Czech Technical University in Prague
Faculty of Civil Engineering
Department of Mechanics

Analysis of Heterogeneous Materials using Efficient Meshless Algorithms: One-dimensional Study

Master's Thesis

by

Jaroslav Vondřejc

May 2009

Analysis of Heterogeneous Materials using Efficient Meshless Algorithms: One-dimensional Study

Master's Thesis

by

Jaroslav Vondřejc

born on the 27th of June 1983 in Opočno

Supervisor:
Jan Zeman

Czech Technical University in Prague
Faculty of Civil Engineering
Department of Mechanics

May 2009

Abstract

This work provide a comprehensive comparison of two algorithms used for modelling heterogenous materials with data provided in the form of pixel or voxel bitmaps. The Fast Fourier Homogenization method was proposed by Moulinec and Suquet in [17] as an efficient homogenization solver for periodical problems. The available studies analyze convergence of the method in multi-dimensional setting without addressing the effect of the unit cell geometry. In this contribution, a complete convergence analysis of the one-dimensional problem is presented, including an explicit expression of the optimal value of the reference stiffness. Next, non-symmetric linear system solved by conjugate gradient algorithm is discussed.

The Gaussian Approximating Function method, proposed for scalar problems by Kanaun and Kochekseraii in [10], is based on the approximation of local fields using Gaussian kernels following the seminal ideas of Maz'ya [13] and Maz'ya and Schmidt [14]. The various approaches of GAF method treating isolated inclusions in an infinite matrix are discussed from both theoretical and numerical point of view.

Keywords: Heterogeneous materials, Fast Fourier transform, Computational homogenization, Gradient methods, Convergence studies, Approximate Approximation.

Contents

List of Figures

List of Tables

Part I
Introduction

Recent experimental and theoretical advances in characterization of microstructure of heterogeneous media open novel possibilities in predictive "bottom-up" modelling of engineering materials. There is currently a variety of techniques being utilized to acquire a comprehensible digital representation of materials' structure, including serial sectioning approaches [22], X-ray tomography [2], depth-sensing indentation methods [3], statistical reconstructing algorithms [27] or microstructure evolution models [26]; see also [6] for a detailed discussion of this topic. With the detailed quantification of microstructure at hand, relevant material constants can be subsequently determined using a suitable numerical technique. With regard to the structure of input data, the pixel- or voxel-based methods appear to be a convenient choice to avoid time consuming mesh generation procedures. Moreover, even though the traditional finite difference and finite element methods were implemented with some success, e.g. [7, 23], it is now well-understood that highly optimized solvers, capable of handling large datasets efficiently, can be developed when combining tools of the theory of heterogeneous materials with the Fourier summation techniques.

In this work, we present a comprehensive comparison of two such algorithms: the Fast Fourier Homogenization (FFH) solver related to periodic heterogeneous materials and the Gaussian Approximating Function (GAF) method treating isolated inclusions in an infinite matrix. The FFH scheme was independently introduced by Bakhvalov and Knyazev [1] on the conceptual level and by Moulinec and Suquet [17, 18] using engineering arguments. Detailed convergence theory of FFH method for linear and non-linear case was presented by Eyre and Milton in [5] together with the multi-grid extension to reduce the computational cost. The successful applications of the algorithm include, among others, simulation of materials with evolving microstructure such as tin/lead solders [4] or hydrating cement paste [26]. The GAF method, proposed for scalar problems by Kanaun and Kochekseraii in [10], is based on the approximation of local fields using Gaussian kernels following the seminal ideas of Maz'ya [13] and Maz'ya and Schmidt [14]. The computational efficiency of the method was further extended by fast multipole expansion techniques [11] or the FFT-based matrix operation [12] to efficiently handle the fully populated matrices resulting from the discretization; see also the work of Novák [19] for the treatment of multi-dimensional linear elasticity. It is worth noting that versatility of the Gaussian approximation approach has been demonstrated by independent applications in wave diffraction simulations [8] or fracture mechanics [9].

Even though the two algorithms were originally developed for the solution of different problems, they share a several common features and suffer from similar convergence difficulties. Therefore, it can be well expected that a unified exposition and a systematic assessment of both methods may provide additional insights into their behavior with the potential to combine the strengths of the numerical schemes. To that end, the essential steps of the investigated methods are introduced in Sections 1 and 3. The convergence

properties are assessed in Sections 2, 4, 5 and 6 by means of analytical and numerical procedures. Finally, Section IV is devoted to the summary of obtained results together with future improvements. In order to minimize technicalities and to obtain the sharpest results possible, the attention is restricted to an one-dimensional elasticity problem (or, equivalently, to a simple laminate subject to forces varying in one direction).

Part II
Gaussian Approximating Function method

1 Theory

In this method, an infinite one dimensional rod with heterogeneities is considered. The material data is described with stiffness function $E(x)$ that is separated into two components

$$E(x) = E_0 + E_1(x) \tag{1}$$

where E_0 is Young's modulus of material of matrix and $E_1(x)$ is complement to real stiffness. The BPM method assumes that $E_1(x)$ function has a nonzero values only on a bounded interval. The rod is loaded at infinity with strain ε_0. For the use of the method, the strain $\frac{du(x)}{dx} = \varepsilon(x)$ is also separated into two components as stated below

$$\varepsilon(x) = \varepsilon_0 + \varepsilon_1(x) \tag{2}$$

where $\varepsilon_1(x)$ strain is complement to real strain. Basic elasticity relation of the rod states that

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(E(x)A(x)\frac{\mathrm{d}u(x)}{\mathrm{d}x}\right) + g(x) = 0 \tag{3}$$

where generalized load function $g(x)$ is equal to zero in this case and function of cross-section area is taken as $A(x) = 1$.
Using (1), (2) and $\frac{\mathrm{d}}{\mathrm{d}x}(E_0\varepsilon_0) = 0$, the basic relation (3) can be rewritten into the following form

$$\frac{\mathrm{d}}{\mathrm{d}x}\Big(E_0\varepsilon_1(x)\Big) = -\frac{\mathrm{d}}{\mathrm{d}x}\Big[E_1(x)\big(\varepsilon_0 + \varepsilon_1(x)\big)\Big] \tag{4}$$

Hence, the right side of Equation (4) can be interpret as generalized load resulting in strain $\varepsilon_1(x)$ within the homogeneous bar with stiffness E_0. It means that it is necessary to solve following linear differential equation

$$E_0\frac{\mathrm{d}\varepsilon_1(x)}{\mathrm{d}x} = f(x)$$

where $f(x) = -\frac{\mathrm{d}}{\mathrm{d}x}\big(E_1(x)\varepsilon(x)\big)$.
It is convenient to solve it with the Fourier transform that is defined as

$$\mathscr{F}\big\{f(x)\big\} = \hat{f}(t) = \int_{-\infty}^{\infty} f(x)\,\mathrm{e}^{-ixt}\,\mathrm{d}x$$

and its inverse Fourier transform then follows

$$\mathscr{F}^{-1}\big\{\hat{f}(t)\big\} = f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)\,\mathrm{e}^{ixt}\,\mathrm{d}t$$

After applying the transform to Equation (4) and using the Fourier transform of function derivative, it leads to

$$\mathscr{F}\{\varepsilon_1(x)\} = \frac{1}{\mathrm{i}E_0 t}\mathscr{F}\{f(x)\} \tag{5}$$

The term $\frac{1}{\mathrm{i}E_0 t}$ can be interpreted as the Fourier transform of some function denoted as $G(x)$ written as $\mathscr{F}\{G(x)\} = \frac{1}{\mathrm{i}E_0 t}$. Using the Fourier transform of a convolution rule, Equation (5) can be rewritten as

$$\mathscr{F}\{\varepsilon_1(x)\} = \mathscr{F}\{G(x)\}\mathscr{F}\{f(x)\} \quad \Leftrightarrow \quad \varepsilon_1(x) = \int_{-\infty}^{\infty} G(x-\xi)f(\xi)d\xi$$

Integration by parts leads to

$$\varepsilon_1(x) \overset{P.P.}{=} \int_{-\infty}^{\infty} \frac{\partial G(x-\xi)}{\partial \xi}F(\xi)d\xi + G(x-\xi)F(\xi)\Big|_{-\infty}^{\infty}$$

where $F(x)$ has to satisfy $F'(x) = f(x)$ and therefore $F(x) = -E_1(x)\varepsilon(x)$. As the second term vanishes due to boundary conditions, we obtain

$$\varepsilon_1(x) = -\int_{-\infty}^{\infty} \frac{\partial G(x-\xi)}{\partial \xi}E_1(\xi)\varepsilon(\xi)d\xi \tag{6}$$

Using the property of the Fourier transform of function derivative

$$\mathscr{F}\{f'(x)\} = \mathrm{i}t\mathscr{F}\{f(x)\}$$

it can be deduced that

$$\frac{\mathrm{d}G(x)}{\mathrm{d}\xi} = \mathscr{F}\left\{\frac{1}{E_0}\right\} = \frac{\delta(x)}{E_0} \tag{7}$$

where $\delta(x)$ is the Dirac distribution. It has to be noted, that the function from Equation (7) is usually called $K^{\infty}(x)$, cf. [10]. Hence, following equation arises

$$\varepsilon_1(x) = \int_{-\infty}^{\infty} \frac{\partial G(x-\xi)}{\partial \xi}F(\xi)d\xi = -\int_{-\infty}^{\infty} \frac{\delta(x-\xi)}{E_0}E_1(\xi)\varepsilon(\xi)d\xi \tag{8}$$

After adding ε_0 to both sides of equation, we finally arrive at

$$\varepsilon(x) + \int_{-\infty}^{\infty} \frac{\delta(x-\xi)}{E_0}E_1(\xi)\varepsilon(\xi)d\xi = \varepsilon_0 \tag{9}$$

2 Solution using Approximate Approximation

This section provides solution of Equation (9) using Approximate Approximation. The method that was proposed in 1991 by Maz'ya [13] and [15], and is based on generating functions representing an approximate partition of unity. This approximation method does not converge if the discretization size tends to zero. In fact, this lack of convergence can be controlled since the resulting error can be chosen less than machine precision. This method uses following approximation formula

$$u(x) \approx \mathcal{A}_h\{u(x)\} = \sum_{m\in\mathbb{Z}} u(hm)\eta\left(x - hm\right) \tag{10}$$

where parameter h sets the size of regular grid and η is a basis function
The choice of basis function $\eta(x)$ depends on various requirements. First of all, the integral of the basis function should be equal to

$$\int_{-\infty}^{\infty} \eta(x)\,\mathrm{d}x = h$$

In the case of generalization into multidimensional spaces, this integral would equal to h^D assuming D-dimensional space.
Next requirement for basis function is to be rapidly decreasing[1] as it is used for integral cubature. It can be shown at following convolution integral where function $f(x)$ is approximated using formula in Equation (10).

$$\int_{-\infty}^{\infty} G(x-\xi)f(\xi)\,\mathrm{d}\xi \approx \int_{-\infty}^{\infty} G(x-\xi)\mathcal{A}_h\{f(\xi)\}\,\mathrm{d}\xi$$

After some algebraic emendations it leads to

$$\int_{-\infty}^{\infty} G(x-\xi)\mathcal{A}_h\{f(\xi)\}\,\mathrm{d}\xi = \int_{-\infty}^{\infty} G(x-\xi)\sum_{m\in\mathbb{Z}} f(hm)\eta\left(\frac{\xi}{h} - m\right)\mathrm{d}\xi \tag{11}$$

$$= \sum_{m\in\mathbb{Z}} f(hm)\int_{-\infty}^{\infty} G(x-\xi)\eta\left(\frac{\xi}{h} - m\right)\mathrm{d}\xi \tag{12}$$

Hence, if the rapidly decaying condition for basis function is satisfied, only limited number of summation in Equation (12) can be considered. The error due to this truncation in summation can be balanced with special choice of basis function in order to calculate integral $\int_{-\infty}^{\infty} G(x-\xi)\eta\left(\frac{\xi}{h} - m\right)\mathrm{d}\xi$ analytically. Note that, in the case of one dimensional

[1] A rapidly decreasing function is in effect a function that goes to zero as $|x| \to \infty$ faster than any inverse power of x, as do all its derivatives.

problem with integral kernel defined by Equation (7) as $\frac{\delta(x)}{E_0}$, the problem does not exist since the Dirac distribution satisfies

$$\int_{-\infty}^{\infty} \delta(x-\xi) f(\xi)\,\mathrm{d}\xi = f(x)$$

Observe that the integral kernel K occurring in Eq. (9) depends on difference $(x-\xi)$ satisfying that the arising matrix possess Toeplitz structure. Hence, it is recommended for basis function to be radial function $f(x) = f(\|x\|)$. In the case of one-dimensional problem, the condition is equivalent to being even function. This condition is important for multiplication $\boldsymbol{Ax}$ in linear system using FFT algorithm as it is described in Section 2.4.
All those requirements are satisfied by basis function derived from Gaussian distribution used in statistics

$$\varphi_H(x) = \frac{1}{\sqrt{\pi H}}\,\mathrm{e}^{-\frac{|x|^2}{Hh^2}} \tag{13}$$

where parameter H is equal to double of the measure of statistical dispersion called variance ($H = 2\sigma^2$). This function is used as the basis function in the following text.

2.1 Standard approach

After general introduction of Approximate approximation, Equation (9)

$$\varepsilon(x) + \int_{-\infty}^{\infty} \frac{\delta(x-\xi)}{E_0} E_1(\xi)\varepsilon(\xi)d\xi = \varepsilon_0$$

will be solved with unknown field $\varepsilon(x)$ expressed with the use of formula in Equation (10). After discretization with regular grid of size h with coordinates $x_i, i = 1, 2, \ldots, N$, the unknown strain can be expressed as

$$\varepsilon(x) \approx \sum_{i=1}^{N} \varphi(x - x_i)\varepsilon(x_i)$$

After substitution, integral equation follows as

$$\sum_{i=1}^{N} \varphi(x_j - x_i)\varepsilon(x_i) + \int_{-\infty}^{\infty} \frac{\delta(x_j-\xi)}{E_0} E_1(\xi) \sum_{i=1}^{N} \varphi(\xi - x_i)\varepsilon(x_i)d\xi = \varepsilon_0$$

$$\sum_{i=1}^{N} \varphi(x_j - x_i)\varepsilon(x_i) + \frac{E_1(x_j)}{E_0} \sum_{i=1}^{N} \varphi(x_j - x_i)\varepsilon(x_i) = \varepsilon_0$$

$$\sum_{i=1}^{N} \varphi(x_j - x_i)\varepsilon(x_i) = \frac{\varepsilon_0 E_0}{E(x_j)}$$

$$\frac{E(x_j)}{E_0}\sum_{i=1}^{N}\varphi(x_j-x_i)\varepsilon(x_i)=\varepsilon_0$$

This equation represent a linear system $\boldsymbol{A}^{\mathrm{S}}\boldsymbol{x}=\boldsymbol{b}$ that approximates the exact solution (at the points of discretization the solution is reproduced exactly). The numerical example for this problem is provided in Section 2.3. Matrix $\boldsymbol{A}^{\mathrm{S}}$ and vectors $\boldsymbol{x},\boldsymbol{b}$ can be expressed as follows

$$\boldsymbol{A}^{\mathrm{S}}=\frac{1}{E_0}\begin{pmatrix} E(x_1)\varphi(x_1-x_1) & E(x_1)\varphi(x_1-x_2) & \dots & E(x_1)\varphi(x_1-x_N)\\ E(x_2)\varphi(x_2-x_1) & E(x_2)\varphi(x_2-x_2) & \dots & E(x_2)\varphi(x_2-x_N)\\ \vdots & \vdots & \ddots & \vdots\\ E(x_N)\varphi(x_N-x_1) & E(x_N)\varphi(x_N-x_2) & \dots & E(x_N)\varphi(x_N-x_N)\end{pmatrix} \tag{14}$$

$$\boldsymbol{x}=\begin{Bmatrix}\varepsilon(x_1) & \varepsilon(x_2) & \dots & \varepsilon(x_N)\end{Bmatrix}^T \tag{15}$$

$$\boldsymbol{b}=\begin{Bmatrix}\varepsilon_0 & \varepsilon_0 & \dots & \varepsilon_0\end{Bmatrix}^T \tag{16}$$

