscr{T}_{\boldsymbol{N}}^d \subseteq \tilde{\mathscr{T}}_{\boldsymbol{N}}^d$ holds. Unfortunately, space $\tilde{\mathscr{T}}_{\boldsymbol{N}}^d$ fails to be real valued $\tilde{\mathscr{T}}_{\boldsymbol{N}}^d \nsubseteq L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ because the Fourier coefficients with frequencies $\boldsymbol{n} \in \underline{\mathbb{Z}}_{\boldsymbol{N}}^d \setminus \mathbb{Z}_{\boldsymbol{N}}^d$ miss the opposite counterpart with frequencies $-\boldsymbol{n}$.

3.2. Fully discrete spaces. In this section, we define fully discrete spaces — the spaces storing the values of the trigonometric polynomials at the nodal points. We show their connection to constant, curl-free, and divergence-free spaces with zero mean introduced in Eq. (2.1).

DEFINITION 3.11 (Fully discrete projections). *Let* $\hat{\mathbf{\Gamma}}^{(i)}(\boldsymbol{n})$ *for* $i = 0, 1, 2$ *and* $\boldsymbol{n} \in \mathbb{Z}^d$ *be the Fourier coefficients of integral kernels from Def. A.1. We define block diagonal matrices* $\hat{\mathbf{G}}^{(0)}$, $\hat{\mathbf{G}}_{\mathbf{0}}^{(i)}$, $\hat{\mathbf{G}}_{\boldsymbol{I}}^{(i)} \in \mathbb{R}^{d\times d\times \boldsymbol{N}\times \boldsymbol{N}}$ *for* $i = 1, 2$ *as*

$$(\hat{\mathbf{G}}^{(0)})^{\boldsymbol{mn}}_{\alpha\beta} = \hat{\mathbf{\Gamma}}^{(0)}_{\alpha\beta}(\boldsymbol{m})\delta_{\boldsymbol{mn}}$$

$$(\hat{\mathbf{G}}_{\mathbf{0}}^{(i)})^{\boldsymbol{mn}}_{\alpha\beta} = \begin{cases} \hat{\mathbf{\Gamma}}^{(i)}_{\alpha\beta}(\boldsymbol{m})\delta_{\boldsymbol{mn}}, \\ 0, \end{cases} \qquad (\hat{\mathbf{G}}_{\boldsymbol{I}}^{(i)})^{\boldsymbol{mn}}_{\alpha\beta} = \begin{cases} \hat{\mathbf{\Gamma}}^{(i)}_{\alpha\beta}(\boldsymbol{m})\delta_{\boldsymbol{mn}}, & \text{for } \boldsymbol{m} \in \mathbb{Z}_{\boldsymbol{N}} \\ \delta_{\alpha\beta}\delta_{\boldsymbol{mn}}, & \text{for } \boldsymbol{m} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}} \setminus \mathbb{Z}_{\boldsymbol{N}} \end{cases}$$

where $\boldsymbol{m}, \boldsymbol{n} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}}$. *Next, we define the matrices without hat as similarity transformation with matrix of DFT* $\mathbf{F}$, *i.e.* $\mathbf{G}^{(0)} = \mathbf{F}^{-1}\hat{\mathbf{G}}^{(0)}\mathbf{F}$, $\mathbf{G}_{\mathbf{0}}^{(i)} = \mathbf{F}^{-1}\hat{\mathbf{G}}_{\mathbf{0}}^{(i)}\mathbf{F}$ *and* $\mathbf{G}_{\boldsymbol{I}}^{(i)} = \mathbf{F}^{-1}\hat{\mathbf{G}}_{\boldsymbol{I}}^{(i)}\mathbf{F}$.

LEMMA 3.12. *Matrices* $\mathbf{G}^{(0)}$, $\mathbf{G}_{\mathbf{0}}^{(i)}$, $\mathbf{G}_{\boldsymbol{I}}^{(i)}$ *for* $i = 1, 2$, *defined in Def. 3.11, are orthogonal projections.*

Proof. From the matrices in Def. 3.11, we deduce

$$\mathbf{I} = \mathbf{G}^{(0)} + \mathbf{G}_{\mathbf{0}}^{(1)} + \mathbf{G}_{\boldsymbol{I}}^{(2)}, \qquad \mathbf{I} = \mathbf{G}^{(0)} + \mathbf{G}_{\boldsymbol{I}}^{(1)} + \mathbf{G}_{\mathbf{0}}^{(2)}.$$

with the help of the Fourier coefficients of integral kernels in Def. A.1. Moreover, a direct calculation shows that both of the triples $\hat{\mathbf{G}}^{(0)}$,$\hat{\mathbf{G}}_{\mathbf{0}}^{(1)}$,$\hat{\mathbf{G}}_{\boldsymbol{I}}^{(2)}$ and $\hat{\mathbf{G}}^{(0)}$,$\hat{\mathbf{G}}_{\boldsymbol{I}}^{(1)}$, $\hat{\mathbf{G}}_{\mathbf{0}}^{(2)}$ are mutually orthogonal projections, see also [17]. □

DEFINITION 3.13 (Finite dimensional subspaces). *With the previously defined projections, we introduce the subspaces of space* $\mathbb{R}^{d\times \boldsymbol{N}}$

$$\mathbb{U}_{\boldsymbol{N}} = \mathbf{G}^{(0)}\mathbb{R}^{d\times \boldsymbol{N}} \qquad \mathbb{E}_{\boldsymbol{N}} = \mathbf{G}_0^{(1)}\mathbb{R}^{d\times \boldsymbol{N}} \qquad \mathbb{J}_{\boldsymbol{N}} = \mathbf{G}_0^{(2)}\mathbb{R}^{d\times \boldsymbol{N}}$$
$$\bar{\mathbb{E}}_{\boldsymbol{N}} = \mathbf{G}_{\boldsymbol{I}}^{(1)}\mathbb{R}^{d\times \boldsymbol{N}} \qquad \bar{\mathbb{J}}_{\boldsymbol{N}} = \mathbf{G}_{\boldsymbol{I}}^{(2)}\mathbb{R}^{d\times \boldsymbol{N}}$$

and their trigonometric relatives

$$\mathscr{U}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}^{-1}[\mathbb{U}_{\boldsymbol{N}}] \qquad \mathscr{E}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}^{-1}[\mathbb{E}_{\boldsymbol{N}}] \qquad \mathscr{J}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}^{-1}[\mathbb{J}_{\boldsymbol{N}}]$$
$$\bar{\mathscr{E}}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}^{-1}[\bar{\mathbb{E}}_{\boldsymbol{N}}] \qquad \bar{\mathscr{J}}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}^{-1}[\bar{\mathbb{J}}_{\boldsymbol{N}}]$$

In Fig. 3.1, we introduce the relation diagram of subspaces based on the following

$$\begin{array}{ccccccccc}
\mathcal{I}_{\boldsymbol{N}}^{-1}[\mathbb{R}^{d\times \boldsymbol{N}}] & = & \bar{\mathscr{T}}^d_{\boldsymbol{N}} & \overset{\text{only for odd } \boldsymbol{N}}{\subsetneq} & L^2_{\text{per}}(\mathcal{Y};\mathbb{R}^d) & \overset{\text{only for odd } \boldsymbol{N}}{\supsetneq} & \bar{\mathscr{T}}^d_{\boldsymbol{N}} & = \mathcal{I}_{\boldsymbol{N}}^{-1}[\mathbb{R}^{d\times \boldsymbol{N}}] \\
\| & & \| & & \| & & \| & \| \\
\mathcal{I}_{\boldsymbol{N}}^{-1}[\mathbb{U}_{\boldsymbol{N}}] & = & \mathscr{U}_{\boldsymbol{N}} & = & \mathscr{U} & = & \mathscr{U}_{\boldsymbol{N}} & = \mathcal{I}_{\boldsymbol{N}}^{-1}[\mathbb{U}_{\boldsymbol{N}}] \\
\oplus^\perp & & \oplus^\perp & & \oplus^\perp & & \oplus^\perp & \oplus^\perp \\
\mathcal{I}_{\boldsymbol{N}}^{-1}[\mathbb{E}_{\boldsymbol{N}}] & = & \mathscr{E}_{\boldsymbol{N}} & \subsetneq & \mathscr{E} & \overset{\text{only for odd } \boldsymbol{N}}{\supsetneq} & \bar{\mathscr{E}}_{\boldsymbol{N}} & = \mathcal{I}_{\boldsymbol{N}}^{-1}[\bar{\mathbb{E}}_{\boldsymbol{N}}] \\
\oplus^\perp & & \oplus^\perp & & \oplus^\perp & & \oplus^\perp & \oplus^\perp \\
\mathcal{I}_{\boldsymbol{N}}^{-1}[\bar{\mathbb{J}}_{\boldsymbol{N}}] & = & \bar{\mathscr{J}}_{\boldsymbol{N}} & \overset{\text{only for odd } \boldsymbol{N}}{\subsetneq} & \mathscr{J} & \supsetneq & \mathscr{J}_{\boldsymbol{N}} & = \mathcal{I}_{\boldsymbol{N}}^{-1}[\mathbb{J}_{\boldsymbol{N}}]
\end{array}$$

FIG. 3.1. *The scheme of subspaces*

two lemmas. The first one is a base for the fully discrete formulation of GAwNI in Sec. 3.3 and especially for a treatment with dual formulation Sec. 3.4.

LEMMA 3.14. *For the subspaces from Def. 3.13, the following three conditions hold:*

(i) Space $\mathbb{R}^{d\times \boldsymbol{N}}$ *can be split into three mutually orthogonal subspaces*

$$\mathbb{R}^{d\times \boldsymbol{N}} = \mathbb{U}_{\boldsymbol{N}} \oplus^{\perp} \mathbb{E}_{\boldsymbol{N}} \oplus^{\perp} \bar{\mathbb{J}}_{\boldsymbol{N}}, \qquad \mathbb{R}^{d\times \boldsymbol{N}} = \mathbb{U}_{\boldsymbol{N}} \oplus^{\perp} \bar{\mathbb{E}}_{\boldsymbol{N}} \oplus^{\perp} \mathbb{J}_{\boldsymbol{N}}. \tag{3.4}$$

(ii) The subspaces with tilde enlarge the original one, i.e. $\mathbb{E}_{\boldsymbol{N}} \subseteq \bar{\mathbb{E}}_{\boldsymbol{N}}$, $\mathbb{J}_{\boldsymbol{N}} \subseteq \bar{\mathbb{J}}_{\boldsymbol{N}}$.

(iii) If $\boldsymbol{N} \in \mathbb{N}^d$ *is odd (for all elements) it simplifies to* $\mathbb{E}_{\boldsymbol{N}} = \bar{\mathbb{E}}_{\boldsymbol{N}}$ *and* $\mathbb{J}_{\boldsymbol{N}} = \bar{\mathbb{J}}_{\boldsymbol{N}}$.

Proof. The proof is based on Lem. 3.12. Since both properties, constitution of identity and orthogonality, hold, all subspaces from the lemma are the subsets of set $\mathbb{R}^{d\times \boldsymbol{N}}$ rather than set $\mathbb{C}^{d\times \boldsymbol{N}}$; decomposition (3.4) thus holds. Finally, if $\boldsymbol{N}$ is odd, the index sets $\underline{\mathbb{Z}}^d_{\boldsymbol{N}}$ and $\mathbb{Z}^d_{\boldsymbol{N}}$ coincide and the same stand for projections, $\mathsf{G}^{(1)}_{\boldsymbol{0}} = \mathsf{G}^{(1)}_{\boldsymbol{I}}$ and $\mathsf{G}^{(2)}_{\boldsymbol{0}} = \mathsf{G}^{(2)}_{\boldsymbol{I}}$, see Def. 3.11. □

LEMMA 3.15. *The scheme stated in Fig. 3.1 holds true.*

Proof. First, the middle column is the Helmholtz decomposition, in our special case provided in Lem. A.2.

Next, we utilize previous Lem. 3.14 and isomorphism of operator $\mathcal{I}_{\boldsymbol{N}}$ in Lem. 3.8 which prove the rest of the columns.

The equality between spaces in the first two and the last two columns comes from Def. 3.13, hence the relations between subspaces in second, third, and fourth column have to be established. Relation $\tilde{\mathscr{T}}^d_{\boldsymbol{N}} \subsetneq L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^{d\times \boldsymbol{N}})$ holding only for odd $\boldsymbol{N}$ is discussed in Rem. 3.10. Equality $\mathscr{U}_{\boldsymbol{N}} = \mathscr{U}$ is trivial as it contains only constant fields — the Fourier representation is exact.

Relation $\mathscr{E}_{\boldsymbol{N}} \subseteq \mathscr{E}$ comes as a consequence of projection $\mathsf{G}^{(1)}_0$ that is deduced from the kernel $\hat{\boldsymbol{\Gamma}}^{(1)}$ of continuous projection defined in Appendix. A; defective frequencies $\boldsymbol{n} \in \underline{\mathbb{Z}}^d_{\boldsymbol{N}} \setminus \mathbb{Z}^d_{\boldsymbol{N}}$ are erased, see Rem. 3.10. The last inclusion $\mathscr{J}_{\boldsymbol{N}} \subseteq \mathscr{J}$ is an analogue. □

REMARK 3.16. *The previous proof yields the alternative characterization of the conforming subspaces:* $\mathscr{E}_{\boldsymbol{N}} = \mathscr{E} \bigcap \mathscr{T}^d_{\boldsymbol{N}}$ *and* $\mathscr{J}_{\boldsymbol{N}} = \mathscr{J} \bigcap \mathscr{T}^d_{\boldsymbol{N}}$.

3.3. FFT-based Finite Element Method. This section provides the overview of the FFT-based FEM. The theory for odd number of discretization points $\boldsymbol{N}$ is described in [31] including convergence results. Here, we extend the situation for non-odd number of discretization points $\boldsymbol{N}$ in a way to provide conforming approximations.

DEFINITION 3.17 (GAwNI). *Let material coefficients* $\boldsymbol{A}$ *have continuos coefficients. Galerkin approximations with numerical integration (GAwNI) of the primal and the dual homogenization problems, Def. 2.3, state as: find discrete homogenized matrices* $\boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff},\boldsymbol{N}}, \boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff,D},\boldsymbol{N}} \in \mathbb{R}^{d\times d}$ *satisfying following relations for arbitrary macroscopic loads* $\boldsymbol{E}, \boldsymbol{J} \in \mathbb{R}^d$

$$\big(\boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff},\boldsymbol{N}}\boldsymbol{E},\boldsymbol{E}\big)_{\mathbb{R}^d} = \inf_{\boldsymbol{e}_{\boldsymbol{N}}\in\mathscr{E}_{\boldsymbol{N}}} \big(\mathcal{Q}_{\boldsymbol{N}}[\boldsymbol{A}(\boldsymbol{E}+\boldsymbol{e}_{\boldsymbol{N}})], \boldsymbol{E}+\boldsymbol{e}_{\boldsymbol{N}}\big)_{L^2_{\mathrm{per}}}, \tag{3.5a}$$

$$\big((\boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff,D},\boldsymbol{N}})^{-1}\boldsymbol{J},\boldsymbol{J}\big)_{\mathbb{R}^d} = \inf_{\boldsymbol{\jmath}_{\boldsymbol{N}}\in\mathscr{J}_{\boldsymbol{N}}} \big(\mathcal{Q}_{\boldsymbol{N}}[\boldsymbol{A}^{-1}(\boldsymbol{J}+\boldsymbol{\jmath}_{\boldsymbol{N}})], \boldsymbol{J}+\boldsymbol{\jmath}_{\boldsymbol{N}}\big)_{L^2_{\mathrm{per}}}. \tag{3.5b}$$

REMARK 3.18. *The scalar products on the right-hand side in* (3.5) *are real valued and hence the discrete homogenized matrices are. Although, functions* $\mathcal{Q}_{\boldsymbol{N}}[\boldsymbol{A}^{-1}(\boldsymbol{E}+$

$\boldsymbol{e}_{\boldsymbol{N}})], \mathcal{Q}_{\boldsymbol{N}}[\boldsymbol{A}^{-1}(\boldsymbol{J}+\boldsymbol{\jmath}_{\boldsymbol{N}})] \in \tilde{\mathcal{T}}_{\boldsymbol{N}}^d$ *generally fail to be only real valued, see Rem. 3.10, the defective Fourier coefficients with frequencies* $\boldsymbol{n} \in \underline{\mathbb{Z}}_{\boldsymbol{N}}^d \setminus \mathbb{Z}_{\boldsymbol{N}}^d$ *are eliminated by the space of test functions* $\mathcal{T}_{\boldsymbol{N}}^d$.

