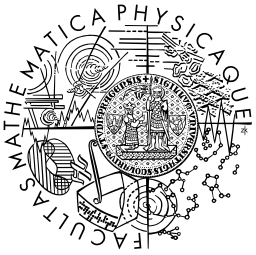


FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Eva Vidličková

Fourier-Galerkin Method for Stochastic Homogenization of Elliptic Partial Differential Equations

Department of Numerical Mathematics

Supervisor of the master thesis: Doc. Ing. Jan Zeman, Ph.D.

Study programme: Mathematics

Study branch: Numerical and Computational Mathematics

Prague 2017

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ..... signature of the author

Title: Fourier-Galerkin Method for Stochastic Homogenization of Elliptic Partial Differential Equations

Author: Eva Vidličková

Department: Department of Numerical Mathematics

Supervisor: Doc. Ing. Jan Zeman, Ph.D., Department of Mechanics, Faculty of Civil Engineering, Czech Technical University in Prague

Abstract: This thesis covers the basics in the stochastic homogenization of elliptic partial differential equations, from underlying theory up to numerical approaches. In particular, we introduce and analyze a combination of the Fourier-Galerkin method in the spatial domain with a collocation method in the stochastic domain. The material coefficients are assumed to depend on a finite number of random variables. We present a comparison of the Monte Carlo method with the full tensor grid and sparse grid collocation method for two applications. The first one is the checkerboard problem with continuous random variables, the other considers the material coefficients to be described in terms of an autocorrelation function.

Keywords: stochastic homogenization, Fourier-Galerkin method, sparse grid, collocation method

I would like to express an honest thanks to my supervisor Doc. Ing. Jan Zeman, Ph.D. for being a greater supervisor than I could possibly wish for, for mentoring me while entering the research world and supporting my ideas. I would also like to thank doc. RNDr. Ivana Pultarová, Ph.D. for her patience, very insightful comments and her willingness to participate in our frequent consultations. I am also grateful to Jaroslav Vondřejc, Ph.D. for the helpful video conference calls and for providing me with FFTHomPy solver. I would like to express my thanks to Prof. Dr. Vincent Heuveline who enabled my first contact with the stochastic collocation method. Last but not least, my thanks go to my family, Dominik Pavlov and Him. This thesis was partially supported by the Czech Science Foundation grants 14-00420S and 17-04150J.

Contents

Introduction

Homogenization is a theory that provides the mathematical basis for describing effective physical properties of heterogeneous materials. It studies partial differential equations with rapidly oscillating coefficients. In many engineering applications the coefficients are affected by a relatively large amount of uncertainty that arises from inaccurate measurements, inaccurate computations or insufficient knowledge. Such uncertainties in the input parameters can be expressed in terms of random variables, which means that the material coefficients depend on both random and spatial variables. Stochastic homogenization is a branch of mathematics that studies homogenization of random media, i.e. materials whose physical properties are modeled with random functions.

The derived theory of stochastic homogenization relies on results from many mathematical disciplines, i.e. mathematical analysis, partial differential equations, probability and ergodic theory. This thesis provides a short summary of the results from deterministic (not dependent on random functions) and stochastic homogenization that are most important for solving engineering applications in the last chapter. The task is to accurately enough approximate the heterogeneous material by a homogeneous one, i.e. approximate the material coefficients defined on non-random and random media by constant coefficients.

The first chapter studies the theory of deterministic homogenization and derives an easy way to solve our problem for periodic media. The second chapter deals with the theory of stochastic homogenization and also comes to an elegant result at the cost of assuming the ergodicity of a specific dynamical system. The main source of the results is [Jikov et al., 1994] and [Cioranescu and Donato, 1999]. The third and the fourth chapter are devoted to adressing the question of numerical computations of the problems of deterministic or stochastic homogenization. In particular, the third chapter introduces the Fourier-Galerkin method, analyzes its rate of convergence and discusses why it is a suitable numerical method for the problems of deterministic homogenization, where the arguments are taken from [Vondřejc et al., 2014]. The fourth chapter shows a way of how to compute the effective coefficients when the material coefficients depend on spatial variables and a finite number of random variables. It introduces the collocation method based on either full-tensor or sparse grid, derives the rates of convergence and presents the Monte-Carlo method as a frame of reference, since that is probably the most widely used method nowadays. The main sources to this chapter were [Babuška et al., 2007] and [Nobile et al., 2008]. The last chapter provides two examples of the stochastic homogenization problem, which reflect some engineering applications. For the computation I employed the collocation method which involves the solution of a deterministic homogenization problem by Fourier-Galerkin method. We compare three methods (Monte Carlo, full tensor grid collocation and sparse grid collocation).

My main contribution consists in introducing the combination of the collocation and the Fourier-Galerkin method as a suitable numerical method for solving the problems of stochastic homogenization. I adjusted the results reported in [Nobile

et al., 2008] to the homogenization setting and provided a consistent coverage of the stochastic homogenization topic from the theory to the numerical approaches. I developed a sparse grid generator in Python and implemented the collocation method that employed a Python Fourier-Galerkin solver [Vondřejc, 2016–2017] for computing the results in Chapter 5.

1. Homogenization of second order elliptic operators

This chapter will be devoted to the mathematical theory of the so called deterministic homogenization. We will assume the heterogeneous material can be described by means of a periodic medium. The first section forms a background which is later used to prove some nice results. The mathematical notion of homogenization is provided in the section 1.2 and in the last section we will prove that solving the problem of homogenization leads to an easy problem defined on a bounded domain.

Most of the formulated results are taken from the books [Jikov et al., 1994] and [Cioranescu and Donato, 1999].

1.1 Preliminaries

In this part we will state some basic facts which are later necessary to derive a proper mathematical description of homogenization. We will give some details on the weak* convergence and introduce some new spaces of periodic functions and periodic boundary value problems. We already assume some basic knowledge in the theory of the weak topology, Lebesgue spaces, Sobolev spaces and boundary value problems.

1.1.1 Weak* convergence

We know that the property of a strong convergence of a sequence is often too strict. One can instead be requiring the weak convergence, a weaker property than the strong convergence. However, there is a big class of sequences which are not weakly convergent. In that case the notion of weak* convergence, as a generalization of the weak convergence is introduced. following the developments in [Jikov et al., 1994].

Since $L^1(Q)$ cannot be characterized as the dual of some Banach space, the classic notion of the weak* convergence is not interesting in this space. However, in our computations we will be looking for some generalizations of the weak limits which preserve some of the features of the classic weak* limits. Therefore we are introducing a new definition of the weak* convergence for functions in $L^1(Q)$, Q being a bounded domain in $\mathbb{R}^m$.

Definition 1.1. *Let Q be a bounded domain in $\mathbb{R}^m$, w^ε, $w^0 \in L^1(Q)$ for all ε. We write $w^\varepsilon \overset{*}{\rightharpoonup} w^0$ and say w^ε is weakly* convergent to w^0, if the sequence w^ε is bounded in $L^1(Q)$ and it holds:*

$$\lim_{\varepsilon\to 0}\int_Q w^\varepsilon\,\varphi\,\mathrm{d}x = \int_Q w^0\,\varphi\,\mathrm{d}x \qquad \forall\varphi\in\mathcal{C}_0^\infty(Q).$$

Corollary. The weak* limit is uniquely defined.

Proof. Let the sequence $w^\varepsilon \in L^1(Q)$ have two weak* limits $w^0, w^1 \in L^1(Q)$. Then

$$\int_Q w^0 \, \varphi \, \mathrm{d}x = \int_Q w^1 \, \varphi, \quad \forall \varphi \in \mathcal{C}_0^\infty(Q).$$

From the theory of distributions we have $w^0 = w^1$ a.e. in Q.

□

Corollary. Any weakly convergent sequence in $L^1(Q)$ is also weakly* convergent.

Proof. Let w^ε be a weakly convergent sequence in $L^1(Q)$. Thanks to the isometric isomorphism of the dual space of $L^1(Q)$ and $L^\infty(Q)$ we have

$$\lim_{\varepsilon \to 0} \int_Q w^\varepsilon \, \varphi \, \mathrm{d}x = \int_Q w^0 \, \varphi \, \mathrm{d}x, \quad \forall \varphi \in L^\infty(Q).$$

Since $\mathcal{C}_0^\infty(Q) \subset L^\infty(Q)$ we obtained what we needed to finish the proof.

□

One often comes across the question if and where the sequence of scalar products $p^\varepsilon \cdot v^\varepsilon$ converges with $p^\varepsilon, \ v^\varepsilon \rightharpoonup 0$ in $L^2(Q)$. Clearly, when one of the sequences strongly converges in $L^2(Q)$ we have $p^\varepsilon \cdot v^\varepsilon \rightharpoonup 0$ in $L^1(Q)$ and therefore $p^\varepsilon \cdot v^\varepsilon \overset{*}{\rightharpoonup} 0$. It is, however, not always possible to provide the strong convergence. In that case the following lemma might be helpful.

Lemma 1.1. *Let $p^\varepsilon, v^\varepsilon$ be vector fields in $L^2(Q)$ such that*

$$p^\varepsilon \rightharpoonup p^0, \ \ v^\varepsilon \rightharpoonup v^0 \qquad \text{in } L^2(Q).$$

In addition, let $p^\varepsilon, v^\varepsilon$ satisfy the conditions:

$$\mathit{curl}\, v^\varepsilon = 0, \quad \operatorname{div} p^\varepsilon \to f^0 \ \text{in } H^{-1}(Q).$$

Then $p^\varepsilon \cdot v^\varepsilon \overset{*}{\rightharpoonup} p^0 \cdot v^0$.

Proof. We can rewrite $p^\varepsilon \cdot v^\varepsilon$ as:

$$p^\varepsilon \cdot v^\varepsilon = (p^\varepsilon - p^0) \cdot (v^\varepsilon - v^0) + p^0 \cdot v^\varepsilon + p^\varepsilon \cdot v^0 - p^0 \cdot v^0.$$

The sum of the last three terms has a weak-* limit $p^0 \cdot v^0$. Therefore, without loss of generality we can assume: $p^0 = v^0 = 0$.

A weak-* convergence is a local property, which means we can assume Q to be simply connected. In that case every irrotational vector field is a potential vector field, i.e. there exists a potential function $u^\varepsilon \in H^1(Q)$ s.t. $v^\varepsilon = \nabla u^\varepsilon$, $\int_Q u^\varepsilon = 0$. Since $v^\varepsilon = \nabla u^\varepsilon \rightharpoonup 0$, thanks to the Poincaré inequality and the Rellich-Kondrachov embedding theorem we have: $u^\varepsilon \rightharpoonup 0$ in $H^1(Q)$ and $u^\varepsilon \to 0$ in $L^2(Q)$.

And now for every $\varphi \in \mathcal{C}_0^\infty(Q)$:

$$\int_Q p^\varepsilon \cdot v^\varepsilon \, \varphi \, \mathrm{d}x = \int_Q p^\varepsilon \cdot \nabla(u^\varepsilon \, \varphi) \, \mathrm{d}x - \int_Q u^\varepsilon \, p^\varepsilon \cdot \nabla\varphi \, \mathrm{d}x$$
$$= -\int_Q \operatorname{div} p^\varepsilon \, \varphi \, u^\varepsilon \, \mathrm{d}x - \int_Q u^\varepsilon \, p^\varepsilon \cdot \nabla\varphi \, \mathrm{d}x \;\to\; 0.$$

The first term $-(\operatorname{div} p^\varepsilon, \varphi \, u^\varepsilon)$ converges to 0, since $\operatorname{div} p^\varepsilon \to 0$ strongly in $H^{-1}(Q)$ and $\varphi \, u^\varepsilon$ is uniformly bounded in $H_0^1(Q)$. The last integral converges to 0 as well, since $u^\varepsilon \;\to 0$ in $L^2(Q)$ and $p^\varepsilon \cdot \nabla\varphi$ is uniformly bounded in $L^2(Q)$.

□

1.1.2 Spaces of periodic functions

The classical theory of homogenization is developed for media with periodic microstructure modeled by partial differential equations with periodic coefficients. Therefore periodic functions and some specific spaces of periodic functions play a crucial role.

Consider measurable functions defined in $\mathbb{R}^m$, $l_1, \ldots, l_2$ are given positive numbers, $\square$ parallelepiped in $\mathbb{R}^m$:

$$\square = [0, l_1] \times \cdots \times [0, l_m], \tag{1.1}$$

Definition 1.2. *Let $\square$ be defined as in (1.1) and f be a measurable function defined a.e. in $\mathbb{R}^m$. The function f is called $\square$ - periodic if*

$$f(x + k \, l_i \, e_i) = f(x) \quad a.e. \ in \ \mathbb{R}^m, \quad \forall k \in \mathbb{Z}, \quad \forall i \in \{1, \ldots, m\},$$

where $\{e_1, \ldots, e_m\}$ is the canonical basis of $\mathbb{R}^m$.

Definition 1.3. *Let f be a $\square$ - periodic function in $\mathbb{R}^m$. We define the mean value of f as:*

$$\langle f \rangle = \frac{1}{|\square|} \int_\square f(x) \, \mathrm{d}x,$$

where $|\square| = l_1 l_2 \cdots l_m$ is the volume of the parallelepiped $\square$.

We shall introduce the Lebesgue space of periodic functions with the norm $\langle |f|^\alpha \rangle^{1/\alpha}$ for $1 \le \alpha < \infty$:

$$L^\alpha(\square) = \{f \text{ is } \square\text{-periodic}, \quad \langle |f|^\alpha \rangle^{1/\alpha} < \infty\}.$$

Without loss of generality we can assume $l_1 = l_2 = \cdots = l_m = 1$ or 2π as the particular values of the periods l_i are often unimportant.

Theorem 1.1 (Mean value property)**.** *Let* $1 \leq \alpha < \infty$*,* $f \in L^\alpha(\square)$*. Set*

$$f_\varepsilon(x) = f\Big(\frac{x}{\varepsilon}\Big) \quad \text{a.e. on } \mathbb{R}^m.$$

Then, as $\varepsilon \to 0$*,*

$$f_\varepsilon \rightharpoonup \langle f \rangle \quad \text{in } L^\alpha(Q),$$

where Q *is an arbitrary bounded domain in* $\mathbb{R}^m$*.*

Proof. Without loss of generality we can assume $Q = s\square$, i.e. Q is a dilatation of the cube $\square$ with ratio $s \geq 1$. A detailed justification of this assumption can be found in [Cioranescu and Donato, 1999, p. 34, 35].

At first let us estimate:

$$\text{For } \varepsilon \leq 1: \quad \int_Q |f_\varepsilon(x)| \, \mathrm{d}x = \varepsilon^m \int_{s\varepsilon^{-1}\square} |f(x)|^\alpha \, \mathrm{d}x \leq \varepsilon^m (\lfloor s\varepsilon^{-1} \rfloor + 1)^m \langle |f|^\alpha \rangle \leq$$
$$\leq c \, \langle |f|^\alpha \rangle, \quad c = c(Q),$$

where $\lfloor s\varepsilon^{-1} \rfloor \rfloor$ stands for the greatest integer not larger than $s\varepsilon^{-1}$. Since $f \in L^\alpha(\square)$, we know there is a trigonometrical polynomial $z(x)$ such that

$$\langle z \rangle = \langle f \rangle, \quad \langle |z - f|^\alpha \rangle \leq \delta,$$

for δ arbitrarily small.
Then for $\varepsilon \leq 1$ we have:

$$\big\| f(\varepsilon^{-1}x) - \langle f \rangle \big\|_{L^\alpha(Q)} \leq \big\| f(\varepsilon^{-1}x) - z(\varepsilon^{-1}x) \big\|_{L^\alpha(Q)} + \big\| z(\varepsilon^{-1}x) - \langle f \rangle \big\|_{L^\alpha(Q)} \leq$$
$$\leq (c\delta)^{1/\alpha} + \big\| z(\varepsilon^{-1}x) - \langle z \rangle \big\|_{L^\alpha(Q)}.$$

The classical Riemann-Lebesgue theorem implies that the second term converges to 0.

□

Definition 1.4. *Let* $\mathcal{C}^\infty(\square)$ *be the subset of* $\mathcal{C}^\infty(\mathbb{R}^m)$ *consisting of all* $\square$ *- periodic functions. We denote by* $H^1(\square)$ *the closure of* $\mathcal{C}^\infty(\square)$ *with respect to the norm* $\|\cdot\|_{H^1(Q)}$*,* $Q = \square$*.*

Remark. Let $u \in H^1(\square)$. Then u has the same trace on the opposite faces of $\square$.

It should be pointed out that $H^1(\square)$ does not coincide with $H^1(Q)$, $Q = \square$. In fact we can state:

Corollary. $H^1(\square) = \{ u \in H^1_{loc}(\mathbb{R}^m), \quad u \text{ is periodic} \}$.

Proof. The proof can be found in [Cioranescu and Donato, 1999, p. 57].

□

To later ensure the uniqueness of the sought solution we can introduce a quotient space $W(\square) = H^1(\square)/\mathbb{R}$:

Remark. A quotient space $W(\square) = H^1(\square)/\mathbb{R}$ is defined as the space of equivalence classes with respect to the relation

$$u \simeq v \iff u - v \text{ is a constant}, \quad \forall u, v \in H^1(\square). \tag{1.2}$$

Every equivalence class can be represented by a function $\bar{u}$ such that $\langle \bar{u} \rangle = 0$. Or in other words, every $u \in H^1(\square)$ can be decomposed as

$$u = c + \bar{u}, \quad c = \langle u \rangle, \quad \langle \bar{u} \rangle = 0. \tag{1.3}$$

Definition 1.5. *We define the space of solenoidal periodic vector fields as*

$$L^2_{sol}(\square) = \{p \in L^2(\square), \quad \operatorname{div} p = 0 \text{ in } \mathbb{R}^m\},$$

where the property $\operatorname{div} p = 0$ *in* $\mathbb{R}^m$ *means*

$$\int_{\mathbb{R}^m} p_i \frac{\partial \phi}{\partial x_i} \,\mathrm{d}x = 0\,, \quad \forall \phi \in \mathcal{C}^\infty_0(\mathbb{R}^m). \tag{1.4}$$

Clearly, $L^2_{sol}(\square)$ is a closed subspace of $L^2(\square)$. The identity (1.4) can be equivalently rewritten as

$$\int_{\square} p \cdot \nabla \phi \,\mathrm{d}x = 0\,, \quad \forall \phi \in \mathcal{C}^\infty(\square). \tag{1.5}$$

A brief proof of this statement can be found in [Jikov et al., 1994, p. 6].