2.2 Regularization approach

In this section Equation (9)

$$\varepsilon(x)+\int_{-\infty}^{\infty}\frac{\delta(x-\xi)}{E_0}E_1(\xi)\varepsilon(\xi)d\xi=\varepsilon_0$$

is solved with different approach in comparison to Standard approach in previous section. Function $E_1(x)\varepsilon(x)$ occurring in integral is approximated using Equation (10) and basis function from Equation (13)

$$E_1(x)\varepsilon(x)\approx\mathcal{A}_h\{E_1(x)\varepsilon(x)\}=\sum_{m\in\mathbb{Z}}\varphi_H(x-hm)E_1(hm)\varepsilon(hm)$$

After substitution into Equation (9), it leads to

$$\varepsilon(x)+\frac{1}{E_0}\sum_{m\in\mathbb{Z}}\varphi_H(x-hm)E_1(hm)\varepsilon(hm)=\varepsilon_0 \tag{17}$$

Since function $E_1(x)\varepsilon(x)$ is compactly supported due to an assumption of heterogeneities being placed to compact area Ω, only finite terms in summation can be considered. Hence, regular grid of size h with coordinates $x_1,x_2,\dots,N$ that covers Ω is assumed. Then, Equation (17) can be written as:

$$\varepsilon(x)+\frac{1}{E_0}\sum_{i=1}^{N}\varphi(x-x_i)E_1(x_i)\varepsilon(x_i)=\varepsilon_0 \tag{18}$$

After simple discretization of this semi-discrete equation, the following linear system arises

$$\varepsilon(x_j)+\frac{1}{E_0}\sum_{i=1}^{N}\varphi(x_j-x_i)E_1(x_i)\varepsilon(x_i)=\varepsilon_0$$

The matrix representation of the linear system $\boldsymbol{A}^{\mathrm{R}}\boldsymbol{x} = \boldsymbol{b}$ can be expressed as follows

$$\boldsymbol{A}^{\mathrm{R}} = \frac{1}{E_0}\begin{pmatrix} 1+E_1(x_1)\varphi(x_1-x_1) & E_1(x_2)\varphi(x_1-x_2) & \dots & E_1(x_N)\varphi(x_1-x_N) \\ E_1(x_1)\varphi(x_2-x_1) & 1+E_1(x_2)\varphi(x_2-x_2) & \dots & E_1(x_N)\varphi(x_2-x_N) \\ \vdots & \vdots & \ddots & \vdots \\ E_1(x_1)\varphi(x_N-x_1) & E_1(x_2)\varphi(x_N-x_2) & \dots & 1+E_1(x_N)\varphi(x_N-x_N) \end{pmatrix} \tag{19}$$

$$\boldsymbol{x} = \begin{Bmatrix} \varepsilon(x_1) & \varepsilon(x_2) & \dots & \varepsilon(x_N) \end{Bmatrix}^T \tag{20}$$

$$\boldsymbol{b} = \begin{Bmatrix} \varepsilon_0 & \varepsilon_0 & \dots & \varepsilon_0 \end{Bmatrix}^T \tag{21}$$

2.3 Numerical examples

In this section, numerical example of Standard and Regularization approach is discussed. In both cases, regular grid about 24 points is placed at interval $\langle 0, 1\rangle$. Stiffnesses of matrix and heterogeinty are $E_0 = 1$ and $E_0 + E_1 = 10$ respectively and heterogeneity is placed only in the middle third of the considered geometry.
First of all, Figures 1 and 2 show a plot of vectors arising from linear system for $H = 1$ and $H = 2$ respectively. Circles in the figures provide information about exact solution and points of discretization. The solid and dot-and-dash line linearly connects points of vector $\boldsymbol{x} = \begin{Bmatrix} \varepsilon(x_1) & \varepsilon(x_2) & \dots & \varepsilon(x_N) \end{Bmatrix}^T$ calculated according to Standard and Regularization approach[2] respectively. In Regularization approach, calculated values immediately correspond to strain along a rod and provide smoothed or regularized variant. In the case of Figures 1 and 2, Regularization approach just connects the values calculated from linear system. In addition to that Figure 3 provide continuous variant calculated from Eq. (18)

$$\varepsilon(x) + \frac{1}{E_0}\sum_{i=1}^{N}\varphi(x-x_i)E_1(x_i)\varepsilon(x_i) = \varepsilon_0$$

Since Gaussian basis function $\varphi(x)$ is infinitely smooth, the calculated strain $\hat{\varepsilon}(x)$ is infinitely smooth as well. The "x" signs provide information about points of discretization. In contrary, in Standard approach, calculated values from linear system do not correspond to exact solution as significant Gibb's effect occur especially for greater value of parameter H. Hence, the values in vector $\boldsymbol{x}$ can not be directly regarded as a solution of the problem. Then, Figure 4 shows continuous strain calculated as

$$\hat{\varepsilon}(x) = \sum_{i=1}^{N}\varphi(x-x_i)\varepsilon(x_i)$$

where $\varepsilon(x_i)$ are components of vector $\boldsymbol{x}$ calculated from linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ as it is described in Section 2.1. It intersect exact solution in points of discretization as it is demonstrated in figure with circles. Finally, it can be noticed that Gibb's effect is greater than in the case of Regularization approach provided in Figure 3 Additional information on the Standard approach is provided in Bachelor's thesis [25].

[2]Section 2.1 and 2.2 resp.

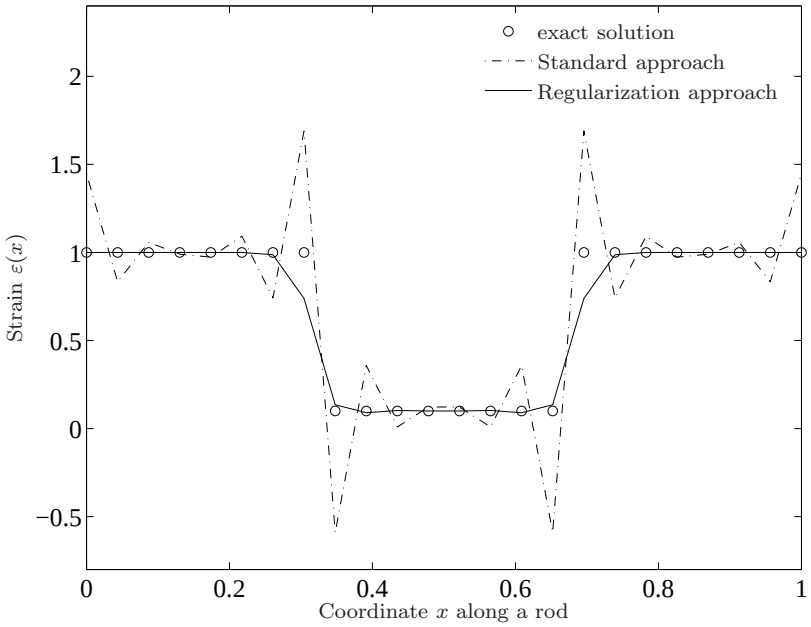

Figure 1: Plot of vectors arising from linear system $H = 1$

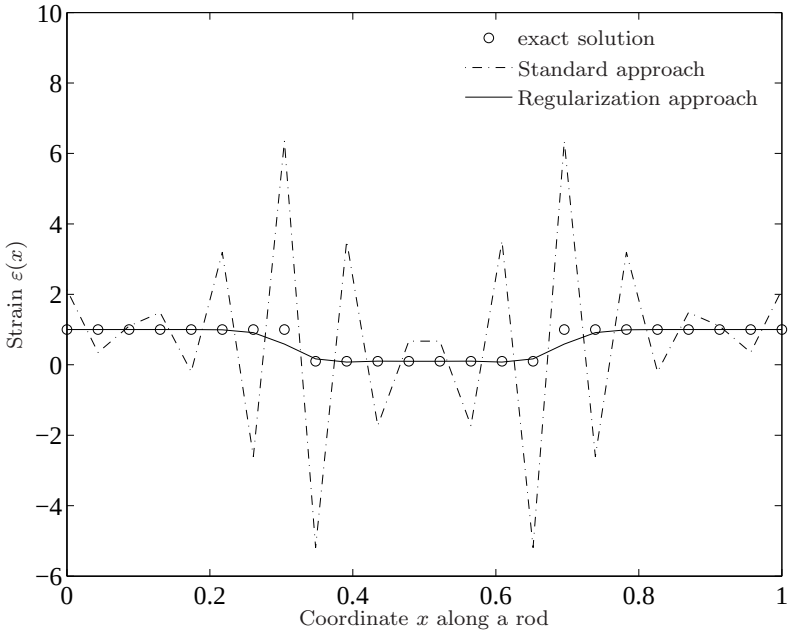

Figure 2: Plot of vectors arising from linear system $H = 2$

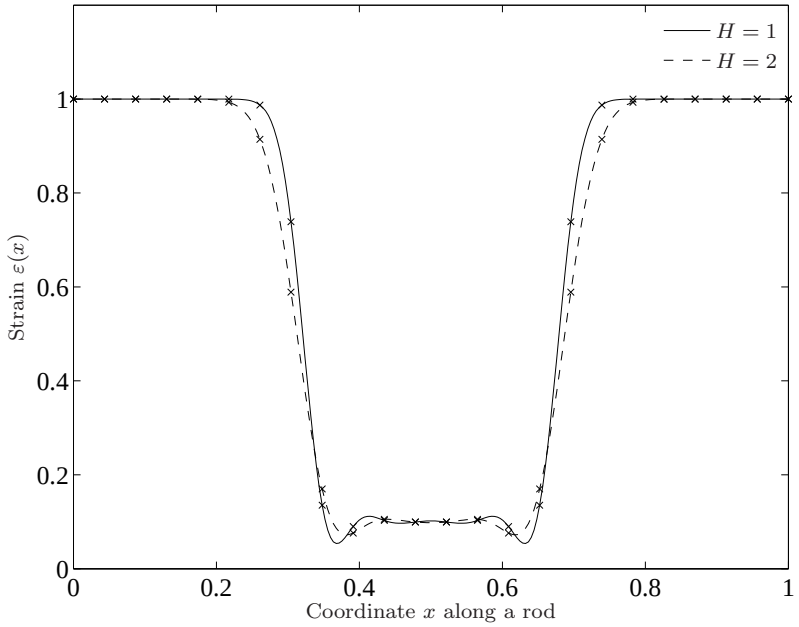

Figure 3: Continuous Regularization approach

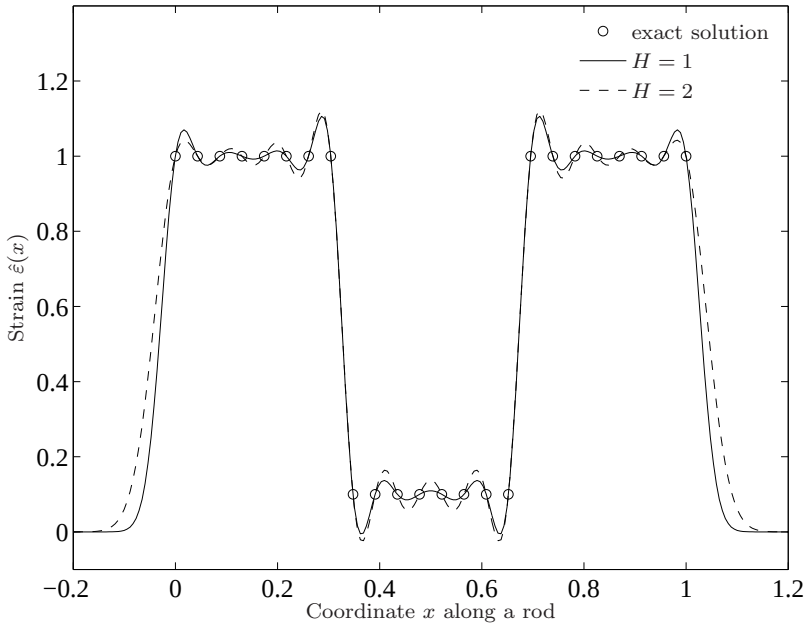

Figure 4: Continuous expression of strain in Standard approach

2.4 Multiplication using FFT

This section describes multiplication $\boldsymbol{Cx}$, where $\boldsymbol{C}$ is a circulant matrix and $\boldsymbol{x}$ is an arbitrary vector, that can be provided using Discrete Fourier Transform or its sped up algorithm Fast Fourier Transform (FFT) making this multiplication in $\mathcal{O}(3n\log(n))$ computational time instead of standard $\mathcal{O}(n^2)$. The matrices arising from GAF method are not exactly circulant but Toeplitz or block Toeplitzed structured in more dimensions. The reordering of Toeplitz matrix into circulant one is also described in this section.
First of all, a Circulant matrix $C \in \mathbb{R}^{n+1\times n+1}$ is defined as

$$\boldsymbol{C} = \begin{pmatrix} c_0 & c_{n-1} & \dots & c_2 & c_1 \\ c_1 & c_0 & c_{n-1} & \dots & c_2 \\ \vdots & c_1 & c_0 & \ddots & \vdots \\ c_{n-1} & \vdots & \ddots & \ddots & c_{n-1} \\ c_n & c_{n-1} & \dots & c_1 & c_0 \end{pmatrix}$$

meaning that it has only $n+1$ degrees of freedom. Hence, following notation is also being used

$$\boldsymbol{C} = \mathcal{C}(\boldsymbol{c})$$

with vector $\boldsymbol{c} = \{c_0, c_1, \dots, c_n\}$ getting as first column of matrix $\boldsymbol{C}$.
Since all elements in individual directions parallel to main diagonal are same, the following expression for circulant matrix can be used

$$\boldsymbol{C} = C_{ij} = c_{(i-j)\div n}$$

where binary operation "$\div$" denotes modulo operator[3]. Then, formulation of multiplication $\boldsymbol{Cx}$ leads to

$$\boldsymbol{Cx} = \sum_{j=1}^{n} C_{ij}x_j = c_{(i-j)\div n}x_j$$

that is a circulant discrete convolution and matrix $\boldsymbol{C}$ acts here as Cyclic Convolution matrix. Using circular convolution theorem, matrix multiplication can be expressed as

$$\boldsymbol{Cx} = \boldsymbol{c} * \boldsymbol{x} = \mathcal{F}_D^{-1}\big\{\mathcal{F}_D\big\{\boldsymbol{c}\big\} \otimes \mathcal{F}_D\big\{\boldsymbol{x}\big\}\big\}$$

where binary operator "$*$" and "$\otimes$" denotes cyclic convolution and pointwise multiplication of vectors respectively. Vector functions $\mathcal{F}_D$ and $\mathcal{F}_D^{-1}$ indicate Discrete Fourier Transform or, in fact, Fast Fourier Transform.
Next, reordering of symmetric Teoplitz matrix $\boldsymbol{T} \in \mathbb{R}^{n\times n}$ into Circulant matrix is discussed

$$\boldsymbol{T} = \begin{pmatrix} t_0 & t_1 & \dots & t_{n-2} & t_{n-1} \\ t_1 & t_0 & t_1 & \dots & t_{n-2} \\ \vdots & t_1 & t_0 & \ddots & \vdots \\ t_{n-2} & \vdots & \ddots & \ddots & t_1 \\ t_{n-1} & t_{n-2} & \dots & t_1 & t_0 \end{pmatrix} = \mathcal{T}(\boldsymbol{t})$$

[3] $a \div q = c \quad \Leftrightarrow \quad \exists z \in \mathbb{Z} : a = zq + c \wedge 0 \leq c < q$

where vector $\boldsymbol{t} = \{t_0, t_1, \ldots, t_{n-1}\}$ is first row or column of matrix $\boldsymbol{T}$. This structure can be observed in matrix $\boldsymbol{\Phi} \in \boldsymbol{T} \in \mathbb{R}^{N-1\times N-1}$ defined as (regular grid of coordinates $x_1, x_2, \ldots, x_N$ is assumed)

$$\boldsymbol{\Phi} = \Phi_{ij} = \varphi(x_i - x_j) = \varphi(h|i-j|), \quad \text{for } i, j = 1, 2, \ldots, N-1 \tag{22}$$

and it can be visualized as

$$\boldsymbol{\Phi} = \begin{pmatrix} \varphi(0) & \varphi(h) & \ldots & \varphi\big(h(N-1)\big) \\ \varphi(h) & \varphi(0) & \ldots & \varphi(h(N-2)) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi(h(N-1)) & \varphi(h(N-2)) & \ldots & \varphi(0) \end{pmatrix}$$

Hence, symmetric Toeplitz matrix $\boldsymbol{T} \in \mathbb{R}^{n\times n}$ can be recast into circulant matrix $\boldsymbol{C} \in \mathbb{R}^{2n-2\times 2n-2}$ as follows

$$\boldsymbol{C} = \mathcal{C}(\boldsymbol{c}) = \mathcal{C}\big(\{t_0, t_1, \ldots, t_{n-2}, t_{n-1}, t_{n-2}, t_{n-3}, \ldots, t_2, t_1\}\big)$$