DEFINITION 3.19 (Fully discrete formulations of GAwNI). *Find* $\tilde{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}^{\mathrm{FFTH}}, \tilde{\boldsymbol{A}}_{\mathrm{eff,D},\boldsymbol{N}}^{\mathrm{FFTH}} \in \mathbb{R}^{d\times d}$ *satisfying following relations for arbitrary macroscopic loads* $\boldsymbol{E}, \boldsymbol{J} \in \mathbb{R}^d$

$$\big(\tilde{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}^{\mathrm{FFTH}}\boldsymbol{E}, \boldsymbol{E}\big)_{\mathbb{R}^d} = \frac{1}{|\boldsymbol{N}|_\Pi} \inf_{\mathbf{e}_{\boldsymbol{N}}\in\mathbb{E}_{\boldsymbol{N}}} \big(\mathbf{A}_{\boldsymbol{N}}(\mathbf{E}_{\boldsymbol{N}}+\mathbf{e}_{\boldsymbol{N}}), \mathbf{E}_{\boldsymbol{N}}+\mathbf{e}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}} \tag{3.6a}$$

$$\big((\tilde{\boldsymbol{A}}_{\mathrm{eff,D},\boldsymbol{N}}^{\mathrm{FFTH}})^{-1}\boldsymbol{J}, \boldsymbol{J}\big)_{\mathbb{R}^d} = \frac{1}{|\boldsymbol{N}|_\Pi} \inf_{\mathbf{j}_{\boldsymbol{N}}\in\mathbb{J}_{\boldsymbol{N}}} \big(\mathbf{A}_{\boldsymbol{N}}^{-1}(\mathbf{J}_{\boldsymbol{N}}+\mathbf{j}_{\boldsymbol{N}}), \mathbf{J}_{\boldsymbol{N}}+\mathbf{j}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}} \tag{3.6b}$$

where $\mathbf{E}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}[\boldsymbol{E}] \in \mathbb{U}_{\boldsymbol{N}}$, $\mathbf{J}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}[\boldsymbol{J}] \in \mathbb{J}_{\boldsymbol{N}}$, *and* $\mathbf{A}_{\boldsymbol{N}} \in \mathbb{R}^{d\times d\times\boldsymbol{N}\times\boldsymbol{N}}$ *with components* $\mathbf{A}_{\boldsymbol{N},\alpha\beta}^{\boldsymbol{mn}} = \boldsymbol{A}_{\alpha\beta}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{m}})\delta_{\boldsymbol{mn}}$ *for* α,β *and* $\boldsymbol{m},\boldsymbol{n} \in \underline{\mathbb{Z}}_{\boldsymbol{N}}^d$.

REMARK 3.20. *Minimizers* $\tilde{\boldsymbol{e}}_{\boldsymbol{N}}^{(\alpha)}$, $\tilde{\boldsymbol{\jmath}}_{\boldsymbol{N}}^{(\alpha)}$, *and* $\tilde{\mathbf{e}}_{\boldsymbol{N}}^{(\alpha)}$, $\tilde{\mathbf{j}}_{\boldsymbol{N}}^{(\alpha)}$ *of both formulations corresponding to unitary macroscopic fields* $\boldsymbol{\epsilon}_\alpha$ *are called discrete unitary minimizers. The existence and uniqueness is provided in following lemma.*

LEMMA 3.21. *The homogenized matrices of both previous formulations, Def. 3.17 and 3.19, coincide* $\boldsymbol{A}_{eff,\boldsymbol{N}}^{\mathrm{FFTH}} = \tilde{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}^{\mathrm{FFTH}}$, $\boldsymbol{A}_{\mathrm{eff,D},\boldsymbol{N}}^{\mathrm{FFTH}} = \tilde{\boldsymbol{A}}_{\mathrm{eff,D},\boldsymbol{N}}^{\mathrm{FFTH}}$. *Moreover, discrete minimizers* $\tilde{\boldsymbol{e}}_{\boldsymbol{N}}^{(\alpha)}$, $\tilde{\boldsymbol{\jmath}}_{\boldsymbol{N}}^{(\alpha)}$ *and* $\tilde{\mathbf{e}}_{\boldsymbol{N}}^{(\alpha)}$, $\tilde{\mathbf{j}}_{\boldsymbol{N}}^{(\alpha)}$ *of both formulations exist, are unique, and are connected to each other* $\mathcal{I}_{\boldsymbol{N}}[\tilde{\boldsymbol{e}}_{\boldsymbol{N}}] = \tilde{\mathbf{e}}_{\boldsymbol{N}}$, $\mathcal{I}_{\boldsymbol{N}}[\tilde{\boldsymbol{\jmath}}_{\boldsymbol{N}}] = \tilde{\mathbf{j}}_{\boldsymbol{N}}$.

Proof. Although, the proof is a generalization of that in [31], it follows the same ideas: the linearity of scalar product with the definition of interpolation operator $\mathcal{Q}_{\boldsymbol{N}}$ and the property of space basis functions $\varphi_{\boldsymbol{N},\boldsymbol{k}}$, and the second property in (3.2b). For completeness, we provide calculation

$$\begin{aligned}
&\big(\mathcal{Q}_{\boldsymbol{N}}[\boldsymbol{A}(\boldsymbol{E}+\tilde{\boldsymbol{e}}_{\boldsymbol{N}})], \boldsymbol{E}+\tilde{\boldsymbol{e}}_{\boldsymbol{N}}\big)_{L^2_{\mathrm{per}}} = \\
&= \Big(\sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} \boldsymbol{A}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{m}})\big[\boldsymbol{E}+\tilde{\boldsymbol{e}}_{\boldsymbol{N}}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{m}})\big]\varphi_{\boldsymbol{N},\boldsymbol{m}}, \sum_{\boldsymbol{n}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} \big[\boldsymbol{E}+\tilde{\boldsymbol{e}}_{\boldsymbol{N}}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{n}})\big]\varphi_{\boldsymbol{N},\boldsymbol{n}}\Big)_{L^2_{\mathrm{per}}} \\
&= \sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d}\sum_{\boldsymbol{n}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} \boldsymbol{A}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{m}})\big[\boldsymbol{E}+\tilde{\boldsymbol{e}}_{\boldsymbol{N}}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{m}})\big]\big[\boldsymbol{E}+\tilde{\boldsymbol{e}}_{\boldsymbol{N}}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{n}})\big]\big(\varphi_{\boldsymbol{N},\boldsymbol{m}}, \varphi_{\boldsymbol{N},\boldsymbol{n}}\big)_{L^2_{\mathrm{per}}} \\
&= \sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d}\sum_{\boldsymbol{n}\in\underline{\mathbb{Z}}_{\boldsymbol{N}}^d} \mathbf{A}_{\boldsymbol{N}}^{\boldsymbol{m}}\big[\mathbf{E}_{\boldsymbol{N}}^{\boldsymbol{m}}+\tilde{\mathbf{e}}_{\boldsymbol{N}}^{\boldsymbol{m}}\big]\big[\mathbf{E}_{\boldsymbol{N}}^{\boldsymbol{n}}+\tilde{\mathbf{e}}_{\boldsymbol{N}}^{\boldsymbol{n}}\big]\frac{\delta_{\boldsymbol{mn}}}{|\boldsymbol{N}|_\Pi} = \frac{1}{|\boldsymbol{N}|_\Pi}\big(\mathbf{A}_{\boldsymbol{N}}[\mathbf{E}_{\boldsymbol{N}}+\tilde{\mathbf{e}}_{\boldsymbol{N}}], \mathbf{E}_{\boldsymbol{N}}+\tilde{\mathbf{e}}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}}
\end{aligned}$$

The existence and the uniqueness of minimizers are provided due to symmetric and positive definite matrix $\mathbf{A}_{\boldsymbol{N}}$. □

REMARK 3.22 (Solution of GAwNI by Conjugate gradients). *According to [33], the minimizers of fully discrete formulations are equivalent to weak formulations, similar to those in Def. 2.5. Moreover, the solutions can be found by the means of Conjugate gradients applied to linear systems* $\mathbf{C}\mathbf{x} = \mathbf{b}$ *and* $\mathbf{C}'\mathbf{x}' = \mathbf{b}'$ *defined for particular* α *as*

$$\underbrace{\mathbf{G}_0^{(1)}\mathbf{A}_{\boldsymbol{N}}}_{\mathbf{C}}\underbrace{\mathbf{e}_{\boldsymbol{N}}^{(\alpha)}}_{\mathbf{x}} = \underbrace{-\mathbf{G}_0^{(1)}\mathbf{A}_{\boldsymbol{N}}\mathbf{E}_{\boldsymbol{N}}^{(\alpha)}}_{\mathbf{b}} \qquad \underbrace{\mathbf{G}_0^{(2)}\mathbf{A}_{\boldsymbol{N}}^{-1}}_{\mathbf{C}'}\underbrace{\mathbf{j}_{\boldsymbol{N}}^{(\alpha)}}_{\mathbf{x}'} = \underbrace{-\mathbf{G}_0^{(2)}\mathbf{A}_{\boldsymbol{N}}^{-1}\mathbf{J}_{\boldsymbol{N}}^{(\alpha)}}_{\mathbf{b}'}$$

for initial approximations $\mathbf{x}_0$, $\mathbf{x}_0'$ *that have to belong to the appropriate subspaces,* $\mathbb{E}_{\boldsymbol{N}}$, $\mathbb{J}_{\boldsymbol{N}}$ *resp. Easily, Conjugate gradients minimizes the quadratic forms in the fully discrete formulations 3.6.*

LEMMA 3.23 (Convergence of discrete minimizers). *Let material coefficients $\boldsymbol{A}$ be continuous and sufficiently regular in order to minimizers $\boldsymbol{e}^{(\alpha)}$, $\boldsymbol{\jmath}^{(\alpha)}$, Def. 2.5, be sufficiently regular — particularly having all weak partial derivatives up to order μ in the space $L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$. Then the sequence of discrete minimizers $\boldsymbol{e}_{\boldsymbol{N}}^{(\alpha)}$, $\boldsymbol{\jmath}_{\boldsymbol{N}}^{(\alpha)}$, Def. 3.17, converge to the minimizers, i.e.*

$$\|\boldsymbol{e}-\boldsymbol{e}_{\boldsymbol{N}}\|_{L^2_{\mathrm{per}}} \leq C\left(\min_{\alpha} N_\alpha\right)^{-\mu} \to 0 \quad \textit{for } \min_{\alpha} N_\alpha \to \infty$$

where constant C is independent of $\boldsymbol{N}$; it depends only on the material coefficients, its positive definiteness, norm, and regularity.

3.4. Connection of primal and dual formulations. This section is dedicated to the connection between the primal and the dual formulations in the fully discrete setting (3.6). General theory for dual problems can be found for example in [7]. We start with the statement of general lemma.

LEMMA 3.24 (Perturbation duality theorem — page 54 in [7]). *Let $\mathbb{V}$ and $\mathbb{X}$ be Hilbert spaces and $\Phi : \mathbb{V}\times\mathbb{X}\to\mathbb{R}$ be a continuous functional, convex, coercive for the first variable ($\lim_{\mathbf{v}\in\mathbb{V},\|\mathbf{v}\|\to\infty}\Phi(\mathbf{v},0)=\infty$), and satisfying following property: there exist $\mathbf{v}_0\in\mathbb{V}$ such that $\Phi(\mathbf{v}_0,\cdot)$ is finite and continuous around $0\in\mathbb{V}$. Then the primal and the dual problems*

$$\min_{\mathbf{v}\in\mathbb{V}}\Phi(\mathbf{v},0) \qquad\qquad \max_{\mathbf{x}^*\in\mathbb{X}} -\Phi^*(0;\mathbf{x}^*) \tag{3.7}$$

have the solutions and the extremal values are equal to one another. Here Φ^* is the usual Fenchel conjugate function defined on $\mathbb{V}\times\mathbb{X}$ as

$$\Phi^*(\mathbf{v}^*;\mathbf{x}^*) := \sup_{\mathbf{v}\in\mathbb{V},\mathbf{x}\in\mathbb{X}}\left[\left(\mathbf{v}^*,\mathbf{v}\right)_{\mathbb{V}}+\left(\mathbf{x}^*,\mathbf{x}\right)_{\mathbb{X}}-\Phi(\mathbf{v},\mathbf{x})\right]. \tag{3.8}$$

The following lemma is an application of previous lemma on the homogenization problem in fully discrete setting; it is sufficiently general to apply for arbitrary number of discretization points. Both the lemma and the proof are analogy to Proposition 2.2 and Corollary 2.3 in [6] that are stated for continuous formulations.

LEMMA 3.25 (Transformation to dual formulation). *Let $\mathbf{A}_{\boldsymbol{N}}\in\mathbb{R}^{d\times d\times\boldsymbol{N}\times\boldsymbol{N}}$ be symmetric positive definite matrix, $\mathbb{V}_{\boldsymbol{N}}$ be a proper nontrivial subspace of $\mathbb{U}_{\boldsymbol{N}}^{\perp}=\{\mathbf{v}\in\mathbb{R}^{d\times\boldsymbol{N}} : \mathbf{v}\cdot\mathbf{u}=0 \textit{ for all } \mathbf{u}\in\mathbb{U}_{\boldsymbol{N}}\}$, and the following primal problem be set: find $\tilde{\boldsymbol{A}}_{\mathrm{eff}}\in\mathbb{R}^d$ satisfying the following relation for arbitrary macroscopic load $\boldsymbol{E}\in\mathbb{R}^d$*

$$\left(\tilde{\boldsymbol{A}}_{\mathrm{eff}}\boldsymbol{E},\boldsymbol{E}\right)_{\mathbb{R}^d}=\frac{1}{|\boldsymbol{N}|_\Pi}\inf_{\mathbf{e}_{\boldsymbol{N}}\in\mathbb{V}_{\boldsymbol{N}}}\left(\mathbf{A}_{\boldsymbol{N}}(\mathbf{E}_{\boldsymbol{N}}+\mathbf{e}_{\boldsymbol{N}}),\mathbf{E}_{\boldsymbol{N}}+\mathbf{e}_{\boldsymbol{N}}\right)_{\mathbb{R}^{d\times\boldsymbol{N}}}$$

where $\mathbf{E}_{\boldsymbol{N}}:=\mathcal{I}_{\boldsymbol{N}}[\boldsymbol{E}]\in\mathbb{U}_{\boldsymbol{N}}$. Then the problem is equivalent to the dual problem: find $\tilde{\boldsymbol{A}}_{\mathrm{eff}}\in\mathbb{R}^d$ satisfying the following relation for arbitrary macroscopic load $\boldsymbol{J}\in\mathbb{R}^d$

$$\left(\tilde{\boldsymbol{A}}_{\mathrm{eff}}^{-1}\boldsymbol{J},\boldsymbol{J}\right)_{\mathbb{R}^d}=\frac{1}{|\boldsymbol{N}|_\Pi}\inf_{\mathbf{j}_{\boldsymbol{N}}\in\mathbb{W}_{\boldsymbol{N}}}\left(\mathbf{A}_{\boldsymbol{N}}^{-1}(\mathbf{J}_{\boldsymbol{N}}+\mathbf{j}_{\boldsymbol{N}}),\mathbf{J}_{\boldsymbol{N}}+\mathbf{j}_{\boldsymbol{N}}\right)_{\mathbb{R}^{d\times\boldsymbol{N}}}$$

where $\mathbf{J}_{\boldsymbol{N}}=\mathcal{I}_{\boldsymbol{N}}[\boldsymbol{J}]\in\mathbb{U}_{\boldsymbol{N}}$ and $\mathbb{W}_{\boldsymbol{N}}=(\mathbb{U}_{\boldsymbol{N}}\oplus\mathbb{V}_{\boldsymbol{N}})^{\perp}$. Moreover, between macroscopic fields $\boldsymbol{E}$, $\boldsymbol{J}$ and between minimizers $\tilde{\mathbf{e}}_{\boldsymbol{N}}^{(\boldsymbol{E})}$, $\tilde{\mathbf{j}}_{\boldsymbol{N}}^{(\boldsymbol{J})}$ of the primal and the dual formulations, the following relations hold

$$\boldsymbol{J}=\tilde{\boldsymbol{A}}_{\mathrm{eff}}\boldsymbol{E}, \qquad\qquad \mathbf{J}_{\boldsymbol{N}}+\tilde{\mathbf{j}}_{\boldsymbol{N}}^{(\boldsymbol{J})}=\mathbf{A}_{\boldsymbol{N}}[\mathbf{E}_{\boldsymbol{N}}+\tilde{\mathbf{e}}_{\boldsymbol{N}}^{(\boldsymbol{E})}]. \tag{3.9}$$