Definition 1.6. *We define the space of periodic potential vector fields as*

$$L^2_{pot}(\square) = \mathbb{R}^m \oplus \nu^2_{pot}(\square),$$

where $\nu^2_{pot}(\square)$ *is the space of periodic potential vector fields with zero mean value*

$$\nu^2_{pot}(\square) = \{\nabla u,\ u \in H^1(\square)\}.$$

Thanks to the Poincaré inequality, $\nu^2_{pot}(\square)$ is a closed subspace of $L^2(\square)$. As a consequence of (1.5) we obtain the following orthogonal representation

$$L^2(\square) = L^2_{sol}(\square) \oplus \nu^2_{pot}(\square). \tag{1.6}$$

Remark. Any vector $v \in L^2_{pot}(\square)$ can be represented in the form

$$v = \langle v \rangle + \nabla u,\ u \in H^1(\square).$$

Similarly, any solenoidal vector field $p \in L^2_{sol}(\square)$ has the form

$$p_j = \langle p_j \rangle + \frac{\partial \alpha_{ij}}{\partial x_i},$$

where α is a skew-symmetrical matrix s.t. $\alpha_{ij} \in H^1(\square), \quad \langle \alpha_{ij} \rangle = 0$. By $\nu^2_{sol}(\square)$ we will denote the space of vector fields $v \in L^2_{sol}(\square)$ with zero mean value, i.e. $\langle v \rangle = 0$. We can then write

$$L^2(\square) = \mathbb{R}^m \oplus \nu^2_{pot}(\square) \oplus \nu^2_{sol}(\square).$$

Lemma 1.2 (Independence property)**.** *Let f, g be vector fields satisfying $f \in L^2_{pot}(\square)$, $g \in L^2_{sol}(\square)$. Then*

$$\langle f \cdot g \rangle = \langle f \rangle \cdot \langle g \rangle.$$

Proof. We know that f can be expressed as

$$f = \langle f \rangle + \nabla u, \quad u \in H^1(\square).$$

Then we have

$$\begin{aligned}
\frac{1}{|\square|}\int_{\square} f \cdot g \,\mathrm{d}x &= \frac{1}{|\square|}\int_{\square} (\langle f \rangle + \nabla u) \cdot g \,\mathrm{d}x \\
&= \langle f \rangle \cdot \frac{1}{|\square|}\int_{\square} g \,\mathrm{d}x \;+\; \frac{1}{|\square|}\int_{\square} \nabla u \cdot g \,\mathrm{d}x \\
&= \langle f \rangle \cdot \langle g \rangle \;+\; \frac{1}{|\square|}\int_{\partial\square} u\, g \cdot \eta \,\mathrm{d}S \;-\; \frac{1}{|\square|}\int_{\square} u \,\mathrm{div}\,(g) \,\mathrm{d}x \\
&= \langle f \rangle \cdot \langle g \rangle,
\end{aligned}$$

where η is the normal vector to the faces (edges) of $\square$.

□

1.1.3 Periodic problem

When deriving the solution of the homogenization problem one has to deal with a special kind of boundary value problem, the periodic problem. The existence and uniqueness of its solution are a consequence of the Lax-Milgram theorem, which can be found in [Evans, 2010, p. 297].

Let $A(x) = \{a_{ij}(x)\}_{i,j=1}^m$ be a matrix (not necessarily symmetric) with $\square$ - periodic bounded measurable elements, satisfying the ellipticity condition, i.e.

$$\begin{aligned}
&a_{ij} \quad \square \text{ - periodic} \\
&a_{ij} \in L^\infty(\mathbb{R}^m) \\
c_A \left\|\xi\right\|^2_{\mathbb{R}^m} \leq (A(x)\xi, \xi)_{\mathbb{R}^m}, &\quad \forall \xi, x \in \mathbb{R}^m, \quad \text{with } c_A > 0.
\end{aligned} \tag{1.7}$$

The matrix $A(x)$ is associated with the following differential operator

$$\frac{\partial}{\partial x_i}\Big(a_{ij}\frac{\partial}{\partial x_j}\Big) = \mathrm{div}\,(A\nabla)$$

and the following bilinear form

$$a(u, \varphi) = \int_Q \nabla\varphi \cdot A\nabla u \,\mathrm{d}x.$$

Let $f_0 \in L^2(\square)$, $f \in L^2(\square)$ be an arbitrary function and vector field. Solving a periodic problem means solving the problem

$$\begin{aligned} -\operatorname{div}(A\nabla u) &= -f_0 + \operatorname{div} f \qquad \text{in } \square \\ u \quad &\square \text{ - periodic} \end{aligned} \tag{1.8}$$

in the sense of the following definition.

Definition 1.7 (Weak solution)**.** *Let $A(x)$ be as in (1.7) and $f_0 \in L^2(\square)$, $f \in L^2(\square)$ be an arbitrary function and vector field. We say that $u \in H^1(\square)$ is a weak solution of the periodic problem (1.8) if the following identity holds*

$$\langle \nabla\varphi \cdot A\nabla u \rangle = \langle f_0\, \varphi \rangle + \langle \nabla\varphi \cdot f \rangle, \quad \forall \varphi \in H^1(\square). \tag{1.9}$$

When assuming $\langle f_0 \rangle = 0$, seeking the solution in the whole $H^1(\square)$ would clearly lead to a non-uniqueness. The appropriate space is

$$V = \{u \in H^1(\square), \quad \langle u \rangle = 0\}$$

and in that case the problem is well-posed according to the Hadamard conditions, as stated in the following theorem.

Theorem 1.2. *Let $A(x)$ be a $\square$ -periodic, c_A -elliptic matrix with bounded elements. Let $f_0 \in L^2(\square)$ be a function, $\langle f_0 \rangle = 0$, $f \in L^2(\square)$, a vector field.*
Then there exists a unique weak solution $\bar{u} \in V$ to the periodic problem (1.8). Moreover,

$$\|\nabla \bar{u}\|_{L^2(\square)} \leq c_A^{-1} \|f\|_{L^2(\square)}. \tag{1.10}$$

Remark. Compare the space V to the space $W(\square)$ mentioned in Remark (1.2). Thanks to the Poincaré-Wirtinger inequality we can introduce a norm in the space $W(\square)$ as

$$\|\bar{u}\|_{W(\square)} = \|\nabla u\|_{L^2(\square)}. \quad \forall u \in \bar{u}, \bar{u} \in W(\square).$$

Therefore the estimate (1.10) is applied to the norm of the equivalence classes in $H^1(\square)$.

Proof. When V being the space of test functions, our problem meets all the assumptions of the Lax-Milgram theorem. Therefore we know there is a unique weak solution $\bar{u} \in V$ satisfying the equality (1.9) for all $\varphi \in V$. Since $\langle f_0 \rangle = 0$, it holds for all $\varphi \in H^1(\square)$.

□

1.2 Setting of the homogenization problem

Let $A(x)$, $x \in \mathbb{R}^m$, be a $\square$ - periodic matrix that satisfies:

$$\begin{array}{l}(A\xi,\xi)_{\mathbb{R}^m} \geq c_A \|\xi\|^2_{\mathbb{R}^m} \\ \|A\xi\|_{\mathbb{R}^m} \leq \beta \|\xi\|_{\mathbb{R}^m}\end{array} \quad \text{for any } \xi \in \mathbb{R}^m \text{ and a.e. in } Q. \tag{1.11}$$

Now set

$$a^\varepsilon_{ij} = a_{ij}\Big(\frac{x}{\varepsilon}\Big) \quad \text{a.e. on } \mathbb{R}^d, \quad \forall i,j = 1,\dots,m$$

and

$$A^\varepsilon(x) = \big(a^\varepsilon_{ij}(x)\big)^m_{i,j=1}, \quad \text{a.e. on } \mathbb{R}^m. \tag{1.12}$$

The homogenization theory allows to describe the asymptotic behaviour as $\varepsilon \to 0$ of partial differential equations of many types. For our purpose we will be dealing with the following Dirichlet problem (1.13)

$$\begin{array}{rl}-\mathrm{div}\,(A^\varepsilon \nabla u^\varepsilon) = f & \text{in } Q \\ u \in H^1_0(Q), & \end{array} \tag{1.13}$$

where $f \in H^{-1}(Q)$ is given.

Definition 1.8. *A constant positive definite matrix A^0 is said to be the homogenized matrix for $A(x)$, if for any bounded domain $Q \subset \mathbb{R}^m$ and any $f \in H^{-1}(Q)$ the solutions u^ε of the Dirichlet problem (1.13) possess the following property of convergence as $\varepsilon \to 0$:*

$$\begin{array}{rl}u^\varepsilon \rightharpoonup u^0 & \text{in } H^1_0(Q), \\ A^\varepsilon \nabla u^\varepsilon \rightharpoonup A^0 \nabla u^0 & \text{in } L^2(Q),\end{array} \tag{1.14}$$

where u^0 is the solution of the Dirichlet problem

$$\begin{array}{rl}-\mathrm{div}(A^0 \nabla u^0) = f & \text{in } Q \\ u^0 \in H^1_0(Q). & \end{array} \tag{1.15}$$

The operator $A = \mathrm{div}\,(A^0\nabla)$ is called a homogenized operator and the equation (1.15) is called a homogenized equation.

The vector fields

$$\begin{array}{l}p^\varepsilon = A^\varepsilon \nabla u^\varepsilon, \\ p^0 = A^0 \nabla u^0\end{array}$$

are called flows.

The homogenized matrix is often referred to as the effective matrix and reflects the physical concept of an effective homogeneous medium.

Remark. The solutions of the sequence of Dirichlet problems satisfy

$$\|u^\varepsilon\|_1 = \|\nabla u^\varepsilon\|_{L^2(Q)} \leq c_A^{-1} \|f\|_{H^{-1}(Q)}. \tag{1.16}$$

The flows satisfy a similar estimate

$$\|p^\varepsilon\|_{L^2(Q)} \leq \beta \|\nabla u^\varepsilon\|_{L^2(Q)} \leq \frac{\beta}{c_A} \|f\|_{H^{-1}(Q)}. \tag{1.17}$$

These estimates will be of great use later, when we discuss the computation of the homogenized matrix.

1.3 The main convergence result

In this section we will state the classical result of the homogenization theory, i.e. that the homogenized matrix can be obtained through solutions of some auxiliary periodic problems in the reference cell $\square$.

1.3.1 Auxiliary periodic problem

Let $A(x)$ be as in (1.7) and let $\lambda \in \mathbb{R}^m$ be an arbitrary constant. Now consider an auxiliary periodic problem

$$\begin{aligned} \operatorname{div}(A\,v) &= 0 \quad \text{in } \square \\ v &\in L^2_{pot}(\square) \\ \langle v \rangle &= \lambda \in \mathbb{R}^m. \end{aligned} \tag{1.18}$$

Lemma 1.3. *The problem (1.18) has a unique weak solution.*

Proof. We know that the vector field $v \in L^2_{pot}(\square)$ can be expressed as

$$v = \lambda + \nabla u, \quad u \in H^1(\square). \tag{1.19}$$

When connecting this with our equation we obtain

$$\operatorname{div}(A\nabla u) = -\operatorname{div}(A\lambda)$$

which has the form of (1.8) with $f_0 = 0, f = -A\lambda$. For the periodic problem we already proved the existence and uniqueness of a weak solution.

□

From (1.19) we can see that v as a solution of an auxialiary periodic problem depends linearly on λ. Since $A\,v$ is a linear transformation with respect to v and $\langle\cdot\rangle$ is a linear operation we know that $\langle A\,v\rangle$ is a linear form with respect to $\lambda \in \mathbb{R}^m$ and therefore can be represented in the from

$$\langle A\,v\rangle = \overline{A}\,\lambda, \qquad \overline{A} \text{ being a constant matrix.} \tag{1.20}$$

Now consider a dual auxiliary periodic problem

$$\begin{aligned} \operatorname{div}(A^\top w) &= 0 \quad \text{in } \square \\ w &\in L^2_{pot}(\square) \\ \langle w \rangle &= \xi \in \mathbb{R}^m, \end{aligned} \tag{1.21}$$

where A is the same as before and $\xi \in \mathbb{R}^m$ is an arbitrary constant.

Analogously to the previous case we can prove the existence of a unique solution, a linear dependency of $\langle A^\top w\rangle$ on ξ and define a constant matrix $\overline{C}$ as

$$\langle A^\top w\rangle = \overline{C^\top}\xi.$$

Corollary. $\overline{A} = \overline{C}$

Proof. For an arbitrary $\lambda, \xi \in \mathbb{R}^m$ it holds:

$$\xi \cdot \overline{A}\lambda = \xi \cdot \langle Av \rangle = \langle w \cdot A\, v \rangle,$$

where the second equality comes from the Independence property (Lemma 1.2). We can continue analogously for the dual operator.

$$\xi \cdot \overline{C}\lambda = \lambda \cdot \overline{C^\top}\xi = \lambda \cdot \langle A^\top w \rangle = \langle v \cdot A^\top w \rangle = \langle w \cdot A\, v \rangle,$$

where the third equality comes again from the Independence property. Since these equations hold for all $\lambda, \xi \in \mathbb{R}^m$, we have what we needed to prove.

□

The equality $\overline{A} = \overline{C}$ also implies that $\overline{A^\top} = \overline{A}^\top$. In particular, if the original matrix $A(x)$ was symmetric, $\overline{A}$ is also symmetric.

Corollary. $\overline{A}$ is elliptic.

Proof. As before, from the fact that $A\, v \in L^2_{sol}(\square)$, $v \in L^2_{pot}(\square)$ and the independence property we obtain

$$\lambda \cdot \overline{A}\lambda = \langle v \cdot A\, v \rangle \geq c_A \langle |v|^2 \rangle \geq c_A \left| \langle v \rangle \right|^2 = c_A \left| \lambda \right|^2, \quad \forall \lambda \in \mathbb{R}^m.$$

□

1.3.2 Fundamental homogenization theorem

The fundamental homogenization theorem explains the relation between the homogenized matrix A^0 and the auxiliary periodic problem, in particular the constant matrix $\overline{A}$.

Theorem 1.3. *Let A be a $\square$ - periodic matrix satisfying the conditions (1.11). Consider the auxiliary periodic problem (1.18) and define a constant matrix $\overline{A}$ by (1.20). Then $\overline{A}$ is a homogenized matrix in the sense of Definition 1.8, i.e. we can set A^0 to be $\overline{A}$.*

Proof. Let u^ε be the solution of the Dirichlet problem (1.13). From the estimates (1.16) and (1.17) we can see that the sequence u^ε is bounded in $H^1_0(Q)$ and the sequence of flows is bounded in $L^2(Q)$. This implies that there is a subsequence u^{ε_1} weakly convergent to u^0 and a subsequence p^{ε_2} weakly convergent to p^0.

Now consider the dual auxiliary periodic problem (1.21). We know that it has a unique weak solution w. For $w^\varepsilon(x) = w(\varepsilon^{-1}x)$ by the property of the mean value we have

$$w^\varepsilon \rightharpoonup \langle w \rangle = \xi,$$
$$A^{\top\varepsilon}\, w^\varepsilon \rightharpoonup \langle A^\top\, w \rangle = \overline{C^\top}\xi.$$

For each ε we can derive that

$$w^\varepsilon \cdot p^\varepsilon = w^\varepsilon \cdot A^\varepsilon \nabla u^\varepsilon = A^{\top^\varepsilon} w^\varepsilon \cdot \nabla u^\varepsilon, \quad x \in Q. \tag{1.22}$$

From the definition of the dual auxiliary problem (1.21) we get

$$\begin{aligned} \operatorname{curl} w^\varepsilon &= 0, \\ \operatorname{div} (A^{\top^\varepsilon} w^\varepsilon) &= 0, \end{aligned}$$

on the other hand, for ∇u^ε and p^ε we have

$$\begin{aligned} \operatorname{curl} \nabla u^\varepsilon &= 0 \\ \operatorname{div} p^\varepsilon &= -f. \end{aligned}$$

Both pairs in the equation (1.22) therefore satisfy the assumption is Lemma 1.1 which means we can pass to the weak* limit. Counting in the uniqueness of the weak* limit we can affirm

$$\xi \cdot p^0 = \overline{C^\top} \xi \cdot \nabla u^0 = \xi \cdot \overline{A} \nabla u^0,$$

which implies

$$p^0 = \overline{A} \nabla u^0. \tag{1.23}$$

Moreover, $\operatorname{div} p^\varepsilon = -f$, and consequently $\operatorname{div} p^0 = -f$. The function u^0 is therefore a solution of the Dirichlet problem (1.13) with $A = \overline{A}$. Up until now u^0 was dependent on the choice of the subsequence u^{ε_1}. By virtue of the uniqueness of the solution of this problem we can affirm that the whole sequence u^ε weakly converges to u^0. From (1.23) we get that the sequence p^ε weakly converges to p^0. The matrix $\overline{A}$ therefore satisfies the definition of a homogenized matrix and we have shown that

$$\overline{A} = A^0.$$

□

2. Stochastic homogenization

As opposed to Chapter 1, this chapter is dealing with a problem where the elements of the matrix defining the operator are random fields, i.e.

$$A = A(x, \omega),$$

where $\omega \in \Omega$, $(\Omega, \mathcal{F}, \mu)$ being a probability space.

The following problem can serve as a motivation for the development of stochastic homogenization ($\varepsilon \ll 1$) :
For almost every $\omega \in \Omega$

$$\begin{aligned} -\text{div}\left(A(\varepsilon^{-1}x, \omega)\nabla u^\varepsilon(x, \omega)\right) &= f(x) \quad \text{in } Q, \\ u^\varepsilon(x, \omega) &= 0 \qquad \text{on } \partial Q. \end{aligned}$$

The rigorous definition of the problem of stochastic homogenization is provided in the section 2.2.