This matrix is represented as

$$\boldsymbol{C} = \left(\begin{array}{ccccccc|ccccc} t_0 & t_1 & t_2 & \ldots & t_{n-3} & t_{n-2} & t_{n-1} & t_{n-2} & t_{n-3} & t_{n-4} & \ldots & t_1 \\ t_1 & t_0 & t_1 & \ldots & t_{n-4} & t_{n-3} & t_{n-2} & t_{n-1} & t_{n-2} & t_{n-3} & \ldots & t_2 \\ t_2 & t_1 & t_0 & \ldots & t_{n-5} & t_{n-4} & t_{n-3} & t_{n-2} & t_{n-1} & t_{n-2} & \ldots & t_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ t_{n-3} & t_{n-4} & t_{n-5} & \ldots & t_0 & t_1 & t_2 & t_3 & t_4 & t_5 & \ldots & t_{n-2} \\ t_{n-2} & t_{n-3} & t_{n-4} & \ldots & t_1 & t_0 & t_1 & t_2 & t_3 & t_4 & \ldots & t_{n-1} \\ t_{n-1} & t_{n-2} & t_{n-3} & \ldots & t_2 & t_1 & t_0 & t_1 & t_2 & t_3 & \ldots & t_{n-2} \\ \hline t_{n-2} & t_{n-1} & t_{n-2} & \ldots & t_3 & t_2 & t_1 & t_0 & t_1 & t_2 & \ldots & t_{n-3} \\ t_{n-3} & t_{n-2} & t_{n-1} & \ldots & t_4 & t_3 & t_2 & t_1 & t_0 & t_1 & \ldots & t_{n-4} \\ t_{n-4} & t_{n-3} & t_{n-2} & \ldots & t_5 & t_4 & t_3 & t_2 & t_1 & t_0 & \ldots & t_{n-5} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ t_1 & t_2 & t_3 & \ldots & t_{n-2} & t_{n-1} & t_{n-2} & t_{n-3} & t_{n-4} & t_{n-5} & \ldots & t_0 \end{array}\right) \tag{23}$$

With the help of operators from Definition 2.1, it can be also expressed as

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{T} & \mathrm{lr}\big(\boldsymbol{T}^{;1,N}\big) \\ \mathrm{ud}\big(\boldsymbol{T}^{1,N;}\big) & \boldsymbol{T}^{1,N;1,N} \end{pmatrix} \tag{24}$$

Definition 2.1. *Let $\boldsymbol{A}$ be a matrix from $\mathbb{R}^{m\times n}$. Next, $r_1, r_2, \ldots, r_p$ be mutually different natural numbers smaller than m and $c_1, c_2, \ldots, c_q$ be also mutually different natural numbers smaller than n in this case.*
Following matrix functions are introduced. $\boldsymbol{A}^{r_1,r_2,\ldots,r_p;c_1,c_2,\ldots,c_q}$ denotes mapping of matrix A into $\mathbb{R}^{m-p\times n-q}$ that removes rows $r_1, r_2, \ldots, r_p$ and columns $c_1, c_2, \ldots, c_q$ from matrix $\boldsymbol{A}$. Next, function $\mathrm{lr}(\boldsymbol{A})$ produces a matrix of the same dimension whose columns are flipped in the left-right direction, that is, about a vertical axis. Analogically, function $\mathrm{ud}(\boldsymbol{A})$ produces a matrix of same dimension whose rows are flipped in the up-down direction, that is, about a horizontal axis.

For calculating product $\boldsymbol{T}\boldsymbol{x}$ using circulant matrix $\boldsymbol{C}$ and FFT algorithm, it is necessary to enlarge vector $\boldsymbol{x}$ as

$$\bar{\boldsymbol{x}} = \begin{Bmatrix} \boldsymbol{x} \\ \boldsymbol{o} \end{Bmatrix}$$

where matrix $\boldsymbol{o}$ is zero matrix from space $\mathbb{R}^{n-2\times 1}$. Hence, using Equation (24) defining matrix $\boldsymbol{C}$ it heads to

$$\boldsymbol{C}\bar{\boldsymbol{x}} = \begin{Bmatrix} \boldsymbol{T}\boldsymbol{x} + \mathrm{lr}\big(\boldsymbol{T}^{;1,N}\big)\boldsymbol{o} \\ \mathrm{ud}\big(\boldsymbol{T}^{1,N;}\big)\boldsymbol{x} + \boldsymbol{T}^{1,N;1,N}\boldsymbol{o} \end{Bmatrix} = \begin{Bmatrix} \boldsymbol{T}\boldsymbol{x} \\ \mathrm{ud}\big(\boldsymbol{T}^{1,N;}\big)\boldsymbol{x} \end{Bmatrix}$$

Thus only the first part of matrix with vector multiplication is considered.
Finally, it is necessary to say that matrices $\boldsymbol{A}^{\mathrm{S}}$ and $\boldsymbol{A}^{\mathrm{R}}$ from linear system of Standard and Regularization approach defined in Equation (14) and (19) respectively are not Toeplitz matrices. Nevertheless, matrix with vector multiplication can be provided using FFT algorithm after some modification. In the case of Standard approach and matrix $\boldsymbol{A}^{\mathrm{S}}$ defined in Equation (14), it can be expressed as

$$\boldsymbol{A}^{\mathrm{S}}\boldsymbol{x} = \frac{1}{E_0}\boldsymbol{e} \otimes \boldsymbol{\Phi}\boldsymbol{x}$$

where $\boldsymbol{e} = \{E(x_1), E(x_2), \ldots, E(x_N)\}^T$ is vector, "$\otimes$" is pointwise multiplication and $\boldsymbol{\Phi}$ is matrix defined in Equation (22) and possessing the Toeplitz structure. Thus $\boldsymbol{\Phi}\boldsymbol{x}$ can be calculated using the FFT algorithm. Similarly, $\boldsymbol{A}^{\mathrm{R}}$ from Regular approach defined in Equation (14) can be multiplied by vector as follows

$$\boldsymbol{A}^{\mathrm{R}}\boldsymbol{x} = \frac{1}{E_0}[\boldsymbol{x} + \boldsymbol{\Phi}(\boldsymbol{e} \otimes \boldsymbol{x})]$$

Finally, it is necessary to emphasise that analogical multiplication can also be used in two or three dimensional space.

Part III
FFH method

3 Theory

The method deals with periodically repeating medium represented by a unit cell. Without the loss of generality, it can be assumed that the unit cell is defined at a finite interval $[0, L]$ with stiffness distribution characterized by function $E(x)$.
Now consider a unit cell subject to an average strain ε_0, leading to the total strain decomposition

$$\varepsilon(x) = \varepsilon_0 + \varepsilon_1(x) \tag{25}$$

with $\varepsilon_1(x)$ begin $[0, L]$ periodic with zero mean

$$\int_0^L \varepsilon_1(x)\,\mathrm{d}x = 0 \tag{26}$$

The governing equations of one-dimensional elasticity read

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(E(x)A(x)\frac{\mathrm{d}u(x)}{\mathrm{d}x}\right) = 0 \tag{27}$$

with the cross-section area set to $A(x) = 1\mathrm{m}^2$ for the sake of simplicity. To proceed, we introduce an auxiliary stiffness E_{ref} and rewrite (27) as

$$\frac{\mathrm{d}}{\mathrm{d}x}\Big[\Big(E_{\mathrm{ref}} - E_{\mathrm{ref}} + E(x)\Big)\varepsilon(x)\Big] = 0 \tag{28}$$

Using the decomposition (25) and noting that $\frac{\mathrm{d}}{\mathrm{d}x}\big(E_H\varepsilon_0\big) = 0$, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}x}\Big(E_{\mathrm{ref}}\varepsilon_1(x)\Big) = \frac{\mathrm{d}}{\mathrm{d}x}\Big[\Big(E_{\mathrm{ref}} - E(x)\Big)\varepsilon(x)\Big] \tag{29}$$

Hence, the original problem with heterogeneous stiffness distribution is transformed into an equivalent problem for homogeneous body with stiffness E_{ref} subject to a generalized load

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x}\Big[\Big(E_{\mathrm{ref}} - E(x)\Big)\varepsilon(x)\Big] \tag{30}$$

3.1 Fourier Transform-Based Solution

Problem (29) can be efficiently solved using the Fourier transform. Thus, the following operator is introduced [20]

$$\mathcal{F}\big\{f(x)\big\} = \hat{f}(n) = \int_0^L f(x)e^{-i\omega_n t}\,\mathrm{d}x, \qquad \omega_n = \frac{2\pi n}{L}, n \in \mathbb{Z} \tag{31}$$

with the inverse counterpart in the form

$$\mathcal{F}^{-1}\big\{\hat{f}(n)\big\} = f(x) = \frac{1}{L}\sum_{-\infty}^{\infty}\hat{f}(n)\,\mathrm{e}^{i\omega_n t}, \qquad \omega_n = \frac{2\pi n}{L}, n \in \mathbb{Z} \tag{32}$$

Employing the identity $\mathcal{F}\big\{f'(x)\big\} = \mathrm{i}\omega_n\mathcal{F}\big\{f(x)\big\}$ and linearity of operator $\mathcal{F}$, Eq. (29) leads to

$$\mathcal{F}\big\{\varepsilon_1(x)\big\} = \frac{1}{\mathrm{i}E_{\mathrm{ref}}\omega_n}\mathcal{F}\big\{f(x)\big\} \tag{33}$$

Introducing a function G via identity

$$\mathcal{F}\big\{G(x)\big\} = \frac{1}{\mathrm{i}E_{\mathrm{ref}}\omega_n} \tag{34}$$

and employing the convolution theorem yields

$$\mathcal{F}\big\{\varepsilon_1(x)\big\} = \mathcal{F}\big\{G(x)\big\}\mathcal{F}\big\{f(x)\big\} \quad \Leftrightarrow \quad \varepsilon_1(x) = \int_0^L G(x-\xi)f(\xi)\,\mathrm{d}\xi \tag{35}$$

The last equation can be integrated by parts, leading to

$$\varepsilon_1(x) \overset{P.P.}{=} \int_0^L \frac{\partial G(x-\xi)}{\partial \xi}F(\xi)\,\mathrm{d}\xi + G(x-\xi)F(\xi)\Big|_0^L \tag{36}$$

where $F(x) = (E_{\mathrm{ref}} - E(x))\varepsilon(x)$ is a primitive function to the generalized load $f(x)$. The second term in Eq. (36) vanishes due to periodicity of the involved fields.
Now, when using Eq. (34) and Fourier coefficients of the derivative, it follows that

$$\varepsilon_1(x) = \mathcal{F}^{-1}\left\{\frac{1}{E_{\mathrm{ref}}}\cdot\mathcal{F}\big\{F(x)\big\}\right\} \tag{37}$$

with the zero mean condition (26) yet to be imposed. To this end, we start with the following chain of identities

$$\int_0^L \varepsilon_1(x)\,\mathrm{d}x = 0 \Leftrightarrow \int_0^L \varepsilon_1\,\mathrm{e}^{-i\omega_n x}\,\mathrm{d}x = 0, \quad \text{for } n = 0 \Leftrightarrow \hat{\varepsilon}_1(0) = 0 \tag{38}$$

and introduce the following kernel

$$\mathcal{F}\big\{K^{\mathrm{per}}(x)\big\} = \hat{K}^{\mathrm{per}}(n) = \begin{cases} 0 & \text{for } n = 0 \\ \frac{1}{E_{\mathrm{ref}}} & \text{for } n \neq 0 \end{cases} \tag{39}$$

to obtain

$$\varepsilon_1(x) = \mathcal{F}^{-1}\big\{\hat{K}^{\mathrm{per}}(n)\hat{F}(n)\big\} \tag{40}$$

where

$$\hat{F}(n) = \mathcal{F}_D\big\{\big(E_{\mathrm{ref}} - E(x)\big)\varepsilon(x)\big\} \tag{41}$$

3.2 Discretization

Until now we have worked with the Fourier coefficients that transfer continuous functions into discrete frequency domain. In practical applications, we deal with functions that are defined at discrete points. Thus, we discretize the observed interval $[0, L]$ into regular grid using N nodes which leads to a sequence of nodal coordinates $\{x_i\}_{i=1}^{N}$, where $x_i = (i-1)\frac{L}{N-1}, i = 1, 2, \ldots, N$. The change of the continuous basis into the discrete one leads to the change from Fourier coefficients into discrete Fourier transform. Thus, we can rewrite Eq. (40) as

$$\varepsilon_1(x_i) = \mathcal{F}_D^{-1}\Big\{\hat{K}^{\mathrm{per}}(n)\mathcal{F}_D\big\{F(x_i)\big\}\Big\} \tag{42}$$

where the discrete Fourier transform is defined as

$$\mathcal{F}_D\big\{x_n\big\} = X_k = \sum_{n=0}^{N-1} x_n \, \mathrm{e}^{\frac{-2\pi \mathrm{i}}{N}kn}, \quad k = 0, 1, 2, \ldots, N-1 \tag{43}$$

and the inverse discrete Fourier transform as:

$$\mathcal{F}_D^{-1}\big\{X_k\big\} = x_n = \frac{1}{N}\sum_{k=0}^{N-1} X_k \, \mathrm{e}^{\frac{2\pi \mathrm{i}}{N}kn}, \quad n = 0, 1, 2, \ldots, N-1 \tag{44}$$

Note that Eq. (42), the periodic Lippman-Schwinger equation, is equivalent to finding a fixed point of a mapping defined by its right hand side. Following the original idea [17], the equation is solved iteratively. This yields following algorithm.

Require: Parameter E_{ref}, $\boldsymbol{E} \in \mathbb{R}^{N\times 1} : \boldsymbol{E} = \big\{E(x_1), E(x_2), \ldots, E(x_N)\big\}^T$ and ε_0
Ensure: $\boldsymbol{\varepsilon} \in \mathbb{R}^{N\times 1} : \boldsymbol{\varepsilon} = \big\{\varepsilon(x_1), \varepsilon(x_2), \ldots, \varepsilon(x_N)\big\}^T$.

1: Set $\hat{\boldsymbol{K}}^{\mathrm{per}} \in \mathbb{R}^{N\times 1} : \hat{\boldsymbol{K}}^{\mathrm{per}} := \big\{0, 1/E_{\mathrm{ref}}, \ldots, 1/E_{\mathrm{ref}}\big\}^T$, $\boldsymbol{\varepsilon} \in \mathbb{R}^{N\times 1} : \boldsymbol{\varepsilon} = \big\{\varepsilon_0, \varepsilon_0, \ldots, \varepsilon_0\big\}^T$
2: **while** until convergence **do**
3: $\quad \boldsymbol{F} := (E_{\mathrm{ref}} - \boldsymbol{E}) \oslash \boldsymbol{\varepsilon}$ {Binary operation "$\oslash$" denotes pointwise division.}
4: $\quad \boldsymbol{\varepsilon} := \mathcal{F}_D^{-1}\Big\{\hat{\boldsymbol{K}}^{\mathrm{per}} \otimes \mathcal{F}_D\big\{\boldsymbol{F}\big\}\Big\} + \varepsilon_0$ {Binary operation "$\otimes$" denotes pointwise multiplication.}
5: **end while**

Algorithm 1: Standard algorithm of FFH method

4 Convergence study

The analysis of the convergence behavior will be based on Algorithm (1) that can be rewritten in the following form

$$\varepsilon_1(x_i, k+1) = \sum_{j=1}^{N} K^{\mathrm{per}}(x_i - x_j)\big(E_{\mathrm{ref}} - E(x_j)\big)\varepsilon(x_j, k) \tag{45}$$

where, e.g. $\varepsilon_1(x_i, k+1)$ denotes the corresponding strain value at the $(k+1)$-th iteration. To express action of the kernel K^{per} in the real space, we start from the identity

$$\mathcal{F}_D^{-1}\left\{\frac{1}{E_{\mathrm{ref}}}\right\} = \frac{\delta(x)}{E_{\mathrm{ref}}} \tag{46}$$

and add an appropriate term corresponding to the requirement $\hat{K}^{\mathrm{per}}(0) = 0$. Now, it is possible to express Eq. (45) explicitly in the form

$$\varepsilon_1(x_j, k+1) = \frac{E_{\mathrm{ref}} - E(x_j)}{E_{\mathrm{ref}}}\varepsilon(x_j, k) - \frac{1}{N}\sum_{i=1}^{N}\frac{E_{\mathrm{ref}} - E(x_i)}{E_{\mathrm{ref}}}\varepsilon(x_i, k), \quad \text{for } j = 1, 2, \ldots, N \tag{47}$$

In following analysis, a binary heterogeneous rod with phase stiffnesses E_0 and $E_0(1+p)$, $p \in (-1, \infty\rangle$ is considered. An arbitrary node with coordinate x_i and stiffness E_0 is denoted as x_{E_0} and, analogically, a node with stiffness $E_0(1+p)$ as $x_{E_0(1+p)}$. The periodic cell is discretized using N nodes and we can assume that stiffness $E_0(1+p)$ occurs in m cases. Hence, it is obvious that stiffness E_0 occurs in $(N-m)$ cases. All used variables related to both sequences are shown in Table 1, the domains of definition of these variables are provided in Table 2.