Proof. First, we define a function $\Phi : \mathbb{V}_{\boldsymbol{N}} \times \mathbb{R}^{d\times \boldsymbol{N}} \to \mathbb{R}$ as

$$\Phi(\mathbf{e}_{\boldsymbol{N}}, \mathbf{x}) := \frac{1}{2}\big(\mathbf{A}_{\boldsymbol{N}}[\mathbf{E}_{\boldsymbol{N}} + \mathbf{e}_{\boldsymbol{N}} + \mathbf{x}], [\mathbf{E}_{\boldsymbol{N}} + \mathbf{e}_{\boldsymbol{N}} + \mathbf{x}]\big)_{\mathbb{R}^{d\times \boldsymbol{N}}}.$$

Since matrix $\mathbf{A}_{\boldsymbol{N}}$ is symmetric positive definite and the undergoing spaces are finite dimensional, the assumptions of previous Lem. 3.25 are satisfied. Hence the primal formulation reformulated with Φ to

$$\big(\boldsymbol{A}_{\text{eff}}\boldsymbol{E}, \boldsymbol{E}\big)_{\mathbb{R}^d} = \frac{2}{|\boldsymbol{N}|_\Pi} \inf_{\mathbf{e}_{\boldsymbol{N}}\in\mathbb{V}_{\boldsymbol{N}}} \Phi(\mathbf{e}_{\boldsymbol{N}}, \mathbf{0}),$$

is equivalent to the dual formulation

$$\big(\boldsymbol{A}_{\text{eff}}\boldsymbol{E}, \boldsymbol{E}\big)_{\mathbb{R}^d} = \frac{2}{|\boldsymbol{N}|_\Pi} \sup_{\mathbf{x}^*\in\mathbb{R}^{d\times \boldsymbol{N}}} -\Phi^*(\mathbf{0}, \mathbf{x}^*)$$

where $\mathbf{0} \in \mathbb{R}^{d\times \boldsymbol{N}}$ is a vector with all components equal to zero and $\Phi^* : \mathbb{V}_{\boldsymbol{N}} \times \mathbb{R}^{d\times \boldsymbol{N}} \to \mathbb{R}$ is the Fenchel conjugate function, see Eq. (3.8).

Using substitution $\mathbf{x}' = \mathbf{E}_{\boldsymbol{N}} + \mathbf{e}_{\boldsymbol{N}} + \mathbf{x}$ where $\mathbf{x}'$ covers the whole space $\mathbb{R}^{d\times \boldsymbol{N}}$, we deduce

$$\begin{aligned}
\big(\boldsymbol{A}_{\text{eff}}\boldsymbol{E}, \boldsymbol{E}\big)_{\mathbb{R}^d} &= \frac{2}{|\boldsymbol{N}|_\Pi} \sup_{\mathbf{x}^*_{\boldsymbol{N}}\in\mathbb{R}^{d\times \boldsymbol{N}}} -\Phi^*(\mathbf{0}, \mathbf{x}^*) \\
&= \frac{2}{|\boldsymbol{N}|_\Pi} \sup_{\mathbf{x}^*\in\mathbb{R}^{d\times \boldsymbol{N}}} \left[- \sup_{\mathbf{e}_{\boldsymbol{N}}\in\mathbb{V}_{\boldsymbol{N}}, \mathbf{x}'\in\mathbb{R}^{d\times \boldsymbol{N}}} \left(\big(\mathbf{x}^*, \mathbf{x}' - \mathbf{E}_{\boldsymbol{N}} - \mathbf{e}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times \boldsymbol{N}}} - \frac{1}{2}\big(\mathbf{A}_{\boldsymbol{N}}\mathbf{x}', \mathbf{x}'\big)_{\mathbb{R}^{d\times \boldsymbol{N}}} \right) \right] \\
&= \frac{2}{|\boldsymbol{N}|_\Pi} \sup_{\mathbf{x}^*\in\mathbb{R}^{d\times \boldsymbol{N}}} \left[\big(\mathbf{x}^*, \mathbf{E}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times \boldsymbol{N}}} + \inf_{\mathbf{e}_{\boldsymbol{N}}\in\mathbb{V}_{\boldsymbol{N}}} \big(\mathbf{x}^*, \mathbf{e}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times \boldsymbol{N}}} \right. \\
&\qquad \left. - \sup_{\mathbf{x}'\in\mathbb{R}^{d\times \boldsymbol{N}}} \left(\big(\mathbf{x}^*, \mathbf{x}'\big)_{\mathbb{R}^{d\times \boldsymbol{N}}} - \frac{1}{2}\big(\mathbf{A}_{\boldsymbol{N}}\mathbf{x}', \mathbf{x}'\big)_{\mathbb{R}^{d\times \boldsymbol{N}}} \right) \right]
\end{aligned}$$

We focus on the supreme in the last equation where the equilibrium point satisfies $\mathbf{A}_{\boldsymbol{N}}\mathbf{x}' = \mathbf{x}^*$. Since $\mathbf{A}_{\boldsymbol{N}}$ is symmetric positive definite, we have $\mathbf{x}' = \mathbf{A}_{\boldsymbol{N}}^{-1}\mathbf{x}^*$. Therefore, the inner supremum is simplified to

$$\sup_{\mathbf{x}'\in\mathbb{R}^{d\times \boldsymbol{N}}} \left(\big(\mathbf{x}^*, \mathbf{x}'\big)_{\mathbb{R}^{d\times \boldsymbol{N}}} - \frac{1}{2}\big(\mathbf{A}_{\boldsymbol{N}}\mathbf{x}', \mathbf{x}'\big)_{\mathbb{R}^{d\times \boldsymbol{N}}} \right) = \frac{1}{2}\big(\mathbf{A}_{\boldsymbol{N}}^{-1}\mathbf{x}^*, \mathbf{x}^*\big)_{\mathbb{R}^{d\times \boldsymbol{N}}}.$$

The inner infimum equals to minus the indicator function of $V_{\boldsymbol{N}}^\perp$, explicitly

$$\inf_{\mathbf{e}_{\boldsymbol{N}}\in\mathbb{V}_{\boldsymbol{N}}} \big(\mathbf{x}^*, \mathbf{e}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times \boldsymbol{N}}} = \begin{cases} 0, & \text{for } \mathbf{x}^* \in V_{\boldsymbol{N}}^\perp \\ -\infty, & \text{otherwise.} \end{cases}$$

Hence, we can omit the inf term while restrict the supremum to space $V_{\boldsymbol{N}}^\perp$. It leads to

$$\big(\boldsymbol{A}_{\text{eff}}\boldsymbol{E}, \boldsymbol{E}\big)_{\mathbb{R}^d} = \frac{2}{|\boldsymbol{N}|_\Pi} \sup_{\mathbf{x}^*\in\mathbb{V}_{\boldsymbol{N}}^\perp} \left[\big(\mathbf{x}^*, \mathbf{E}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times \boldsymbol{N}}} - \frac{1}{2}\big(\mathbf{A}_{\boldsymbol{N}}^{-1}\mathbf{x}^*, \mathbf{x}^*\big)_{\mathbb{R}^{d\times \boldsymbol{N}}} \right].$$

Since the following inclusion holds $\mathbb{U}_{\boldsymbol{N}} \subsetneq V_{\boldsymbol{N}}^\perp$, space $V_{\boldsymbol{N}}^\perp$ can be decomposed $V_{\boldsymbol{N}}^\perp = \mathbb{U}_{\boldsymbol{N}} \oplus^\perp \mathbb{W}_{\boldsymbol{N}}$ where $\mathbb{W}_{\boldsymbol{N}} = \{\mathbf{w} \in V_{\boldsymbol{N}}^\perp : \mathbf{w}\cdot\mathbf{u} = 0 \text{ for all } \mathbf{u} \in \mathbb{U}_{\boldsymbol{N}}\}$. Hence, we state

$$\big(\boldsymbol{A}_{\text{eff}}\boldsymbol{E}, \boldsymbol{E}\big)_{\mathbb{R}^d} = \frac{2}{|\boldsymbol{N}|_\Pi} \sup_{\mathbf{J}_{\boldsymbol{N}}\in\mathbb{U}_{\boldsymbol{N}}} \left[\big(\mathbf{J}_{\boldsymbol{N}}, \mathbf{E}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times \boldsymbol{N}}} - \inf_{\mathbf{j}_{\boldsymbol{N}}\in W_{\boldsymbol{N}}} \frac{1}{2}\big(\mathbf{A}_{\boldsymbol{N}}^{-1}(\mathbf{J}_{\boldsymbol{N}} + \mathbf{j}_{\boldsymbol{N}}), \mathbf{J}_{\boldsymbol{N}} + \mathbf{j}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times \boldsymbol{N}}} \right]$$

where $\mathsf{x}^* = \mathsf{J}_{\boldsymbol{N}} + \mathsf{j}_{\boldsymbol{N}}$ such that $\mathsf{J}_{\boldsymbol{N}} \in \mathbb{U}_{\boldsymbol{N}}$, $\mathsf{j}_{\boldsymbol{N}} \in \mathbb{W}_{\boldsymbol{N}}$. Then we define matrix $\boldsymbol{B}_{\rm eff} \in \mathbb{R}^{d\times d}$ satisfying for arbitrary $\boldsymbol{J} \in \mathbb{R}^d$

$$\big(\boldsymbol{B}_{\rm eff}\boldsymbol{J},\boldsymbol{J}\big)_{\mathbb{R}^d} = \frac{1}{|\boldsymbol{N}|_\Pi} \inf_{\mathsf{j}_{\boldsymbol{N}}\in V_{\boldsymbol{N}}} \big(\mathsf{A}_{\boldsymbol{N}}^{-1}(\mathsf{J}_{\boldsymbol{N}}+\mathsf{j}_{\boldsymbol{N}}),\mathsf{J}_{\boldsymbol{N}}+\mathsf{j}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}}$$

where $\mathsf{J}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}[\boldsymbol{J}]$.

Now, we show $\boldsymbol{B}_{\rm eff} = \boldsymbol{A}_{\rm eff}^{-1}$, the first identity in Eq. (3.9). With obvious identity

$$\frac{1}{|\boldsymbol{N}|_\Pi}\big(\mathsf{J}_{\boldsymbol{N}},\mathsf{E}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}} = \big(\boldsymbol{J},\boldsymbol{E}\big)_{\mathbb{R}^d},$$

the dual problem becomes

$$\big(\boldsymbol{A}_{\rm eff}\boldsymbol{E},\boldsymbol{E}\big)_{\mathbb{R}^d} = \sup_{\boldsymbol{J}\in\mathbb{R}^d}\big[2\big(\boldsymbol{J},\boldsymbol{E}\big)_{\mathbb{R}^d} - \big(\boldsymbol{B}_{\rm eff}\boldsymbol{J},\boldsymbol{J}\big)_{\mathbb{R}^d}\big]$$

and comply with the equilibrium state $\boldsymbol{B}_{\rm eff}\boldsymbol{J} = \boldsymbol{E}$. The matrix $\boldsymbol{B}_{\rm eff}$ is symmetric positive definite as $\mathsf{A}_{\boldsymbol{N}}^{-1}$ is (see e.g. [4]), whence substition of $\boldsymbol{J} = \boldsymbol{B}_{\rm eff}^{-1}\boldsymbol{E}$ into the dual formulation leads to required identity.

The second identity in Eq. (3.9) follows from equations obtained during the proof, particularly $\mathsf{A}_{\boldsymbol{N}}(\mathsf{E}_{\boldsymbol{N}}+\mathsf{v}+\mathsf{x}) = \mathsf{x}^* = \mathsf{J}_{\boldsymbol{N}}+\mathsf{j}_{\boldsymbol{N}}$, and the fact that the primal formulation is obtained for $\mathsf{x} = 0$. □

Corollary 3.26 (Special case of the odd number of discretization points $\boldsymbol{N}$). *Let $\boldsymbol{N} \in \mathbb{N}^d$ be odd, and the fully discrete formulations 3.6 be defined. Then:*

(i) Both the primal and the dual homogenized matrices coincide $\boldsymbol{A}_{{\rm eff},\boldsymbol{N}}^{\rm FFTH} = \boldsymbol{A}_{{\rm eff,D},\boldsymbol{N}}^{\rm FFTH}$.

(ii) Primal and dual discrete minimizers $\tilde{\mathsf{e}}_{\boldsymbol{N}}^{(\alpha)}$, $\tilde{\mathsf{j}}_{\boldsymbol{N}}^{(\alpha)}$ are related

$$\boldsymbol{\epsilon}_\beta + \tilde{\mathsf{j}}_{\boldsymbol{N}}^{(\beta)} = \mathsf{A}_{\boldsymbol{N}}\sum_\alpha E_\alpha(\boldsymbol{\epsilon}_\alpha + \tilde{\mathsf{e}}_{\boldsymbol{N}}^{(\alpha)}) \tag{3.10}$$

where $\boldsymbol{E} = (\boldsymbol{A}_{{\rm eff},\boldsymbol{N}}^{\rm FFTH})^{-1}\boldsymbol{\epsilon}_\beta$.