As the main sources we shall mention [Jikov et al., 1994], [Anantharaman et al., 2011] and [Bourgeat and Piatnitski, 2004].

2.1 Preliminaries

As mentioned before, in order to obtain some valuable homogenization results we need to state some assumptions on the random field $A(x, \omega)$. This section collects all necessary background concepts required in our coverage of stochastic homogenization.

2.1.1 Probabilistic setting

In what follows, $(\Omega, \mathcal{F}, \mu)$ will be the considered probability space.

Definition 2.1. *We define the spaces $L^\alpha(\Omega)$, $\alpha \geq 1$ by*

$$\begin{aligned} L^\alpha(\Omega) &= \{f : \Omega \to \mathbb{X},\ f \text{ measurable },\ \int_\Omega \|f\|_{\mathbb{X}}^\alpha \ \mathrm{d}\mu < \infty\} \\ L^\infty(\Omega) &= \{f : \Omega \to \mathbb{X},\ f \text{ measurable },\ f \text{ essentially bounded }\}, \end{aligned}$$

where $\mathbb{X}$ denotes either $\mathbb{R}$ or $\mathbb{R}^m$, depending on the context.

One of the needed assumptions will be that of the stationarity of a random field.

Definition 2.2. *Let $G : \mathbb{R}^m \times \Omega \to \mathbb{R}$ be a random field. We say that G is stationary if for any finite collection of points $x_i \in \mathbb{R}^m$, $i = 1, \ldots, k$ and any $h \in \mathbb{R}^m$ the joint distribution of the random $k-$dimensional vector $(G(x_1 + h, \omega), \ldots, G(x_k + h), \omega)^T$ is the same as that of $(G(x_1, \omega), \ldots, G(x_k), \omega)^T$.*

2.1.2 Ergodic theory

Some of the concepts from ergodic theory are crucial to the discussion that follows in the rest of the thesis.

Definition 2.3. *Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be measure spaces. A transformation $T : \Omega_1 \to \Omega_2$ is called measure preserving if it is measurable, i.e.*

$$\forall E \in \mathcal{F}_2 \quad T^{-1}(E) \in \mathcal{F}_1,$$

and satisfies

$$\mu_1(T^{-1}(E)) = \mu_2(E), \quad \forall E \in \mathcal{F}_2.$$

Definition 2.4 (Dynamical system)**.** *We define an m-dimensional measure-preserving dynamical system T on Ω as a family of measurable mappings $T(x) : \Omega \to \Omega$, parametrized by $x \in \mathbb{R}^m$, which satisfies the following conditions:*

1. *The group property:*

$$\begin{aligned} T(0) &= I \\ T(x+y) &= T(x)\,T(y), \quad \forall x, y \in \mathbb{R}^m, \end{aligned} \tag{2.1}$$

   where I stands for the identity.

2. *T preserves the measure μ on Ω:*

$$\forall x \in \mathbb{R}^m,\ \forall E \in \mathcal{F} : \quad \mu(T^{-1}(E)) = \mu(E).$$

3. *$T(x)$ is a measurable mapping from $\mathbb{R}^m \times \Omega$ to Ω, where $\mathbb{R}^m \times \Omega$ is equipped with the product σ-algebra $\mathcal{B} \times \mathcal{F}$ and $\mathcal{B}$ is the Borel σ-algebra in $\mathbb{R}^m$.*

Remark. It can be shown that if the random field G from the Definition 2.2 can be written in the form

$$G(x, \omega) = g(T(x)\omega),$$

where $g : \Omega \to \Omega$ is a measurable function and T is a measure-preserving dynamical system, then G is stationary.

Now we will explain the notion of an invariant function, invariant set and an ergodic dynamical system.

Definition 2.5 (Invariant function, invariant set)**.** *Let $T(x)$, $x \in \mathbb{R}^m$ be a measure-preserving dynamical system. A measurable function f defined on Ω is called T-invariant if*

$$\forall x \in \mathbb{R}^m : \quad f\big(T(x)\omega\big) = f(\omega), \quad \text{for a.e. } \omega \in \Omega. \tag{2.2}$$

A measurable set $E \in \mathcal{F}$ is called T-invariant if its characteristic function $\mathbb{1}_E$ is T-invariant.

Definition 2.6 (Ergodic dynamical system)**.** *Let $T(x),\ x \in \mathbb{R}^m$ be a measure-preserving dynamical system. T is called ergodic if for every T-invariant set $E \in \mathcal{F}$ we have $\mu(E) = 0$ or 1.*

We shall mention an equivalent characterization of an ergodic dynamical system which is often referred to as the definition of an ergodic dynamical system.

Corollary. Let $T(x),\ x \in \mathbb{R}^m$ be a measure-preserving dynamical system. Then the following is equivalent:

1. T is ergodic.
2. Every invariant function is constant almost everywhere in Ω.

Proof. The proof can be found in [Dajani and Dirksin, 2008, p. 23,24].

□

Birkhoff Ergodic Theorem

The following section is dedicated to show the relation of a function defined on Ω to its corresponding function defined on $\mathbb{R}^m$. The Birkhoff ergodic theorem at the end then shows that under some assumptions we have a strong connection of a stochastic homogenization problem to a deterministic problem.

Definition 2.7 (realization of f)**.** *Let $T(x),\ x \in \mathbb{R}^m$ be a measure-preserving dynamical system. Corresponding to a measurable function f defined on Ω we define the function f_T defined on $\mathbb{R}^m \times \Omega$ by*

$$f_T(x, \omega) = f(T(x)\omega), \quad x \in \mathbb{R}^m,\ \omega \in \Omega. \tag{2.3}$$

For a fixed $\omega \in \Omega$ the function $f_T(\cdot, \omega)$ defined on $\mathbb{R}^m$ is called a realization of f.

Lemma 2.1. *If $f \in L^\alpha(\Omega)$ then almost all its realizations belong to $L^\alpha_{loc}(\mathbb{R}^m)$.*

Proof. The Lebesgue-measurability of the function $f_T(\cdot, \omega)$ is a direct consequence of the third condition in the definition of a measure-preserving dynamical system (Definition 2.4).
We still need to prove that

$$\text{for a.e. } \omega \in \Omega \text{ we have: } \int_{|x| \le t} \left| f\big(T(x)\omega\big) \right|^\alpha \, \mathrm{d}x < \infty, \quad \forall t \in \mathbb{R}.$$

With a help of the Fubini theorem we obtain:

$$\begin{aligned} \int_\Omega \Big(\int_{|x| \le t} \left| f\big(T(x)\omega\big) \right|^\alpha \, \mathrm{d}x \Big) \mathrm{d}\mu &= \int_{|x| \le t} \Big(\int_\Omega \left| f\big(T(x)\omega\big) \right|^\alpha \mathrm{d}\mu \Big) \, \mathrm{d}x \\ &= \gamma_m t^m \left\| f \right\|^\alpha_{L^\alpha(\Omega)} < \infty. \end{aligned}$$

From the Chebyshev theorem we know that the finiteness of a Lebesgue integral of $|f|$ implies finiteness of f almost everywhere, which finishes the proof.

□

Corollary. Convergence in $L^\alpha(\Omega)$ implies convergence for a subsequence in $L^\alpha_{loc}(\mathbb{R}^m)$ for almost all corresponding realizations.

Proof. The proof is analogous to the previous one. Again, we can derive:

$$\int_\Omega \Big(\int_{|x|\le t} \big| f_k\big(T(x)\omega\big) - f\big(T(x)\omega\big)\big|^\alpha \, \mathrm{d}x \Big) \mathrm{d}\mu$$
$$= \int_{|x|\le t} \Big(\int_\Omega \big| f_k\big(T(x)\omega\big) - f\big(T(x)\omega\big)\big|^\alpha \, \mathrm{d}\mu \Big) \, \mathrm{d}x = \gamma_m t^m \, \|f_k - f\|^\alpha_{L^\alpha(\Omega)} \to 0.$$

The L^1 convergence implies an almost everywhere pointwise convergence for a subsequence. Which in our case means that there exists a subsequence f_{n_k} for which we have:

$$\int_{|x|\le t} \big| f_{n_k}\big(T(x)\omega\big) - f\big(T(x)\omega\big)\big|^\alpha \, \mathrm{d}x \to 0, \quad \forall t \in \mathbb{R}. \tag{2.4}$$

Or in other words, $\|f_{n_k} - f\|_{L^\alpha_{loc}(\mathbb{R}^m)} \to 0$.

□

Definition 2.8. *Let $f \in L^1(\mathbb{R}^m)$. We call M_f the mean value of f if for every Lebesgue measurable bounded set $K \subset \mathbb{R}^m$ it holds*

$$\lim_{\varepsilon\to\infty} \int_K f(\varepsilon^{-1}x) \, \mathrm{d}x = |K| \, M_f, \tag{2.5}$$

where $|K|$ denotes the Lebesgue measure of K.

There are many ways of how to equivalently express the definition of the mean value of $f(x) \in L^1(\mathbb{R}^m)$. Some of them require some additional assumptions. Here we provide two more for a better geometric understanding:

Remark. Let $f(x) \in \text{Ł}^1(\mathbb{R}^m)$. We call $M_f \in \mathbb{R}$ the mean value of f if it holds

$$\lim_{t\to\infty} \frac{1}{t^m \ |K|} \int_{K_t} f(x) \, \mathrm{d}x = M_f, \ |K| \neq 0, \tag{2.6}$$

for any Lebesgue measurable bounded set $K \subset \mathbb{R}^m$, where $K_t = \{x \in \mathbb{R}^m, t^{-1}x \in K\}$, $t > 0$.

Remark. Let the family of functions $f(\varepsilon^{-1}\,x)$ be bounded in $L^\alpha_{loc}(\mathbb{R}^m)$ for some $\alpha \ge 1$. Then the mean value $M_f \in \mathbb{R}$ can be defined as

$$f(\varepsilon^{-1}x) \rightharpoonup M_f \quad \text{in} \ \ L^\alpha_{loc}(\mathbb{R}^m). \tag{2.7}$$

Theorem 2.1 (Birkhoff Ergodic Theorem)**.** *Let $T(x)$, $x \in \mathbb{R}^m$ be a measure-preserving dynamical system on Ω, let $f \in L^\alpha(\Omega)$, $\alpha \ge 1$. Then for almost all $\omega \in \Omega$ the realization $f_T(x,\omega)$, as defined in (2.3) possesses a mean value in the sense of (2.7). Moreover, the mean value $M_f(\omega)$ is T-invariant, i.e.*

$$M_f(T(x)\omega) = M_f(\omega) \quad \forall x \in \mathbb{R}^m, \ \mu - a.e.$$

Also,

$$E(f) \stackrel{\text{def}}{\equiv} \int_\Omega f(\omega)\,\mathrm{d}\mu = \int_\Omega M_f(\omega)\,\mathrm{d}\mu. \tag{2.8}$$

In particular, if the system $T(x)$ is ergodic then the mean value M_f is constant a.e. and is given by

$$M_f = E(f).$$

Proof. Proof can be found in [Dunford and Schwartz, 1988].

□

Let's formulate the result for an ergodic measure-preserving dynamical system in one Corollary.

Corollary. Let $T(x),\ x \in \mathbb{R}^m$ be a measure-preserving and ergodic dynamical system on Ω. Let $f \in L^\alpha(\Omega),\ \alpha \geq 1$. Define $f_T^\varepsilon(x,\omega) = f_T(\varepsilon^{-1}x,\omega)$ for $\varepsilon > 0$. Then, for almost every $\omega \in \Omega$

$$f_T^\varepsilon(\cdot,\omega) \rightharpoonup E(f) \quad \text{in } L^\alpha_{loc}(\mathbb{R}^m),\ \text{as } \varepsilon \to 0.$$

2.1.3 Spaces of random vector fields

The aim of this section is to bring out an important decomposition of the space of vector fields from $L^2(\Omega)$ - the Weyl decomposition. This result will be later used for the stochastic homogenization of random elliptic operators.

Firstly, let's recall some of the basic notions of spaces for vector fields $f = \{f_1,\dots,f_m\},\ f_i \in L^2_{loc}(\mathbb{R}^m)$.

Definition 2.9. *A vector field $f \in L^2_{loc}(\mathbb{R}^m)$ is called vortex-free in $\mathbb{R}^m$ if*

$$\int_{\mathbb{R}^m} \big(f_i \frac{\partial\varphi}{\partial x_j} - f_j \frac{\partial\varphi}{\partial x_i}\big)\,\mathrm{d}x = 0, \quad \forall\varphi \in \mathcal{C}_0^\infty(\mathbb{R}^m). \tag{2.9}$$

A vector field $f \in L^2_{loc}(\mathbb{R}^m)$ is called soleniodal in $\mathbb{R}^m$ if

$$\int_{\mathbb{R}^m} f_i \frac{\partial\varphi}{\partial x_i}\,\mathrm{d}x = 0, \quad \forall\varphi \in \mathcal{C}_0^\infty(\mathbb{R}^m). \tag{2.10}$$

Remark. Since we are dealing with the whole domain $\mathbb{R}^m$, the property of (2.9) is equivalent to the potentiality of the vector field, i.e.

$$f = \nabla u, \quad u \in H^1_{loc}(\mathbb{R}^m). \tag{2.11}$$

Now, let us consider vector fields defined on Ω, i.e. random vector fields.

Definition 2.10. *Let $T(x),\ x \in \mathbb{R}^m$ be a measure-preserving dynamical system on Ω. A vector field $f \in L^2(\Omega)$ is called potential (resp. solenoidal) if almost all its realizations $f_T(\cdot, \omega)$ are potential (resp. solenoidal) in $\mathbb{R}^m$. The space of potential and solenoidal vector fields is denoted by $L^2_{pot}(\Omega)$ and $L^2_{sol}(\Omega)$, respectively.*

Corollary. $L^2_{pot}(\Omega)$ and $L^2_{sol}(\Omega)$ are closed subsets of $L^2(\Omega)$.

Proof. This property is a direct consequence of the second corollary in the Section Birkhoff Ergodic Theorem.

□

Definition 2.11. *We define the space of potential and solenoidal vector fields with zero mean value:*

$$\nu^2_{pot}(\Omega) = \{f \in L^2_{pot}(\Omega),\ E(f) = 0\}$$
$$\nu^2_{sol}(\Omega) = \{f \in L^2_{sol}(\Omega),\ E(f) = 0\}.$$

Therefore we can decompose $L^2_{pot}(\Omega)$ and $L^2_{sol}(\Omega)$ as:

$$L^2_{pot}(\Omega) = \nu^2_{pot}(\Omega) \oplus \mathbb{R}^m$$
$$L^2_{sol}(\Omega) = \nu^2_{sol}(\Omega) \oplus \mathbb{R}^m.$$

The following lemma implies that $\nu^2_{pot}(\Omega)$ and $\nu^2_{sol}(\Omega)$ are mutually orthogonal subspaces of $L^2(\Omega)$, assuming $T(x),\ x \in \mathbb{R}^m$ is ergodic.

Lemma 2.2. *Let $T(x),\ x \in \mathbb{R}^m$ be a measure-preserving, ergodic dynamical system. Let $f \in L^2_{pot}(\Omega)$, $g \in L^2_{sol}(\Omega)$. Then*

$$E(f \cdot g) = E(f) \cdot E(g).$$

Proof. Let's consider a realizations of f and g, $f(x) = f_T(x, \omega), g(x) = g_T(x, \omega)$ for a fixed $\omega \in \Omega$. Thanks to the Birkhoff ergodic theorem and the ergodicity assumption we know that

$$f(\varepsilon^{-1}x) \rightharpoonup E(f)$$
$$g(\varepsilon^{-1}x) \rightharpoonup E(q)$$

and also

$$f(\varepsilon^{-1}x) \cdot g(\varepsilon^{-1}x) \rightharpoonup E(f \cdot g) \quad \text{in } L^1_{loc}(\mathbb{R}^m).$$

On the other hand, from the definition of $L^2_{pot}(\Omega)$, $L^2_{sol}(\Omega)$ we know that $f(x) \in L^2_{pot}(\mathbb{R}^m)$, $g(x) \in L^2_{sol}(\mathbb{R}^m)$ and from the Lemma 1.1 we get:

$$f(\varepsilon^{-1}x) \cdot g(\varepsilon^{-1}x) \overset{*}{\rightharpoonup} E(f) \cdot E(g).$$

Both of these limits are uniquely determined and that finishes the proof.

□

Theorem 2.2 (Weyl Decomposition)**.** *If the measure-preserving dynamical system $T(x)$, $x \in \mathbb{R}^m$ is ergodic then following orthogonal decomposition of $L^2(\Omega)$ holds:*

$$L^2(\Omega) = \nu^2_{pot}(\Omega) \oplus \nu^2_{sol}(\Omega) \oplus \mathbb{R}^m = \nu^2_{pot}(\Omega) \oplus L^2_{sol}(\Omega). \tag{2.12}$$

Proof. A detailed proof is provided in [Jikov et al., 1994, p. 231, 232, 233]. □

2.2 Homogenization

This section is dedicated to the mathematical description of stochastic homogenization. We will define a homogenized matrix in a stochastic setting, state needed assumptions and suggest a way how to compute it.

Firstly, let us introduce the problem we are dealing with.