Table 1: Sequences and related variables

studied sequence	coordinate	iteration	stiffness $E(x)$	nodes	frequency	volume fraction
$\varepsilon_1(x_{E_0}, k)$	x_{E_0}	k	E_0	N	$N-m$	$\frac{N-m}{N} = 1-c$
$\varepsilon_1(x_{E_0(1+p)}, k)$	$x_{E_0(1+p)}$	k	$E_0(1+p)$	N	m	$\frac{m}{N} = c$

Table 2: Domains of definition

variables	E_0, $E_0(1+p), E_{\mathrm{ref}}$	p	x	N , m , $(N-m)$	c	k
def. domain	$\mathbb{R}^+$	$(-1, \infty\rangle$	$\langle a, b\rangle$	$\mathbb{N}$	$\langle 0, 1\rangle$	$\mathbb{N}_0$

Now, we can take a look at ε_1 strain in the characteristic node x_{E_0}. Using Eq. (47), the recurrence relation for the strain ε_1 as the $(k+1)^{\mathrm{th}}$ iteration can be written as follows

$$\begin{aligned}\varepsilon_1(x_{E_0}, k+1) = {} & \frac{E_{\mathrm{ref}} - E_0}{E_{\mathrm{ref}}}\varepsilon(x_{E_0}, k) - \\ & - \frac{1}{N}\left[(N-m)\frac{E_{\mathrm{ref}} - E_0}{E_{\mathrm{ref}}}\varepsilon(x_{E_0}, k) + m\frac{E_{\mathrm{ref}} - E_0(1+p)}{E_{\mathrm{ref}}}\varepsilon(x_{E_0(1+p)}, k)\right]\end{aligned} \tag{48}$$

Subsequently, we can use condition that average strain of $\varepsilon_1(x,k)$ is equal to zero, yielding

$$\sum_{i=1}^{N}\varepsilon_1(x_i,k)=0 \quad \Rightarrow \quad (N-m)\varepsilon_1(x_{E_0},k)+m\varepsilon_1(x_{E_0(1+p)},k)=0 \quad \Leftrightarrow \tag{49}$$

$$\Leftrightarrow \quad \varepsilon_1(x_{E_0(1+p)},k)=\left(1-\frac{N}{m}\right)\varepsilon_1(x_{E_0},k) \tag{50}$$

It is necessary to note that even the initial strain $\varepsilon_1(x,0)$ has to satisfy condition (50). It can be noticed that both Eq. (48) and (50) can be rewritten using the volume fractions

$$c=\frac{m}{N} \tag{51}$$

where $c\in\langle 0,1\rangle$. In particular, substituting (51) into (48) and using (50) leads to

$$\begin{aligned}\varepsilon_1(x_{E_0},k+1)=\frac{E_{\text{ref}}-E_0}{E_{\text{ref}}}\big[\varepsilon_0+\varepsilon_1(x_{E_0},k)\big]-(1-c)\frac{E_{\text{ref}}-E_0}{E_{\text{ref}}}\big[\varepsilon_0+\varepsilon_1(x_{E_0},k)\big]+\\ +c\cdot\frac{E_{\text{ref}}-E_0(1+p)}{E_{\text{ref}}}\left[\varepsilon_0+\frac{1-c}{c}\varepsilon_1(x_{E_0},k)\right]\end{aligned} \tag{52}$$

In the following text, $\varepsilon_1(x_{E_0},k)$ will be abbreviated to $\varepsilon_1(k)$. Hence, after several algebraic manipulations, Eq. (52) leads to

$$\varepsilon_1(k+1)=\frac{E_{\text{ref}}-E_0(1+p-cp)}{E_{\text{ref}}}\cdot\varepsilon_1(k)+\frac{E_0\varepsilon_0 cp}{E_{\text{ref}}} \tag{53}$$

Employing the following substitutions

$$a=\frac{E_{\text{ref}}-E_0(1+p-cp)}{E_{\text{ref}}},\quad b=\frac{E_0\varepsilon_0 cp}{E_{\text{ref}}} \tag{54}$$

Eq. (53) yields

$$\varepsilon_1(k+1)=a\cdot\varepsilon_1(k)+b \tag{55}$$

which is a linear inhomogeneous recurrence relation with constant coefficients. To solve this equation, we first convert the recurrence relation into the homogeneous form. We start with writing a formula for the $(k+2)^{\text{th}}$ member

$$\varepsilon_1(k+2)=a\cdot\varepsilon_1(k+1)+b \tag{56}$$

Subtracting Eq. (55) from (56) leads to homogeneous form of linear recurrence relation

$$\varepsilon_1(k+2)-(a+1)\varepsilon_1(k+1)+a\varepsilon_1(k)=0 \tag{57}$$

which admits a solution in the form

$$\varepsilon_1(k)=rt^k \tag{58}$$

with $r \in \mathbb{R}$ and $t \in \mathbb{C}$. In order to find nontrivial solution, t has to verify the quadratic equation

$$t^2 - (a+1)t + a = 0 \tag{59}$$

from which we obtain

$$t_1 = 1, \quad t_2 = a \tag{60}$$

and the following expression for the k^{th} member of the strain sequence

$$\varepsilon_1(k) = r_1 + r_2 a^k, \quad r_1, r_2 \in \mathbb{R} \tag{61}$$

In order to determine r_1 and r_2 coefficients, we have to satisfy two linearly independent equations obtained for two members of the studied sequence

$$\varepsilon_1(0) = r_1 + r_2 \qquad a\varepsilon_1(0) + b = r_1 + r_2 a \tag{62}$$

leading to

$$r_1 = \frac{b}{1-a} \qquad r_2 = \varepsilon_1(0) + \frac{b}{a-1} \tag{63}$$

(notice that $1-a > 0$, so we do not divide by zero). Using (61), (60) and (54) in Eq. (63), the formula for the searched sequence can be written as

$$\varepsilon_1(k) = \frac{\varepsilon_0 cp}{1+p-cp} + \left(\varepsilon_1(0) - \frac{\varepsilon_0 cp}{1+p-cp}\right)\left(1 - \frac{E_0(1+p-cp)}{E_{\text{ref}}}\right)^k \tag{64}$$

Now, we investigate the sequence behavior in respect to E_{ref} parameter. First, we consider the following situation

$$\varepsilon_1(0) = \frac{\varepsilon_0 cp}{1+p-cp} \tag{65}$$

which implies

$$\varepsilon_1(k) = \varepsilon_1(0) = \frac{\varepsilon_0 cp}{1+p-cp} \tag{66}$$

thereby representing the situation when the initial solution is chosen such that is coincides with the true solution.

In all other cases, the convergence behavior is governed by the limit

$$\lim_{k\to\infty} \varepsilon_1(k) \tag{67}$$

which converges if and only if

$$\left|1 - \frac{E_0(1+p-cp)}{E_{\text{ref}}}\right| < 1 \quad \Leftrightarrow \quad 0 < \frac{E_0(1+p-cp)}{E_{\text{ref}}} < 2 \tag{68}$$

Since the first inequality in the last relation is always valid, we finally obtain

$$\frac{E_0}{2}(1+p-cp) < E_{\text{ref}} \tag{69}$$

The sequence that converges to the limit

$$\lim_{k\to\infty} \varepsilon_1(k) = \frac{\varepsilon_0 cp}{1+p-cp} \tag{70}$$

Finally, it can be noticed that if the following condition is satisfied

$$1 - \frac{E_0(1+p-cp)}{E_{\text{ref}}} = 0 \quad \Leftrightarrow \quad E_{\text{ref}} = E_0(1+p-cp) \tag{71}$$

then the algorithm converges in one step regardless of the choice of the reference media, see also [24] where an identical result was presented for $c = 50\%$ without a proof. Note that the optimal choice of the reference media is generally different from the homogenized modulus of elasticity

$$E_H = \left(\frac{c}{E_0} + \frac{1-c}{E_0(1+p)}\right)^{-1} = \frac{1+p}{1+cp}E_0 \tag{72}$$

To close the discussion on the subject, we briefly comment on behavior of the sequence $\varepsilon_1(x_{E_0(1+p)}, k)$ related to an arbitrary point with $E_0(1+p)$ stiffness. It follows from Eq. (49) that at each iteration step k the average strain of ε_1, is equal to zero. Therefore, the domain of convergence is identical to the case studied above.

For the sake of clarity, all results obtained in the current section are summarized in Table 3.

Table 3: Summary of convergence properties of FFH algorithm

E_{ref}	$\varepsilon_1(x_{E_0}, k)$	$\varepsilon_1(x_{E_0(1+p)}, k)$
$\frac{E_0}{2}(1+p-cp) < E_{\text{ref}}$	$\lim_{k\to\infty} \varepsilon_1(x_{E_0}, k) = \varepsilon_0 \frac{cp}{1+p-cp}$	$\lim_{k\to\infty} \varepsilon_1(x_{E_0(1+p)}, k) = \varepsilon_0 \frac{p(c-1)}{1+p-cp}$
$\frac{E_0}{2}(1+p-cp) \geq E_{\text{ref}}$	$\lim_{k\to\infty} \varepsilon_1(x_{E_0}, k)$ does not exist	$\lim_{k\to\infty} \varepsilon_1(x_{E_0(1+p)}, k)$ does not exist
$E_0(1+p-cp) = E_{\text{ref}}$	$\forall k > 0: \quad \varepsilon_1(k) = \varepsilon_0 \frac{cp}{1+p-cp}$	$\forall k > 0: \quad \varepsilon_1(k) = \frac{p(1-c)}{1+p-cp}$

5 Numerical examples

After the general discussion on the convergence properties of FFH, in this Section the influence of the choice of the E_{ref} parameter to the performance of the algorithm in graphical format is illustrated. To that end, we introduce the following non-dimensional measure of stiffness

$$\eta = \frac{E_{\text{ref}} - E_0}{pE_0} \tag{73}$$

Figure 5 provides information on the optimal value of E_{ref} parameter satisfying the convergence in one step represented with a solid line. The rest of the lines presents the limit value (69) ensuring the convergence of the method for different values of p. In particular, the right half-planes in the $c - \eta$ coordinate system specify the domains of convergence of the method, the complementary parts (including the limit lines) collect the choices for which the method diverges or oscillates.

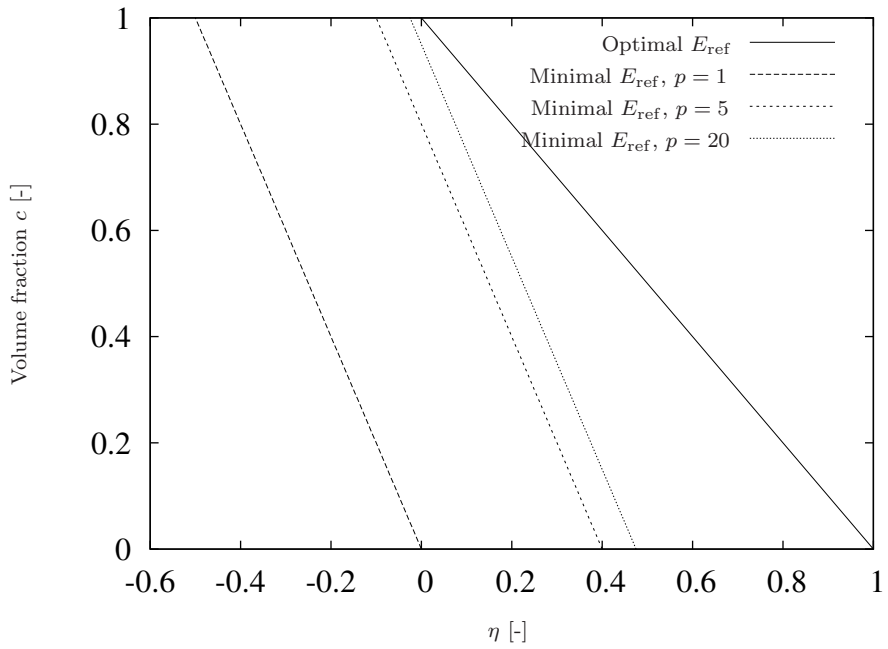

Figure 5: A domain of convergence as a function of c and E_{ref} parameters.

Next, Figure 6 shows the influence of the choice of the reference medium and volume fraction on the number of iterations. Note that the Algorithm 1 is terminated when the relative difference between the exact and approximate solution in ℓ^2 norm decreased to a value of 10^{-6}. The optimal choice of the reference medium $E_{\text{ref}} = E_0(1 + p - cp)$, leading to a one-step convergence, corresponds to the value $\eta = 1 - c$. It can be observed that increasing E_{ref} leads to a linear increase of the required number of iterations, whereas smaller values result in a super-linear increase in the number of iterations, leading to non-convergence for $\eta \leq \frac{1}{2}\left(1 - c - \frac{1}{p}\right)$.

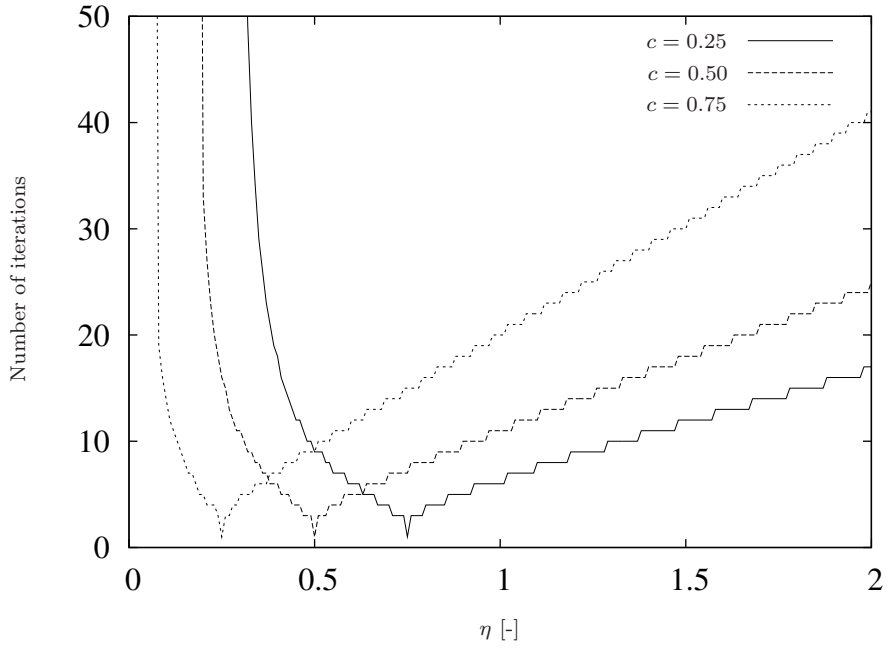

Figure 6: Number of iterations as a function of c; $p = 20$.

Finally, Figures 7 and 8 plot the number of iteration as a function of scaled E_{ref} parameter and phase modulus contrast p for two values of volume fractions c. The general trend displayed by these figures is identical to the previous case, the only difference is that higher values of p lead to notably higher increase of the number of iterations, indicating that non-optimal choice leads to an increased computational cost.