Proof. The proof is the consequence of Lem. 3.25 for $\mathbb{V}_{\boldsymbol{N}} = \mathbb{E}_{\boldsymbol{N}}$, $\mathbb{W}_{\boldsymbol{N}} = \mathbb{J}_{\boldsymbol{N}}$, and decomposition $\mathbb{R}^{d\times\boldsymbol{N}} = \mathbb{U}_{\boldsymbol{N}} \oplus \mathbb{E}_{\boldsymbol{N}} \oplus \mathbb{J}_{\boldsymbol{N}}$ stated in Lem. 3.14. □

Corollary 3.27 (General case of the arbitrary number of discretization points $\boldsymbol{N}$). *Let we have following homogenization problems: find $\tilde{\boldsymbol{A}}_{{\rm eff},\boldsymbol{N}}^{\rm FFTH}$, $\tilde{\boldsymbol{A}}_{{\rm eff,D},\boldsymbol{N}}^{\rm FFTH} \in \mathbb{R}^{d\times d}$ such that*

$$\big((\tilde{\boldsymbol{A}}_{{\rm eff},\boldsymbol{N}}^{\rm FFTH})^{-1}\boldsymbol{J},\boldsymbol{J}\big)_{\mathbb{R}^d} = \frac{1}{|\boldsymbol{N}|_\Pi}\inf_{\mathsf{j}_{\boldsymbol{N}}\in\mathbb{J}_{\boldsymbol{N}}}\big(\mathsf{A}_{\boldsymbol{N}}^{-1}(\mathsf{J}_{\boldsymbol{N}}+\mathsf{j}_{\boldsymbol{N}}),\mathsf{J}_{\boldsymbol{N}}+\mathsf{j}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}} \tag{3.11a}$$

$$\big(\tilde{\boldsymbol{A}}_{{\rm eff,D},\boldsymbol{N}}^{\rm FFTH}\boldsymbol{E},\boldsymbol{E}\big)_{\mathbb{R}^d} = \frac{1}{|\boldsymbol{N}|_\Pi}\inf_{\mathsf{e}_{\boldsymbol{N}}\in\mathbb{E}_{\boldsymbol{N}}}\big(\mathsf{A}_{\boldsymbol{N}}(\mathsf{E}_{\boldsymbol{N}}+\mathsf{e}_{\boldsymbol{N}}),\mathsf{E}_{\boldsymbol{N}}+\mathsf{e}_{\boldsymbol{N}}\big)_{\mathbb{R}^d} \tag{3.11b}$$

and let $\tilde{\mathsf{e}}_{\boldsymbol{N}}^{(\alpha)}$ and $\tilde{\mathsf{j}}_{\boldsymbol{N}}^{(\alpha)}$ be their approximate unitary minimizers. Then the following holds:

(i) The fully discrete primal and dual formulations, Eq. 3.6a and 3.6b in Def. 3.19 are equivalent to those here in Eq. (3.11a) *and* (3.11b) *resp. in the sense the homogenized matrices coincide $\boldsymbol{A}_{{\rm eff},\boldsymbol{N}}^{\rm FFTH} = \tilde{\boldsymbol{A}}_{{\rm eff},\boldsymbol{N}}^{\rm FFTH}$ and $\boldsymbol{A}_{{\rm eff,D},\boldsymbol{N}}^{\rm FFTH} = \tilde{\boldsymbol{A}}_{{\rm eff,D},\boldsymbol{N}}^{\rm FFTH}$.*

(ii) The discrete unitary minimizers $\tilde{\mathbf{e}}_{\boldsymbol{N}}^{(\beta)}$, $\tilde{\mathbf{j}}_{\boldsymbol{N}}^{(\alpha)}$ *of Eq. 3.6 can be expressed as a linear combination of these minimizers* $\bar{\mathbf{e}}_{\boldsymbol{N}}^{(\alpha)}$, $\bar{\mathbf{j}}_{\boldsymbol{N}}^{(\alpha)}$ *of Eq.* (3.11) *as*

$$\boldsymbol{\epsilon}_\beta + \tilde{\mathbf{e}}_{\boldsymbol{N}}^{(\beta)} = \mathbf{A}_{\boldsymbol{N}}^{-1} \sum_\alpha J_\alpha(\boldsymbol{\epsilon}_\alpha + \bar{\mathbf{j}}_{\boldsymbol{N}}^{(\alpha)}) \in \mathbb{E}_{\boldsymbol{N}} \tag{3.12a}$$

$$\boldsymbol{\epsilon}_\beta + \tilde{\mathbf{j}}_{\boldsymbol{N}}^{(\beta)} = \mathbf{A}_{\boldsymbol{N}} \sum_\alpha E_\alpha(\boldsymbol{\epsilon}_\alpha + \bar{\mathbf{e}}_{\boldsymbol{N}}^{(\alpha)}) \in \mathbb{J}_{\boldsymbol{N}} \tag{3.12b}$$

where the macroscopic quantities are set to $\boldsymbol{E} := (\boldsymbol{A}_{\mathrm{eff,D},\boldsymbol{N}}^{\mathrm{FFTH}})^{-1}\boldsymbol{\epsilon}_\beta$ *and* $\boldsymbol{J} := \boldsymbol{A}_{\mathrm{eff},\boldsymbol{N}}^{\mathrm{FFTH}}\boldsymbol{\epsilon}_\beta$.

(iii) The primal and the dual homogenized matrices can be compared

$$\boldsymbol{A}_{\mathrm{eff,D},\boldsymbol{N}}^{\mathrm{FFTH}} \preceq \boldsymbol{A}_{\mathrm{eff},\boldsymbol{N}}^{\mathrm{FFTH}}. \tag{3.13}$$

Proof. The proof is mainly the consequence of Lem. 3.25 and 3.14: for the primal fully discrete formulation with $\mathbb{V}_{\boldsymbol{N}} = \mathbb{E}_{\boldsymbol{N}}$, $\mathbb{W}_{\boldsymbol{N}} = \bar{\mathbb{J}}_{\boldsymbol{N}}$ and for the dual one with $\mathbb{V}_{\boldsymbol{N}} = \mathbb{J}_{\boldsymbol{N}}$, $\mathbb{W}_{\boldsymbol{N}} = \bar{\mathbb{E}}_{\boldsymbol{N}}$ — see Fig. 3.1 with the scheme of the subspaces. The inequality in 3.13 comes from relation $\mathbb{E}_{\boldsymbol{N}} \subseteq \bar{\mathbb{E}}_{\boldsymbol{N}}$, see Lem. 3.14, and a relation

$$\frac{1}{|\boldsymbol{N}|_\Pi} \inf_{\mathbf{e}_{\boldsymbol{N}} \in \bar{\mathbb{E}}_{\boldsymbol{N}}} \big(\mathbf{A}_{\boldsymbol{N}}(\mathbf{E}_{\boldsymbol{N}}+\mathbf{e}_{\boldsymbol{N}}), \mathbf{E}_{\boldsymbol{N}}+\mathbf{e}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}} \leq \frac{1}{|\boldsymbol{N}|_\Pi} \inf_{\mathbf{e}_{\boldsymbol{N}} \in \mathbb{E}_{\boldsymbol{N}}} \big(\mathbf{A}_{\boldsymbol{N}}(\mathbf{E}_{\boldsymbol{N}}+\mathbf{e}_{\boldsymbol{N}}), \mathbf{E}_{\boldsymbol{N}}+\mathbf{e}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d\times\boldsymbol{N}}}$$

holding for arbitrary fixed $\mathbf{E}_{\boldsymbol{N}} \in \mathbb{U}_{\boldsymbol{N}}$. □

REMARK 3.28. *The effective matrices* $\boldsymbol{A}_{\mathrm{eff},\boldsymbol{N}}^{\mathrm{FFTH}}$ *and* $\boldsymbol{A}_{\mathrm{eff,D},\boldsymbol{N}}^{\mathrm{FFTH}}$ *can be compared with relation* $\preceq$*, in the sense of quadratic norms, to none of the matrices* $\boldsymbol{A}_{\mathrm{eff}}$ $\underline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}$*, and* $\overline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}$ *as is numerically shown in Sec. 4.1.*

REMARK 3.29. *In engineering literature relating FFT-based homogenization, e.g. [19, 20, 16], the criterion for the numerical convergence of the primal approximate minimizers* $\mathbf{e}_{\boldsymbol{N}}$ *is based on an equilibrium condition of the dual fields* $\mathbf{A}_{\boldsymbol{N}}\mathbf{e}_{\boldsymbol{N}}$ *controlling to be divergence-free. However, this criterion is reasonable only for the odd-number of discretization points, cf. Eq.* (3.10)*, as observed in [20]; they also offer a remedy that exactly corresponds to the formulation in Eq.* (3.11a) *— the dual fields* (3.12a) *are then, if a convergence is reached, in appropriate subspace* $\mathscr{J}_{\boldsymbol{N}}$*, the space of divergence-free fields.*

3.5. Calculation of upper-lower bounds. The calculation of the upper-lower bounds of the homogenized matrix consists of the integral evaluation of type $\big(\boldsymbol{A}\boldsymbol{e}_{\boldsymbol{N}}^{(\alpha)}, \boldsymbol{e}_{\boldsymbol{N}}^{(\beta)}\big)_{L^2_{\mathrm{per}}}$ occurring in Def. 2.8. Generally, the integral cannot be evaluated in a closed form because of non-specific material coefficients. The idea is to adjust material coefficients to calculate the integrals accurately and efficiently and simultaneously keep the upper-lower bounds structure.

For an easier orientation among various homogenized matrices, we refer to their scheme in Fig. 3.2. The matrices $\boldsymbol{A}_{\mathrm{eff}}$, $\underline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}$, $\overline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}$, $\boldsymbol{A}_{\mathrm{eff},\boldsymbol{N}}$, and $\boldsymbol{D}_{\boldsymbol{N}}$ introduced in Def. 2.3, 2.8, 2.11 are in no relation to matrices $\boldsymbol{A}_{\mathrm{eff},\boldsymbol{N}}^{\mathrm{FFTH}}$, $\boldsymbol{A}_{\mathrm{eff,D},\boldsymbol{N}}^{\mathrm{FFTH}}$ from Def. 3.17, see Rem. 3.28.

In this section, we introduce approximations of upper-lower bounds $\tilde{\boldsymbol{A}}_{\mathrm{eff,D},\boldsymbol{N}}^{\mathrm{lin},\boldsymbol{M}}$, $\tilde{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}^{\mathrm{lin},\boldsymbol{M}}$ based on piecewise bilinear material coefficients and homogenized matrices $\underline{\underline{A}}_{\mathrm{eff},\boldsymbol{N}}^{\mathrm{lin},\boldsymbol{M}}$, $\overline{\overline{A}}_{\mathrm{eff},\boldsymbol{N}}^{\mathrm{lin},\boldsymbol{M}}$ based on piecewise constant material coefficients defined in a way to guaranty bounds. All four introduced matrices can be calculated efficiently by FFT algorithm, see Lem. 3.31 and Lem. 3.32.

$$
\begin{array}{ccccccccccc}
 & & & & \tilde{\boldsymbol{A}}^{\mathrm{lin},M}_{\mathrm{eff},\mathrm{D},\boldsymbol{N}} & & & & \tilde{\boldsymbol{A}}^{\mathrm{lin},M}_{\mathrm{eff},\boldsymbol{N}} & & \\
 & & & & \wr & & & & \wr & & \\
0 & \preceq & \underline{\underline{\boldsymbol{A}}}^{\mathrm{con},M}_{\mathrm{eff},\boldsymbol{N}} & \preceq & \underline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}} & \preceq & \boldsymbol{A}_{\mathrm{eff}} & \preceq & \overline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}} & \preceq & \overline{\overline{\boldsymbol{A}}}^{\mathrm{con},M}_{\mathrm{eff},\boldsymbol{N}} \\
 & & & & \| & & & & \| & & \\
 & & & & \boldsymbol{A}_{\mathrm{eff},\boldsymbol{N}} - \boldsymbol{D}_{\boldsymbol{N}} & \preceq & \boldsymbol{A}_{\mathrm{eff},\boldsymbol{N}} & \preceq & \boldsymbol{A}_{\mathrm{eff},\boldsymbol{N}} + \boldsymbol{D}_{\boldsymbol{N}} & & \\
 & & & & \nwarrow\!\!\!\!\!\!\!\!\nwarrow & & & & \nwarrow\!\!\!\!\!\!\!\!\nwarrow & & \\
 & & & & \boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff},\mathrm{D},\boldsymbol{N}} & \underset{\preceq}{\text{equality if } \boldsymbol{N} \text{ is odd}} & & & \boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff},\boldsymbol{N}} & &
\end{array}
$$

FIG. 3.2. *The overview of homogenized material bounds*

LEMMA 3.30 (Sufficient condition for the upper-lower bounds). *Let* $\boldsymbol{A} \in L^\infty_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^{d\times d}_{\mathrm{spd}})$ *be material coefficients and* $\overline{\boldsymbol{A}}, \underline{\boldsymbol{A}} \in L^\infty_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^{d\times d})$ *its upper and lower approximations satisfying*

$$\underline{\boldsymbol{A}}(\boldsymbol{x}) \preceq \boldsymbol{A}(\boldsymbol{x}) \preceq \overline{\boldsymbol{A}}(\boldsymbol{x}), \quad \text{for almost all } \boldsymbol{x} \in \mathcal{Y}. \tag{3.14}$$

Let $\tilde{\boldsymbol{e}}^{(\alpha)}_{\boldsymbol{N}} \in \mathscr{E}_{\boldsymbol{N}}$ *and* $\tilde{\boldsymbol{\jmath}}^{(\alpha)}_{\boldsymbol{N}} \in \mathscr{J}_{\boldsymbol{N}}$ *be unitary minimizers for material coefficients* $\boldsymbol{A}$*, cf. Def. 2.5. Then matrices* $\overline{\overline{\boldsymbol{A}}}_{\mathrm{eff}}, \underline{\underline{\boldsymbol{A}}}_{\mathrm{eff}} \in \mathbb{R}^{d\times d}$*, defined as*

$$(\overline{\overline{\boldsymbol{A}}}_{\mathrm{eff},\boldsymbol{N}})_{\alpha\beta} = \big(\overline{\boldsymbol{A}}(\boldsymbol{\epsilon}_\beta + \tilde{\boldsymbol{e}}^{(\beta)}_{\boldsymbol{N}}), \boldsymbol{\epsilon}_\alpha + \tilde{\boldsymbol{e}}^{(\alpha)}_{\boldsymbol{N}}\big)_{L^2_{\mathrm{per}}}, \tag{3.15a}$$

$$(\underline{\underline{\boldsymbol{A}}}^{-1}_{\mathrm{eff},\boldsymbol{N}})_{\alpha\beta} = \big(\underline{\boldsymbol{A}}^{-1}(\boldsymbol{\epsilon}_\beta + \tilde{\boldsymbol{\jmath}}^{(\beta)}_{\boldsymbol{N}}), \boldsymbol{\epsilon}_\alpha + \tilde{\boldsymbol{\jmath}}^{(\alpha)}_{\boldsymbol{N}}\big)_{L^2_{\mathrm{per}}}, \tag{3.15b}$$

comply with the upper-lower bound structure, i.e.

$$\underline{\underline{\boldsymbol{A}}}_{\mathrm{eff},\boldsymbol{N}} \preceq \underline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}} \preceq \boldsymbol{A}_{\mathrm{eff}} \preceq \overline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}} \preceq \overline{\overline{\boldsymbol{A}}}_{\mathrm{eff},\boldsymbol{N}}.$$

Proof. The inner inequalities $\underline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}} \preceq \boldsymbol{A}_{\mathrm{eff}} \preceq \overline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}$ are already proven in Lem. 2.10; the rest easily arise from assumed inequality (3.14) that is kept under integration. □

Next lemma provides a way for the calculation of the homogenized matrices by the FFT routine, Lem. 3.32. It requires the material coefficients to be expressed as a linear combination of some basis functions concentrated on the set of nodal points.