As before, let $(\Omega, \mathcal{F}, \mu)$ be a probability space, let the physical domain be given by a bounded open set $Q \in \mathbb{R}^m$, $f \in H^{-1}(Q)$ be a deterministic source term. The material properties of a medium with random microstructure are specified by a matrix valued function

$$A(\cdot, \omega) : \mathbb{R}^m \to \mathbb{R}^{m \times m}.$$

For simplicity, we again assume the boundary conditions to be homogeneous Dirichlet. The problem is then given by:

For almost every $\omega \in \Omega$

$$\begin{aligned} -\text{div}\left(A(\varepsilon^{-1}x, \omega)\nabla u^\varepsilon(x, \omega)\right) &= f(x) \quad \text{in } Q, \\ u^\varepsilon(x, \omega) &= 0 \qquad \text{on } \partial Q. \end{aligned} \tag{2.13}$$

The goal is to find a constant matrix A^0 such that for every $\omega \in \Omega$ the solution u^0 of the problem (2.14) given bellow provides a reasonable approximation of the limit u^ε as $\varepsilon \to 0$, u^ε being the solution of the problem (2.13).

$$\begin{aligned} -\text{div}\left(A^0 \nabla u^0\right) &= f \quad \text{in } Q, \\ u^0 &= 0 \quad \text{on } \partial Q. \end{aligned} \tag{2.14}$$

Definition 2.12 (admits homogenization)**.** *Let $A : \mathbb{R}^m \to \mathbb{R}^{m \times m}$ be a uniformly bounded and positive definite matrix valued function. We say that A admits homogenization if there exists a constant matrix A^0 satisfying the Definition 1.8.*

Remark. In the case of a periodic media the existence of a homogenized matrix is proved in Chapter 1.

In what follows we will assume:

1. A can be expressed in terms of a measure-preserving dynamical system:

$$A(x,\omega) = \mathcal{A}(T(x)\omega), \quad \forall x \in \mathbb{R}^m,\, \omega \in \Omega, \tag{2.15}$$

   where $T(x)$, $x \in \mathbb{R}^m$ is an $m-$dimensional measure-preserving ergodic dynamical system.

2. $\mathcal{A} : \Omega \to \mathbb{R}^{m\times m}$ is a measurable function.

3. $\mathcal{A}$ is uniformly bounded and positive definite: $0 < c_A \le C_A$

$$c_A \left|\xi\right|^2 \le \xi \cdot \mathcal{A}(\omega)\xi \le C_A \left|\xi\right|^2, \quad \forall \xi \in \mathbb{R}^m, \text{ for a.e. } \omega \in \Omega.$$

2.2.1 Auxiliary equation problem

In this section we introduce an auxiliary problem defined on Ω and show its equivalency to a series of problems defined on $\mathbb{R}^m$.

For a fixed $\xi \in \mathbb{R}^m$ consider the following problem:

$$\begin{gathered} \text{Find } v_\xi \in \nu^2_{pot}(\Omega) \text{ s.t.} \\ \int_\Omega \varphi(\omega) \cdot \mathcal{A}(\omega)\left(\xi + v_\xi(\omega)\right) \mathrm{d}\mu = 0, \quad \forall \varphi \in \nu^2_{pot}(\Omega). \end{gathered} \tag{2.16}$$

Thanks to the estimate $\int_\Omega v(\omega) \cdot \mathcal{A}(\omega)\, v(\omega))\, \mathrm{d}\mu = E(v \cdot \mathcal{A}v) \ge c_A \left\|v\right\|^2_{L^2(\Omega)}$ we can use the Lax-Milgram theorem which proves the existence of a unique solution.

The problem (2.16) can be rewritten as:

$$\begin{gathered} \text{Find } v_\xi \in \nu^2_{pot}(\Omega) \text{ s.t.} \\ \mathcal{A}\left(\xi + v_\xi\right) \in Ł^2_{sol}(\Omega). \end{gathered} \tag{2.17}$$

Now, let's translate this problem to $\mathbb{R}^m$ by considering realization for a fixed $\omega \in \Omega$. Let $u_\xi(x)$ be the potential function for the realization of the solution $\bar{v_\xi}(x) = v_\xi(T(x)\omega)$ and $\bar{\mathcal{A}}(x) = \mathcal{A}(T(x)\omega)$. Then

$$\operatorname{div}\left(\bar{\mathcal{A}}(x)\left(\xi + \nabla u_\xi\right)\right) = 0. \tag{2.18}$$

Hence, for a realization the equation (2.16) is reduced to an elliptic equation in $\mathbb{R}^m$, which is of the same form as the auxiliary equation in the periodic case (1.18) from Chapter 1.

From the other side, assume that a vector field $v \in \nu^2_{pot}(\Omega)$ satisfies the equation (2.18) for almost all its realizations. Then almost all realizations of $\mathcal{A}(\omega)\left(\xi + v(\omega)\right)$ are solenoidal vector fields, which means (2.17) holds.

Similarly as before, we can see that $E(\bar{\mathcal{A}}(\xi + v_\xi))$ is a linear form with respect to ξ. Therefore we can define a constant matrix $\overline{A}$ by

$$\overline{A}\,\xi = E(\bar{\mathcal{A}}(\xi + v_\xi)). \tag{2.19}$$

Now, for a fixed $\varsigma \in \mathbb{R}^m$ consider the dual auxiliary problem:

$$\begin{aligned} &\text{Find } w_\varsigma \in L^2_{pot}(\Omega) \text{ s.t.} \\ &A^\top w_\varsigma \in L^2_{sol}(\Omega), \\ &E(w_\varsigma) = \varsigma. \end{aligned} \tag{2.20}$$

We proceed with a similar train of thoughts. The solution w_ς depends linearly on $\varsigma \in \mathbb{R}^m$, which means $E(A^\top w_\varsigma)$ is also linear with respect to ς and therefore we can define a constant matrix $\overline{C}$ s.t.

$$E(A^\top w_\varsigma) = \overline{C}^\top \varsigma.$$

Corollary. $\overline{C} = \overline{A}$.

Proof. From the definition of the problems (2.16), (2.20) and the Weyl decomposition (2.12) we observe the following orthogonality properties

$$\begin{aligned} A(v_\xi + \xi) &\perp w_\varsigma - \varsigma \\ A^\top w_\varsigma &\perp v_\xi. \end{aligned}$$

From which we can derive

$$\begin{aligned} \varsigma \cdot \overline{A}\xi &= \varsigma \cdot E(A(v_\xi + \xi)) = E(\varsigma \cdot A(v_\xi + \xi)) \\ &= E(w_\varsigma \cdot A(v_\xi + \xi)) = E(A^\top w_\varsigma \cdot \xi) = E(A^\top w_\varsigma) \cdot \xi = \overline{C}^\top \varsigma \cdot \xi = \varsigma \cdot \overline{C}\xi, \end{aligned}$$

for arbitrarily chosen $\varsigma, \xi \in \mathbb{R}^m$.

□

Corollary. The matrix $\overline{\mathcal{A}}$ is positive definite.

Proof. Let us derive

$$\begin{aligned} \xi \cdot \overline{A}\xi &= E(\xi \cdot A(v_\xi + \xi)) = E((v_\xi + \xi)A(v_\xi + \xi)) \\ &\geq c_A E(|v_\xi + \xi|^2) \geq c_A \left|E(v_\xi + \xi)\right|^2 = c_A \left|\xi\right|^2, \end{aligned}$$

where we used the fact that A is elliptic with the constant of ellipticity c_A and the orthogonality $A(v_\xi + \xi) \perp v_\xi$.

□

2.2.2 The main theorem

This section states the main theorem of stochastic homogenization, explaining the meaning of the ergodicity assumption.

Theorem 2.3. *Let $\mathcal{A} : \Omega \to \mathbb{R}^{m\times m}$ meet both the second and the third assumption from section 4.1.1. Moreover, assume that $T(x)$, $x \in \mathbb{R}^m$ is a measure-preserving ergodic dynamical system. Then, for almost all $\omega \in \Omega$ the realization $A(x,\omega) = \mathcal{A}(T(x)\omega)$ admits homogenization. The homogenized matrix A^0 is then independent of ω and is equal to the matrix $\overline{A}$ defined in (2.19).*

Proof. The proof will be held in a similar fashion as the proof of the Theorem 1.3. Let u^ε be the solution of the Dirichlet problem

$$\begin{aligned} \operatorname{div}\,(A^\varepsilon(x)\nabla u^\varepsilon(x)) &= f, \\ u^\varepsilon &\in H^1_0(Q), \end{aligned}$$

with $Q \subset \mathbb{R}^m$ bounded, $f \in H^{-1}(Q)$, both arbitrarily chosen and $A(x)$ is a realization of $A(T(x)\omega_1)$ for a specific $\omega_1 \in \Omega$.

We know that the sequence u^ε is bounded in $H^1_0(Q)$ and the sequence of flows $A^\varepsilon\nabla u^\varepsilon$ is bounded in $L^2(Q)$, which implies there are a weakly convergent subsequences u^{ε_1} and p^{ε_1} with weak limits u^0, p^0, respectively.

Consider the auxiliary problem (2.20) and set

$$\begin{aligned} w_\varsigma(x) = w_\varsigma(T(x)\omega_1), &\quad w^\varepsilon_\varsigma(x) = w_\varsigma(\varepsilon^{-1}x), \\ q_\varsigma(x) = A^\top(x)w_\varsigma(x), &\quad q^\varepsilon_\varsigma(x) = q_\varsigma(\varepsilon^{-1}x). \end{aligned}$$

The Birkhoff ergodic theorem implies that

$$\begin{aligned} w^\varepsilon_\varsigma &\rightharpoonup \varsigma, \\ q^\varepsilon_\varsigma &\rightharpoonup \overline{A}^\top\varsigma. \end{aligned}$$

Both of these problems are related through the identity

$$w^\varepsilon_\varsigma \cdot p^\varepsilon = w^\varepsilon_\varsigma \cdot A^\varepsilon\nabla u^\varepsilon = \nabla u^\varepsilon \cdot A^{\varepsilon\top} w^\varepsilon_\varsigma = \nabla u^\varepsilon \cdot q^\varepsilon_\varsigma. \tag{2.21}$$

We would like to use the Lemma 1.1 again, for which we need to verify its assumptions. From the definitions of both of these problems we attain

$$\begin{aligned} \operatorname{curl} w^\varepsilon_\varsigma = 0, &\quad \operatorname{div} p^\varepsilon = 0, \\ \operatorname{curl} \nabla u^\varepsilon = 0, &\quad \operatorname{div} q^\varepsilon_\varsigma = 0, \end{aligned}$$

which means we can pass to the weak* limit and by virtue of the uniqueness of the weak* limit we can affirm

$$\varsigma \cdot p^0 = \nabla u^0 \cdot \overline{A}^\top\varsigma = \varsigma \cdot \overline{A}\nabla u^0.$$

Since ς was chosen arbitrarily, it follows that $p^0 = \overline{A}\nabla u^0$. Moreover, $\operatorname{div} p^\varepsilon = f$ and consequently $\operatorname{div} p^0 = f$. Therefore, u^0 is a solution of the homogenized Dirichlet problem

$$\operatorname{div} \overline{A}\nabla u^0 = f, \quad u^0 \in H^1_0(Q). \tag{2.22}$$

In the beginning, u^0 and p^0 were dependent on the choice of the subsequence of u^ε and p^ε. However, thanks to the uniqueness of the solution of the problem

(2.22) we can state $u^\varepsilon \rightharpoonup u^0$ and $p^\varepsilon \rightharpoonup p^0$. Since the matrix $\overline{A}$ satisfies all of the requirements for the homogenized matrix A^0, we can proclaim

$$\overline{A} = A^0.$$

In addition, $\overline{A}$ is independent of the choice of $\omega_1 \in \Omega$.

□

2.3 Approximations of effective coefficients

This section follows results stated in [Bourgeat and Piatnitski, 2004].

Thanks to the Theorem 2.3 we know that in order to find the homogenized matrix we need to solve (2.17). This problem is stated in the whole space $\mathbb{R}^m$ and therefore it does not allow for any direct approximation procedures.

However, according to the Birkhoff theorem 2.1, A^0 can be approximated by spatial averages:

$$A^0 = \lim_{\rho\to\infty} \frac{1}{\rho^m} \int_{S_\rho} \mathcal{A}(T(x)\omega)\big(\xi + v_\xi(T(x)\omega)\big)\,\mathrm{d}x,$$

where $S_\rho = [-\frac{\rho}{2}, \frac{\rho}{2}]^m$.

This motivates the use of some approximation models which are only dealing with a deterministic problem on a bounded domain. In the next section we will introduce one of them - the periodic model and prove its convergence.

2.3.1 Periodic model

The process is such that the material coefficients $A(x,\omega) = A(T(x)\omega)$ are firstly restricted onto the domain S_ρ and then S_ρ-periodically extended to the whole domain $\mathbb{R}^m$. The new periodic media is denoted by $A^\rho_{per}(x,\omega)$

$$A^\rho_{per}(x,\omega) = \mathcal{A}(T(\mathrm{x} \bmod S_\rho)\,\omega).$$

Now consider $\omega \in \Omega$ fixed. The family of operators

$$A^{\rho,\varepsilon} = \big(A^\rho_{per,ij}(\frac{x}{\varepsilon},\omega)\big)^m_{i,j=1}$$

has periodic coefficients which means we returned to the problem studied in Chapter 1 where we proved the existence of a homogenized matrix and suggested an easy way of how to obtain it:

$$\widetilde{A}^\rho\lambda = \frac{1}{\rho^m} \int_{S_\rho} A(x,\omega)\,(\lambda + \nabla u^\rho_\lambda)\,\mathrm{d}x, \tag{2.23}$$

where u_λ^ρ is the solution of:

$$\begin{aligned} &\text{Find } u_\lambda^\rho \in H^1(S_\rho) \text{ s.t.} \\ &-\text{div } \left(A_{per}^\rho(x,\omega)(\nabla u_\lambda^\rho + \lambda)\right) = 0 \\ &\langle u_\lambda^\rho \rangle = 0. \end{aligned} \tag{2.24}$$

We have to realize that $A_{per}^\rho(x,\omega)$ is no longer ergodic which means the obtained homogenized matrix $\widetilde{A}^\rho$ is dependent on $\omega \in \Omega$.

2.3.2 Convergence result

Theorem 2.4. *Let $\widetilde{A}^\rho$ be the effective matrix of the periodic approximation model obtained in (2.23). Then the following convergence holds*

$$\forall i,j = 1,\dots,m \qquad \lim_{\rho\to\infty} \widetilde{A}_{ij}^\rho = A_{ij}^0 \quad a.s.$$

Proof. We start with the case $S_1 = [0,1]^m$. Consider an auxiliary problem

$$\begin{aligned} -\text{div } (A_{per}^\rho(\rho x,\omega)\nabla u^\rho) = f(x), \quad x \in S_1, \\ u^\rho \in H_0^1(S_1). \end{aligned} \tag{2.25}$$

Since $A_{per}^\rho(\rho x,\omega) = A(\rho x,\omega)$ for $x \in S_1$, the problem (2.25) is a particular problem from the family of problems (1.13) for $Q = S_1$ and $\varepsilon = 1/\rho$. This means that the homogenized matrix obtained from the problem (2.25) as $\rho \to \infty$ coincides with A^0.

By setting

$$u_{\lambda,0}^\rho(x) = \frac{1}{\rho} u_\lambda^\rho(\rho x)$$

we rescale the problem (2.24) into the unit cube S_1, i.e. $u_{\lambda,0}^\rho$ is a $[0,1]^m$-periodic function. By u_I^ρ we will denote the vector-function $\left(u_{0,e_1}^\rho,\dots,u_{0,e_m}^\rho\right)$.

We know that the solution is bounded in $H^1(S_1)$ norm, i.e.

$$\|u_I^\rho\|_{H^1(S_1)} \le C.$$

As a consequence, there is a subsequence of u_I^ρ weakly convergent to u_I^∞ as $\rho \to \infty$. From the Theorem 5.2. in [Jikov et al., 1994] we know that u_I^∞ satisfies

$$\begin{aligned} &u_I^\infty \ S_1\text{-periodic} \\ &-\text{div } \left(A^0(\nabla u_I^\infty + I)\right) = 0 \quad \text{in } S_1, \\ &\langle u_I^\infty \rangle = 0. \end{aligned}$$

By uniqueness of the solution of this problem we can affirm $u_I^\infty = 0$. This weak limit does not depend on the choice of the weakly convergent subsequence, which implies that the whole sequence u_I^ρ is weakly convergent a.s., as $\rho \to \infty$.

Moreover, by the same Theorem in [Jikov et al., 1994], the flows also weakly converge a.s.

$$A(\rho x, \omega)\big(\nabla u_I^\rho\big) \rightharpoonup A^0\big(\nabla u_I^\infty + I\big) = A^0 \quad \text{in } L^2(S_1)^{m\times m}. \tag{2.26}$$

By integrating both sides over S_1, i.e. applying a functional on each entry of the matrices (2.26) and using the formula (2.23) we can write

$$\tilde{A}^\rho = \int_{S_1} A(\rho x, \omega)\big(\nabla u_I^\rho + I\big)\,\mathrm{d}x \xrightarrow{\rho\to\infty} \int_{S_1} A^0\,\mathrm{d}x = A^0 \quad \text{a.s.} \tag{2.27}$$

□

3. Fourier-Galerkin method

This chapter is devoted to the numerical solution of the problem introduced in Chapter 1 and follows the results stated in [Vondřejc et al., 2014]. As we could have seen in Chapter 2, under some assumptions even the problem of stochastic homogenization is translated into the problem of deterministic homogenization with S_ρ-periodic material coefficients. In the following we will assume

$$\square = S_\rho = \big[-\frac{\rho}{2}, \frac{\rho}{2}\big]^m.$$

3.1 Preliminaries

In this chapter we will specify the dimension of the space where the image of our functions lies. In Chapter 1 we used $L^p(\square)$ for both real-valued periodic functions or periodic vector fields because it was easy to distinguish them from the context. To keep the notation compact when introducing some concepts defined regardless of the dimension, we will denote $\mathbb{R}$, $\mathbb{R}^m$ or $\mathbb{R}^{m\times m}$ by $\mathbb{X}$ and $\mathbb{C}$, $\mathbb{C}^m$ or $\mathbb{C}^{m\times m}$ by $\hat{\mathbb{X}}$.

Let's agree on the new notation

$$L^p(\square, \mathbb{X}) = \{f \in L^p_{loc}(\mathbb{R}^m, \mathbb{X}) : \ f \text{ is } \square\text{-periodic}\}.$$

and analogically for $\nu^2_{pot}(\square, \mathbb{X}), \nu^2_{sol}(\square, \mathbb{X})$.