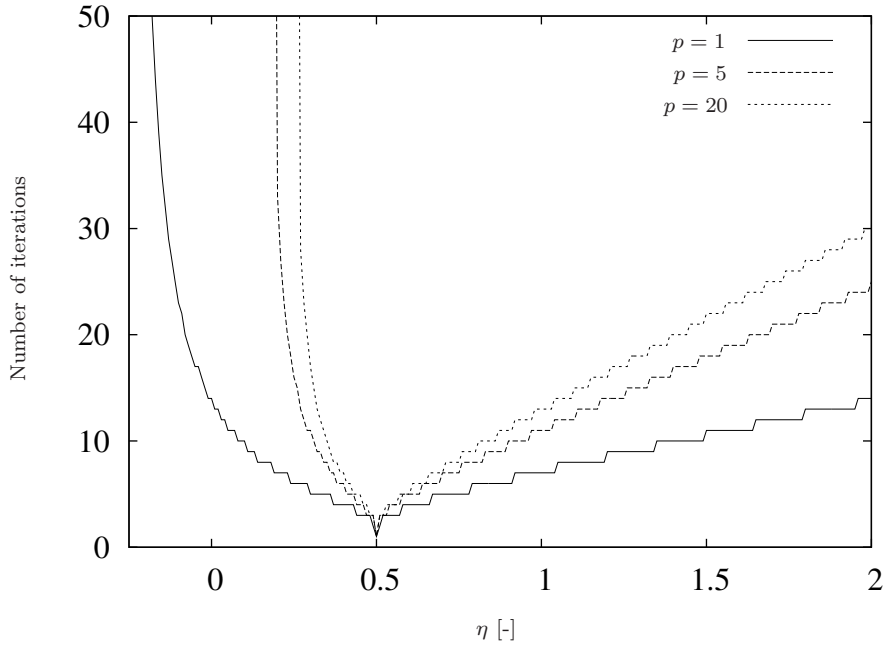

Figure 7: Number of iterations for $c = 0.5$

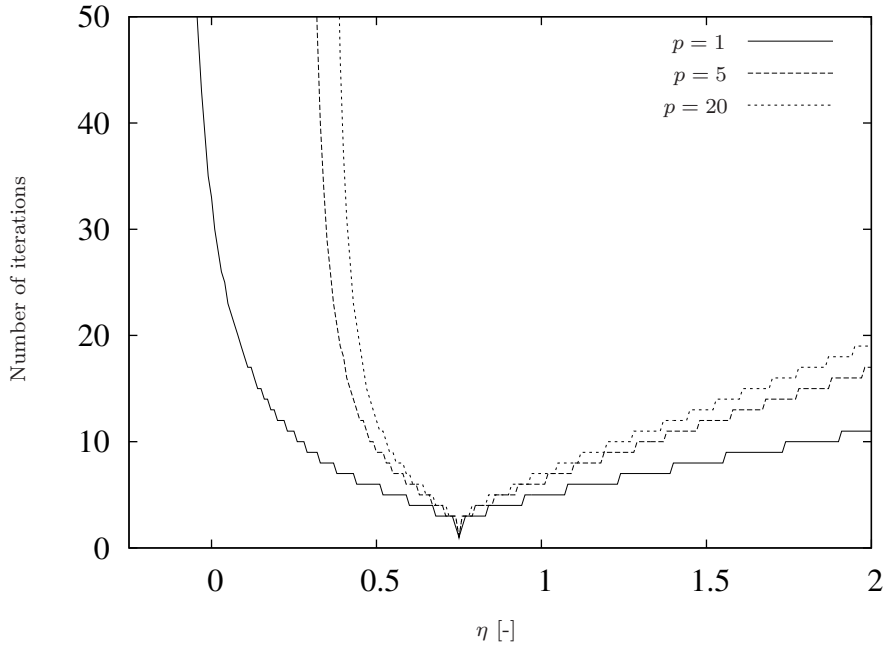

Figure 8: Number of iterations for $c = 0.25$

6 Solution using linear system

Contrary to Algorithm 1 that uses iterations of one dimensional Discrete Fourier Transform, this section deals with matrix formulation of the method. First, Section 6.1 introduces the linear system and compares it with the standard formulation of FFH method. Next, Section 6.2 describes solution of linear system using conjugate gradient methods. The application of Conjugate gradient method instead of Biconjugate gradient method for nonsymmetric linear system is discussed. Finally, we propose that two phase medium can be solved using the Conjugate gradient method in just one iteration.

6.1 Matrix formulation

The reformulation of Discrete Fourier Transform into algebraic form provided in Section 4 is utilized in this section to introduce linear system of equations. To remind the algebraic formulation of FFH method, Equation (47) is again rewritten with addition of term ε_0 to both sides of equation:

$$\varepsilon(x_j, k+1) = \frac{E_{\text{ref}} - E(x_j)}{E_{\text{ref}}}\varepsilon(x_j, k) - \frac{1}{N}\sum_{i=1}^{N}\frac{E_{\text{ref}} - E(x_i)}{E_{\text{ref}}}\varepsilon(x_i, k) + \varepsilon_0 \tag{74}$$

where $j = 1, 2, \ldots, N$.
For the subsequent analysis, the following notation is being used

$$\begin{array}{lll} \boldsymbol{x} \in \mathbb{R}^{N\times 1} & : & \boldsymbol{x} = \big\{\varepsilon(x_1), \varepsilon(x_2), \ldots, \varepsilon(x_N)\big\}^T \\ \boldsymbol{b} \in \mathbb{R}^{N\times 1} & : & \boldsymbol{b} = \big\{\varepsilon_0, \varepsilon_0, \ldots, \varepsilon_0\big\}^T \end{array} \tag{75}$$

Hence, Equation (74) can be written in matrix form as

$$\boldsymbol{x}_{k+1} = \boldsymbol{B}\boldsymbol{x}_k + \boldsymbol{b} \tag{76}$$

where matrix $\boldsymbol{B} \in \mathbb{R}^{N\times N}$ can be expressed as

$$\boldsymbol{B} = B_{ij} = \frac{E_{\text{ref}} - E(x_i)}{E_{\text{ref}}}\delta_{ij} + \frac{E(x_j) - E_{\text{ref}}}{NE_{\text{ref}}}$$

Next, we can write the linear system in form $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. Since the algebraic formulation of FFH method has been provided, it is obvious that matrix $\boldsymbol{A}$ has the following relation to matrix $\boldsymbol{B}$

$$\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{B}$$

where $\boldsymbol{I} = I_{ij} = \delta_{ij}$. Hence, matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ can be represented as

$$\boldsymbol{A} = a_{ij} = \frac{E(x_i)}{E_{\text{ref}}}\delta_{ij} + \frac{E_{\text{ref}} - E(x_j)}{NE_{\text{ref}}} \tag{77}$$

Require: Parameter E_{ref}, $\boldsymbol{E} \in \mathbb{R}^{N\times 1} : \boldsymbol{E} = \{E(x_1), E(x_2), \ldots, E(x_N)\}^T$ and ε_0
Ensure: $\boldsymbol{x} \in \mathbb{R}^{N\times 1} : \boldsymbol{x} = \{\varepsilon(x_1), \varepsilon(x_2), \ldots, \varepsilon(x_N)\}^T$

1: Set $\boldsymbol{B} \in \mathbb{R}^{N\times N} : \boldsymbol{B} := B_{ij} = \frac{E_{\text{ref}} - E(x_i)}{E_{\text{ref}}}\delta_{ij} + \frac{E(x_j) - E_{\text{ref}}}{NE_{\text{ref}}}$ and $\boldsymbol{b} \in \mathbb{R}^{N\times 1} : \boldsymbol{b} := b_i = \varepsilon_0$.
2: Set $\boldsymbol{x}_0 = \{\varepsilon_0, \varepsilon_0, \ldots, \varepsilon_0\}^T$
3: **while** until convergence **do**
4: $\quad \boldsymbol{x}_{k+1} = \boldsymbol{B}\boldsymbol{x}_k + \boldsymbol{b}$
5: **end while**

Algorithm 2: Matrix algorithm of FFH

or it can be visualized as

$$\boldsymbol{A} = \begin{pmatrix} \frac{E(x_1)}{E_{\text{ref}}} & 0 & \cdots & 0 \\ 0 & \frac{E(x_2)}{E_{\text{ref}}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{E(x_N)}{E_{\text{ref}}} \end{pmatrix} + \\ + \frac{1}{NE_{\text{ref}}} \begin{pmatrix} E_{\text{ref}} - E(x_1) & E_{\text{ref}} - E(x_2) & \cdots & E_{\text{ref}} - E(x_N) \\ E_{\text{ref}} - E(x_1) & E_{\text{ref}} - E(x_2) & \cdots & E_{\text{ref}} - E(x_N) \\ \vdots & \vdots & \ddots & \vdots \\ E_{\text{ref}} - E(x_1) & E_{\text{ref}} - E(x_2) & \cdots & E_{\text{ref}} - E(x_N) \end{pmatrix} \tag{78}$$

Finally, Equation (76) can be expressed as

$$\boldsymbol{x}_{k+1} = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x}_k + \boldsymbol{b} \tag{79}$$

Thus, the vector $\boldsymbol{\varepsilon}$ that appears at line 4 of Algorithm 1 correspond to left multiplication with matrix $\boldsymbol{B}$ and addition of vector $\boldsymbol{b}$ and it can be regarded as a basic iteration method for solution of linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. The matrix form of FFH method is provided in Algorithm 2. Its drawback is covered in necessity of storage matrix $\boldsymbol{B}$ instead of individual vectors $\boldsymbol{E}, \boldsymbol{\varepsilon}$ in Algorithm 1.

6.2 Gradient methods

In this section, the solution using gradient methods for system defined in Section 6.1 is discussed. The most popular methods are Conjugate Gradient method (CG) and Biconjugate Gradient method (BiCG). Both methods belong to projection techniques into Krylov subspace and in contrary to residual methods such as Generalized Minimal Residual Method (GMRES) they are based on short recurrence relation, hence only the last iteration vectors are necessary to store. While the CG method is suitable and converge for symmetric and positive definite matrices, BiCG method is a generalization about non-symmetric matrices. Primarily, both CG and BiCG methods will be defined and the main properties identified. Next, Theorem 6.6 shows that the nonsymmetric system defined in Section 6.1 can be solved using CG algorithm.

6.2.1 Conjugate and Biconjugate Gradient Method

Both methods cited from Saad [21] are provided in Algorithms 3 and 4. The methods themselves are constructed in order to satisfy the following orthogonality properties

$$\boldsymbol{r}_j^T \boldsymbol{r}_i = 0 \quad \text{for } i \neq j \tag{80}$$

$$\boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{p}_i = 0 \quad \text{for } i \neq j \tag{81}$$

$$\boldsymbol{r}_j^{*T} \boldsymbol{r}_i^* = 0 \quad \text{for } i \neq j \tag{82}$$

$$\boldsymbol{p}_j^{*T} \boldsymbol{A} \boldsymbol{p}_i^* = 0 \quad \text{for } i \neq j \tag{83}$$

that will be widely used in following proofs. Vectors $\boldsymbol{r}_j := \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_j$ are called residuum vectors of vectors $\boldsymbol{x}_j$, vectors $\boldsymbol{p}_j$ are conjugate directions in which the solution is searched. The starred vectors $\boldsymbol{r}_i^*$ and $\boldsymbol{p}_i^*$ arise from dual system $\boldsymbol{A}^T x^* = b^*$ and possess the same meaning as vectors $\boldsymbol{r}$ and $\boldsymbol{p}$ respectively.

Require: $\boldsymbol{A} \in \mathbb{R}^{N \times N}, \boldsymbol{b} \in \mathbb{R}^{N \times 1}, \boldsymbol{x}_0 \in \mathbb{R}^{N \times 1}$
Ensure: $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.

1: Compute $\boldsymbol{r}_0 := \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0$, $\boldsymbol{p}_0 := \boldsymbol{r}_0$
2: **for** $j = 0, 1, 2, \ldots,$ until convergence **do**
3: $\quad \alpha_j := \frac{\boldsymbol{r}_j^T \boldsymbol{r}_j}{\boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{p}_j}$
4: $\quad \boldsymbol{x}_{j+1} := \boldsymbol{x}_j + \alpha_j \boldsymbol{p}_j$
5: $\quad \boldsymbol{r}_{j+1} := \boldsymbol{r}_j - \alpha_j \boldsymbol{A} \boldsymbol{p}_j$
6: $\quad \beta_j := \frac{\boldsymbol{r}_{j+1}^T \boldsymbol{r}_{j+1}}{\boldsymbol{r}_j^T \boldsymbol{r}_j}$
7: $\quad \boldsymbol{p}_{j+1} := \boldsymbol{r}_{j+1} + \beta_j \boldsymbol{p}_j$
8: **end for**

Algorithm 3: Conjugate Gradient Method (CG) [21]

Lemma 6.1 (cited from Saad [21] without proof)**.** *The vectors produced by the Biconjugate Gradient algorithm satisfy the following orthogonality properties*

$$\boldsymbol{r}_j^T \boldsymbol{r}_i^* = 0, \quad \textit{for } i \neq j \tag{84}$$

$$\boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{p}_i^* = 0, \quad \textit{for } i \neq j \tag{85}$$

Proof. Throughout the proof the duality property of Biconjugate gradient method is fully utilized, hence it is sufficient to assume that $j > i$. Proof is provided using mathematical induction.
Base case: First, Eq. (84) will be proved for $j = 1$ and $i = 0$ noting that the second case for $j = 0$ and $j = 0$ is naturally satisfied from the duality property as stated above. Hence, the proof begins with substituting into vector $\boldsymbol{r}_1$ the expression from line 6 of BiCG Algorithm 4.

$$\boldsymbol{r}_1^T \boldsymbol{r}_0^* = (\boldsymbol{r}_0 - \alpha_0 \boldsymbol{A} \boldsymbol{p}_0)^T \boldsymbol{r}_0^* = \boldsymbol{r}_0^T \boldsymbol{r}_0^* - \alpha_0 \boldsymbol{p}_0^T \boldsymbol{A}^T \boldsymbol{r}_0^* = \boldsymbol{r}_0^T \boldsymbol{r}_0^* - \alpha_0 \boldsymbol{p}_0^T \boldsymbol{A}^T \boldsymbol{r}_0^* \tag{86}$$

Require: $\boldsymbol{A} \in \mathbb{R}^{N\times N}, \boldsymbol{b} \in \mathbb{R}^{N\times 1}, \boldsymbol{x}_0 \in \mathbb{R}^{N\times 1}$
Ensure: $\boldsymbol{Ax} = \boldsymbol{b}$.

1: Compute $\boldsymbol{r}_0 := \boldsymbol{b} - \boldsymbol{Ax}_0$. Choose $\boldsymbol{r}_0^*$ such that $\boldsymbol{r}_0^T\boldsymbol{r}_0^* \neq 0$
2: Set $\boldsymbol{p}_0 := \boldsymbol{r}_0$, $\boldsymbol{p}_0^* := \boldsymbol{r}_0^*$
3: **for** $j = 0, 1, 2, \ldots,$ until convergence **do**
4: $\quad \alpha_j := \frac{\boldsymbol{r}_j^T\boldsymbol{r}_j^*}{\boldsymbol{p}_j^T\boldsymbol{A}^T\boldsymbol{p}_j^*}$
5: $\quad \boldsymbol{x}_{j+1} := \boldsymbol{x}_j + \alpha_j\boldsymbol{p}_j$
6: $\quad \boldsymbol{r}_{j+1} := \boldsymbol{r}_j - \alpha_j\boldsymbol{Ap}_j$
7: $\quad \boldsymbol{r}_{j+1}^* := \boldsymbol{r}_j^* - \alpha_j\boldsymbol{A}^T\boldsymbol{p}_j^*$
8: $\quad \beta_j := \frac{\boldsymbol{r}_{j+1}^T\boldsymbol{r}_{j+1}^*}{\boldsymbol{r}_j^T\boldsymbol{r}_j^*}$
9: $\quad \boldsymbol{p}_{j+1} := \boldsymbol{r}_{j+1} + \beta_j\boldsymbol{p}_j$
10: $\quad \boldsymbol{p}_{j+1}^* := \boldsymbol{r}_{j+1}^* + \beta_j\boldsymbol{p}_j^*$
11: **end for**

Algorithm 4: Biconjugate Gradient Method (BiCG) [21]

Now, vector $\boldsymbol{r}_0^*$ is equal to $\boldsymbol{p}_0^*$ as stated at line 2 and finally, the substitution into term α_0 from line 4 of the algorithm guarantee that Eq. (86) heads to

$$\boldsymbol{r}_0^T\boldsymbol{r}_0^* - \alpha_0\boldsymbol{p}_0^T\boldsymbol{A}^T\boldsymbol{r}_0^* = \boldsymbol{r}_0^T\boldsymbol{r}_0^* - \frac{\boldsymbol{r}_0^T\boldsymbol{r}_0^*}{\boldsymbol{p}_0^T\boldsymbol{A}^T\boldsymbol{p}_0^*}\boldsymbol{p}_0^T\boldsymbol{A}^T\boldsymbol{p}_0^* = 0$$