LEMMA 3.31 (Calculation of homogenized matrices). *Let* $\boldsymbol{u}_{\boldsymbol{N}}, \boldsymbol{v}_{\boldsymbol{N}} \in \mathscr{T}^d_{\boldsymbol{N}}$ *be trigonometric polynomials and* $\boldsymbol{A}_{\boldsymbol{M}} \in L^\infty_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^{d\times d})$ *for* $\boldsymbol{M} \in \mathbb{N}^d$ *be function explicitly expressed as*

$$\boldsymbol{A}_{\boldsymbol{M}}(\boldsymbol{x}) = \sum_{\boldsymbol{n}\in\mathbb{Z}^d_{\boldsymbol{M}}} \psi(\boldsymbol{x} + \boldsymbol{x}^{\boldsymbol{n}}_{\boldsymbol{M}})\mathsf{A}^{\boldsymbol{n}}_{\boldsymbol{M}}, \quad \boldsymbol{x} \in \mathcal{Y}$$

where $\psi \in L^\infty_{\mathrm{per}}(\mathcal{Y};\mathbb{R})$ *is some basis function and* $\mathsf{A}_{\boldsymbol{M}} \in \mathbb{R}^{d\times d\times \boldsymbol{M}}$*. Then the integrals of the type occurring in Eq.* (3.15) *can be calculated as*

$$\big(\boldsymbol{A}_{\boldsymbol{M}}\boldsymbol{u}_{\boldsymbol{N}}, \boldsymbol{v}_{\boldsymbol{N}}\big)_{L^2_{\mathrm{per}}} = \frac{1}{|\mathcal{Y}|_d} \sum_{\alpha,\beta} \sum_{\boldsymbol{m}\in\mathbb{Z}^d_{2\boldsymbol{N}}} w(\boldsymbol{m})\widehat{\mathsf{u}}\mathsf{v}^{\boldsymbol{m}}_{\boldsymbol{N},\alpha,\beta}\hat{\mathsf{A}}^{\boldsymbol{m}}_{\alpha\beta} \tag{3.16}$$

where integration weight $w(\boldsymbol{m})$ *is defined as* $w(\boldsymbol{m}) := \int_{\mathcal{Y}} \psi(\boldsymbol{x})\varphi_{\boldsymbol{m}}(\boldsymbol{x})$ *and factors*

$\widehat{\mathsf{uv}}^{\boldsymbol{m}}_{\boldsymbol{N},\beta,\alpha}$, $\hat{\mathsf{A}}^{\boldsymbol{m}}_{\alpha\beta}$ *are defined as*

$$\widehat{\mathsf{uv}}^{\boldsymbol{m}}_{\boldsymbol{N},\beta,\alpha} = \frac{1}{2|\boldsymbol{N}|_\Pi} \sum_{\boldsymbol{k}\in\underline{\mathbb{Z}}^d_{2\boldsymbol{N}}} \boldsymbol{u}_{\boldsymbol{N},\beta}(\boldsymbol{x}^{\boldsymbol{k}}_{2\boldsymbol{N}})\boldsymbol{v}_{\boldsymbol{N},\alpha}(\boldsymbol{x}^{\boldsymbol{k}}_{2\boldsymbol{N}})\omega^{-\boldsymbol{mk}}_{2\boldsymbol{N}} \tag{3.17a}$$

$$\hat{\mathsf{A}}^{\boldsymbol{m}}_{\alpha\beta} = \sum_{\boldsymbol{n}\in\underline{\mathbb{Z}}^d_{\boldsymbol{M}}} \mathsf{A}^{\boldsymbol{n}}_{\alpha\beta}\omega^{-\boldsymbol{mn}}_{\boldsymbol{M}} \tag{3.17b}$$

Proof. First, we note that $\int_{\mathcal{Y}} \psi(\boldsymbol{x}+\boldsymbol{x}^{\boldsymbol{n}}_{\boldsymbol{M}})\varphi_{\boldsymbol{m}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_{\mathcal{Y}} \psi(\boldsymbol{x})\varphi_{\boldsymbol{m}}(\boldsymbol{x}-\boldsymbol{x}^{\boldsymbol{n}}_{\boldsymbol{M}})\mathrm{d}\boldsymbol{x} = w(\boldsymbol{m})\omega^{-\boldsymbol{mn}}_{\boldsymbol{M}}$ where $\boldsymbol{m},\boldsymbol{n}\in\mathbb{Z}^d$. Since $u_{\boldsymbol{N},\beta}, v_{\boldsymbol{N},\alpha} \in \mathscr{T}_{\boldsymbol{N}}$, for their multiplication holds $u_{\boldsymbol{N},\beta}v_{\boldsymbol{N},\alpha} \in \mathscr{T}_{2\boldsymbol{N}}$. Thus, it can be expressed as interpolation through $2\boldsymbol{N}$ nodal points

$$\boldsymbol{u}_{\boldsymbol{N},\beta}\boldsymbol{v}_{\boldsymbol{N},\alpha} = \sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}^d_{2\boldsymbol{N}}} \widehat{\mathsf{uv}}^{\boldsymbol{m}}_{\boldsymbol{N},\beta,\alpha}\varphi_{\boldsymbol{m}} \tag{3.18}$$

Then, the direct calculation finish the proof

$$\begin{aligned}
\big(\boldsymbol{A}_{\boldsymbol{M}}\boldsymbol{u}_{\boldsymbol{N}},\boldsymbol{v}_{\boldsymbol{N}}\big)_{L^2_{\mathrm{per}}} &= \frac{1}{|\mathcal{Y}|_d}\sum_{\alpha,\beta}\int_{\mathcal{Y}} A_{\boldsymbol{M},\alpha\beta}(\boldsymbol{x})\boldsymbol{u}_{\boldsymbol{N},\beta}(\boldsymbol{x})\boldsymbol{v}_{\boldsymbol{N},\alpha}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{|\mathcal{Y}|_d}\sum_{\alpha,\beta}\sum_{\boldsymbol{n}\in\underline{\mathbb{Z}}^d_{\boldsymbol{M}}}\sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}^d_{2\boldsymbol{N}}} \mathsf{A}^{\boldsymbol{n}}_{\alpha\beta}\widehat{\mathsf{uv}}^{\boldsymbol{m}}_{\boldsymbol{N},\beta,\alpha}\int_{\mathcal{Y}}\psi(\boldsymbol{x}+\boldsymbol{x}^{\boldsymbol{n}}_{\boldsymbol{M}})\varphi_{\boldsymbol{m}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{|\mathcal{Y}|_d}\sum_{\alpha,\beta}\sum_{\boldsymbol{m}\in\underline{\mathbb{Z}}^d_{2\boldsymbol{N}}} \widehat{\mathsf{uv}}^{\boldsymbol{m}}_{\boldsymbol{N},\beta,\alpha}w(\boldsymbol{m})\sum_{\boldsymbol{n}\in\underline{\mathbb{Z}}^d_{\boldsymbol{M}}}\mathsf{A}^{\boldsymbol{n}}_{\alpha\beta}\omega^{-\boldsymbol{mn}}_{\boldsymbol{M}}.
\end{aligned}$$

□

Lemma 3.32 (Homogenized matrices by FFT algorithm). *Let the assumptions from the previous lemma be satisfied and, in addition, let* $\boldsymbol{M}\in\mathbb{N}^d$ *be such that* $M_\alpha \geq 2N_\alpha$ *then the formula in Eq.* (3.16) *can be calculated using the FFT algorithm of size* $2\boldsymbol{N}$ *and* $\boldsymbol{M}$ *resp.*

Proof. In Eq. (3.17a), the function values of $\boldsymbol{u}_{\boldsymbol{N},\beta}$, $\boldsymbol{v}_{\boldsymbol{N},\alpha}$ at nodal points $\boldsymbol{x}^{\boldsymbol{k}}_{2\boldsymbol{N}}$ can be calculated with inverse DFT of size $2\boldsymbol{N}$ — the Fourier coefficients with frequencies $\underline{\mathbb{Z}}^d_{2\boldsymbol{N}} - \mathbb{Z}^d_{\boldsymbol{N}}$ are completed with nils. Then Eq. (3.17) is, up to a constant, the DFT on space $\mathbb{R}^{2\boldsymbol{N}}$ and $\mathbb{R}^{\boldsymbol{M}}$ resp. — compare it to DFT matrix $\mathsf{F}^d_{\boldsymbol{N}}$ in Def. 3.1 acting on space $\mathbb{R}^{d\times\boldsymbol{N}}$. The Fourier coefficients of $\widehat{\mathsf{uv}}^{\boldsymbol{m}}_{\boldsymbol{N},\beta,\alpha}$ equal to zero for $\boldsymbol{m}\in\mathbb{Z}^d\setminus\mathbb{Z}^d_{2\boldsymbol{N}}$, it reveals requirement $M_\alpha \geq 2N_\alpha$. □

Up to now, we have shown that the homogenized properties can be calculated effectively using the FFT algorithm if the material coefficients are expressed as the linear combination of basis functions concentrated on the set of nodal points. Now, we show, in Def. 3.33, some examples of basis functions, to be utilized in Lem. 3.31, that can be used to calculate the homogenized matrices; the choice depends on the possibility to analytically express the integral weights, cf. Lem. 3.34.

Definition 3.33 (Constant and bilinear basis functions). *Let* $\boldsymbol{M}\in\mathbb{R}^d$ *be a parameter such that* $M_\alpha > 1$. *We say that functions* $\chi_{\boldsymbol{M}}, \mathrm{tri}_{\boldsymbol{M}} \in L^\infty_{\mathrm{per}}(\mathcal{Y};\mathbb{R})$, *defined on* $\mathcal{Y}$ *as*

$$\chi_{\boldsymbol{M}}(\boldsymbol{x}) = \begin{cases} 1, & |x_\alpha| < \frac{Y_\alpha}{M_\alpha} \text{ for all } \alpha \\ 0, & \text{otherwise} \end{cases}, \quad \mathrm{tri}_{\boldsymbol{M}}(\boldsymbol{x}) = \prod_\alpha \max\{1-|\frac{x_\alpha M_\alpha}{2Y_\alpha}|, 0\},$$

are, one by one, a constant and a bilinear basis functions.

LEMMA 3.34 (Weights of numerical integration). *Let* $\chi_{\boldsymbol{M}}$ *and* $\mathrm{tri}_{\boldsymbol{M}}$ *be the basis functions from Def. 3.33. Then for* $\boldsymbol{m} \in \mathbb{Z}^d$ *we state*

$$w^0_{\boldsymbol{M}}(\boldsymbol{m}) := \int_{\mathcal{Y}} \chi_{\boldsymbol{M}}(\boldsymbol{x})\varphi_{\boldsymbol{m}}(\boldsymbol{x})\mathrm{d}\mathrm{x} = \prod_\alpha \frac{2\mathrm{Y}_\alpha}{\mathrm{M}_\alpha}\,\mathrm{sinc}\left(\frac{\mathrm{m}_\alpha}{\mathrm{M}_\alpha}\right)$$
$$w^1_{\boldsymbol{M}}(\boldsymbol{m}) := \int_{\mathcal{Y}} \mathrm{tri}_{\boldsymbol{M}}(\boldsymbol{x})\varphi_{\boldsymbol{m}}(\boldsymbol{x})\mathrm{d}\mathrm{x} = \prod_\alpha \frac{2\mathrm{Y}_\alpha}{\mathrm{M}_\alpha}\,\mathrm{sinc}^2\left(\frac{\mathrm{m}_\alpha}{\mathrm{M}_\alpha}\right)$$

where $\mathrm{sinc}(x) := \begin{cases} 1, & x = 0 \\ \frac{\sin(\pi x)}{\pi x}, & x \neq 0 \end{cases}$.

Proof. For $\boldsymbol{m} \in \mathbb{Z}^d$ we calculate

$$\int_{\mathcal{Y}} \chi_{\boldsymbol{M}}\varphi_{\boldsymbol{m}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \prod_\alpha \int_{|x_\alpha|<\frac{Y_\alpha}{M_\alpha}} \exp(\mathrm{i}\pi\frac{x_\alpha m_\alpha}{Y_\alpha})\mathrm{d}\mathrm{x}_\alpha = \prod_\alpha \left[\frac{\mathrm{Y}_\alpha}{\mathrm{i}\pi\mathrm{m}_\alpha}\exp(\mathrm{i}\pi\frac{\mathrm{x}_\alpha\mathrm{m}_\alpha}{\mathrm{Y}_\alpha})\right]_{-\frac{\mathrm{Y}_\alpha}{\mathrm{M}_\alpha}}^{\frac{\mathrm{Y}_\alpha}{\mathrm{M}_\alpha}}$$
$$= \prod_\alpha \left[\frac{2Y_\alpha}{M_\alpha}\frac{\sin(\frac{\pi m_\alpha}{M_\alpha})}{\frac{\pi m_\alpha}{M_\alpha}}\right] = \prod_\alpha \frac{2Y_\alpha}{M_\alpha}\,\mathrm{sinc}\left(\frac{m_\alpha}{M_\alpha}\right)$$

Integral weights $w^1_{\boldsymbol{M}}(\boldsymbol{m})$ for bilinear basis functions $\mathrm{tri}_{\boldsymbol{M}}$ are calculated accordingly. □

REMARK 3.35. *Since we realize that the integral weights can be calculated for the basis functions shifted by* $\boldsymbol{h} \in \mathbb{R}^d$*, e.g.*

$$\int_{\mathcal{Y}} \chi_{\boldsymbol{N}}(\boldsymbol{x}+\boldsymbol{h})\varphi_{\boldsymbol{m}}(\boldsymbol{x})\mathrm{d}\mathrm{x} = \mathrm{w}^0_{\boldsymbol{N}}(\boldsymbol{m})\varphi_{\boldsymbol{m}}(-\boldsymbol{h}),$$

we can calculate the upper and lower bounds exactly if the material coefficients are expressed as some linear combination of the shifted basis functions from Def. 3.33. We note that the shifts by particular $\boldsymbol{h}$ *enable to calculate the bounds using the FFT algorithm, cf. Lem. 3.31 and Lem. 3.32.*

Next, the basis functions stated in Def. 3.33 are not the only suitable ones, for example the circle basis function defined for $\boldsymbol{x} \in \mathcal{Y}$ *as*

$$\mathrm{circ}_r(\boldsymbol{x}) = \begin{cases} 1, & \|\boldsymbol{x}\|_2 \leq r \\ 0, & \textit{otherwise} \end{cases}$$

produces the integral weight with the Bessel function of the first kind B_1*, i.e. for* $\xi_\alpha(\boldsymbol{m}) = \frac{m_\alpha}{Y_\alpha}$

$$w^{\mathrm{circ}}_{\boldsymbol{N}}(\boldsymbol{m}) := \int_{\mathcal{Y}} \mathrm{circ}_r(\boldsymbol{x})\varphi_{\boldsymbol{m}}(\boldsymbol{x})\mathrm{d}\mathrm{x} = \begin{cases} \pi r^2, & \boldsymbol{m} = \boldsymbol{0} \\ r^2\frac{B_1(2\pi r\|\boldsymbol{\xi}(\boldsymbol{m})\|_2)}{r\|\boldsymbol{\xi}(\boldsymbol{m})\|_2}, & \textit{otherwise} \end{cases}.$$

Now, we will define approximations of the guaranteed bounds with the basis functions stated in Def. 3.33; a piecewise constant approximation is defined in a way to still provide the guaranteed bounds while a piecewise bilinear approximation only approximate these bounds, see Lem. 3.30.

DEFINITION 3.36 (Constant and bilinear approximation of material coefficients). *Let $\boldsymbol{A} \in L^\infty_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^{d\times d}_{\mathrm{spd}})$ be material coefficients and $\boldsymbol{M} \in \mathbb{N}$ a parameter. Then we define functions $\overline{\boldsymbol{A}}^{\mathrm{con},\boldsymbol{M}}$, $\underline{\boldsymbol{B}}^{\mathrm{con},\boldsymbol{M}}$, $\boldsymbol{A}^{\mathrm{lin},\boldsymbol{M}}$, $\boldsymbol{B}^{\mathrm{lin},\boldsymbol{M}} \in L^\infty_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^{d\times d})$ for $\boldsymbol{x} \in \mathcal{Y}$ as*

$$\begin{aligned}
\overline{\boldsymbol{A}}^{\mathrm{con},\boldsymbol{M}}(\boldsymbol{x}) &= \sum_{\boldsymbol{n}\in\mathbb{Z}^d_{\boldsymbol{M}}} \chi_{\boldsymbol{M}}(\boldsymbol{x}+\boldsymbol{x}^{\boldsymbol{n}}_{\boldsymbol{M}})\overline{\mathsf{A}_{\boldsymbol{M}}}^{\mathrm{con},\boldsymbol{M},\boldsymbol{n}},\\
\underline{\boldsymbol{B}}^{\mathrm{con},\boldsymbol{M}}(\boldsymbol{x}) &= \sum_{\boldsymbol{n}\in\mathbb{Z}^d_{\boldsymbol{M}}} \chi_{\boldsymbol{M}}(\boldsymbol{x}+\boldsymbol{x}^{\boldsymbol{n}}_{\boldsymbol{M}})\underline{\mathsf{B}}^{\mathrm{con},\boldsymbol{M},\boldsymbol{n}},\\
\boldsymbol{A}^{\mathrm{lin},\boldsymbol{M}}(\boldsymbol{x}) &= \sum_{\boldsymbol{n}\in\mathbb{Z}^d_{\boldsymbol{M}}} \mathrm{tri}_{\boldsymbol{M}}(\boldsymbol{x}+\boldsymbol{x}^{\boldsymbol{n}}_{\boldsymbol{M}})\boldsymbol{A}(\boldsymbol{x}^{\boldsymbol{n}}_{\boldsymbol{M}}),\\
\boldsymbol{B}^{\mathrm{lin},\boldsymbol{M}}(\boldsymbol{x}) &= \sum_{\boldsymbol{n}\in\mathbb{Z}^d_{\boldsymbol{M}}} \mathrm{tri}_{\boldsymbol{M}}(\boldsymbol{x}+\boldsymbol{x}^{\boldsymbol{n}}_{\boldsymbol{M}})\boldsymbol{A}^{-1}(\boldsymbol{x}^{\boldsymbol{n}}_{\boldsymbol{M}}),
\end{aligned}$$

where matrices $\overline{\mathsf{A}_{\boldsymbol{M}}}^{\mathrm{con},\boldsymbol{M},\boldsymbol{n}}, \underline{\mathsf{B}}^{\mathrm{con},\boldsymbol{M},\boldsymbol{n}} \in \mathbb{R}^{d\times d}$ are defined as

$$\overline{\mathsf{A}_{\boldsymbol{M}}}^{\mathrm{con},\boldsymbol{M},\boldsymbol{n}} = \|\boldsymbol{A}(\boldsymbol{x}+\boldsymbol{x}^{\boldsymbol{n}}_{\boldsymbol{M}})\|_{L^\infty(\Omega_{\boldsymbol{M}};\mathbb{R}^{d\times d}_{\mathrm{spd}})}\boldsymbol{I}, \quad \underline{\mathsf{B}}^{\mathrm{con},\boldsymbol{M},\boldsymbol{n}} = \|\boldsymbol{A}^{-1}(\boldsymbol{x}+\boldsymbol{x}^{\boldsymbol{n}}_{\boldsymbol{M}})\|_{L^\infty(\Omega_{\boldsymbol{M}};\mathbb{R}^{d\times d}_{\mathrm{spd}})}\boldsymbol{I},$$

noting that L^∞_{per}-norm is restricted on $\Omega_{\boldsymbol{M}} = \prod_\alpha \left(-\frac{Y_\alpha}{M_\alpha}, \frac{Y_\alpha}{M_\alpha}\right)$. The matrices are called, one by one, upper constant, lower constant, upper bilinear, and lower bilinear approximation of material coefficients.