When $p = 2$, $L^2(\square, \mathbb{X})$ forms a Hilbert space with the scalar product

$$\big(u,\, v\big)_{L^2(\square,\mathbb{X})} = \frac{1}{|\square|}\int_\square \big(u(x), v(x)\big)_{\mathbb{X}}\,\mathrm{d}x.$$

We are introducing the space of weakly s-differentiable periodic functions which will be later needed for estimating the error of our numerical solution

$$W^{s,p}(\square, \mathbb{R}^m) = \{f \in W^{s,p}_{loc}(\mathbb{R}^m, \mathbb{R}^m) : \ f \text{ is } \square\text{-periodic}\}, \quad s \geq 1,\, p \geq 1$$
$$H^s(\square, \mathbb{R}^m) = W^{s,p}(\square, \mathbb{R}^m) \text{ for } p = 2.$$

3.1.1 Fourier transform

We will recall some well-known facts from the theory of Fourier transform.

Let's consider functions

$$\varphi_k(x) = exp\big(2\pi i \frac{k\cdot x}{\rho}\big), \quad x \in \square,\ k \in \mathbb{Z}^m.$$

The set of functions $\{\varphi_k(x)\}_{k\in\mathbb{Z}^m}$ forms an orthonormal basis of the space $L^2(\square, \mathbb{C})$. To deal with only real-valued functions we have to ensure the condition

$$c_k = \overline{c_{-k}}, \quad \forall k \in \mathbb{Z}^m.$$

The space $L^2(\square, \mathbb{R})$ can be equivalently expressed as

$$L^2(\square, \mathbb{R}) = \Big\{ \sum_{k\in\mathbb{Z}^m} c_k\, \varphi_k,\ c_k \in \mathbb{R},\ \sum_{k\in\mathbb{Z}^m} \|c_k\|^2 < \infty,\ c_k = \overline{c_{-k}} \Big\}.$$

The Fourier transform of $f \in L^2(\square, \mathbb{X})$ is given by

$$\widehat{f}(k) = \overline{\widehat{f}(-k)} = \frac{1}{|\square|} \int_{\square} f(x)\, \varphi_{-k}(x)\, \mathrm{d}x \ \in \hat{\mathbb{X}} \quad k \in \mathbb{Z}^m.$$

Every function $f \in L^2(\square, \mathbb{X})$ can be expressed in the form

$$f(x) = \sum_{k\in\mathbb{Z}^m} \widehat{f}(k)\varphi_k(x), \quad x \in \square$$

and for $f, g \in L^2(\square, \mathbb{X})$ their scalar product becomes

$$\big(f,\, g\big)_{L^2(\square,\mathbb{X})} = \sum_{k\in\mathbb{Z}^m} \big(\widehat{f}(k),\, \widehat{g}(k)\big)_{\hat{\mathbb{X}}}. \tag{3.1}$$

The Fourier transform enables an easy differentiation

$$\frac{\partial f}{\partial x_j} = \sum_{k\in\mathbb{Z}^m} \widehat{f}(k)\, k_j\, \varphi_k(x)\, \frac{2i\pi}{\rho}. \tag{3.2}$$

3.1.2 Auxiliary operator

As explained in Chapter 1, the space $L^2(\square, \mathbb{R}^m)$ admits an orthogonal decomposition in the form

$$L^2(\square, \mathbb{R}^m) = \mathbb{R}^m \oplus \nu^2_{pot}(\square, \mathbb{R}^m) \oplus \nu^2_{sol}(\square, \mathbb{R}^m).$$

Since the solution of our problem (3.3) and the test functions are sought in the space $\nu^2_{pot}(\square, \mathbb{R}^m)$, we will need a more convenient way of how to characterize this space.

Definition 3.1. *We define an operator* $\mathcal{G} : L^2(\square, \mathbb{R}^m) \to \nu^2_{pot}(\square, \mathbb{R}^m)$ *as*

$$\mathcal{G}[f](x) = \int_{\square} \Gamma(x-y) f(y)\, \mathrm{d}y = \sum_{k\in\mathbb{Z}^m} \widehat{\Gamma}(k)\widehat{f}(k)\varphi_k(x),$$

where Γ *is expressed by means of its Fourier transform*

$$\widehat{\Gamma}(k) = \begin{cases} 0 \otimes 0, & \text{for } k = 0, \\ \frac{k\otimes k}{k\cdot k}, & \text{for } k \in \mathbb{Z}^m \setminus \{0\}. \end{cases}$$

In the following lemma we prove the correctness of the Definition 3.1.

Lemma 3.1. *The operator* $\mathcal{G}$ *defined in the Definition 3.1 satisfies following*

1. *$\mathcal{G}$ is an operator $L^2(\square, \mathbb{R}^m) \to L^2(\square, \mathbb{R}^m)$.*
2. *$\mathcal{G}$ is a projection.*
3. *$\mathcal{G}$ is a projection to $\nu^2_{pot}(\square, \mathbb{R}^m)$.*
4. *$\mathcal{G}$ is a projection onto $\nu^2_{pot}(\square, \mathbb{R}^m)$.*
5. *$\mathcal{G}$ is self-adjoint.*

Proof.

1. Let's show that $\mathcal{G}$ maps a real-valued input to a real-valued output. This property is equivalent to the following for all $k \in \mathbb{Z}^m$:

$$\overline{\widehat{\mathcal{G}[f]}(k)} = \widehat{\Gamma}(k)\overline{\widehat{f}(k)} \overset{*}{=} \widehat{\Gamma}(-k)\widehat{f}(-k) = \widehat{\mathcal{G}[f]}(-k),$$

   where $*$ holds thanks to the fact that f is real-valued.

2. We can easily see that

$$\begin{aligned}\mathcal{G}[\mathcal{G}[f]] &= \sum_{k\in\mathbb{Z}^m} \widehat{\Gamma}(k)\widehat{\mathcal{G}[f]}(k)\varphi_k \\ &= \sum_{k\in\mathbb{Z}^m} \widehat{\Gamma}(k)\widehat{\Gamma}(k)\widehat{f}(k)\varphi_k \\ &= \sum_{k\in\mathbb{Z}^m} \widehat{\Gamma}(k)\widehat{f}(k)\varphi_k = \mathcal{G}[f],\end{aligned}$$

   for all $f \in L^2(\square, \mathbb{R}^m)$.

3. Since the domain $\square$ is simply connected, the property of $\mathcal{G}[f]$ being potential is equivalent to the property of $\mathcal{G}[f]$ being irrotational, i.e. $\nabla \times \mathcal{G}[f] = 0$. Thanks to (3.2) we have

$$\nabla \times \mathcal{G}[f] = \frac{2i\pi}{\rho} \sum_{k\in\mathbb{Z}^m} k \times \frac{kk^\top}{k^\top k}\widehat{f}(k)\varphi_k.$$

   The operation $\frac{kk^\top}{k^\top k}\widehat{f}(k)$ projects $\widehat{f}(k)$ onto the space spanned by k. Therefore, $k \times \frac{kk^\top}{k^\top k}\widehat{f}(k) = 0$ and so $\nabla \times \mathcal{G}[f] = 0$.

4. We will show $\mathcal{G}[f] = f$ for all $f \in \nu^2_{pot}(\square, \mathbb{R}^m)$. By definition we know that there is a function $u \in H^1(\square, \mathbb{R})$ s.t. $f = \nabla u$. Then we have

$$\mathcal{G}[f] = \mathcal{G}[\nabla u] = \frac{2i\pi}{\rho} \sum_{k\in\mathbb{Z}^m} \frac{kk^\top}{k^\top k} k\widehat{u}(k)\varphi_k = \frac{2i\pi}{\rho} \sum_{k\in\mathbb{Z}^m} k\widehat{u}(k)\varphi_k = f.$$

5. We need to prove that $\big(\mathcal{G}[f],\, g\big)_{L^2(\square,\mathbb{R}^m)} = \big(f,\, \mathcal{G}[g]\big)_{L^2(\square,\mathbb{R}^m)}$ for all

$f, g \in L^2(\square, \mathbb{R}^m)$. By means of 3.1 we derive:

$$\begin{aligned}
\big(\mathcal{G}[f],\, g\big)_{L^2(\square,\mathbb{R}^m)} &= \sum_{k\in\mathbb{Z}^m} \big(\widehat{\mathcal{G}[f]}(k),\, \widehat{g}(k)\big)_{\mathbb{C}^m} \\
&= \sum_{k\in\mathbb{Z}^m} \big(\widehat{\Gamma}(k)\widehat{f}(k),\, \widehat{g}(k)\big)_{\mathbb{C}^m} \\
&\overset{*}{=} \sum_{k\in\mathbb{Z}^m} \big(\widehat{f}(k),\, \widehat{\Gamma}(k)\widehat{g}(k)\big)_{\mathbb{C}^m} \\
&= \sum_{k\in\mathbb{Z}^m} \big(\widehat{f}(k),\, \widehat{\mathcal{G}[g]}(k)\big)_{\mathbb{C}^m} = \big(f,\, \mathcal{G}[g]\big)_{L^2(\square,\mathbb{R}^m)},
\end{aligned}$$

where $*$ holds thanks to the fact that $\widehat{\Gamma}(k)$ is a symmetric matrix.

This completes the proof.

□

Remark. A projection is orthogonal if and only if it is self-adjoint. Therefore altogether we can state that $\mathcal{G}$ is an orthogonal projection from $L^2(\square, \mathbb{R}^m)$ onto the space $\nu^2_{pot}(\square, \mathbb{R}^m)$.

3.2 Problem setting

Let us recall the problem we are dealing with.

Let A be a $\square$-periodic matrix $A(x) = \{a_{ij}(x)\}_{i,j=1}^m$ satisfying

- $A \in L^\infty(\square, \mathbb{R}^{m\times m})$
- $c_A \|\lambda\|^2_{\mathbb{R}^m} \le \big(A(x)\lambda,\, \lambda\big)_{\mathbb{R}^m} \le C_A \|\lambda\|^2_{\mathbb{R}^m}$, a.e. in $\square$, $\forall \lambda \in \mathbb{R}^m$, where $0 < c_A \le C_A < \infty$.

We are trying to find a vector field $v \in L^2_{pot}(\square, \mathbb{R}^m)$ s.t.

$$\begin{aligned}
\operatorname{div}(A\, v) &= 0 \quad \text{in } \square \\
\langle v \rangle &= \lambda \in \mathbb{R}^m.
\end{aligned} \tag{3.3}$$

According to Lemma 1.3 there is a unique weak solution to the problem (3.3), i.e. there is a unique $u \in H^1(\square, \mathbb{R})$ s.t.

$$\begin{aligned}
v &= \lambda + \nabla u \\
\big(A\nabla u,\, \nabla w\big)_{L^2(\square,\mathbb{R}^m)} &= -\big(A\lambda,\, \nabla w\big)_{L^2(\square,\mathbb{R}^m)}, \quad \forall w \in H^1(\square, \mathbb{R}).
\end{aligned} \tag{3.4}$$

Then the sought effective matrix can be obtained by

$$A^0\lambda = \langle A\, v \rangle.$$

As we can see, there is no need to explicitly compute $u \in H^1(\square, \mathbb{R})$ since later we are only working with $v \in L^2_{pot}(\square, \mathbb{R}^m)$.

To that purpose, let us define a bilinear form $a : L^2(\square, \mathbb{R}^m) \times L^2(\square, \mathbb{R}^m) \to \mathbb{R}$ and a linear form $l : L^2(\square, \mathbb{R}^m) \to \mathbb{R}$ associated to the problem (3.3) as

$$\begin{aligned} a(v, w) &= \big(Av,\, w\big)_{L^2(\square, \mathbb{R}^m)} \\ l(w) &= -\big(A\lambda,\, w\big)_{L^2(\square, \mathbb{R}^m)}. \end{aligned}$$

Then an equivalent weak formulation of the problem (3.3) states

$$\begin{aligned} &\text{Find } v \in \nu^2_{pot}(\square, \mathbb{R}^m) \text{ s.t.} \\ &a(v, w) = l(w), \quad \forall w \in \nu^2_{pot}(\square, \mathbb{R}^m). \end{aligned} \tag{3.5}$$

Thanks to the assumptions (3.2) on matrix A, the bilinear form a and the linear form l meet the standard conditions of coercivity and boundedness, and so prove the existence of a unique solution v.

Lemma 3.1 justifies that the equation in (3.5) can be equivalently expressed as

$$a\big(\, v,\, \mathcal{G}[w]\,\big) = l\big(\, \mathcal{G}[w]\,\big), \quad \forall w \in L^2(\square, \mathbb{R}^m).$$

3.3 Discretization

The goal of this section is to introduce an appropriate finite-dimensional space for computing the approximate solution of (3.5).

Definition 3.2. *We define the space of m-dimensional real-valued trigonometric polynomials of order N as*

$$\mathcal{T}_N(\square, \mathbb{R}^m) = \Big\{ \sum_{k \in \mathbb{Z}^m_N} c_k\, \varphi_k,\ c_k \in \mathcal{C}^m,\ c_k = \overline{c_{-k}} \Big\} \subset L^2(\square, \mathbb{R}^m),$$

where $N = [N_1, \ldots, \mathbb{N}_m]$ is a discretization parameter and

$$\mathbb{Z}^m_N = \Big\{ k \in \mathbb{Z}^m : -\frac{N_i}{2} \le k_i < \frac{N_i}{2},\ i = 1, \ldots, m \Big\}.$$

The condition $c_k = \overline{c_{-k}}$ ensures that functions from $\mathcal{T}_N$ are real-valued. For this purpose we will always assume N_i odd for all $i = 1, \ldots, m$. We shall notice that $\mathcal{T}_N(\square, \mathbb{R}^m) \subset C^\infty(\square, \mathbb{R}^m)$. By $|N|$ we will denote the value $N_1 \cdot \cdots \cdot N_m$ and let h denote

$$h = \max_i h_i, \quad h_i = \frac{\rho}{N_i}, \quad i = 1, \ldots, m.$$

Definition 3.3. *We define the truncation operator* $\mathcal{P}_N : L^2(\square, \mathbb{R}^m) \to \mathcal{T}_N(\square, \mathbb{R}^m)$ *as*

$$\mathcal{P}_N[f](x) = \sum_{k \in \mathbb{Z}_N^m} \widehat{f}(k)\varphi_k(x), \quad x \in \square.$$

Lemma 3.2. *The operator* $\mathcal{P}_N : L^2(\square, \mathbb{R}^m) \to \mathcal{T}_N(\square, \mathbb{R}^m)$ *is an orthogonal projection in the scalar product of* $L^2(\square, \mathbb{R}^m)$.

Proof. Clearly, $\mathcal{P}_N[\mathcal{P}_N[f]] = \mathcal{P}_N[f]$ for all $f \in L^2(\square, \mathbb{R}^m)$.

Next we compute

$$\big(f - \mathcal{P}_N[f],\, g\big)_{L^2(\square,\mathbb{R}^m)} = \big(\sum_{k \in \mathbb{Z}^m \setminus \mathbb{Z}_N^m} \widehat{f}(k)\varphi_k,\, \sum_{k \in \mathbb{Z}_N^m} \widehat{g}(k)\varphi_k\big)_{L^2(\square,\mathbb{R}^m)} = 0,$$

for all $g \in \mathcal{T}_N(\square, \mathbb{R}^m)$ as $\{\varphi_k\}_{k \in \mathbb{Z}^m}$ form an orthonormal basis of the space $L^2(\square, \mathbb{R})$.

□

Lemma 3.3. *For* $f \in L^2(\square, \mathbb{R}^m)$

$$\lim_{N \to \infty} \|f - \mathcal{P}_N[f]\|_{L^2(\square,\mathbb{R}^m)} \to 0.$$

If in addition $f \in H^s(\square, \mathbb{R}^m)$ *with* $s > r \geq 0$ *we get*

$$\|f - \mathcal{P}_N[f]\|_{H^r(\square,\mathbb{R}^m)} \leq C_1 h^{s-r} \, \|f\|_{H^s(\square,\mathbb{R}^m)}.$$

Proof. Proof can be found in [Vondřejc et al., 2014].

□

An essential advantage of trigonometrical polynomials is that they allow us to construct structure-preserving conforming finite-dimensional approximations of spaces $\nu^2_{pot}(\square, \mathbb{R}^m)$, $\nu^2_{sol}(\square, \mathbb{R}^m)$ in a transparent way. We simply set

$$\nu^2_{pot,N}(\square, \mathbb{R}^m) = \nu^2_{pot}(\square, \mathbb{R}^m) \cap \mathcal{T}_N(\square, \mathbb{R}^m) = \mathcal{P}_N[\nu^2_{pot}(\square, \mathbb{R}^m)]$$
$$\nu^2_{sol,N}(\square, \mathbb{R}^m) = \nu^2_{sol}(\square, \mathbb{R}^m) \cap \mathcal{T}_N(\square, \mathbb{R}^m) = \mathcal{P}_N[\nu^2_{sol}(\square, \mathbb{R}^m)].$$

Then $\nu^2_{pot,N}(\square, \mathbb{R}^m)$, and $\nu^2_{sol,N}(\square, \mathbb{R}^m)$ collect all of the zero-mean, potential and divergence-free trigonometric polynomials with the degree N, respectively.

Moreover, an analogous variant of the Weyl decomposition holds:

$$\mathcal{T}_N(\square, \mathbb{R}^m) = \mathbb{R}^m \oplus \nu^2_{pot,N}(\square, \mathbb{R}^m) \oplus \nu^2_{sol,N}(\square, \mathbb{R}^m).$$

Remark. The space $\nu^2_{pot,N}(\square, \mathbb{R}^m)$ can be also very easily expressed by means of our auxiliary operator $\mathcal{G}$ as

$$\nu^2_{pot,N}(\square, \mathbb{R}^m) = \mathcal{G}[\mathcal{T}_N(\square, \mathbb{R}^m)].$$

3.4 Galerkin approximation

As we specified the finite-dimensional spaces we will work with in the last section, we are proceeding to discretize the problem by a Galerkin method in a standard way.