Next, it is necessary to prove that $\boldsymbol{p}_1^T\boldsymbol{A}^T\boldsymbol{p}_0^* = 0$. First, vector $\boldsymbol{p}_1$ is expressed using line 9 of BiCG algorithm while vector $\boldsymbol{A}^T\boldsymbol{p}_0^*$ using line 7. Thus

$$\boldsymbol{p}_1^T\boldsymbol{A}^T\boldsymbol{p}_0^* = (\boldsymbol{r}_1 + \beta_j\boldsymbol{r}_0)^T\frac{\boldsymbol{r}_0^* - \boldsymbol{r}_1^*}{\alpha_0}$$

After cross multiplication, it is possible to utilize already proven property $\boldsymbol{r}_1^T\boldsymbol{r}_0^* = 0$ and $\boldsymbol{r}_0^T\boldsymbol{r}_1^* = 0$ respectively. Hence

$$(\boldsymbol{r}_1 + \beta_j\boldsymbol{r}_0)^T\frac{\boldsymbol{r}_0^* - \boldsymbol{r}_1^*}{\alpha_0} = \frac{1}{\alpha_0}\big(\beta_0\boldsymbol{r}_0^T\boldsymbol{r}_0^* - \boldsymbol{r}_1^T\boldsymbol{r}_1^*\big)$$

and finally, the substitution into β_0 from line 8 heads to the end of base case part.

$$\frac{1}{\alpha_0}\big(\beta_0\boldsymbol{r}_0^T\boldsymbol{r}_0^* - \boldsymbol{r}_1^T\boldsymbol{r}_1^*\big) = \frac{1}{\alpha_0}\left(\frac{\boldsymbol{r}_1^T\boldsymbol{r}_1^*}{\boldsymbol{r}_0^T\boldsymbol{r}_0^*}\boldsymbol{r}_0^T\boldsymbol{r}_0^* - \boldsymbol{r}_1^T\boldsymbol{r}_1^*\right) = 0$$

Induction hypothesis: Now, it is assumed that following statements hold for some $j \in \mathbb{N}$

$$\begin{aligned}\forall i \in \mathbb{N}, i < j: \quad & \boldsymbol{r}_j^T\boldsymbol{r}_i^* = 0 \wedge \boldsymbol{r}_i^T\boldsymbol{r}_j^* = 0 \\ & \boldsymbol{p}_j^T\boldsymbol{A}^T\boldsymbol{p}_i^* = 0 \wedge \boldsymbol{p}_i^T\boldsymbol{A}^T\boldsymbol{p}_j^* = 0\end{aligned}$$

Induction step: Now, following equations will be proved for $i < j + 1$

$$\boldsymbol{r}_{j+1}^T \boldsymbol{r}_i^* = 0 \tag{87}$$

$$\boldsymbol{p}_{j+1}^T \boldsymbol{A}^T \boldsymbol{p}_i^* = 0 \tag{88}$$

noting that $\boldsymbol{r}_i^T \boldsymbol{r}_{j+1}^* = 0$ and $\boldsymbol{p}_i^T \boldsymbol{A}^T \boldsymbol{p}_{j+1}^* = 0$ for $i < j + 1$ will be satisfied due to duality. Then, the proof of Eq. (87) begins with substituting into vector $\boldsymbol{r}_{j+1}$ from line 7 of BiCG algorithm.

$$\boldsymbol{r}_{j+1}^T \boldsymbol{r}_i^* = (\boldsymbol{r}_j - \alpha \boldsymbol{A}\boldsymbol{p}_j)^T \boldsymbol{r}_i^* = \boldsymbol{r}_j^T \boldsymbol{r}_i^* - \alpha \boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{r}_i^*$$

Next, $\boldsymbol{r}_i^*$ is expressed using line 10 of BiCG. Using induction hypothesis, we obtain

$$\begin{aligned} \boldsymbol{r}_j^T \boldsymbol{r}_i^* - \alpha_j \boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{r}_i^* = \boldsymbol{r}_j^T \boldsymbol{r}_i^* - \alpha_j \boldsymbol{p}_j^T \boldsymbol{A}^T (\boldsymbol{p}_i^* - \beta_{i-1} \boldsymbol{p}_{i-1}^*) = \\ = \boldsymbol{r}_j^T \boldsymbol{r}_i^* - \alpha_j \boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{p}_i^* - \alpha_j \beta_{i-1} \boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{p}_{i-1}^* = \boldsymbol{r}_j^T \boldsymbol{r}_i^* - \alpha_j \boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{p}_i^* \end{aligned}$$

Assuming $i < j$, it is immediately equal to zero as induction hypothesis holds.

$$\boldsymbol{r}_j^T \boldsymbol{r}_i^* - \alpha_j \boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{p}_i^* = 0$$

Otherwise $i = j$ and koeficient α_j can be expressed using line 4 of BiCG Algorithm. Hence

$$\boldsymbol{r}_j^T \boldsymbol{r}_j^* - \alpha_j \boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{p}_j^* = \boldsymbol{r}_j^T \boldsymbol{r}_j^* - \frac{\boldsymbol{r}_j^T \boldsymbol{r}_j^*}{\boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{p}_j^*} \boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{p}_j^* = 0$$

Next, Eq. (88) will be proved. First, vector $\boldsymbol{p}_{j+1}$ is expressed using line 9 of BiCG algorithm.

$$\boldsymbol{p}_{j+1}^T \boldsymbol{A}^T \boldsymbol{p}_i^* = (\boldsymbol{r}_{j+1} + \beta_j \boldsymbol{p}_j)^T \boldsymbol{A}^T \boldsymbol{p}_i^* = \boldsymbol{r}_{j+1}^T \boldsymbol{A}^T \boldsymbol{p}_i^* + \beta_j \boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{p}_i^*$$

Term $\boldsymbol{A}^T \boldsymbol{p}_i^*$ can be expressed using line 7 of BiCG Algorithm and β_j can represented as in line 8. Hence

$$\boldsymbol{r}_{j+1}^T \boldsymbol{A}^T \boldsymbol{p}_i^* + \beta_j \boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{p}_i^* = \boldsymbol{r}_{j+1}^T \frac{\boldsymbol{r}_i^* - \boldsymbol{r}_{i+1}^*}{\alpha_i} + \frac{\boldsymbol{r}_{j+1}^T \boldsymbol{r}_{j+1}^*}{\boldsymbol{r}_j^T \boldsymbol{r}_j^*} \boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{p}_i^*$$

If it is assumed that $i < j$ than previous equation immediately equals to zero due to induction hypothesis. Otherwise $i = j$ and it also leads to the fact that it is equal to zero as the induction hypothesis is used and α_i is substituted

$$\boldsymbol{r}_{j+1}^T \frac{\boldsymbol{r}_i^* - \boldsymbol{r}_{i+1}^*}{\alpha_i} + \frac{\boldsymbol{r}_{j+1}^T \boldsymbol{r}_{j+1}^*}{\boldsymbol{r}_j^T \boldsymbol{r}_j^*} \boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{p}_i^* = -\frac{\boldsymbol{r}_{j+1}^T \boldsymbol{r}_{i+1}^*}{\alpha_i} + \frac{\boldsymbol{r}_{j+1}^T \boldsymbol{r}_{j+1}^*}{\boldsymbol{r}_j^T \boldsymbol{r}_j^*} \boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{p}_i^* =$$

$$= -\frac{\boldsymbol{r}_{j+1}^T \boldsymbol{r}_{i+1}^*}{\boldsymbol{r}_j^T \boldsymbol{r}_j^*} \boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{p}_i^* + \frac{\boldsymbol{r}_{j+1}^T \boldsymbol{r}_{j+1}^*}{\boldsymbol{r}_j^T \boldsymbol{r}_j^*} \boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{p}_i^* = 0$$

□

Definition 6.2. *Square matrices* $\boldsymbol{S}, \boldsymbol{W} \in \mathbb{R}^{N\times N}$ *are defined as*

$$\boldsymbol{S} = S_{ij} = \frac{E(x_i)}{E_{\text{ref}}}\delta_{ij} \tag{89}$$

$$\boldsymbol{W} = W_{ij} = \frac{E_{\text{ref}} - E(x_j)}{NE_{\text{ref}}} \tag{90}$$

.

It is obvious that matrices $\boldsymbol{A}, \boldsymbol{S}$ and $\boldsymbol{W}$ satisfy the following relation

$$\boldsymbol{A} = \boldsymbol{S} + \boldsymbol{W}$$

.

Definition 6.3. *Subspaces* $\mathcal{E}^N$ *and* $\mathcal{E}_0^N$ *of vector space* $\mathbb{R}^{N\times 1}$ *are defined as*

$$\mathcal{E}^N = \left\{ x \in \mathbb{R}^{N\times 1}; \frac{1}{N}\sum_{i=1}^{N} x_i = \varepsilon_0 \right\}$$

$$\mathcal{E}_0^N = \left\{ x \in \mathbb{R}^{N\times 1}; \frac{1}{N}\sum_{i=1}^{N} x_i = 0 \right\}$$

Lemma 6.4. *Let* $\boldsymbol{S}$ *and* $\boldsymbol{W}$ *be matrices defined in Equations* (89) *and* (90). *Then following relation holds*

$$\boldsymbol{W}^T\boldsymbol{S} + \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{W}^T \tag{91}$$

Proof. The proof is executed by simple algebraic emendations that are follows

$$W^TS + W^TW = \frac{E(x_j)[E_{\text{ref}} - E(x_i)]}{NE_{\text{ref}}^2} + \frac{[E_{\text{ref}} - E(x_i)][E_{\text{ref}} - E(x_j)]}{NE_{\text{ref}}^2} =$$

$$= \frac{E(x_j)E_{\text{ref}} - E(x_i)E(x_j) + E_{\text{ref}}^2 - E_{\text{ref}}E(x_j) - E_{\text{ref}}E(x_i) + E(x_i)E(x_j)}{NE_{\text{ref}}^2} =$$

$$= \frac{E_{\text{ref}}^2 - E_{\text{ref}}E(x_i)}{NE_{\text{ref}}^2} = \frac{E_{\text{ref}} - E(x_i)}{NE_{\text{ref}}} = W^T$$

□

Lemma 6.5. *Let* $\boldsymbol{W} \in \mathbb{R}^{N\times N}$ *be a matrix defined in Equation* (90) *and* $\boldsymbol{x} \in \mathbb{R}^{N\times 1}$ *be a vector from* $\mathcal{E}_0$. *Then*

$$\boldsymbol{x}^T\boldsymbol{W} = \boldsymbol{o}^T$$

where $\boldsymbol{o} \in \mathbb{R}^{N\times 1}$ *is zero vector (all of the components are equal to zero).*

Proof. Firstly, components of vector $\boldsymbol{x}$ are marked as x_j for $j = 1, 2, \ldots, N$. Hence, it is possible to write

$$\boldsymbol{r}^T\boldsymbol{W} = \sum_{i=1}^{N} r_i W_{ij} = \sum_{i=1}^{N} r_i \frac{E_{\text{ref}} - E(x_j)}{NE_{\text{ref}}} = \frac{E_{\text{ref}} - E(x_j)}{NE_{\text{ref}}} \sum_{i=1}^{N} r_i$$

Here, it is possible to apply the property of space $\mathcal{E}_0$ that $\frac{1}{N}\sum_{i=1}^{N} r_i = 0$. Hence, it immediately follows that

$$\boldsymbol{r}^T\boldsymbol{W} = \boldsymbol{o}^T$$

□

Theorem 6.6. *Let $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ be a linear system with matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ defined in Equation (77) and with vector $\boldsymbol{b} \in \mathbb{R}^{N\times 1}$ defined in Equation (75). Next, $\boldsymbol{x}_0$ be a vector from space $\mathcal{E}^N$. Then solution of linear system using Biconjugate Gradient method defined in Algorithm 4 is equivalent to Conjugate Gradient method defined in Algorithm 3. In addition, the following equations are satisfied*

$$\boldsymbol{r}_j\boldsymbol{r}_j^* = \boldsymbol{r}_j\boldsymbol{r}_j \tag{92}$$

$$\boldsymbol{p}_j^T\boldsymbol{A}^T\boldsymbol{p}_j^* = \boldsymbol{p}_j^T\boldsymbol{A}^T\boldsymbol{p}_j \tag{93}$$

Proof. Since vector $\boldsymbol{r}_0^*$ can be choosen arbitrary such that $\boldsymbol{r}_0^T\boldsymbol{r}_0^*$, it is possible to take $\boldsymbol{r}_0^*$ equal to $\boldsymbol{r}_0$. The equivalence of BiCG and CG algorithms are satisfied if coefficients from BiCG algorithm α_j and β_j are equal to coefficients of CG algorithm. In this case it naturally arises that Equations (92) and (93) have to be satisfied. Again, the proof will be based on mathematical induction.
Base case: Firstly, the basis of mathematical induction is established. It consists of showing that Equations (92) and (93) hold for $j = 0$. It is simply satisfied since

$$\boldsymbol{r}_0 = \boldsymbol{r}_0^* = \boldsymbol{p}_0 = \boldsymbol{p}_0^*$$

Next, it is necessary to show that coefficient β_0 is same in both BiCG and CG algorithms. Since Equation (92) holds for $j = 0$ it still require to prove it for $j = 1$. We investigate the difference of following inner products

$$\boldsymbol{r}_1^T\boldsymbol{r}_1 - \boldsymbol{r}_1^T\boldsymbol{r}_1^* = \boldsymbol{r}_1^T(\boldsymbol{r}_1 - \boldsymbol{r}_1^*) \tag{94}$$

Since the equations for the first residuum are as follows

$$\boldsymbol{r}_1 = \boldsymbol{r}_0 - \alpha_0\boldsymbol{A}\boldsymbol{p}_0 = \boldsymbol{r}_0 - \alpha_0\boldsymbol{A}\boldsymbol{r}_0 \tag{95}$$

$$\bar{\boldsymbol{r}}_1 = \boldsymbol{r}_0 - \alpha_0\boldsymbol{A}^T\boldsymbol{p}_0^* = \boldsymbol{r}_0 - \alpha_0\boldsymbol{A}^T\boldsymbol{r}_0 \tag{96}$$

Equation (94) can be recast as

$$\boldsymbol{r}_1^T(\boldsymbol{r}_1 - \boldsymbol{r}_1^*) = \boldsymbol{r}_1^T(\alpha_0\boldsymbol{A}^T\boldsymbol{r}_0 - \alpha_0\boldsymbol{A}\boldsymbol{r}_0) = \alpha_0\boldsymbol{r}_1^T(\boldsymbol{A}^T - \boldsymbol{A})\boldsymbol{r}_0 \tag{97}$$

Using expression for vector $\boldsymbol{r}_1$ as in Equation (95) and the property that $\boldsymbol{x}^T(\boldsymbol{A}-\boldsymbol{A}^T)\boldsymbol{x}$ is equal to zero for arbitrarily taken vector $\boldsymbol{x}$, Equation (97) leads to

$$\alpha_0\boldsymbol{r}_1^T(\boldsymbol{A}^T-\boldsymbol{A})\boldsymbol{r}_0=\alpha_0(\boldsymbol{r}_0-\alpha_0\boldsymbol{A}\boldsymbol{r}_0)^T(\boldsymbol{A}^T-\boldsymbol{A})\boldsymbol{r}_0=\alpha_0^2\boldsymbol{r}_0^T(\boldsymbol{A}^T\boldsymbol{A}-\boldsymbol{A}^2)\boldsymbol{r}_0 \tag{98}$$

Next, matrix $(\boldsymbol{A}^T\boldsymbol{A}-\boldsymbol{A}^2)$ can be expressed using matrices S and W as follows

$$\begin{aligned}&\boldsymbol{A}^T\boldsymbol{A}-\boldsymbol{A}^2=(\boldsymbol{S}+\boldsymbol{W})^T(\boldsymbol{S}+\boldsymbol{W})+(\boldsymbol{S}+\boldsymbol{W})^2=\\&=\boldsymbol{S}^2+\boldsymbol{S}\boldsymbol{W}+\boldsymbol{W}^T\boldsymbol{S}+\boldsymbol{W}^T\boldsymbol{W}-\boldsymbol{S}^2-\boldsymbol{S}\boldsymbol{W}-\boldsymbol{W}\boldsymbol{S}-\boldsymbol{W}^2=\boldsymbol{W}^T\boldsymbol{S}+\boldsymbol{W}^T\boldsymbol{W}-\boldsymbol{W}\boldsymbol{S}-\boldsymbol{W}^2\end{aligned} \tag{99}$$

With the help of previous Equation (99) that can be substituted into Equation (98), the proof heads to

$$\begin{aligned}\alpha_0^2\boldsymbol{r}_0^T(\boldsymbol{A}^T\boldsymbol{A}-\boldsymbol{A}^2)\boldsymbol{r}_0=\alpha_0^2\boldsymbol{r}_0^T(\boldsymbol{W}^T\boldsymbol{S}+\boldsymbol{W}^T\boldsymbol{W}-\boldsymbol{W}\boldsymbol{S}-\boldsymbol{W}^2)\boldsymbol{r}_0=\\=\alpha_0^2\boldsymbol{r}_0^T(\boldsymbol{W}^T\boldsymbol{S}+\boldsymbol{W}^T\boldsymbol{W})\boldsymbol{r}_0-\alpha_0^2\boldsymbol{r}_0^T\boldsymbol{W}(\boldsymbol{S}+\boldsymbol{W})\boldsymbol{r}_0\end{aligned} \tag{100}$$

Since $\boldsymbol{r}_0$ is expressed as $\boldsymbol{r}_0=\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}_0$ then $\boldsymbol{r}_0$ is from the space $\mathcal{E}_0^N$ because

$$\langle\boldsymbol{r}_0\rangle=\langle\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}_0\rangle=\langle\boldsymbol{b}\rangle-\langle\boldsymbol{A}\boldsymbol{x}_0\rangle=\varepsilon_0-\varepsilon_0=0$$

and assumptions of Lemma 6.5 are satisfied. Hence, the second term of Equation (100) is equal to zero as there is a multiplication with zero vector. Next, first term in Equation (100) can be manipulated using Lemma 6.4 to obtain

$$\alpha_0^2\boldsymbol{r}_0^T(\boldsymbol{W}^T\boldsymbol{S}+\boldsymbol{W}^T\boldsymbol{W})\boldsymbol{r}_0-\alpha_0^2\boldsymbol{r}_0^T\boldsymbol{W}(\boldsymbol{S}+\boldsymbol{W})\boldsymbol{r}_0=\alpha_0^2\boldsymbol{r}_0^T\boldsymbol{W}^T\boldsymbol{r}_0 \tag{101}$$

Another use of Lemma 6.5 leads to

$$\alpha_0^2\boldsymbol{r}_0^T\boldsymbol{W}^T\boldsymbol{r}_0=\alpha_0^2\boldsymbol{r}_0^T\boldsymbol{o}=0$$

and the base case of the proof has been finished.