LEMMA 3.37 (Constant approximation of material coefficients). *The constant approximations of the material coefficients, Def. 3.36, satisfy the sufficient condition in Lem. 3.30 for guaranteeing bounds.*

Proof. It is necessary to show

$$\boldsymbol{A}(\boldsymbol{x}) \preceq \overline{\boldsymbol{A}}^{\mathrm{con},\boldsymbol{M}}(\boldsymbol{x}), \quad \boldsymbol{A}^{-1}(\boldsymbol{x}) \preceq \underline{\boldsymbol{B}}^{\mathrm{con},\boldsymbol{M}}(\boldsymbol{x}), \quad \left(\underline{\boldsymbol{B}}^{\mathrm{con},\boldsymbol{M}}\right)^{-1}(\boldsymbol{x}) \preceq \boldsymbol{A}(\boldsymbol{x}) \preceq \overline{\boldsymbol{A}}^{\mathrm{con},\boldsymbol{M}}(\boldsymbol{x}),$$

however, it is a direct consequence of the definition of the approximated material coefficients. □

DEFINITION 3.38 (Bounds and its approximation of homogenized material coefficients). *Let $\tilde{\boldsymbol{e}}^{(\alpha)}_{\boldsymbol{N}} \in \mathscr{E}_{\boldsymbol{N}}$ and $\tilde{\boldsymbol{\jmath}}^{(\alpha)}_{\boldsymbol{N}} \in \mathscr{J}_{\boldsymbol{N}}$ be approximations of unitary minimizers, e.g. from Def. 2.5, and $\overline{\boldsymbol{A}}^{\mathrm{con},\boldsymbol{M}}$, $\underline{\boldsymbol{B}}^{\mathrm{con},\boldsymbol{M}}$, $\boldsymbol{A}^{\mathrm{lin},\boldsymbol{M}}, \boldsymbol{B}^{\mathrm{lin},\boldsymbol{M}}$ for $\boldsymbol{M} \in \mathbb{N}^d$ be approximations of material coefficients from Def. 3.36. Then we define the bounds $\overline{\overline{A}}^{\mathrm{con},\boldsymbol{M}}_{\mathrm{eff},\boldsymbol{N}}$, $\underline{\underline{A}}^{\mathrm{con},\boldsymbol{M}}_{\mathrm{eff},\boldsymbol{N}}$, $\tilde{\boldsymbol{A}}^{\mathrm{lin},\boldsymbol{M}}_{\mathrm{eff},\boldsymbol{N}}$, $\tilde{\boldsymbol{A}}^{\mathrm{lin},\boldsymbol{M}}_{\mathrm{eff,D},\boldsymbol{N}} \in \mathbb{R}^{d\times d}$ of the homogenized matrix as*

$$\overline{\overline{A}}^{\mathrm{con},\boldsymbol{M}}_{\mathrm{eff},\boldsymbol{N},\alpha\beta} = \left(\overline{\boldsymbol{A}}^{\mathrm{con},\boldsymbol{M}}\boldsymbol{e}^{(\beta)}_{\boldsymbol{N}}, \boldsymbol{e}^{(\alpha)}_{\boldsymbol{N}}\right)_{L^2_{\mathrm{per}}}, \quad \left(\underline{\underline{\boldsymbol{A}}}^{\mathrm{con},\boldsymbol{M}}_{\mathrm{eff},\boldsymbol{N}}\right)^{-1}_{\alpha\beta} = \left(\underline{\boldsymbol{B}}^{\mathrm{con},\boldsymbol{M}}\boldsymbol{\jmath}^{(\beta)}_{\boldsymbol{N}}, \boldsymbol{\jmath}^{(\alpha)}_{\boldsymbol{N}}\right)_{L^2_{\mathrm{per}}}, \tag{3.19a}$$

$$\tilde{\boldsymbol{A}}^{\mathrm{lin},\boldsymbol{M}}_{\mathrm{eff},\boldsymbol{N},\alpha\beta} = \left(\boldsymbol{A}^{\mathrm{lin},\boldsymbol{M}}\boldsymbol{e}^{(\beta)}_{\boldsymbol{N}}, \boldsymbol{e}^{(\alpha)}_{\boldsymbol{N}}\right)_{L^2_{\mathrm{per}}}, \quad \left(\tilde{\boldsymbol{A}}^{\mathrm{lin},\boldsymbol{M}}_{\mathrm{eff,D},\boldsymbol{N}}\right)^{-1}_{\alpha\beta} = \left(\boldsymbol{B}^{\mathrm{lin},\boldsymbol{M}}\boldsymbol{\jmath}^{(\beta)}_{\boldsymbol{N}}, \boldsymbol{\jmath}^{(\alpha)}_{\boldsymbol{N}}\right)_{L^2_{\mathrm{per}}}. \tag{3.19b}$$

THEOREM 3.39 (Guaranteed bounds for piecewise constant material coefficients). *The bounds $\underline{\underline{\boldsymbol{A}}}^{\mathrm{con},\boldsymbol{M}}_{\mathrm{eff},\boldsymbol{N}}, \overline{\overline{\boldsymbol{A}}}^{\mathrm{con},\boldsymbol{M}}_{\mathrm{eff},\boldsymbol{N}} \in \mathbb{R}^{d\times d}$ from the previous definition are the guaranteed bounds, i.e.*

$$\underline{\underline{\boldsymbol{A}}}^{\mathrm{con},\boldsymbol{M}}_{\mathrm{eff},\boldsymbol{N}} \preceq \boldsymbol{A}_{\mathrm{eff}} \preceq \overline{\overline{\boldsymbol{A}}}^{\mathrm{con},\boldsymbol{M}}_{\mathrm{eff},\boldsymbol{N}}, \tag{3.20}$$

and can be calculated with the FFT algorithm.

Proof. Relation (3.20) is a corollary of Lem. 3.37. The possibility to caclulate it using the FFT algorithm is a consequence of formula (3.16) in Lem. 3.31 and discussion in Lem. 3.32. □

REMARK 3.40 (Homogenized bounds). *The bounds obtained with the bilinear approximations of the material coefficients are only approximation of the bounds from Def. 2.8, explicitly* $\tilde{\boldsymbol{A}}^{\mathrm{lin},\boldsymbol{M}}_{\mathrm{eff},\boldsymbol{N}} \approx \overline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}$, $\tilde{\boldsymbol{A}}^{\mathrm{lin},\boldsymbol{M}}_{\mathrm{eff,D},\boldsymbol{N}} \approx \underline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}$. *Nevertheless, it is still possible to calculate it using the FFT algorithm, cf. Lem. 3.32.*

4. Numerical experiments. This section is dedicated to numerical experiments. We discuss the practical aspects of computing the homogenized matrices, and then verify the theoretical results: the upper-lower bounds structure of homogenized matrices in Section 4.1, and the rate of convergence, particularly convergence of approximate solutions to continuous one and the convergence of the homogenized matrices, Sec. 4.2.

Finally, the numerical experiments are compared to the p-version of the Finite Element Method provided in [6], see Sec. 4.3.

REMARK 4.1. *The calculations are provided for a 2-dimensional problem with* $(2,2)$*-periodic material coefficients defined on periodic unit cell* $\mathcal{Y} = (-1,1)\times(-1,1) \subset \mathbb{R}^2$ *as*

$$\boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{I}[1 + \rho f(\boldsymbol{x})], \quad \boldsymbol{x} \in \mathcal{Y},$$

where $\boldsymbol{I} \in \mathbb{R}^{d\times d}_{\mathrm{spd}}$ *is the identity matrix,* ρ *corresponds to the phase ratio and is taken as* 10 *except Sec. 4.3 where it is taken as* $(10^{-3}-1)$*, and* $f : \mathcal{Y} \to \mathbb{R}$ *is a scalar nonnegative function defined explicitly for particular experiment — it controls the shape of inclusions and the regularity of material coefficients.*

Moreover, discrete minimizers $\boldsymbol{e}^{(\alpha)}_{\boldsymbol{N}} \in \mathscr{E}_{\boldsymbol{N}}$ *and* $\boldsymbol{\jmath}^{(\alpha)}_{\boldsymbol{N}} \in \mathscr{J}_{\boldsymbol{N}}$ *from Def. 3.17 are obtained for both odd and even number of discretization points, namely* $N = (n,n)$ *where either* $n \in \{5\cdot 3^\alpha | \alpha \in \mathbb{N}_0, 0 \le \alpha \le 6\}$ *or* $n \in \{2^\alpha | \alpha \in \mathbb{N}, 2 \le \alpha \le 10\}$*; both the sets of numbers are geometric series carrying ratios between succesive terms* λ *equal to either* 3 *or* 2*.*

LOGARITHM 4.2 (Calculation of guaranteed bounds of homogenized material properties). *Let* $\boldsymbol{A}_{\boldsymbol{M}} \in L^\infty_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^{d\times d})$ *be material coefficients. The algorithm for calculation of the homogenized matrices is composed of several steps:*

(i) Set the number of discretization points $\boldsymbol{N}$ *and assemble matrices* $\mathbf{A}_{\boldsymbol{N}}$, $\mathbf{A}^{-1}_{\boldsymbol{N}}$, $\hat{\mathbf{G}}^{(1)}_{\mathbf{0}}$, $\hat{\mathbf{G}}^{(2)}_{\mathbf{0}} \in \mathbb{R}^{d\times d\times \boldsymbol{N}\times \boldsymbol{N}}$ *defined in Def. 3.19 and 3.11. Since they are block diagonal, only the diagonals are stored and the matrix by vector multiplication is provided as element-wise multiplication.*

(ii) For α*, find discrete minimizers* $\tilde{\mathbf{e}}^{(\alpha)}_{\boldsymbol{N}} \in \mathbb{E}_{\boldsymbol{N}}$, $\tilde{\mathbf{j}}^{(\alpha)}_{\boldsymbol{N}} \in \mathbb{J}_{\boldsymbol{N}}$ *as the solutions of linear systems described in Rem. 3.22 for unitary macroscopic loads* $\boldsymbol{\epsilon}_\alpha$*. The convergence criterion is based on the norm of residuum* $\|\mathbf{r}\|_{\mathrm{CG}} \le \varepsilon$ *where* $\|\mathbf{r}\|^2_{\mathrm{CG}} := |\boldsymbol{N}|^{-1}_\Pi \langle \mathbf{r},\mathbf{r}\rangle_{\mathbb{R}^{d\times\boldsymbol{N}}}$*; the yielding value is set as* $\varepsilon = 10^{-10}$ *and initial approximate vectors of CG are set as zeros.*

(iii) Calculate, if possible as stated in Rem. 3.35, the exact upper and lower bounds $\overline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}$, $\underline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}$*, see Def. 2.8. Otherwise, evaluate the approximations of the upper-lower bounds* $\overline{\overline{A}}^{\mathrm{lin},\boldsymbol{M}}_{\mathrm{eff},\boldsymbol{N}}$, $\underline{\underline{A}}^{\mathrm{lin},\boldsymbol{M}}_{\mathrm{eff},\boldsymbol{N}}$, $\tilde{\boldsymbol{A}}^{\mathrm{con},\boldsymbol{M}}_{\mathrm{eff},\boldsymbol{N}}$, $\tilde{\boldsymbol{A}}^{\mathrm{con},\boldsymbol{M}}_{\mathrm{eff,D},\boldsymbol{N}}$*, Def. 3.38, according to Lem. 3.31 for some sufficiently large* $\boldsymbol{M} \in \mathbb{R}^d$*, cf. Lem. 3.32.*

REMARK 4.3 (Convergence criterion). *The norm for residuum* $\|\mathbf{r}\|_{\mathrm{CG}}$*, due to Plancherel's theorem, equals to* $\|\mathcal{I}_{\boldsymbol{N}}^{-1}[\mathbf{r}]\|_{L^2_{\mathrm{per}}}$*. The yielding value is set as small as possible to diminish an error caused by an inaccuracy in the solution of the linear systems.*

REMARK 4.4 (Avoiding the solution of dual formulation). *If* $\boldsymbol{N}$ *is odd, the dual discrete minimizers* $\tilde{\mathbf{j}}_{\boldsymbol{N}}^{(\alpha)}$ *can be obtained from Eq. 3.10 based on the assumption that the original minimizers* $\tilde{\mathbf{e}}_{\boldsymbol{N}}^{(\alpha)}$ *are the exact solutions of the corresponding linear systems, see Rem. 3.22. In reality, the linear systems are solved only approximately, thus it fails the dual minimizers to be from appropriate subspace* $\tilde{\mathbf{j}}_{\boldsymbol{N}}^{(\alpha)} \notin \mathbb{J}_{\boldsymbol{N}}$*. It can be saved with projection operator* $\mathbf{G}_{\mathbf{0}}^{(2)}$ *and, in case* $\mathbf{A}_{\boldsymbol{N}}$ *is badly conditioned, by providing couple of iterations of the dual formulation* (3.6b).

REMARK 4.5 (Interpolation operator $\mathcal{Q}_{\boldsymbol{N}}$ for non-continuous functions). *The discrete formulations, Def. 3.17, require the interpolation operator* $\mathcal{Q}_{\boldsymbol{N}}$ *to be well defined; it takes the function values and thus, originally, the operator is well defined, for example, on continuous functions.*

We assume the piecewise constant material coefficients and define the discrete formulations as it states in Def. 3.19. Generally, it still provides the upper-lower bounds, however, it can fail to converge. Alternatively, some regularization of material coefficients, see [31], can be performed to obtain convergence, nevertheless, generally arbitrary slow.