Definition 3.4. *A vector field v_N is a weak solution of the Galerkin approximation of the original problem (3.5) if it satisfies:*

$$\begin{aligned} &v_N \in \nu^2_{pot,N}(\square, \mathbb{R}^m) \\ &a(v_N, w) = l(w), \quad \forall w \in \nu^2_{pot,N}(\square, \mathbb{R}^m). \end{aligned} \tag{3.6}$$

Lemma 3.4. *Let A satisfy the assumptions stated in (3.2). Let v be the solution of the original problem (3.5). Then there is a unique weak solution v_N as defined in Definition 3.4 satisfying*

$$\lim_{N\to\infty} \|v - v_N\|_{L^2(\square,\mathbb{R}^m)} = 0.$$

If, in addition, $v \in H^s(\square, \mathbb{R}^m)$ for $s > 0$, we get

$$\|v - v_N\|_{L^2(\square,\mathbb{R}^m)} \leq C\, h^s \, \|v\|_{H^s(\square\mathbb{R}^m)}\,.$$

Proof. The existence and uniqueness of the solution come from the Lax-Milgram theorem as the estimates (3.2) still hold. From Céa lemma we obtain

$$\begin{aligned} \|v - v_N\|_{L^2(\square,\mathbb{R}^m)} &\leq \frac{C_A}{c_A} \inf_{w_N \in \nu^2_{pot,N}(\square,\mathbb{R}^m)} \|v - v_N\|_{L^2(\square,\mathbb{R}^m)} \\ &\leq \frac{C_A}{c_A} \|v - \mathcal{P}_N[v]\|_{L^2(\square,\mathbb{R}^m)}\,. \end{aligned}$$

The rest of the statement is now a direct consequence of the estimates from Lemma 3.3.

□

After all the preparation we have provided, there is a new task arising: choosing an appropriate basis of the space $\mathcal{T}_N(\square, \mathbb{R}^m)$ in order to obtain a matrix with convenient properties. The straightforward suggestion would be to take the functions $\{\varphi_k\}_{k\in\mathbb{Z}^m_N}$. This set of basis functions, however, brings along a set of unfortunate consequences:

- no obvious way of an exact integration
- the resulting matrix does not have a sparse representation.

When not ensuring an exact integration, the estimations from Lemma 3.4 are not valid anymore.

3.4.1 Fundamental trigonometric polynomials

To each $k \in \mathbb{Z}_N^m$ we shall assign a point x_N^k

$$x_N^k = [k_1 h_1, \ldots, k_m h_m], \quad h_i = \frac{\rho}{N_i}. \tag{3.7}$$

In what follows we will use some concepts from the theory of the discrete Fourier transform (DFT). For $f \in \mathcal{T}_N(\square, \mathbb{R}^m)$ the inverse and the forward DFT are given by

- the inverse DFT
$$f(x_N^k) = \sum_{l \in \mathbb{Z}_N^m} \widehat{f}(l) \omega_N^{lk},$$
where $\omega_N^{kl} = \varphi_l(x_N^k), \quad l, k \in \mathbb{Z}_N^m,$
- the forward DFT
$$\widehat{f}(k) = \frac{1}{|N|} \sum_{l \in \mathbb{Z}_N^m} f(x_N^l) \omega_N^{-kl}.$$

Definition 3.5. *We define the set of fundamental trigonometric polynomials* $\{\varphi_{N,l}(x)\}_{l \in \mathbb{Z}_N^m}$ *as*

$$\varphi_{N,l}(x) = \frac{1}{|N|} \sum_{k \in \mathbb{Z}_N^m} \omega_N^{-kl} \varphi_k(x).$$

Corollary. The fundamental trigonometric polynomials satisfy

$$\varphi_{N,l}(x_N^k) = \delta_{lk}, \qquad \big(\varphi_{N,k}, \varphi_{N,l}\big)_{L^2(\square,\mathbb{R})} = \frac{1}{|N|} \delta_{kl}.$$

From the following computations we can see that every trigonometric polynomial can be expressed as a linear combination of fundamental trigonometric polynomials.

Let $f \in \mathcal{T}_N(\square, \mathbb{R}^m)$. Then

$$\begin{aligned} f(x) &= \sum_{l \in \mathbb{Z}_N^m} \widehat{f}(l)\, \varphi_l(x) = \\ &\overset{DFT}{=} \sum_{l \in \mathbb{Z}_N^m} \Big(\frac{1}{|N|} \sum_{k \in \mathbb{Z}_N^m} f(x_N^k) \omega_N^{-lk} \Big) \varphi_l(x) = \\ &= \sum_{k \in \mathbb{Z}_N^m} \Big(\frac{1}{|N|} \sum_{l \in \mathbb{Z}_N^m} \omega_N^{-lk} \varphi_l(x) \Big) f(x_N^k) = \\ &= \sum_{k \in \mathbb{Z}_N^m} \varphi_{N,k}(x) f(x_N^k). \end{aligned}$$

Thanks to the properties from the previous corollary there is an easy way of computing the $L^2(\square, \mathbb{R}^m)$ scalar product of two functions from $\mathcal{T}_N(\square, \mathbb{R}^m)$.

Corollary. Let $f,\, g \in \mathcal{T}_N(\square, \mathbb{R}^m)$. Then the $L^2(\square, \mathbb{R}^m)$ scalar product satisfies

$$\big(f,\, g\big)_{L^2(\square, \mathbb{R}^m)} = \frac{1}{|N|} \sum_{k \in \mathbb{Z}_N^m} \big(f(x_N^k),\, g(x_N^k)\big)_{\mathbb{R}^m}.$$

Definition 3.6. *We define the trigonometric interpolation operator* $Q_N : C(\square, \mathbb{R}^m) \to \mathcal{T}_N(\square, \mathbb{R}^m)$ *as*

$$Q_N[f](x) = \sum_{k \in \mathbb{Z}_N^m} f(x_N^k) \varphi_{N,k}(x).$$

This operator is a projection, but not an orthogonal one. It does have some nice approximation properties.

Lemma 3.5. *Let* $f \in H^s(\square, \mathbb{R}^m), \quad s > \frac{m}{2}$. *Then*

$$\|f - Q_N[f]\|_{H^r(\square, \mathbb{R}^m)} \leq c_{r,s} h^{s-r} \|f\|_{H^s(\square, \mathbb{R}^m)}.$$

Proof. The proof can be found in [Vondřejc et al., 2014].

□

Thanks to all of these properties the fundamental trigonometric polynomials become an appropriate basis of the $\mathcal{T}_N(\square, \mathbb{R}^m)$ for the Galerkin method. The coefficients corresponding to this basis are function values at the grid points $x_N^k,\ k \in \mathbb{Z}_N^m$. Therefore these coefficients will become the sought solution of the correspondent algebraic system.

3.4.2 Galerkin approximation with numerical integration

After choosing the finite-dimensional approximation space and its basis functions we shall evaluate the equation introduced in the Definition 3.4. Numerically evaluating an $L^2(\square, \mathbb{R}^m)$-scalar product means numerically integrating a product of two functions.

As mentioned before, for $f, g \in \mathcal{T}_N(\square, \mathbb{R}^m)$ we can exactly evaluate their $L^2(\square, \mathbb{R}^m)$-scalar product as

$$\big(f,\, g\big)_{L^2(\square, \mathbb{R}^m)} = \frac{1}{|N|} \sum_{k \in \mathbb{Z}_N^m} \big(f(x_N^k),\, g(x_N^k)\big)_{\mathbb{R}^m}.$$

As for $f, g \in C(\square, \mathbb{R}^m)$, we will approximate the integral using the trapezoidal rule:

$$\int_{\square} \big(f(x),\, g(x)\big)_{\mathbb{R}^m} \mathrm{d}x \approx \frac{|\square|}{|N|} \sum_{k \in \mathbb{Z}_N^m} \big(f(x_N^k),\, g(x_N^k)\big)_{\mathbb{R}^m}.$$

Therefore their approximated $L^2(\square, \mathbb{R}^m)$-scalar product then becomes

$$\big(f,\, g\big)_{L^2(\square,\mathbb{R}^m)} \approx \frac{1}{|N|} \sum_{k\in\mathbb{Z}_N^m} \big(f(x_N^k),\, g(x_N^k)\big)_{\mathbb{R}^m} = \big(Q_N[f],\, Q_N[g]\big)_{L^2(\square,\mathbb{R}^m)}. \quad (3.8)$$

This results in parameter-dependent forms $a_N : \mathcal{T}_N(\square, \mathbb{R}^m) \times \mathcal{T}_N(\square, \mathbb{R}^m) \to \mathbb{R}$ and $l_N : \mathcal{T}_N(\square, \mathbb{R}^m) \to \mathbb{R}$ given by

$$\begin{aligned} a_N(v_N,\, w_N) &= \big(Q_N[Av_N],\, w_N\big)_{L^2(\square,\mathbb{R}^m)} \\ l_N(w_N) &= -\big(Q_N[A\lambda],\, w_N\big)_{L^2(\square,\mathbb{R}^m)}. \end{aligned} \quad (3.9)$$

Definition 3.7. *A vector field v_N is a weak solution of the Galerkin approximation of the original problem (3.5) with numerical integration if it satisfies:*

$$\begin{aligned} &v_N \in \nu^2_{pot,N}(\square, \mathbb{R}^m) \\ &a_N(v_N, w_N) = l(w_N), \quad \forall w_N \in \nu^2_{pot,N}(\square, \mathbb{R}^m). \end{aligned}$$

Note that due to the involvement of the operator Q_N, the matrix needs to satisfy

$$A \in C(\square, \mathbb{R}^{m\times m}). \quad (3.10)$$

Lemma 3.6. *Under the assumption (3.10), there is a unique weak solution of the Galerkin approximation of the original problem (3.5) with numerical integration. If, in addition, $A \in W^{s,\infty}(\square, \mathbb{R}^{m\times m})$ with $s > \frac{m}{2}$, we obtain the following estimate*

$$\|v - v_N\|_{L^2(\square,\mathbb{R}^m)} \le c_\rho h^s \, \|A\|_{W^{s,\infty}(\square,\mathbb{R}^{m\times m})} \, \|v\|_{H^s(\square,\mathbb{R}^m)} \, . \quad (3.11)$$

Proof. Firstly, let us verify that the forms (3.9) meet the assumptions from the Lax-Milgram lemma.
For arbitrary $u_N, w_N \in \mathcal{T}_N(\square, \mathbb{R}^m)$ we obtain

$$\begin{aligned} a_N(w_N, w_N) &= \big(Q_N[Aw_N],\, w_N\big)_{L^2(\square,\mathbb{R}^m)} = \frac{1}{|N|} \sum_{k\in\mathbb{Z}_N^m} \big(A(x_N^k)w_N(x_N^k),\, w_N(x_N^k)\big)_{\mathbb{R}^m} \\ &\ge \frac{c_A}{|N|} \sum_{k\in\mathbb{Z}_N^m} \big(w_N(x_N^k),\, w_N(x_N^k)\big)_{\mathbb{R}^m} = c_A \, \|w_N\|^2_{L^2(\square,\mathbb{R}^m)} \end{aligned}$$

$$\begin{aligned} a_N(u_N, w_N) &= \big(Q_N[Au_N],\, w_N\big)_{L^2(\square,\mathbb{R}^m)} = \frac{1}{|N|} \sum_{k\in\mathbb{Z}_N^m} \big(A(x_N^k)u_N(x_N^k),\, w_N(x_N^k)\big)_{\mathbb{R}^m} \\ &\le \frac{C_A}{|N|} \sum_{k\in\mathbb{Z}_N^m} \big(u_N(x_N^k),\, w_N(x_N^k)\big)_{\mathbb{R}^m} = C_A \, \|u_N\|_{L^2(\square,\mathbb{R}^m)} \, \|w_N\|_{L^2(\square,\mathbb{R}^m)} \end{aligned}$$

The existence and the uniqueness of the weak solution as defined in the Definition 3.7 are consequences of the Lax-Milgram lemma.

By virtue of the second Strang lemma we can estimate the convergence rate:

$$\begin{aligned}\|v - v_N\|_{L^2(\square,\mathbb{R}^m)} \leq \frac{1}{c_A} \sup_{u_N \in \nu^2_{pot,N}(\square,\mathbb{R}^m)} \frac{|l(u_N) - l_N(u_N)|}{\|u_N\|_{L^2(\square,\mathbb{R}^m)}} \\ + \inf_{w_N \in \nu^2_{pot,N}(\square,\mathbb{R}^m)} \Big[(1 + \frac{C_A}{c_A}) \|v - w_N\|_{L^2(\square,\mathbb{R}^m)} \\ + \frac{1}{c_A} \sup_{u_N \in \nu^2_{pot,N}(\square,\mathbb{R}^m)} \frac{|a(w_N, u_N) - a_N(w_N, u_N)|}{\|u_N\|_{L^2(\square,\mathbb{R}^m)}} \Big].\end{aligned} \tag{3.12}$$

The differences between the forms hold

$$\begin{aligned}|l(u_N) - l_N(u_N)| &\leq \|A\lambda - Q_N[A\lambda]\|_{L^2(\square,\mathbb{R}^m)} \|u_N\|_{L^2(\square,\mathbb{R}^m)} \\ |a(w_N, u_N) - a_N(w_N, u_N)| &\leq \|Aw_N - Q_N[Aw_N]\|_{L^2(\square,\mathbb{R}^m)} \|u_N\|_{L^2(\square,\mathbb{R}^m)}.\end{aligned}$$

Now set $w_N = \mathcal{P}_N[v]$. Thanks to the relations from Lemma 3.3 and Lemma 3.5 we obtain

$$\begin{aligned}\|v - \mathcal{P}_N[v]\|_{L^2(\square,\mathbb{R}^m)} &\leq Ch^s \|v\|_{H^2(\square,\mathbb{R}^m)}, \\ \|A\mathcal{P}_N[v] - Q_N[A\mathcal{P}_N[v]]\|_{L^2(\square,\mathbb{R}^m)} &\leq C_0 h^s \|A\mathcal{P}_N[v]\|_{H^s(\square,\mathbb{R}^m)} \\ &\leq C_0 h^s \|A\|_{W^{s,\infty}(\square,\mathbb{R}^{m\times m})} \|v\|_{H^s(\square,\mathbb{R}^m)}, \\ \|A\lambda - Q_N[A\lambda]\|_{L^2(\square,\mathbb{R}^m)} &\leq C_1 h^s \|A\lambda\|_{H^s(\square,\mathbb{R}^m)} \\ &\leq C_1 h^s \|A\|_{W^{s,\infty}(\square,\mathbb{R}^{m\times m})} \|\lambda\|_{\mathbb{R}^m}\end{aligned}$$

Filling these estimates into the (3.12) completes the proof. Let us note that the required regularity of the solution $v \in H^s(\square,\mathbb{R}^m)$ is justified by $A \in W^{s,\infty}(\square,\mathbb{R}^{m\times m})$.

□

3.5 Algebraic system

The present section is dedicated to the analysis of the fully discrete version of the Fourier-Galerkin method with numerical integration.

Let $k_1, \ldots, k_{|N|}$ be an arbitrarily ordered set $\mathbb{Z}^m_N$.

Definition 3.8. *We define an operator* $s_N : C(\square,\mathbb{R}^m) \to \mathbb{R}^{|N|m}$ *as*

$$s_N[u_N] = \begin{pmatrix} u_N(x_N^{k_1}) \\ \vdots \\ u_N(x_N^{k_{|N|}}) \end{pmatrix} \in \mathbb{R}^{|N|m}.$$

$$a_N(v_N,\, w_N) = \sum_{k\in\mathbb{Z}_N^m} \big(A(x_N^k)v_N(x_N^k),\, w_N(x_N^k)\big)_{\mathbb{R}^m} = \bar{w}_N^\top \bar{A}_N \bar{v}_N$$
$$l_N(w_N) = -\sum_{k\in\mathbb{Z}_N^m} \big(A(x_N^k)\lambda,\, w_N(x_N^k)\big)_{\mathbb{R}^m} = -\bar{w}_N^\top \bar{A}_N \bar{\lambda}_N,$$

where

$$\bar{v}_N = s_N[v_N] = \begin{pmatrix} v_N(x_N^{k_1}) \\ \vdots \\ v_N(x_N^{k_{|N|}}) \end{pmatrix} \in \mathbb{R}^{|N|m}$$

$$\bar{w}_N = s_N[w_N] = \begin{pmatrix} w_N(x_N^{k_1}) \\ \vdots \\ w_N(x_N^{k_{|N|}}) \end{pmatrix} \in \mathbb{R}^{|N|m}$$

$$\bar{A}_N = \begin{pmatrix} A(x_N^{k_1}) & 0 & \cdots & 0 \\ 0 & A(x_N^{k_2}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A(x_N^{k_{|N|}}) \end{pmatrix} \in \mathbb{R}^{|N|m\times|N|m}$$

$$\bar{\lambda}_N = s_N[\lambda] = \begin{pmatrix} \lambda \\ \vdots \\ \lambda \end{pmatrix} \in \mathbb{R}^{|N|m}$$

Definition 3.9. *We define a discrete analogue of the space* $\nu^2_{pot}(\square, \mathbb{R}^m)$ *as:*

$$\mathbb{V}^2_{pot,N} = \{s_N[w_N];\, w_N \in \nu^2_{pot}(\square, \mathbb{R}^m)\}.$$

The discrete version of the Fourier-Galerkin method with numerical integration is then formulated as

$$\begin{aligned} &\text{Find } \bar{v}_N \in \mathbb{V}^2_{pot,N} \\ &\big(\bar{A}_N\bar{v}_N,\, \bar{w}_N\big)_{\mathbb{R}^{|N|m}} = -\big(\bar{A}_N\bar{\lambda}_N,\, \bar{w}_N\big)_{\mathbb{R}^{|N|m}}, \quad \forall \bar{w}_N \in \mathbb{V}^2_{pot,N}. \end{aligned} \tag{3.13}$$

In the following we will try to conveniently characterize the space $\mathbb{V}^2_{pot,N}$. We will proceed similarly to the subsection 3.1.2.