Induction hypothesis: Now, it is assumed that the algorithm is satisfied for iteration number j. It means that following formulas hold up

$$\boldsymbol{r}_j^T\boldsymbol{r}_j=\boldsymbol{r}_j^T\boldsymbol{r}_j^* \tag{102}$$

$$\boldsymbol{r}_{j+1}^T\boldsymbol{r}_{j+1}=\boldsymbol{r}_{j+1}^T\boldsymbol{r}_{j+1}^* \tag{103}$$

$$\boldsymbol{p}_j^T\boldsymbol{A}^T\boldsymbol{p}_j=\boldsymbol{p}_j^T\boldsymbol{A}^T\boldsymbol{p}_j^* \tag{104}$$

Inductive step In this part the following formulas are being proved using induction hypothesis.

$$\boldsymbol{p}_{j+1}^T\boldsymbol{A}^T\boldsymbol{p}_{j+1}=\boldsymbol{p}_{j+1}^T\boldsymbol{A}^T\boldsymbol{p}_{j+1}^* \tag{105}$$

$$\boldsymbol{r}_{j+2}^T\boldsymbol{r}_{j+2}=\boldsymbol{r}_{j+2}^T\boldsymbol{r}_{j+2}^* \tag{106}$$

First, Equation (105) will be proved. Hence

$$\boldsymbol{p}_{j+1}^T\boldsymbol{A}^T\boldsymbol{p}_{j+1} - \boldsymbol{p}_{j+1}^T\boldsymbol{A}^T\boldsymbol{p}_{j+1}^* = \boldsymbol{p}_{j+1}^T\boldsymbol{A}^T\boldsymbol{p}_{j+1} - \frac{\boldsymbol{r}_{j+1}^T\boldsymbol{r}_{j+1}^*}{\alpha_{j+1}}$$

as the second term has been expressed using line 4 of BiCG algorithm. Next, it is possible to use Equation (103) from induction hypothesis. Vector $\boldsymbol{p}_{j+1}$ can be expressed using line 9 of BiCG. With the help of A-conjugacy of $\boldsymbol{p}_j$ vectors, it heads to

$$\boldsymbol{p}_{j+1}^T\boldsymbol{A}^T\boldsymbol{p}_{j+1} - \frac{\boldsymbol{r}_{j+1}^T\boldsymbol{r}_{j+1}^*}{\alpha_{j+1}} = \boldsymbol{p}_{j+1}^T\boldsymbol{A}^T(\boldsymbol{r}_{j+1} + \beta_j\boldsymbol{p}_j) - \frac{\boldsymbol{r}_{j+1}^T\boldsymbol{r}_{j+1}}{\alpha_{j+1}} = \boldsymbol{p}_{j+1}^T\boldsymbol{A}^T\boldsymbol{r}_{j+1} - \frac{\boldsymbol{r}_{j+1}^T\boldsymbol{r}_{j+1}}{\alpha_{j+1}}$$

Next, term $\boldsymbol{A}\boldsymbol{p}_{j+1}$ can be expressed using line 6 of BiCG as

$$\boldsymbol{A}\boldsymbol{p}_{j+1} = \frac{\boldsymbol{r}_{j+1} - \boldsymbol{r}_{j+2}}{\alpha_{j+1}}$$

After substituting and using orthogonality properties of $\boldsymbol{r}_j$ vectors, we obtain

$$\boldsymbol{p}_{j+1}^T\boldsymbol{A}^T\boldsymbol{r}_{j+1} - \frac{\boldsymbol{r}_{j+1}^T\boldsymbol{r}_{j+1}}{\alpha_{j+1}} = \left(\frac{\boldsymbol{r}_{j+1} - \boldsymbol{r}_{j+2}}{\alpha_{j+1}}\right)^T\boldsymbol{r}_{j+1} - \frac{\boldsymbol{r}_{j+1}^T\boldsymbol{r}_{j+1}}{\alpha_{j+1}} = 0$$

Next, Equation (106) will be proved.

$$\boldsymbol{r}_{j+2}^T\boldsymbol{r}_{j+2} - \boldsymbol{r}_{j+2}^T\boldsymbol{r}_{j+2}^* = \boldsymbol{r}_{j+2}^T(\boldsymbol{r}_{j+2} - \boldsymbol{r}_{j+2}^*) = (\boldsymbol{r}_{j+1} - \alpha_{j+1}\boldsymbol{A}\boldsymbol{p}_{j+1})^T(\boldsymbol{r}_{j+2} - \boldsymbol{r}_{j+2}^*) =$$

$$= \boldsymbol{r}_{j+1}^T(\boldsymbol{r}_{j+2} - \boldsymbol{r}_{j+2}^*) - \alpha_{j+1}\boldsymbol{p}_{j+1}^T\boldsymbol{A}^T(\boldsymbol{r}_{j+2} - \boldsymbol{r}_{j+2}^*)$$

Since the orthogonality properties of vectors $\boldsymbol{r}_i$ and Lemma 6.1 hold, the first term equals to zero. With substitution of $\boldsymbol{r}_{j+2}$ and $\boldsymbol{r}_{j+2}$ as in the lines 9 and 10 of Algorithm 4, we arrive at

$$\boldsymbol{r}_{j+1}^T(\boldsymbol{r}_{j+2} - \boldsymbol{r}_{j+2}^*) - \alpha_{j+1}\boldsymbol{p}_{j+1}^T\boldsymbol{A}^T(\boldsymbol{r}_{j+2} - \boldsymbol{r}_{j+2}^*) =$$

$$= -\alpha_{j+1}\boldsymbol{p}_{j+1}^T\boldsymbol{A}^T[(\boldsymbol{p}_{j+2} - \beta_{j+1}\boldsymbol{p}_{j+1}) - (\boldsymbol{p}_{j+2}^* - \beta_{j+1}\boldsymbol{p}_{j+1}^*)]$$

Next, use of Lemma 6.1 and A-conjugacy of $\boldsymbol{p}_j$ vectors with some algebraic emendations leads to

$$-\alpha_{j+1}\boldsymbol{p}_{j+1}^T\boldsymbol{A}^T[(\boldsymbol{p}_{j+2} - \beta_{j+1}\boldsymbol{p}_{j+1}) - (\boldsymbol{p}_{j+2}^* - \beta_{j+1}\boldsymbol{p}_{j+1}^*)] = \alpha_{j+1}\beta_{j+1}(\boldsymbol{p}_{j+1}^T\boldsymbol{A}^T\boldsymbol{p}_{j+1} - \boldsymbol{p}_{j+1}^T\boldsymbol{A}^T\boldsymbol{p}_{j+1}^*)$$

and finally, the utilization of Equation (105) proved in previous step heads to

$$\alpha_{j+1}\beta_{j+1}(\boldsymbol{p}_{j+1}^T\boldsymbol{A}^T\boldsymbol{p}_{j+1} - \boldsymbol{p}_{j+1}^T\boldsymbol{A}^T\boldsymbol{p}_{j+1}) = 0$$

□

Corollary 6.7. *Let $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ be a linear system with matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ defined in Equation (77) and with vector $\boldsymbol{b} \in \mathbb{R}^{N\times 1}$ defined in Equation (75). If initial solution $\boldsymbol{x}_0$ is choosen from space $\mathcal{E}^N$ then CG algorithm converges to a solution of the linear system.*

6.2.2 Two phase medium

In this part, the special case of two phase medium and initial vector $\boldsymbol{x}_0$ taken as vector $\boldsymbol{b}$ is considered. Numerical solution gets convergence in one iteration and it will be compared with the results in Section 4 Convergence Study.

Lemma 6.8. *Let $\boldsymbol{A}$ be a matrix defined in Equation (77) and $\boldsymbol{b} = b_i = \varepsilon_0$ vector defined in Equation (75), then following equations hold:*

$$\boldsymbol{A}^T\boldsymbol{b} = \boldsymbol{b} \tag{107}$$

$$\boldsymbol{b}^T\boldsymbol{A}^k\boldsymbol{b} = \boldsymbol{b}^T\boldsymbol{b} = N\varepsilon_0^2 \tag{108}$$

Proof. First, Equation (107) will be proved.

$$\boldsymbol{A}^T\boldsymbol{b} = \sum_{i=1}^{N} A_{ij}b_i = \sum_{i=1}^{N}\left(\frac{E(x_i)}{E_{\text{ref}}}\delta_{ij} + \frac{E_{\text{ref}} - E(x_j)}{NE_{\text{ref}}}\right)\varepsilon_0 = \\ = \varepsilon_0\left(\frac{E(x_j)}{E_{\text{ref}}} + \frac{E_{\text{ref}} - E(x_j)}{E_{\text{ref}}}\right) = \varepsilon_0 = b_i = \boldsymbol{b}$$

Next, proof of Equation (108) is provided. Using simple transposition and k times application of previously proven Equation (107), it is possible to write:

$$\boldsymbol{b}^T\boldsymbol{A}^k\boldsymbol{b} = (\boldsymbol{A}^{kT}\boldsymbol{b})^T\boldsymbol{b} = \boldsymbol{b}^T\boldsymbol{b} = \sum_{i=1}^{N}\varepsilon_0^2 = N\varepsilon_0^2$$

□

Lemma 6.9. *Let $\boldsymbol{A}$ be a matrix defined in Equation (77) and $\boldsymbol{b} = b_i = \varepsilon_0$ vector defined in Equation (75), then following quadratic form can be expressed as:*

$$\boldsymbol{b}^T(\boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{A}^T)\boldsymbol{b} = \frac{\varepsilon_0^2 N E_0^2 c p^2}{E_{\text{ref}}^2}[1-c] = cN(1-c)\left(\frac{\varepsilon_0 E_0 p}{E_{\text{ref}}}\right)^2 \tag{109}$$

It is necessary to note that variables c and p were already defined in Section 4. Just for clarification, variable c is volume fraction of inclusions and p is a ratio between the perturbing part of stiffness E_1 and stiffness of matrix E_0: $p = \frac{E_1}{E_0}$.

Proof. Using expression for matrix $\boldsymbol{A} = \boldsymbol{S} + \boldsymbol{W}$ and Equation (91) in Lemma 6.4 leads to:

$$\boldsymbol{b}^T(\boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{A}^T)\boldsymbol{b} = \boldsymbol{b}^T(\boldsymbol{S}^2 + \boldsymbol{S}\boldsymbol{W} + \boldsymbol{W}^T\boldsymbol{S} + \boldsymbol{W}^T\boldsymbol{W} - \boldsymbol{S} - \boldsymbol{W}^T)\boldsymbol{b} = \\ = \boldsymbol{b}^T(\boldsymbol{S}^2 + \boldsymbol{S}\boldsymbol{W} - \boldsymbol{S})\boldsymbol{b} \tag{110}$$

Next, individual quadratic forms $\boldsymbol{b}^T\boldsymbol{S}\boldsymbol{b}$, $\boldsymbol{b}^T\boldsymbol{S}^2\boldsymbol{b}$ and $\boldsymbol{b}^T\boldsymbol{S}\boldsymbol{W}\boldsymbol{b}$ will be calculated.

$$\boldsymbol{b}^T\boldsymbol{S}\boldsymbol{b} = \sum_{i=1}^{N}\varepsilon_0^2\frac{E(x_i)}{E_{\text{ref}}} = \frac{\varepsilon_0^2}{E_{\text{ref}}}\sum_{i=1}^{N}E(x_i) \tag{111}$$

The sum of stiffness in individual points can be divided into two sums, one related to stiffness of matrix E_0 and the other one to stiffness of inclusion $E_0(1+p)$:

$$\sum_{i=1}^{N} E(x_i) = \sum_{i=1}^{N(1-c)} E_0 + \sum_{i=1}^{Nc} E_0(1+p) = NE_0(1+cp)$$

After substituting this into Equation (111), it leads to:

$$\boldsymbol{b}^T\boldsymbol{S}\boldsymbol{b} = \frac{\varepsilon_0^2 NE_0}{E_{\text{ref}}}(1+cp) \tag{112}$$

Analogically to previous quadratic form, it is possible to express:

$$\boldsymbol{b}^T\boldsymbol{S}^2\boldsymbol{b} = \frac{\varepsilon_0^2}{E_{\text{ref}}^2}\sum_{i=1}^{N} E^2(x_i) = \frac{\varepsilon_0^2 NE_0^2}{E_{\text{ref}}^2}[1+cp(2+p)] \tag{113}$$

Since $\boldsymbol{W}\boldsymbol{b} = \sum_{j=1}^{N} \varepsilon_0 \frac{E_{\text{ref}}-E(x_j)}{E_{\text{ref}}} = \varepsilon_0\left[1-\frac{E_0}{E_{\text{ref}}}(1+cp)\right]$, the last quadratic form can be written as:

$$\boldsymbol{b}^T\boldsymbol{S}\boldsymbol{W}\boldsymbol{b} = \frac{\varepsilon_0^2 NE_0}{E_{\text{ref}}}\left[1-\frac{E_0}{E_{\text{ref}}}(1+cp)\right](1+cp) = \frac{\varepsilon_0^2 NE_0}{E_{\text{ref}}}(1+cp) - \frac{\varepsilon_0^2 NE_0^2}{E_{\text{ref}}^2}(1+cp)^2 \tag{114}$$

Finally, it is possible to continue with initial quadratic form stated in Equation (110). Hence

$$\boldsymbol{b}^T\boldsymbol{S}^2\boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{S}\boldsymbol{W}\boldsymbol{b} - \boldsymbol{b}^T\boldsymbol{S}\boldsymbol{b} = \frac{\varepsilon_0^2 NE_0^2}{E_{\text{ref}}^2}[1+cp(2+p)] - \frac{\varepsilon_0^2 NE_0^2}{E_{\text{ref}}^2}(1+cp)^2 = \frac{\varepsilon_0^2 NE_0^2 cp^2}{E_{\text{ref}}^2}[1-c]$$

□

Lemma 6.10. *Let $\boldsymbol{A}$ be a matrix defined in Equation (77) and $\boldsymbol{b} = b_i = \varepsilon_0$ vector defined in Equation (75), then following quadratic form can be expressed as:*

$$\boldsymbol{b}^T(\boldsymbol{A}^{2T}\boldsymbol{A} - \boldsymbol{A}^{2T})\boldsymbol{b} = \frac{\varepsilon_0^2 NE_0^2 cp^2}{E_{\text{ref}}^2}(1-c) - \frac{\varepsilon_0^2 NE_0^3 cp^2}{E_{\text{ref}}^3}(1-c)(1+p-cp) \tag{115}$$