4.1. The behavior of the bounds of the homogenized matrix. In this section, we validate the properties of the upper-lower bounds of the homogenized matrices for two phase materials with three types of inclusions characterized by scalar functions f, see Rem. 4.1; explicitly, they are defined as

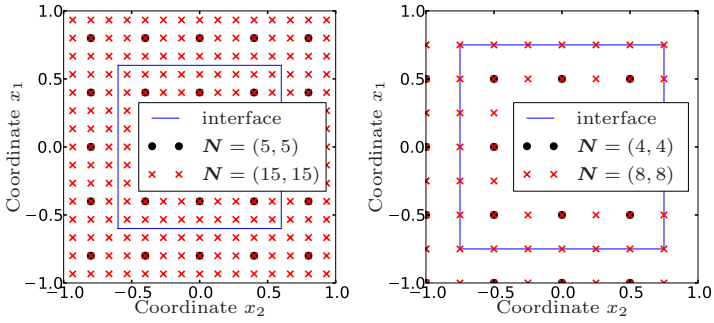
- square (S): $f(\boldsymbol{x}) = \begin{cases} 1, & \|\boldsymbol{x}\|_\infty < \frac{3}{5} \\ 0, & \text{otherwise} \end{cases}$,
- square (S1): $f(\boldsymbol{x}) = \begin{cases} 1, & \|\boldsymbol{x}\|_\infty < \frac{3}{4} \\ 0, & \text{otherwise} \end{cases}$,
- square (S2): $f(\boldsymbol{x}) = \begin{cases} 1, & \|\boldsymbol{x}\|_\infty \leq \frac{3}{4} \\ 0, & \text{otherwise} \end{cases}$.

Fig. 4.1 depicts the interface between phases and the nodal points sets, $\{\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{n}} \in \mathcal{Y} : \boldsymbol{n} \in \underline{\mathbb{Z}}_{\boldsymbol{N}}^d\}$, for particular $\boldsymbol{N}$. The squares (S1) and (S2) differ only at the interface having the 2-dimensional Lebesque's measure equal to zero, thus insignificant for the upper and lower bounds, see Def. 2.8.

The square (S) receive the interface exactly between the nodal points — it approximate well an inclusion contrary to the squares (S1) and (S2) representing the extreme cases of an inclusion approximation. The nodal points lie exactly on the boundary causing the difference in the discrete formulation and consequently in discrete minimizers $\boldsymbol{e}_{\boldsymbol{N}}^{(\alpha)}, \boldsymbol{\jmath}_{\boldsymbol{N}}^{(\alpha)}$ and upper-lower bounds $\overline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}, \underline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}$.

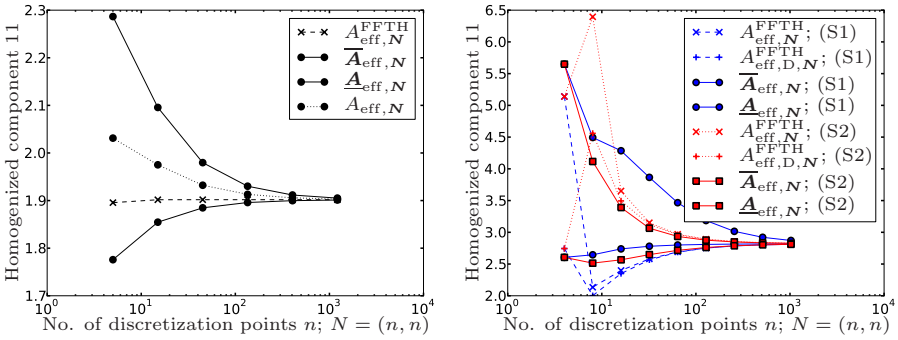
In Fig. 4.1, we demonstrate the properties of the homogenized matrices for their particular diagonal component. Inequality $\underline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}} \preceq \overline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}$ stated in Lem. 2.10 is satisfied and the error, difference between them, is approaching zero supporting Lem. 2.14.

For odd $\boldsymbol{N}$ in Fig. 4.1(a), the primal and the dual approximate homogenized matrices coincide $\boldsymbol{A}_{\mathrm{eff},\boldsymbol{N}}^{\mathrm{FFTH}} = \boldsymbol{A}_{\mathrm{eff,D},\boldsymbol{N}}^{\mathrm{FFTH}}$ as stated in Theorem 3.26. Moreover, it approximates properly the real homogenized coefficients $\boldsymbol{A}_{\mathrm{eff}}$ even for small $\boldsymbol{N}$ compared to



(a) square (S) (b) square (S1) and (S2) resp.

FIG. 4.1. *Periodic unit cell with nodal points and interfaces between phases*



(a) square (S); n is odd (b) square (S1) and (S2); n is even

FIG. 4.2. *The upper-lower bounds of the homogenized matrix for both the odd and the even number of discretization points*

the mean value $\boldsymbol{A}_{\mathrm{eff},\boldsymbol{N}} = \frac{1}{2}(\underline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}} + \overline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}})$ that overestimates.

For even $\boldsymbol{N}$ in Fig. 4.1(b), the approximate homogenized matrices satisfy a sharp inequality $\boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff,D},\boldsymbol{N}} \prec \boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff},\boldsymbol{N}}$ for both squares (S1) and (S2) confirming (3.13) in Theorem 3.27. Both the homogenized matrices $\boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff},\boldsymbol{N}}$, $\boldsymbol{A}^{\mathrm{FFTH}}_{\mathrm{eff,D},\boldsymbol{N}}$ either overestimates or underestimates even over the upper or the lower bounds. Exception is a case $\boldsymbol{N} = (4, 4)$ when the material coefficients coincide at the nodal points for both squares (S1) and (S2).

However, the primal and the dual homogenized matrices, in Eq. (3.6a) and (3.11b), differ substantially for small $\boldsymbol{N}$ alike subspaces $\mathscr{E}_{\boldsymbol{N}}$ and $\bar{\mathscr{E}}_{\boldsymbol{N}}$ do. Hence, the mean of the upper-lower bounds $\boldsymbol{A}_{\mathrm{eff},\boldsymbol{N}}$ provides the more accurate result. Albeit, from the design perspective, either the lower or the upper bound is chosen as the most reliable homogenized property depending on a design demand: resistance or conductivity.

4.2. Rate of convergence. In this section, we focus on the rate of convergence (RoC) depending primarily on the regularity of material coefficient. Thus, according to Rem. 4.1, we define material coefficients through scalar functions f as

- circle (C): $f(\boldsymbol{x}) = \begin{cases} 1, & \|\boldsymbol{x}\|_2 < \frac{3}{5} \\ 0, & \text{otherwise} \end{cases}$,

- cone (E): $f(\boldsymbol{x}) = \begin{cases} 1 - \|\boldsymbol{x}\|_2, & \|\boldsymbol{x}\|_2 < 1 \\ 0, & \text{otherwise} \end{cases}$,
- hummock (H): $f(\boldsymbol{x}) = \begin{cases} 1 - 2\|\boldsymbol{x}\|_2^2, & \|\boldsymbol{x}\|_2 < \frac{1}{2} \\ 2(1 - \|\boldsymbol{x}\|_2)^2, & \frac{1}{2} \leq \|\boldsymbol{x}\|_2 < 1 \\ 0, & \text{otherwise} \end{cases}$,
- standard mollifier (M): $f(\boldsymbol{x}) = \begin{cases} \exp(1 - \frac{1}{1-\|\boldsymbol{x}\|_2}), & \|\boldsymbol{x}\|_2 < 1 \\ 0, & \text{otherwise} \end{cases}$;

the cut through the periodic unit cell, for $x_2 \equiv 0$, can be observed in Fig. 4.3. The regularity of the material coefficients are based on the regularity of scalar functions f; circle (C) is piecewise constant, cone (E) has piecewise constant the first derivative, hummock (H) has piecewise constant the second derivative while standard mollifier (M) is infinitely smooth, having continuous all derivatives.

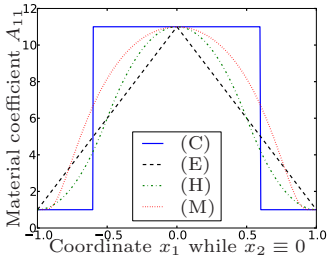


FIG. 4.3. *The material coefficients at PUC for $x_2 \equiv 0$ showing their smoothness*

First, we verify the RoC of $\|\boldsymbol{e}^{(\alpha)} - \boldsymbol{e}^{(\alpha)}_{\boldsymbol{N}}\|_{L^2_{\text{per}}} \leq C(\frac{n}{2})^{-\mu}$ from Lem. 3.23 for $\alpha = 1$. Unfortunately, the exact solution $\boldsymbol{e}^{(1)}$ is possible to obtain only in a very special cases, hence, we approximate it by a solution calculated with a very fine grid, i.e. $\boldsymbol{e}^{(1)} \approx \boldsymbol{e}^{(1)}_{\bar{\boldsymbol{N}}}$ where we have chosen $\bar{\boldsymbol{N}} = (2005, 2005)$. It enables to calculate the approximations of order μ, particularly

$$\mu_{\boldsymbol{N}} = \log\left(\frac{\|\boldsymbol{e}^{(1)}_{\bar{\boldsymbol{N}}} - \boldsymbol{e}^{(1)}_{\boldsymbol{N}}\|_{L^2_{\text{per}}}}{\|\boldsymbol{e}^{(1)}_{\bar{\boldsymbol{N}}} - \boldsymbol{e}^{(1)}_{\lambda\boldsymbol{N}}\|_{L^2_{\text{per}}}}\right)(\log\lambda)^{-1}, \tag{4.1}$$

where the factors λ equals to 2 or 3, cf. Rem. 4.1. Norms $\|\cdot\|_{L^2_{\text{per}}}$ in (4.1) can be calculated exactly using Plancherel's theorem as all functions are the trigonometric polynomials.

TABLE 4.1
The rate of convergence $\mu_{\boldsymbol{N}}$ for the odd number of discretization points

RoC	$f \backslash n$	5	15	45	135	405	limit
$\mu_{\boldsymbol{N}}$	(C)	0.568	0.579	0.511	0.544	0.565	0.5
$\mu_{\boldsymbol{N}}$	(E)	1.083	1.410	1.522	1.527	1.557	1.5
$\mu_{\boldsymbol{N}}$	(H)	2.504	2.554	2.476	2.478	2.499	2.5
$\mu_{\boldsymbol{N}}$	(M)	1.491	3.009	6.777	10.266	6.091	—

The rate of convergence can also be studied from the upper-lower bounds of the homogenized matrix. According to Lem. 2.14, it converges with rate 2μ that represents the smallest value of following rates $\|\boldsymbol{e}^{(\alpha)} - \boldsymbol{e}^{(\alpha)}_{\boldsymbol{N}}\|^2_{L^2_{\text{per}}}$ and $\|\boldsymbol{\jmath}^{(\alpha)} - \boldsymbol{\jmath}^{(\alpha)}_{\boldsymbol{N}}\|^2_{L^2_{\text{per}}}$ for all α.

TABLE 4.2
The rate of convergence $\mu_{\boldsymbol{N}}$ for the even number of discretization points

RoC	$f \setminus n$	4	8	16	32	64	128	256	512	limit
$\mu_{\boldsymbol{N}}$	(C)	-0.139	1.111	0.315	0.484	0.553	0.578	0.467	0.603	0.5
$\mu_{\boldsymbol{N}}$	(E)	1.067	1.210	1.293	1.407	1.483	1.515	1.522	1.525	1.5
$\mu_{\boldsymbol{N}}$	(H)	2.326	2.964	2.652	2.468	2.443	2.452	2.469	2.486	2.5
$\mu_{\boldsymbol{N}}$	(M)	1.475	1.710	2.714	4.812	7.246	8.895	12.656	4.858	—

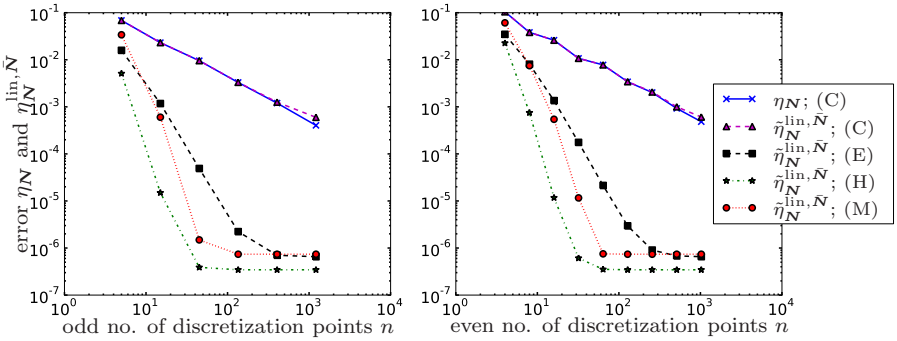


FIG. 4.4. *Progress in error $\eta_{\boldsymbol{N}}$ for increasing in the number of discretization points $\boldsymbol{N} = (n, n)$*

Hence, we define

$$\bar{\mu}_{\boldsymbol{N}} = \frac{1}{2} \log \left(\frac{\operatorname{tr} \boldsymbol{D}_{\boldsymbol{N}}}{\operatorname{tr} \boldsymbol{D}_{\lambda \boldsymbol{N}}} \right) (\log \lambda)^{-1}, \tag{4.2}$$

where $\boldsymbol{D}_{\boldsymbol{N}}$ is the error from Def. 2.11 and λ is taken as 2 or 3.

Nevertheless, it can be calculated only for special material coefficients $\boldsymbol{A}$, e.g. for problem (C), cf. Rem. 3.35. Thus, we define the alternative

$$\tilde{\mu}_{\boldsymbol{N}}^{\mathrm{lin},\bar{\boldsymbol{N}}} = \frac{1}{2} \log \left(\frac{\operatorname{tr} \tilde{\boldsymbol{D}}_{\boldsymbol{N}}^{\mathrm{lin},\bar{\boldsymbol{N}}}}{\operatorname{tr} \tilde{\boldsymbol{D}}_{\lambda \boldsymbol{N}}^{\mathrm{lin},\bar{\boldsymbol{N}}}} \right) (\log \lambda)^{-1}, \tag{4.3}$$

where $\tilde{\boldsymbol{D}}_{\boldsymbol{N}}^{\mathrm{lin},\bar{\boldsymbol{N}}} = \frac{1}{2} \left| \tilde{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}^{\mathrm{lin},\bar{\boldsymbol{N}}} - \tilde{\boldsymbol{A}}_{\mathrm{eff,D},\boldsymbol{N}}^{\mathrm{lin},\bar{\boldsymbol{N}}} \right|$ and $\bar{\boldsymbol{N}} = (2005, 2005)$.

Analogically, we define the normalized errors

$$\eta_{\boldsymbol{N}} := \frac{\operatorname{tr} \boldsymbol{D}_{\boldsymbol{N}}}{\operatorname{tr} \boldsymbol{A}_{\mathrm{eff},\boldsymbol{M}}} \tag{4.4a}$$

$$\tilde{\eta}_{\boldsymbol{N}}^{\mathrm{lin},\bar{\boldsymbol{N}}} := \frac{\operatorname{tr} \tilde{\boldsymbol{D}}_{\boldsymbol{N}}^{\mathrm{lin},\bar{\boldsymbol{N}}}}{\operatorname{tr} \boldsymbol{A}_{\mathrm{eff},\boldsymbol{M}}} \tag{4.4b}$$

where $\boldsymbol{M} = (1215, 1215)$ or $\boldsymbol{M} = (1024, 1024)$ depending on $\boldsymbol{N}$ being odd or even, and $\bar{\boldsymbol{N}}$ is taken as previously.

Now, we will discuss and compare the results. First, there is no significant observation between even and odd number of discretization points $\boldsymbol{N}$ in the rates of convergence, compare Tab. 4.1, 4.3 versus Tab. 4.2, 4.4, and see also almost straight lines of the normalized errors in Fig. 4.4.

The rates of convergence suffer from an inaccuracy for large $\boldsymbol{N}$; the rates of $\|\boldsymbol{e}^{(1)} - \boldsymbol{e}_{\boldsymbol{N}}^{(1)}\|_{L^2_{\mathrm{per}}}$ stated in Tab. 4.1 and 4.2 are depreciated due to approximation of

TABLE 4.3
The rate of convergence (RoC) from the guaranteed bounds for odd $\boldsymbol{N} = (n, n)$

RoC	$f \setminus n$	5	15	45	135	405	theory
$\bar{\mu}_{\boldsymbol{N}}$	(C)	0.495	0.402	0.484	0.460	0.496	0.5
$\tilde{\mu}_{\boldsymbol{N}}^{\mathrm{lin},\bar{\boldsymbol{N}}}$	(C)	0.495	0.401	0.483	0.441	0.338	0.5
$\tilde{\mu}_{\boldsymbol{N}}^{\mathrm{lin},\bar{\boldsymbol{N}}}$	(E)	1.185	1.446	1.406	0.526	0.032	1.5
$\tilde{\mu}_{\boldsymbol{N}}^{\mathrm{lin},\bar{\boldsymbol{N}}}$	(H)	2.655	1.662	0.054	0.000	0.000	2.5
$\tilde{\mu}_{\boldsymbol{N}}^{\mathrm{lin},\bar{\boldsymbol{N}}}$	(M)	1.836	2.732	0.319	0.000	0.000	???