Recall the definition of $\widehat{\Gamma}$:

$$\widehat{\Gamma}(k) = \begin{cases} 0 \otimes 0, & \text{for } k = 0, \\ \frac{k\otimes k}{k\cdot k}, & \text{for } k \in \mathbb{Z}^m \setminus \{0\}. \end{cases}$$

The following matrix $\bar{\widehat{\Gamma}}_N \in \mathbb{R}^{|N|m \times |N|m}$ can be used to project the whole space $\mathbb{R}^{|\mathbb{N}|m}$ to $\mathbb{V}^2_{pot,N}$

$$\bar{\widehat{\Gamma}}_N = \begin{pmatrix} \widehat{\Gamma}(k_1) & 0 & \cdots & 0 \\ 0 & \widehat{\Gamma}(k_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \widehat{\Gamma}(k_{|N|}) \end{pmatrix} \in \mathbb{R}^{|N|m \times |N|m}.$$

The projection operator is then obtained by transferring $\bar{\widehat{\Gamma}}_N$ to the real space by means of matrices F_N, F_N^{-1} implementing the forward and inverse discrete Fourier transform (recall the section 3.4.1)

$$F_N = \frac{1}{|N|} \begin{pmatrix} \omega_N^{-k_1 k_1} I_m & \omega_N^{-k_2 k_1} I_m & \omega_N^{-k_3 k_1} I_m & \dots & \omega_N^{-k_{|N|} k_1} I_m \\ \omega_N^{-k_1 k_2} I_m & \omega_N^{-k_2 k_2} I_m & \omega_N^{-k_3 k_2} I_m & \dots & \omega_N^{-k_{|N|} k_2} I_m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \omega_N^{-k_1 k_{|N|}} I_m & \omega_N^{-k_2 k_{|N|}} I_m & \omega_N^{-k_3 k_{|N|}} I_m & \dots & \omega_N^{-k_{|N|} k_{|N|}} I_m \end{pmatrix} \in \mathbb{C}^{|N|m \times |N|m}$$

$$F_N^{-1} = \begin{pmatrix} \omega_N^{k_1 k_1} I_m & \omega_N^{k_2 k_1} I_m & \omega_N^{k_3 k_1} I_m & \dots & \omega_N^{k_{|N|} k_1} I_m \\ \omega_N^{k_1 k_2} I_m & \omega_N^{k_2 k_2} I_m & \omega_N^{k_3 k_2} I_m & \dots & \omega_N^{k_{|N|} k_2} I_m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \omega_N^{k_1 k_{|N|}} I_m & \omega_N^{k_2 k_{|N|}} I_m & \omega_N^{k_3 k_{|N|}} I_m & \dots & \omega_N^{k_{|N|} k_{|N|}} I_m \end{pmatrix} \in \mathbb{C}^{|N|m \times |N|m}.$$

Definition 3.10. *We define the operator $G_N : \mathbb{R}^{|N|m} \to \mathbb{V}^2_{pot,N}$ as*

$$G_N = F_N^{-1} \bar{\widehat{\Gamma}}_N F_N. \tag{3.14}$$

Lemma 3.7. *Following statements hold:*

1. *G_N is an operator $\mathbb{R}^{|N|m} \to \mathbb{R}^{|N|m}$.*
2. *G_N is a projection.*
3. *G_N is a projection to $\mathbb{V}^2_{pot,N}$.*
4. *G_N is a projection onto $\mathbb{V}^2_{pot,N}$.*
5. *G_N is self-adjoint, i.e. G_N is a Hermitian matrix.*

Proof. The proof is a direct analogue to the proof of Lemma 3.1.

□

The discrete problem defined in (3.13) then becomes

$$\begin{aligned} &\text{Find } \bar{v}_N \in \mathbb{V}^2_{pot,N} \\ &\big(\bar{A}_N \bar{v}_N,\, G_N \bar{w}_N\big)_{\mathbb{R}^{|N|m}} = -\big(\bar{A}_N \bar{\lambda}_N,\, G_N \bar{w}_N\big)_{\mathbb{R}^{|N|m}}, \quad \forall \bar{w}_N \in \mathbb{R}^{|N|m} \end{aligned} \tag{3.15}$$

Using the projection properties of G_N, we proceed

$$(3.15) \iff \left(G_N\bar{A}_N\bar{v}_N,\, \bar{w}_N\right)_{\mathbb{R}^{|N|m}} = -\left(G_N\bar{A}_N\bar{\lambda}_N,\, \bar{w}_N\right)_{\mathbb{R}^{|N|m}}, \quad \forall \bar{w}_N \in \mathbb{R}^{|N|m}$$
$$\iff G_N\bar{A}_N\bar{v}_N = -G_N\bar{A}_N\bar{\lambda}_N. \tag{3.16}$$

The first suggestion for the solution of (3.16) could be $\bar{v}_N = -\bar{\lambda}_N$. We have to, however, keep in mind the condition $\bar{v}_N \in \mathbb{V}^2_{pot,N}$, which is for $-\bar{\lambda}_N$ not fulfilled.

The following Corollary states an important fact of the Fourier-Galerkin method, i.e. the spectrum of the final algebraic matrix does not depend on the discretization parameter.

Corollary. It holds

$$c_A \left\|x\right\|^2_{\mathbb{R}^{|N|m}} \leq \left(G_N\bar{A}_N x,\, x\right)_{\mathbb{R}^{|N|m}} \leq C_A \left\|x\right\|^2_{\mathbb{R}^{|N|m}}, \quad \forall x \in \mathbb{V}^2_{pot,N}.$$

Proof. Thanks to Lemma 3.7 we have

$$\left(G_N\bar{A}_N x,\, x\right)_{\mathbb{R}^{|N|m}} = \left(\bar{A}_N x,\, G_N x\right)_{\mathbb{R}^{|N|m}} = \left(\bar{A}_N x,\, x\right)_{\mathbb{R}^{|N|m}}.$$

The rest is a direct consequence of the assumptions in 3.2.

□

Remark. Let us notice the special structure of $G_N\bar{A}_N$. As we can see, $G_N\bar{A}_N = F_N^{-1}\widehat{\widehat{\Gamma}}_N F_N \bar{A}_N$, where $\widehat{\widehat{\Gamma}}_N, \bar{A}_N$ are block-diagonal, i.e. their matrix-vector multiplication is $O(|N|)$. F_N^{-1}, F_N have a special structure in such a way that their matrix-vector multiplication is $O(|N|\, log\, |N|)$. Altogether the operation of matrix-vector multiplication for $G_N\bar{A}_N$ is $O(|N|\, log\, |N|)$.

3.5.1 Solution of the linear system

Lemma 3.8. *For $A(x)$ symmetric for all $x \in \square$ the system (3.16) can be solved by the conjugate gradient algorithm for an arbitrary initial solution $\bar{v}^0_N \in \mathbb{V}^2_{pot,N}$.*

Proof. We define the i-th Krylov subspace as

$$\mathbb{K}_i = \text{span}\left\{r^0_N, G_N\bar{A}_N r^0_N, \ldots, (G_N\bar{A}_N)^{i-1} r^0_N\right\}, \tag{3.17}$$

where the vector r^0_N is the residual correspondent to the initial solution $\bar{v}^0_N$, i.e.

$$r^0_N = G_N\bar{A}_N(\bar{v}^0_N + \bar{\lambda}_N).$$

We will look for the solution in the form $\bar{v}_{N,i} = \bar{v}^0_N + x_i$ with $x_i \in \mathbb{K}_i$. Since the initial solution $\bar{v}^0_N \in \mathbb{V}^2_{pot,N}$ and also $\mathbb{K}_i \subset \mathbb{V}^2_{pot,N}$, we know that all of the iterates satisfy $\bar{v}_{N,i} \in \mathbb{V}^2_{pot,N}$ which was one of the required condition for the solution.

In each iteration we are solving a problem of an orthogonal projection on the subspace $\mathbb{K}_i$

$$\left(G_N\bar{A}_N x_i,\, w_N\right)_{\mathbb{R}^{|N|m}} = -\left(G_N\bar{A}_N(\bar{v}^0_N + \bar{\lambda}_N), w_N\right)_{\mathbb{R}^{|N|m}} \quad \forall w_N \in \mathbb{K}_i. \tag{3.18}$$

Since G_N is a self-adjoint, $\mathbb{V}^2_{pot,N}$-invariant operator (recall Lemma 3.7), (3.18) is equivalent to

$$\big(\bar{A}_N x_i,\, w_N\big)_{\mathbb{R}^{|N|m}} = -\big(\bar{A}_N(\bar{v}^0_N + \bar{\lambda}_N), w_N\big)_{\mathbb{R}^{|N|m}} \quad \forall w_N \in \mathbb{K}_i. \tag{3.19}$$

Within the subspace $\mathbb{K}_i$ $\bar{A}_N$ is a positive-definite, symmetric matrix, therefore (3.19) represent convergent iterations of the conjugate gradient algorithm.

□

The Fourier-Galerkin method is a suitable numerical method for solving the problem of homogenization if the required assumptions are fulfilled. To obtain the derived rate of convergence we need the matrix A to be $A \in W^{s,\infty}(\square, \mathbb{R}^{m\times m})$, which might be limiting. However, in many applications the coefficient matrix can be accurately approximated by $A \in W^{s,\infty}(\square, \mathbb{R}^{m\times m})$.

The method itself brings along many advantages:

- The operation of matrix-vector multiplication for the final algebraic matrix $G_N\bar{A}_N$ is $O(|N|\, log\, |N|)$.
- The condition number of the final algebraic matrix $G_N\bar{A}_N$ does not depend on the discretization.
- The method benefits from the fact that we are only interested in the quantity $v = \nabla u$ and not u.
- The method is suitable for regular domains which works in our case since our domain is $[-\frac{\rho}{2}; \frac{\rho}{2}]^m$.

4. Stochastic collocation method

As in Chapter 2, we are concerned with a problem, for a. e. $\omega \in \Omega$

$$\begin{aligned} -\text{div}\left(A(x,\omega)\nabla u(x,\omega)\right) &= f(x) \quad \text{in } Q \\ u(x,\omega) &= 0 \quad \text{on } \partial Q, \end{aligned}$$

where the matrix $A(x,\omega)$ describes the material properties of our heterogeneous material.

We are assuming that $A(x,\omega)$ can be expressed as $A(x,\omega) = \mathcal{A}(T(x)\omega)$ where $T(x)$ is an m-dimensional ergodic, dynamical system.

The problem of homogenization can be approximated by the periodic model (2.3.1), thanks to which we obtain a periodic medium described by

$$A^{\rho}_{per}(x,\omega) = \mathcal{A}(T(x \, mod \, S_{\rho})\omega). \tag{4.1}$$

For ω fixed we come to a deterministic problem that can be solved by the approach from Chapter 1. We obtain $\tilde{A}^{\rho}$, which is still dependent on the variable ω, because by the periodization (4.1) we lost the ergodicity. The dependence on ω will be denoted by $\tilde{A}^{\rho}_{\omega}$.

In Chapter 2 we were able to show that

$$\lim_{\rho\to\infty} \widetilde{A}^{\rho}_{\omega} = A^0 \quad \text{a.s. in } \Omega, \tag{4.2}$$

which directly leads to the algorithm, pick any $\omega \in \Omega$, compute $\tilde{A}^{\rho}_{\omega}$ for increasing ρ and watch where it converges. This procedure, however, faces several problems:

- it might be hard or even impossible to find the dynamical system $T(x)$ and the domain Ω,
- which ω to choose, since we have just the almost sure convergence.

Convergence almost surely by definition means

$$\mu\big(\omega \in \Omega | \lim_{\rho\to\infty} \widetilde{A}^{\rho}_{\omega} = A^0\big) = 1.$$

The use of the numerical methods discussed in the rest of this chapter is justified by the following lemma.

Lemma 4.1. *Let $X, Y, X_n : \Omega \to \mathbb{R}$, $n \in \mathbb{N}$ be random variables that satisfy*

$$\begin{aligned} &\lim_{n\to\infty} X_n = X, \quad \textit{a.s.} \\ &|X_n| < Y \\ &E(Y) < \infty. \end{aligned}$$

Then X_n converge to X in mean i.e. $\lim_{n\to\infty} E(|X_n - X|) = 0$ *which implies* $\lim_{n\to\infty} E(X_n) = E(X)$.

A^0 is a constant matrix, therefore its expectation value $E(A^0) = A^0$. After applying Lemma 4.1 we can see that for each ρ we are trying to obtain $E(\widetilde{A}^\rho)$

$$E(\widetilde{A}^\rho) = \int_\Omega \widetilde{A}^\rho_\omega \, \mathrm{d}\mu. \tag{4.3}$$

In what follows we will denote S_ρ by $\square$.

4.1 Problem Setting

As explained in Chapter 2, the problem of finding $\widetilde{A}^\rho_\omega$ for almost all $\omega \in \Omega$ is transformed to the following problem:

For a fixed $\lambda \in \mathbb{R}^m$ find a random function $v : \Omega \times \overline{\square} \to \mathbb{R}^m$ s.t. μ-almost everywhere in Ω the following equation holds:

$$\begin{gathered} -\mathrm{div}\Big(A(\cdot,\omega)\big(v(\cdot,\omega) + \lambda\big)\Big) = 0 \quad \text{in } \square \\ v(\cdot,\omega) \in \nu^2_{pot}(\square). \end{gathered} \tag{4.4}$$

The main sources for this chapter were [Babuška et al., 2007], [Nobile et al., 2008].

4.1.1 Assumptions

We need to state the following assumptions:

1. $A(x,\omega)$ is uniformly bounded from below, i.e. there exists $a_{min} > 0$ s.t.

$$\mu(\omega \in \Omega : \ \xi \cdot A(x,\omega)\xi > a_{min} \ \ \forall x \in \overline{\square}, \ \forall \xi \in \mathbb{R}^m) = 1. \tag{4.5}$$

2. $a_{i,j}(x,\omega)$ is square integrable w.r.t μ for all $i,j = 1,\ldots,m$, i.e.,

$$\int_\square E\big(a^2_{ij}\big) \, \mathrm{d}x < \infty. \tag{4.6}$$

3. Finite number of random variables:

$$A(x,\omega) = A(x, Y_1(\omega), \ldots, Y_M(\omega)) \text{ on } \Omega \times \overline{\square}, \tag{4.7}$$

where $M \in \mathbb{N}_+$ and $\{Y_n\}_{n=1}^M$ are real-valued random variables with zero mean value and unit variance.

To characterize an appropriate space for v dependent on both spatial and random variables we introduce the following Hilbert spaces.

Definition 4.1. *We define the space V_μ as*

$$\begin{aligned} V_\mu &= L^2_\mu(\Omega) \otimes \nu^2_{pot}(\square), \quad \textit{with the norm} \\ \|v\|^2_{V_\mu} &= \frac{1}{\square} \int_\square \int_\Omega |v|^2 \, \mathrm{d}\mu \, \mathrm{d}x = \frac{1}{\square} \int_\square E(|v|^2) \, \mathrm{d}x \end{aligned} \tag{4.8}$$

and the space $V_{\mu,A}$ as

$$\begin{aligned} V_{\mu,A} &= \{v \in V_\mu : \ \frac{1}{\square} \int_\square E(v \cdot Av) \, \mathrm{d}x < \infty\}, \quad \textit{with the norm} \\ \|v\|^2_{\mu,A} &= \frac{1}{\square} \int_\square E(v \cdot Av) \, \mathrm{d}x. \end{aligned} \tag{4.9}$$

Note that in this section, for the sake of brevity, we are no longer using the notation $\nu^2_{pot}(\square, \mathbb{R}^m)$, only $\nu^2_{pot}(\square)$, like in Chapter 1.

Remark. Under the assumptions (4.1.1), the space $V_{\mu,A}$ is continuously embedded in V_μ and

$$\|v\|_\mu \leq \frac{1}{\sqrt{a_{min}}} \|v\|_{\mu,A} . \tag{4.10}$$

The weak formulation of the problem (4.4) states:

$$\begin{aligned} &\text{Find } v \in V_{\mu,A} \text{ s.t.} \\ &\left\langle E(w \cdot Av) \right\rangle = - \left\langle E(w \cdot A\lambda) \right\rangle \quad \forall w \in V_{\mu,A} \end{aligned} \tag{4.11}$$

and the following lemma summarizes the well-posedness of problem (4.11)

Lemma 4.2. *Under the assumptions (4.5) and (4.6), the problem (4.11) has a unique weak solution $v \in V_{\mu,A}$ which satisfies the estimate*

$$\|v\|_{V_\mu} \leq \frac{1}{a_{min}} \|A\lambda\|_{V_\mu} . \tag{4.12}$$

Proof. The lemma is a direct consequence of the Lax-Milgram theorem. We need to therefore verify the necessary assumptions.
For the right-hand side we estimate

$$\begin{aligned} \left| \left\langle E(w \cdot A\lambda) \right\rangle \right| &= \frac{1}{\square} \int_\square E(w \cdot A\lambda) \, \mathrm{d}x \leq \|A\lambda\|_{V_\mu} \, \|w\|_{V_\mu} \\ &\leq \frac{1}{\sqrt{a_{min}}} \|A\lambda\|_{V_\mu} \, \|w\|_{V_{\mu,A}} , \quad \forall w \in V_{\mu,A}. \end{aligned}$$

The left-hand side bilinear form satisfies for all $w, u \in V_{\mu,A}$

$$\begin{aligned} \left\langle E(w \cdot Au) \right\rangle &= \frac{1}{\square} \int_\square E(w \cdot Au) \, \mathrm{d}x \leq \|w\|_{V_{\mu,A}} \, \|u\|_{V_{\mu,A}} \\ \left\langle E(w \cdot Aw) \right\rangle &= \frac{1}{\square} \int_\square E(w \cdot Aw) \, \mathrm{d}x = \|w\|^2_{V_{\mu,A}} . \end{aligned}$$

The unique existence is proved by the Lax-Milgram theorem and the estimate (4.12) follows from all of the computations above.