Proof. At the beginning matrix $\boldsymbol{A}$ will be expressed using matrices $\boldsymbol{S}$ and $\boldsymbol{W}$ as in the previous proof.

$$\boldsymbol{b}^T(\boldsymbol{A}^{2T}\boldsymbol{A} - \boldsymbol{A}^{2T})\boldsymbol{b} = \boldsymbol{b}^T(\boldsymbol{S}^3 + 2\boldsymbol{S}^2\boldsymbol{W} + \boldsymbol{W}^T\boldsymbol{S}\boldsymbol{W} - \boldsymbol{S}^2 - \boldsymbol{W}^T\boldsymbol{S})\boldsymbol{b} \tag{116}$$

Next, expression for quadratic forms $\boldsymbol{b}^T\boldsymbol{S}^2\boldsymbol{b}$ and $\boldsymbol{b}^T\boldsymbol{W}^T\boldsymbol{S}\boldsymbol{b}$ can be used from previous proof while $\boldsymbol{b}^T\boldsymbol{S}^3\boldsymbol{b}$,$\boldsymbol{b}^T\boldsymbol{S}^2\boldsymbol{W}\boldsymbol{b}$ and $\boldsymbol{b}^T\boldsymbol{W}^T\boldsymbol{S}\boldsymbol{W}\boldsymbol{b}$ will be calculated.

$$\boldsymbol{b}^T\boldsymbol{S}^3\boldsymbol{b} = \frac{\varepsilon_0^2}{E_{\text{ref}}^3}\sum_{i=1}^{N} E^3(x_i) = \frac{\varepsilon_0^2 NE_0^3}{E_{\text{ref}}^3}[1+3cp+3cp^2+cp^3]$$

$$\boldsymbol{b}^T\boldsymbol{S}^2\boldsymbol{W}\boldsymbol{b} = \frac{\varepsilon_0^2 N E_0^2}{E_{\text{ref}}^2}\left[1 - \frac{E_0}{E_{\text{ref}}}(1+cp)\right][1+cp(2+p)] =$$
$$= \frac{\varepsilon_0^2 N E_0^2}{E_{\text{ref}}^2}[1+cp(2+p)] - \frac{\varepsilon_0^2 N E_0^3}{E_{\text{ref}}^3}(1+cp)[1+cp(2+p)]$$

$$\boldsymbol{b}^T\boldsymbol{W}^T\boldsymbol{S}\boldsymbol{W}\boldsymbol{b} = \frac{\varepsilon_0^2 N E_0}{E_{\text{ref}}}\left[1 - \frac{E_0}{E_{\text{ref}}}(1+cp)\right]^2(1+cp) =$$
$$= \frac{\varepsilon_0^2 N E_0}{E_{\text{ref}}}(1+cp) - 2\frac{\varepsilon_0^2 N E_0^2}{E_{\text{ref}}^2}(1+cp)^2 + \frac{\varepsilon_0^2 N E_0^3}{E_{\text{ref}}^3}(1+cp)^3$$

Finally, it can be substituted into Equation (116). After subtraction of some terms and several algebraic emendations, it leads to:

$$\boldsymbol{b}^T(\boldsymbol{A}^{2T}\boldsymbol{A} - \boldsymbol{A}^{2T})\boldsymbol{b} = \frac{\varepsilon_0^2 N E_0^2 cp^2}{E_{\text{ref}}^2}(1-c) - \frac{\varepsilon_0^2 N E_0^3 cp^2}{E_{\text{ref}}^3}(1-c)(1+p-cp)$$

□

Theorem 6.11. *For the two phase heterogeneous rod and vector $\boldsymbol{x}_0$ equal to vector $\boldsymbol{b}$, CG method converge in one step and coefficient α_0 has following expression*

$$\alpha_0 = \frac{E_{\text{ref}}}{E_0(1+p-cp)}$$

Proof. Since vector $\boldsymbol{p}_0$ is equal to $\boldsymbol{r}_0$ and $\boldsymbol{x}_0$ to $\boldsymbol{b}$, then

$$\boldsymbol{p}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x_0} = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{b}$$

and the first iteration vector $\boldsymbol{x}_1$ can be calculated as

$$\boldsymbol{x}_1 = \boldsymbol{x}_0 + \alpha_0\boldsymbol{p}_0 = \boldsymbol{b} + \alpha_0(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{b}$$

Next, coefficient α_0 will be investigated. Hence, it is possible to express it as

$$\alpha_0 = \frac{\boldsymbol{r}_0^T\boldsymbol{r}_0}{\boldsymbol{p}_0^T\boldsymbol{A}^T p_0} = \frac{\boldsymbol{b}^T(\boldsymbol{I}-\boldsymbol{A})^T(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{b}}{\boldsymbol{b}^T(\boldsymbol{I}-\boldsymbol{A})^T\boldsymbol{A}^T(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{b}} = \frac{\boldsymbol{b}^T(\boldsymbol{I}-\boldsymbol{A}^T-\boldsymbol{A}+\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{b}}{\boldsymbol{b}^T(\boldsymbol{A}^T-\boldsymbol{A}^T\boldsymbol{A}-\boldsymbol{A}^{2T}+\boldsymbol{A}^{2T}\boldsymbol{A})\boldsymbol{b}} =$$

$$= \frac{\boldsymbol{b}^T(\boldsymbol{A}^T\boldsymbol{A}-\boldsymbol{A}^T)\boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{b} - \boldsymbol{b}^T\boldsymbol{A}\boldsymbol{b}}{\boldsymbol{b}^T(\boldsymbol{A}^{2T}\boldsymbol{A}-\boldsymbol{A}^{2T})\boldsymbol{b} - \boldsymbol{b}^T(\boldsymbol{A}^T\boldsymbol{A}-\boldsymbol{A})\boldsymbol{b}} = \frac{\boldsymbol{b}^T(\boldsymbol{A}^T\boldsymbol{A}-\boldsymbol{A})\boldsymbol{b}}{\boldsymbol{b}^T(\boldsymbol{A}^{2T}\boldsymbol{A}-\boldsymbol{A}^{2T})\boldsymbol{b} - \boldsymbol{b}^T(\boldsymbol{A}^T\boldsymbol{A}-\boldsymbol{A}^T)\boldsymbol{b}}$$

After substituting from Lemmas 6.9 and 6.10 it heads to:

$$\alpha_0 = \frac{\boldsymbol{b}^T(\boldsymbol{A}^T\boldsymbol{A}-\boldsymbol{A})\boldsymbol{b}}{\boldsymbol{b}^T(\boldsymbol{A}^{2T}\boldsymbol{A}-\boldsymbol{A}^{2T})\boldsymbol{b} - \boldsymbol{b}^T(\boldsymbol{A}^T\boldsymbol{A}-\boldsymbol{A}^T)\boldsymbol{b}} = \frac{\frac{\varepsilon_0^2 N E_0^2 cp^2}{E_{\text{ref}}^2}[1-c]}{\frac{\varepsilon_0^2 N E_0^3 cp^2}{E_{\text{ref}}^3}(1-c)(1+p-cp)} =$$
$$= \frac{E_{\text{ref}}}{E_0(1+p-cp)}$$

Next, it is possible to use results obtained in Section 4 dealing with convergence behaviour of two phase heterogenous rod. The basic result provide Equation (64)

$$\varepsilon_1(k) = \frac{\varepsilon_0 cp}{1+p-cp} + \left(\varepsilon_1(0) - \frac{\varepsilon_0 cp}{1+p-cp}\right)\left(1 - \frac{E_0(1+p-cp)}{E_{\mathrm{ref}}}\right)^k$$

that describes behaviour of perturbating strain ε_1 in points x_{E_0} with stiffness E_0. This sequence is obtained from the iteration algorithm with matrix $(\boldsymbol{I}-\boldsymbol{A})$ as it is described in Equation (79). The multiplication with this matrix also correspond to projection vector $\boldsymbol{p}_0 = (\boldsymbol{I}-\boldsymbol{A})\boldsymbol{b}$. Hence, its components that relates to stiffness E_0 can be marked as $p_0(x_{\mathrm{E}_0})$ and expressed as:

$$p_0(x_{\mathrm{E}_0}) = \varepsilon_1(1) - \varepsilon_1(0)$$

After substituting into sequence defined in Equation (64) for $k = 1$ and considering that $\varepsilon_1(0) = 0$ as initial vector $\boldsymbol{x}_0$ is taken as $\boldsymbol{b}$, it can be expressed as:

$$p_0(x_{\mathrm{E}_0}) = \frac{\varepsilon_0 cp}{1+p-cp}\frac{E_0(1+p-cp)}{E_{\mathrm{ref}}}$$

Thus, it is possible to express strain relating to stiffness E_0 as in line 4 of CG algorithm $(\boldsymbol{x}_1 = \boldsymbol{x}_0 + \alpha_0\boldsymbol{p}_0)$.

$$\varepsilon(x_{\mathrm{E}_0}) = \varepsilon_0 + \alpha_0 p_0(x_{\mathrm{E}_0})$$

after substitution into this equation, it heads to:

$$\varepsilon(x_{\mathrm{E}_0}) = \varepsilon_0 + \frac{\varepsilon_0 cp}{1+p-cp}$$

that is exact solution for stiffness E_0 as it is described in Section 4. It says that it converges in one step for points with stiffness E_0. Next, behaviour of points with stiffness $E_0(1+p)$ can be investigated from the condition of compatibility saying that average perturbing strain is equal to zero

$$\sum_{i=1}^{N}\varepsilon_1(x_i) = 0 \quad \Rightarrow \quad (N-m)\varepsilon_1(x_{E_0}) + m\varepsilon_1(x_{E_0(1+p)}) = 0$$

Since the condition has to be satisfied as $\boldsymbol{p}_0 \in \mathcal{E}_0^N$ and first iteration of $\varepsilon(x_{\mathrm{E}_0})$ is equal to exact solution, then $\varepsilon(x_{\mathrm{E}_0(1+\mathrm{p})})$ is necessarily equal to exact solution as well. □

Corollary 6.12. *When optimal value for $E_{\mathrm{ref}} = E_0(1+p-cp)$ is taken, coefficient α_0 is equal to one.*

Proof. Since $\alpha_0 = \frac{E_{\mathrm{ref}}}{E_0(1+p-cp)}$, it is obvious. □

Finally, the iteration method with matrix $(\boldsymbol{I}-\boldsymbol{A})$ or simple iteration by FFT respectively versus matrix solution by CG algorithm is compared. In the first case, parameter E_{ref} plays a crucial role for convergence while "poor" choice of E_{ref} for CG algorithm is compansated with inner coefficent of CG algorithm α_k.

Part IV
Comparison and conclusion

This section provides conclusion and comparison of GAF and FFH methods that are used for modeling heterogeneous materials for discrete data provided in bitmaps. First method is being used for compactly located heterogeneities placed in an infinity matrix while the second method deals with periodically repeating medium. The summarization of the results is provided in the following list.

GAF method

- Two approaches using Approximate Approximation (Standard and Regularization approach) were investigated.
- The discrete values resulting from the Standard approach represent poorly the exact solution as it suffer from serious Gibb's effect especially for larger values of parameter H.
- The discrete values resulting from Regularization approach represent well smoothed or regularized exact solution. Parameter H regulating Gaussian basis function has direct influence on the rate of smoothing or regularization.
- Fast Fourier Transform can be used for matrix with vector mutliplication resulting in sped up.

FFH method

- The optimal choice of the reference medium depends on the problem geometry.
- In one-dimensional situation, the optimal choice of the reference medium exists, for which the algorithm converges in a single step.
- This value is generally different from $\eta = \frac{1}{2}$, which is the choice reported in available convergence studies, e.g. [16, 24]
- The optimal value is different from the homogenized stiffness.
- Linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ arising from this method is non-symmetric and full.
- The standard algorithm of FFH method corresponds to basic iteration method $\boldsymbol{x}_{k+1} = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x}_k + \boldsymbol{b}$ for the solution of the linear system.
- Linear system can be efficiently solved using Conjugate Gradient Algorithm as it produces the same sequence of vectors as in Biconjugate Gradient Algorithm.
- Linear system from the problem of two phase medium is solved by Conjugate Gradient Algorithm in just one iteration.

Extention of these conclusions into the multi-dimensional setting remains an open problem, which is currently being studied. The totally unsolved remains Conjugate gradient method convergence in the case of GAF method.

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] 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.

[3] G. Constantinides, F. Ulm, and K. Van Vliet, *On the use of nanoindentation for cementitious materials*, Materials and Structures **36** (2003), no. 3, 191–196.

[4] W. Dreyer and W. H. Müller, *A study of the coarsening in tin/lead solders*, International Journal of Solids and Structures **37** (2000), no. 28, 3841–3871.

[5] 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.

[6] K. Forstová, *Characterization and reconstruction of microstructure of cement based composites*, vol. 12, Czech Technical University in Prague, Prague, Czech Republic, 2008.

[7] E.J. Garboczi, *Finite element and finite difference programs for computing the linear electric and elastic properties of digital images of random materials*, Tech. Report NISTIR 6269, Building and Fire Research Laboratory, National Institute of Standards and Technology, Gaithesburg, Maryland 2089, 1998, `http://ciks.cbt.nist.gov/~garbocz/manual` [September 3, 2009].

[8] S. Kanaun, *A numerical method for the solution of electromagnetic wave diffraction problems on perfectly conducting screens*, Journal of Computational Physics **176** (2002), no. 1, 170–195.

[9] ———, *Fast solution of 3D-elasticity problem of a planar crack of arbitrary shape*, International Journal of Fracture **148** (2007), no. 4, 435–442.

[10] S. Kanaun and S. B. Kochekseraii, *A numerical method for the solution of thermo- and electro-static problems for a medium with isolated inclusions*, Journal of Computational Physics **192** (2003), no. 2, 471–493.

[11] ———, *A numerical method for the solution of 3D-integral equations of electrostatic theory based on Gaussian approximating functions*, Applied Mathematics and Computation **184** (2007), no. 2, 754–768.

[12] ———, *Effective conductive and dielectric properties of matrix composites with inclusions of arbitrary shapes*, International Journal of Engineering Science **46** (2008), no. 2, 147–163.

[13] V. Maz'ya, *A new approximation method and its applications to the calculation of volume potentials, boundary point method*, 3.DFG-Kolloquium des DFG-Forschungsschwerpunktes Randelementmethoden (30.9.-05.10.1991), 8, Schloss Reisenburg (1991).

[14] V. Maz'ya and G. Schmidt, *On approximate approximations using Gaussian kernels*, IMA Journal of Numerical Analysis **16** (1996), no. 1, 13–29.

[15] ———, *Approximate approximations*, vol. 361p., American Mathematical Society, 2007.

[16] J. C. Michel, H. Moulinec, and P. Suquet, *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.

[17] 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.

[18] ———, *A numerical method for computing the overall response of nonlinear composites with complex microstructure*, Computer Methods in Applied Mechanics and Engineering **157** (1998), no. 1–2, 69–94.

[19] J. Novák, *Calculation of elastic stresses and strains inside a medium with multiple isolated inclusions*, Proceedings of the Sixth International Conference on Engineering Computational Technology (Stirlingshire, Scotland) (M. Papadrakakis and B. H. V. Topping, eds.), Civil-Comp Press, 2008, paper No. 127.

[20] W. Rudin, *Fourier analysis on groups*, Wiley Classics Library, Wiley-Interscience, 1990.

[21] Y. Saad, *Iterative methods for sparse linear systems*, second edition with corrections ed., 2000.

[22] J. Spowart, *Automated serial sectioning for 3-D analysis of microstructures*, Scripta Materialia **55** (2006), no. 1, 5–10.

[23] K. Terada, T. Miura, and N. Kikuchi, *Digital image-based modeling applied to the homogenization analysis of composite materials*, Computational Mechanics **20** (1997), no. 4, 331–346.

[24] 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.

[25] Jaroslav Vondřejc, *Preliminaries to Analysis of Heterogeneous Materials Using Meshless Methods.* Bachelor's thesis, Department of Mechanics, Faculty of Civil Engineering, Czech Technical University, Czech Republic, Prague, May 2008, Supervisor: Zeman, J.

[26] 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.

[27] C. L. Y. Yeong and S. Torquato, *Reconstructing random media. II. Three-dimensional media from two-dimensional cuts*, Physical Review E **58** (1998), no. 1, 224–233.

Acknowledgements

Financial support of this work provided by the Grant Agency of the Czech Republic, projects no. GAČR 103/09/1748 and no. 103/09/P490, is gratefully acknowledged.