TABLE 4.4
The rate of convergence (RoC) from the guaranteed bounds for even $\boldsymbol{N} = (n, n)$

RoC	$f \setminus n$	4	8	16	32	64	128	256	512	limit
$\bar{\mu}_{\boldsymbol{N}}$	(C)	0.727	0.276	0.641	0.230	0.596	0.379	0.561	0.469	0.5
$\tilde{\mu}_{\boldsymbol{N}}^{\mathrm{lin},\bar{\boldsymbol{N}}}$	(C)	0.727	0.276	0.641	0.231	0.593	0.363	0.524	0.360	0.5
$\tilde{\mu}_{\boldsymbol{N}}^{\mathrm{lin},\bar{\boldsymbol{N}}}$	(E)	1.055	1.284	1.475	1.520	1.427	0.853	0.207	0.026	1.5
$\tilde{\mu}_{\boldsymbol{N}}^{\mathrm{lin},\bar{\boldsymbol{N}}}$	(H)	2.456	2.999	2.133	0.396	0.015	0.000	0.000	0.000	2.5
$\tilde{\mu}_{\boldsymbol{N}}^{\mathrm{lin},\bar{\boldsymbol{N}}}$	(M)	1.511	1.892	2.774	1.975	0.013	0.000	-0.000	0.000	???

minimizers $\boldsymbol{e}^{(1)} \approx \boldsymbol{e}_{\bar{\boldsymbol{N}}}^{(1)}$ and the rates of the upper-lower bounds, see Tab. 4.3 and 4.3, due to approximations $\overline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}} \approx \tilde{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}^{\mathrm{lin},\bar{\boldsymbol{N}}}$ and $\underline{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}} \approx \tilde{\boldsymbol{A}}_{\mathrm{eff},\boldsymbol{N}}^{\mathrm{lin},\bar{\boldsymbol{N}}}$. The exception is the rate $\bar{\mu}_{\boldsymbol{N}}$ in (4.2), calculated for circle (C), that can be compared with its approximation $\tilde{\mu}_{\boldsymbol{N}}^{\mathrm{lin},\bar{\boldsymbol{N}}}$ in (4.3), see Tab. 4.3 and 4.4. Both the rates coincide for two digits up to $\boldsymbol{N} = (45, 45)$ for the odd case and up to $\boldsymbol{N} = (64, 64)$ for the even case. It can also be compared in terms of errors (4.4) shown in Fig. 4.4 — the corresponding lines differ significantly only for $\boldsymbol{N} = (1215, 1215)$ and $\boldsymbol{N} = (1215, 1215)$ resp.

Further, the errors in Fig. 4.4 — observed for problems (E), (H), and (M) — reach a limit state about the order 10^{-6}. It is primarily caused by an inaccuracy in approximation $\boldsymbol{e}^{(1)} \approx \boldsymbol{e}_{\bar{\boldsymbol{N}}}^{(1)}$. It also depreciate the rates in Tab. 4.3, 4.4. However, the rates give evidence in the parts where errors in Fig. 4.4 produces the straight lines.

Concluding, the rates of convergence, calculated as in Eq. (4.1), (4.2), and (4.3), are comparable and depend on the regularity of material coefficients. Particularly, problem (C) reaches the rate $\frac{1}{2}$, while (E) $\frac{3}{2}$, and the last (H) and (M) even higher – about $\frac{5}{2}$ – however, this value is highly influenced by the inaccuracy in its determination.

4.3. Comparison with p-version of FEM. In this section, we compare the upper-lower bounds calculated with FFT-based FEM to p-version of FEM by [6]. The comparison is made for material coefficients defined according to Rem. 4.1 through scalar functions:

- circle (C): $f(\boldsymbol{x}) = \begin{cases} 1, & \|\boldsymbol{x}\|_2 < \pi^{-\frac{1}{2}} \\ 0, & \text{otherwise} \end{cases}$
- square (S): $f(\boldsymbol{x}) = \begin{cases} 1, & \|\boldsymbol{x}\|_\infty \leq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$
- rectangle (R): $f(\boldsymbol{x}) = \begin{cases} 1, & |x_1| \leq \frac{\sqrt{2}}{4} \bigwedge |x_2| \leq \frac{\sqrt{2}}{2} \\ 0, & \text{otherwise} \end{cases}$

Contrary to the previous examples, the conductivity of circle, square, and rectangle inclusion is set to 10^{-3} that intends to model void. The number of degrees of freedom

for the p-version of FEM reaches 959 for circle and 511 for square and rectangle while the number of discretization points for FFT-based FEM was taken as $\boldsymbol{N} = (1215, 1215)$, hence substantially larger.

However, p-version of FEM is still significantly better in the terms of errors η, see Tab. 4.5. It is mainly caused by better approximation properties of the p-version of FEM, the shapes of inclusions are well suited for this method. Although, p-version of FEM only approximates circle inclusion contrary to the FFT-based method. Moreover, the FFT-based FEM is influenced by variational crime caused by the numerical integration in Def. 3.19.

TABLE 4.5
The comparison of the homogenized matrices between the p-version of FEM and the FFT-based FEM

method	p-FEM	p-FEM	p-FEM	FFTH	FFTH	FFTH
problem\property	$A^{\mathrm{FEM}}_{\mathrm{eff},11}$	$A^{\mathrm{FEM}}_{\mathrm{eff},22}$	η^{FEM}	$A_{\mathrm{eff},\boldsymbol{N},11}$	$A_{\mathrm{eff},\boldsymbol{N},22}$	$\eta_{\boldsymbol{N}}$
(C)	0.600	0.600	1.166582e-04	0.588	0.588	2.676e-02
(S)	0.578	0.578	3.611941e-03	0.476	0.476	2.149e-01
(R)	0.425	0.671	6.403170e-03	0.346	0.569	1.951e-01

Nevertheless, the p-version of FEM with divergence-free subspaces is mostly suitable to 2-dimensional problems. It is an opportunity for the FFT-based FEM that manage higher dimensional problems without any additional effort, especially for data provided as voxel images.

Acknowledgments. The authors are thankfull to Jaroslav Haslinger for pointing out the works of Jan Dvořák, [5, 6].

Appendix A. Continuous projections on solenoidal and curl-free spaces.

DEFINITION A.1. *For $i \in \{0,1,2\}$, we define operators $\mathcal{G}^{(i)}[\cdot] : L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d) \to L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ as convolution*

$$\mathcal{G}^{(i)}[\boldsymbol{v}](\boldsymbol{x}) := \int_{\mathcal{Y}} \boldsymbol{\Gamma}^{(i)}(\boldsymbol{x}-\boldsymbol{y})\boldsymbol{v}(\boldsymbol{y})\mathrm{d}\boldsymbol{y} = \sum_{\boldsymbol{n}\in\mathbb{Z}^{\mathrm{d}}} \hat{\boldsymbol{\Gamma}}^{(\mathrm{i})}(\boldsymbol{n})\hat{\boldsymbol{v}}(\boldsymbol{n})\varphi_{\boldsymbol{n}}(\boldsymbol{x})$$

where $\hat{v}_\alpha := \big(v_\alpha, \varphi_{\boldsymbol{n}}\big)_{L^2_{\mathrm{per}}}$ denotes the Fourier coefficients for $\varphi_{\boldsymbol{n}}(\boldsymbol{x}) = \exp(\mathrm{i}\pi \sum_{\alpha=1}^d \frac{k_\alpha x_\alpha}{Y_\alpha})$.

Integral kernels $\boldsymbol{\Gamma}^{(i)}$ are easily expressed in the Fourier space; the matrices $\hat{\boldsymbol{\Gamma}}^{(i)}(\boldsymbol{n}) \in \mathbb{R}^{d\times d}$ of the Fourier coefficients reads

$$\hat{\boldsymbol{\Gamma}}^{(0)}(\boldsymbol{n}) = \begin{cases} \boldsymbol{I} \\ \boldsymbol{0} \end{cases} \qquad \hat{\boldsymbol{\Gamma}}^{(1)}(\boldsymbol{n}) = \begin{cases} \boldsymbol{0} \\ \frac{\boldsymbol{\xi}(\boldsymbol{n})\otimes\boldsymbol{\xi}(\boldsymbol{n})}{\boldsymbol{\xi}(\boldsymbol{n})\cdot\boldsymbol{\xi}(\boldsymbol{n})} \end{cases} \qquad \hat{\boldsymbol{\Gamma}}^{(2)}(\boldsymbol{n}) = \begin{cases} \boldsymbol{0}, & \text{for } \boldsymbol{n} = \boldsymbol{0} \\ \boldsymbol{I} - \frac{\boldsymbol{\xi}(\boldsymbol{n})\otimes\boldsymbol{\xi}(\boldsymbol{n})}{\boldsymbol{\xi}(\boldsymbol{n})\cdot\boldsymbol{\xi}(\boldsymbol{n})}, & \text{for } \boldsymbol{n} \in \mathbb{Z}^d \setminus \{\boldsymbol{0}\} \end{cases}$$

where $\boldsymbol{I} \in \mathbb{R}^{d\times d}$ is the identity matrix, $\boldsymbol{0}$ denotes either a vector or a matrix with zero components, and $\xi_\alpha(\boldsymbol{n}) = \frac{n_\alpha}{Y_\alpha}$ for period $\boldsymbol{Y}$ being consistent with periodic unit cell $\mathcal{Y} = \prod_\alpha (-Y_\alpha, Y_\alpha) \subset \mathbb{R}^d$.

LEMMA A.2. *Operators $\mathcal{G}^{(i)}$ from previous definition are mutually orthogonal projections on, step-by-step, $\mathscr{U}$,$\mathscr{E}$, and $\mathscr{J}$ — the subspaces of $L^2_{\mathrm{per}}(\mathcal{Y};\mathbb{R}^d)$ defined in Eq. (2.1).*

Proof. In [31], we show in detail that $\mathcal{G}^{(1)}$ is a projection onto $\mathscr{E}$, the other cases are analogical. It is based on mutual orthogonality of $\hat{\boldsymbol{\Gamma}}^{(i)}(\boldsymbol{n})$ for particular $\boldsymbol{n} \in \mathbb{Z}^d$ that can be found for example in [17]. □

REFERENCES

[1] A. Bensoussan, G. Papanicolau, and J. Lions, *Asymptotic analysis for periodic structures*, vol. 5, North Holland, 1978.

[2] S. Brisard and L. Dormieux, *FFT-based methods for the mechanics of composites: A general variational framework*, Computational Materials Science, 49 (2010), pp. 663–671.

[3] ———, *Combining Galerkin approximation techniques with the principle of Hashin and Shtrikman to derive a new FFT-based numerical method for the homogenization of composites*, Computer Methods in Applied Mechanics and Engineering, (2012).

[4] D. Cioranescu and P. Donato, *An Introduction to Homogenization*, Oxford Lecture Series in Mathematics and Its Applications, Oxford University Press, 1999.

[5] J. Dvořák, *Optimization of composite materials*, Master's thesis, The Charles University in Prague, June 1993.

[6] ———, *A reliable numerical method for computing homogenized coefficients*, tech. rep., available at CiteSeerX `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.45.1190`, 1995.

[7] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, SIAM, 1976.

[8] 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 (1999), pp. 41–47.

[9] Z. Hashin and S. Shtrikman, *On some variational principles in anisotropic and nonhomogeneous elasticity*, Journal of the Mechanics and Physics of Solids, 10 (1962), pp. 335–342.

[10] ———, *A variational approach to the theory of the effective magnetic permeability of multiphase materials*, Journal of Applied Physics, 33 (1962), pp. 3125–3131.

[11] ———, *A variational approach to the theory of the elastic behaviour of multiphase materials*, Journal of the Mechanics and Physics of Solids, 11 (1963), pp. 127–140.

[12] H. Hoang-Duc and G. Bonnet, *Effective properties of viscoelastic heterogeneous periodic media: an approximate solution accounting for the distribution of heterogeneities.*, Mechanics of Materials, (2012).

[13] V. Jikov, S. Kozlov, and O. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, 1994.

[14] S. Lee, R. Lebensohn, and A. Rollett, *Modeling the viscoplastic micromechanical response of two-phase materials using Fast Fourier Transforms*, International Journal of Plasticity, 27 (2011), pp. 707–727.

[15] J. Li, X. Tian, and R. Abdelmoula, *A damage model for crack prediction in brittle and quasi-brittle materials solved by the FFT method*, International journal of fracture, (2012), pp. 1–12.

[16] J. C. Michel, H. Moulinec, and P. Suquet, *A computational method based on augmented Lagrangians and fast Fourier transforms for composites with high contrast*, CMES-Computer Modeling in Engineering & Sciences, 1 (2000), pp. 79–88.

[17] G. W. Milton, *The Theory of Composites*, vol. 6 of Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, UK, 2002.

[18] V. Monchiet and G. Bonnet, *A polarization-based FFT iterative scheme for computing the effective properties of elastic composites with arbitrary contrast*, International Journal for Numerical Methods in Engineering, 89 (2012), pp. 1419–1436.

[19] H. Moulinec and P. Suquet, *A fast numerical method for computing the linear and nonlinear mechanical properties of composites*, Comptes rendus de l'Académie des sciences. Série II, Mécanique, physique, chimie, astronomie, 318 (1994), pp. 1417–1423.

[20] ———, *A numerical method for computing the overall response of nonlinear composites with complex microstructure*, Computer Methods in Applied Mechanics and Engineering, 157 (1997), pp. 69–94.

[21] J. Novák, A. Kučerová, and J. Zeman, *Microstructural enrichment functions based on stochastic wang tilings*, arXiv preprint arXiv:1110.4183, (2011).

[22] J. Němeček, V. Králík, and J. Vondřejc, *Micromechanical analysis of heterogeneous structural materials*, Cement and Concrete Composites, (2012).

[23] J. Němeček, V. Králík, and J. Vondřejc, *A two-scale micromechanical model for aluminium foam based on results from nanoindentation*, (2012). Submitted.

[24] J. Oden, T. Belytschko, I. Babuška, and T. Hughes, *Research directions in computational mechanics*, Computer Methods in Applied Mechanics and Engineering, 192 (2003), pp. 913–922.

[25] A. Reuss and Z. Angnew, *A calculation of the bulk modulus of polycrystalline materials*, Math Meth, 9 (1929), p. 55.

[26] J. Saranen and G. Vainikko, *Periodic Integral and Pseudodifferential Equations with Numerical Approximation*, Springer Monographs Mathematics, 2000.
[27] P. Suquet, *A dual method in homogenization: application to elastic media*, J. Mech. Theor. Appl, 79 (1982), p. 98.
[28] P. Suquet, H. Moulinec, O. Castelnau, M. Montagnat, N. Lahellec, F. Grennerat, P. Duval, and R. Brenner, *Multi-scale modeling of the mechanical behavior of polycrystalline ice under transient creep*, Procedia IUTAM, 3 (2012), pp. 64–78.
[29] W. Voigt, *Lehrbuch der kristallphysik*, vol. 34, BG Teubner, 1910.
[30] J. Vondřejc, J. Zeman, and I. Marek, *Analysis of a Fast Fourier Transform based method for modeling of heterogeneous materials*, Lecture Notes in Computer Science, 7116 (2012), pp. 512–522.
[31] ———, *FFT-based finite element method for homogenization*, (2013). In preparation.
[32] Z. Wieçkowski, *Dual finite element methods in mechanics of composite materials*, Journal of Theoretical and Applied Mechanics, 2 (1995), pp. 233–252.
[33] 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 (2010), pp. 8065–8071.