□

Following derivations are based above all on the third assumption (4.7).

Let $\Gamma_n \equiv Y_n(\Omega)$ denote the image of Y_n and $\Gamma = \prod_{n=1}^{M} \Gamma_n$. By δ we will denote the joint probability density function of variables $[Y_1, Y_2, \dots, Y_M]$, where $\delta : \Gamma \to \mathbb{R}_+,\ \delta \in L^\infty(\Gamma)$.

Now let's transfer our problem from the sample space to Γ. The stochastic variational problem (4.11) can be equivalently expressed in a "deterministic" scenario:

$$\begin{gathered} \text{Find } v \in V_{\delta,A} \text{ s.t.} \\ \int_\Gamma (Av, w)_{L^2(\square)} \delta \,\mathrm{d}y = \int_\Gamma (A\lambda, w)_{L^2(\square)} \delta \,\mathrm{d}y, \quad \forall w \in V_{\delta,A}, \end{gathered} \tag{4.13}$$

where the space $V_{\delta,A}$ is an analogue of $V_{\mu,A}$ with $(\Omega, \mathcal{F}, \mu)$ replaced by $(\Gamma, \mathcal{B}^M, \delta \,\mathrm{d}y)$, i.e.

$$V_{\delta,A} = \{v \in L^2_{\delta\,\mathrm{d}y}(\Gamma) \otimes \nu^2_{pot}(\square) : \ \frac{1}{\square} \int_\square \int_\Gamma v \cdot Av \, \delta \,\mathrm{d}y \,\mathrm{d}x < \infty\}.$$

The solution of (4.13) clearly has the form $v(x, \omega) = v(x, Y_1(\omega), \dots, Y_M(\omega))$ and is also the solution of (4.11). The stochastic boundary value problem (4.4) now becomes a deterministic boundary value problem for an elliptic PDE with an M-dimensional parameter. We will consider the solution v as a function $v : \Gamma \to \nu^2_{pot}(\square)$ and by $v(y)$ we will denote the dependence of v on $y \in \Gamma$. Similarly for $A(y)$.

Then (4.4) is equivalent to

$$\frac{1}{\square} \int_\square w \cdot A(y)v(y) \,\mathrm{d}x = \frac{1}{\square} \int_\square w \cdot A(y)\lambda \,\mathrm{d}x \quad \forall w \in \nu^2_{pot}(\square), \quad \delta - \text{a.e. in } \Gamma. \tag{4.14}$$

Let's suppose that A can be smoothly extended on the $\delta \,\mathrm{d}y$-zero measure sets, so that (4.14) can be considered a.e. in Γ w.r.t the Lebesgue measure.

4.2 Monte Carlo Method

As was explained at the beginning of this chapter, particularly in 4.3, the quantity we are eventually interested in is the expected value of $\widetilde{A}^\rho$

$$E(\widetilde{A}^\rho) = \int_\Omega \widetilde{A}^\rho_\omega \,\mathrm{d}\mu = \int_\Gamma \widetilde{A}^\rho(y)\delta(y) \,\mathrm{d}y,$$

which leads us to the problem of numerical computation of a multi-dimensional integral. When the dimension of the domain Ω is large, the Monte Carlo method offers us a big advantage that the error of the mean does not depend on the

dimension whilst some other methods show even an exponential dependency on the dimension.

The algorithm is very easy: we sample points with respect to the probability distribution determined by δ, obtain $y_1, \ldots, y_L$ and approximate the expected value by

$$E(\widetilde{A}^\rho) \approx Q_L(\widetilde{A}^\rho) \equiv \frac{1}{L} \sum_{i=1}^{L} \widetilde{A}^\rho(y_i). \tag{4.15}$$

By the virtue of the Law of large numbers we know that

$$\lim_{L\to\infty} Q_L(\widetilde{A}^\rho) = E(\widetilde{A}^\rho).$$

The rate of convergence of such an approximation will be discussed in the section 4.4.1 below.

4.3 Collocation Method

As opposed to the Monte Carlo method, where the specific points in the domain Γ are chosen randomly according to their probability distribution, the collocation methods suggest a deterministic way to pick the collocation points.

We are seeking a numerical approximation of the exact solution of (4.13) in a finite-dimensional subspace specified for each method individually. To describe such a subspace properly, we introduce some standard approximation subspaces, namely

- $\nu^2_{pot,N}(\square) \subset \nu^2_{pot}(\square)$ is a finite-dimensional space with the dimension $|N|$, which contains the numerical solution of the Galerkin approximation of the deterministic (only spatial-dependent) problem (4.14) for a fixed parameter $y \in \Gamma$, N being a discretization parameter,
- $\mathcal{P}_p(\Gamma) \subset L^2_\delta(\Gamma)$ is the span of tensor product polynomials with the degree at most $p = (p_1, \ldots, p_M)$

$$\mathcal{P}_{p_n}(\Gamma_n) = \operatorname{span}(y_n^m, m = 0, \ldots, p_n), \quad n = 1, \ldots, M$$
$$\mathcal{P}_p(\Gamma) = \bigotimes_{n=1}^{M} \mathcal{P}_{p_n}(\Gamma_n).$$

  Hence the dimension of $\mathcal{P}_p$ is $M_p = \Pi_{n=1}^{M}(p_n + 1)$.

The algorithm is such that:

1. We firstly choose the collocation points $\{y_k\}_{k\in\mathcal{K}} \subset \Gamma$.

2. We treat them as the M-dimensional parameters and then project (4.14) onto the subspace $\nu^2_{pot,N}(\square)$, for each of the collocation points $y_k,\ k \in \mathcal{K}$, i.e.

$$\frac{1}{\square}\int_{\square} w_N \cdot A(y_k) v_N(y_k)\,\mathrm{d}x = \frac{1}{\square}\int_{\square} w_N \cdot A(y_k)\lambda\,\mathrm{d}x \qquad \forall w_N \in \nu^2_{pot,N}(\square) \quad \forall k \in \mathcal{K}. \tag{4.16}$$

By π_N we will denote the projection of $v \in \nu^2_{pot}(\square)$ onto its finite-dimensional approximation $v_N \in \nu^2_{pot,N}(\square)$,i.e. $\pi_N v = v_N$.

3. The final step is interpolating in y the collocated solutions and thus building the discrete solution $v_{p,N}$

$$v_{p,N}(x,y) = \sum_{k\in\mathcal{K}} v_N(x,y_k)\, l^p_k(y),$$

where the functions l^p_k can be taken as, for instance, the Lagrange polynomials.

Now, let's discuss several ways of how to approach the problem of choosing the collocation points and the interpolation.

4.3.1 Full Tensor Collocation Method

Firstly we briefly recall interpolation based on Lagrange polynomials in a 1-dimensional case ($M = 1$). Let $i \in \mathbb{N}_+$ denote a level of approximation, m_i the number of collocation points and $\{y^i_1, \ldots, y^i_{m_i}\} \subset \Gamma^1$ the set of collocation points or abscissas for the interpolation. In this case $\Gamma^1 = \Gamma$.

Consider $v \in C(\Gamma^1; \nu^2_{pot}(\square))$. We introduce a sequence of one-dimensional interpolation operators $\mathcal{U}^i : C(\Gamma^1; \nu^2_{pot}(\square)) \to V_{m_i}(\Gamma^1; \nu^2_{pot}(\square))$ with standard Lagrange interpolation

$$\mathcal{U}^i(v)(y) = \sum_{k=1}^{m_i} v(y^i_k)\cdot l^i_k(y), \quad \forall v \in C^0(\Gamma^1; \nu^2_{pot}(\square)), \tag{4.17}$$

where $l^i_k \in \mathcal{P}_{m_i-1}(\Gamma^1)$ are Lagrange polynomials of degree $p_i = m_i - 1$, i.e.

$$l^i_k(y) = \prod_{\substack{j=1\\ j\neq k}}^{m_i} \frac{(y - y^i_j)}{(y^i_k - y^i_j)}$$

and

$$V_m(\Gamma^1; \nu^2_{pot}(\square)) = \Big\{ v \in C^0(\Gamma^1; \nu^2_{pot}(\square)) : v(x,y) = \sum_{k=1}^{m} \tilde{v}_k(x)\, l_k(y),\ \{\tilde{v}_k\}_{k=1}^m \in \nu^2_{pot}(\square) \Big\}.$$

The integral of v over Γ^1 is then computed by the quadrature rule

$$E_\delta(v) \equiv \int_{\Gamma^1} v(y)\delta(y)\,\mathrm{d}y = \sum_{k=1}^{m_i} v(y_k^i) \int_{\Gamma_1} l_k^i(y)\delta(y)\,\mathrm{d}y. \tag{4.18}$$

The extension to the multi-dimensional ($M > 1$) case is quite straightforward. Consider a multi-index $i = (i_1, \ldots, i_M) \in \mathbb{N}_+^M$ and its correspondent vector of number of abscissas in each dimension $(m_{i_1}, \ldots, m_{i_M})$. The full tensor product interpolation operator is defined as

$$\begin{aligned}\mathcal{I}_i v(y) &\equiv \big(\mathcal{U}^{i_1} \otimes \cdots \otimes \mathcal{U}^{i_M}\big)(v)(y) \\ &= \sum_{k_1=1}^{m_{i_1}} \cdots \sum_{k_M=1}^{m_{i_M}} u\big(y_{k_1}^{i_1}, \ldots, y_{k_M}^{i_M}\big) \cdot \big(l_{k_1}^{i_1} \otimes \cdots \otimes l_{k_M}^{i_M}\big), \quad \forall v \in C^0(\Gamma; \nu_{pot}^2(\square)).\end{aligned} \tag{4.19}$$

The set of the abscissas $[y_{k_1}^{i_1}, \ldots, y_{k_M}^{i_M}],\ 1 \le k_n \le m_{i_n}$ is referred to as the grid, or in this case, the full-tensor grid.

Clearly, the final fully discrete solution $u_{p,N}(x, y)$ belongs to the space $\mathcal{P}_p(\Gamma) \otimes \nu_{pot,N}^2(\square)$, where $p = (m_{i_1} - 1, \ldots, m_{i_M} - 1)$.

The expected value of v can be approximated by

$$E_\delta(v) \approx \sum_{k_1=1}^{m_{i_1}} \cdots \sum_{k_M=1}^{m_{i_M}} v\big(y_{k_1}^{i_1}, \ldots, y_{k_M}^{i_M}\big) \int_\Gamma \big(l_{k_1}^{i_1} \otimes \cdots \otimes l_{k_M}^{i_M}\big)(y)\delta(y)\,\mathrm{d}y. \tag{4.20}$$

The formula (4.17) and the quadrature rule (4.18) is exact for all polynomials of degree less than m_i. We know that we can improve this by using the Gaussian quadrature rule. This scenario will be investigated in the following section.

Gaussian formulas

Before we continue, let us introduce you an auxiliary probability density $\hat{\delta} : \Gamma \to \mathbb{R}^+$ such that

$$\left\| \frac{\delta}{\hat{\delta}} \right\|_{L^\infty(\Gamma)} < \infty \qquad \text{and} \qquad \hat{\delta}(y) = \Pi_{n=1}^M \hat{\delta}_n(y_n) \quad \forall y \in \Gamma. \tag{4.21}$$

With the introduction of the auxiliary density function we obtain the independency of the random variables $Y_n,\ n = 1, \ldots, M$ w.r.t the probability density $\hat{\delta}$. This feature has direct consequences on the choice of the collocation points, since now we can determine them on the basis of the marginal density functions $\hat{\delta}_n,\ n = 1, \ldots, M$.

We start again with the 1-dimensional case, i.e. having the level $i \in \mathbb{N}_+$ and its correspondent number of abscissas m_i.

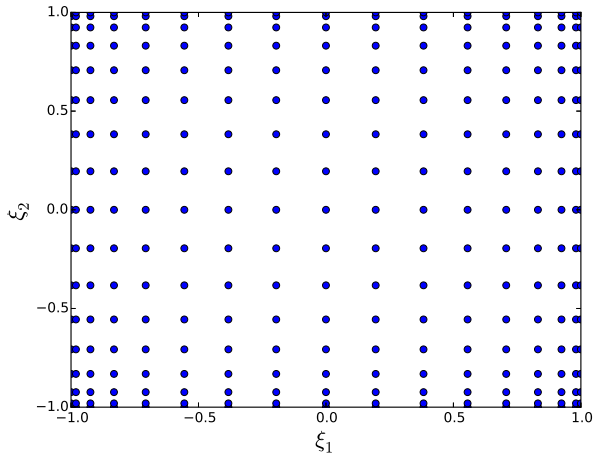


Figure 4.1: Full tensor grid $M = 2, m_{i_1} = m_{i_2} = 17$

Consider the orthogonal polynomial q_{m_i} with respect to the weight δ of the degree m_i. Such polynomial satisfies

$$\int_{\Gamma^1} q_{m_i}(y)v(y)\delta(y)\,\mathrm{d}y = 0, \quad \forall v \in \mathcal{P}_{m_i-1}(\Gamma^1).$$

The set of the collocation points is the set of m_i roots of the polynomial q_{m_i} denoted by $\{y_1^i, \ldots, y_{m_i}^i\}$. The formula (4.17) and the quadrature rule (4.18) is exact for all polynomials of degree $2m_i - 1$.

The multi-dimensional case is where the auxiliary density (4.21) plays its big role. We start again with the multiindex $i = (i_1, \ldots, i_M)$ and its vector of number of abscissas $(m_{i_1}, \ldots, m_{i_M})$. For each dimension $n = 1, \ldots, M$ consider the orthogonal polynomial $q_{m_{i_n}}$ with respect to the weight $\hat{\delta}_n$ of the degree m_{i_n}. Such polynomial satisfies

$$\int_{\Gamma_n} q_{m_{i_n}}(y)v(y)\hat{\delta}_n(y)\,\mathrm{d}y = 0, \quad \forall v \in \mathcal{P}_{m_{i_n}-1}(\Gamma_n).$$

The set of collocation points in the domain Γ is then the full tensor product of $\{y_{j_n}^{i_n}\}$, $1 \le j_n \le m_{i_n}$, i.e. the set of points

$$[y_{j_1}^{i_1}, \ldots, y_{j_M}^{i_M}] \in \Gamma, \quad 1 \le j_n \le m_{i_n}. \tag{4.22}$$

An example of a full tensor grid for $M = 2$ random variables, each uniformly distributed in $[-1; 1]$ with 17 points in each dimension.

Remark. There is a 3-recurrence relation for any arbitrary weight function $\hat{\delta}_n$. However, standard choices of the probability density functions lead to some well-known orthogonal polynomials whose roots are tabulated (see Table 4.1) and thus do not need to be computed.

Distribution	Pdf	Polynomials	Weights	Support
Uniform	$1/2$	Legendre	1	$[-1;1]$
Gaussian	$(1/\sqrt{(2\pi)})exp(-x^2/2)$	Hermite	$\exp(-x^2/2)$	$[-\infty;\infty]$
Exponential	$exp(-x)$	Laguerre	$exp(-x)$	$[0;\infty]$
Beta	$\frac{(1-x)^\alpha(1+x)^\beta}{B(\alpha,\beta)}$	Jacobi	$(1-x)^\alpha(1+x)^\beta$	$[-1;1]$

Table 4.1: Standard distributions with correspondent orthogonal polynomials

The interpolation is done in the same way as in (4.19), however, a major improvement is done when computing the expected value (4.20) due to the auxiliary density function.

Since the random variables are independent w.r.t. $\hat{\delta}(y) = \prod_{n=1}^{M} \hat{\delta}_n(y_n)$, we can derive:

$$\begin{aligned} E_\delta(v) &\approx \sum_{k_1=1}^{m_{i_1}} \cdots \sum_{k_M=1}^{m_{i_M}} v\big(y_{k_1}^{i_1}, \ldots, y_{k_M}^{i_M}\big) \int_\Gamma \big(l_{k_1}^{i_1} \otimes \cdots \otimes l_{k_M}^{i_M}\big)(y)\delta(y)\,\mathrm{d}y \\ &= \sum_{k_1=1}^{m_{i_1}} \cdots \sum_{k_M=1}^{m_{i_M}} \Big[\big(\frac{\delta}{\hat{\delta}} v\big)(y_{k_1}^{i_1}, \ldots, y_{k_M}^{i_M})\big(\prod_{n=1}^{M} \int_{\Gamma^n} l_{k_n}^{i_n}(y_n)\hat{\delta}(y_n)\,\mathrm{d}y_n\big)\Big], \end{aligned} \tag{4.23}$$

for any $v \in C^0(\Gamma; \nu_{pot}^2(\square))$, assuming $\delta/\hat{\delta}$ is a smooth function. Otherwise the expected value should be computed with a quadrature rule suitable for possible discontinuities of $\delta/\hat{\delta}$. Each of the quantities $\int_{\Gamma^n} l_{k_n}^{i_n}(y_n)\hat{\delta}(y_n)\,\mathrm{d}y_n$ is easily computable and for well-known weight functions $\hat{\delta}_n$ are even tabulated.

The number of collocation points is $m_{i_1} \cdot m_{i_2} \cdot \cdots \cdot m_{i_M}$. Therefore this approach can be computationally expensive if the number M of random variables needed to describe the input data is large. This phenomenon is often referred to as the curse of dimensionality - exponential growth in the required work with respect to the number of random variables. The next section is trying to address this problem with a way of sparsifying the grid.

4.3.2 Sparse Grid Collocation Method

The construction of a sparse grid is done by the Smolyak algorithm. The idea behind it is to build linear combinations of product formulas (4.19) so that only products with relatively small number of knots are used.

For $i \in \mathbb{N}_+$ let's define the operator Δ^i

$$\Delta^i = \mathcal{U}^i - \mathcal{U}^{i-1},$$

where $\mathcal{U}^0 = 0$.

With $|i| = i_1 + \cdots + i_M$ for $i = (i_1, \ldots, i_M) \in \mathbb{N}_+^M$ we define the Smolyak isotropic formula $\mathcal{A}(q, M)$ by

$$\mathcal{A}(q, M) = \sum_{|i| \le q} \big(\Delta^{i_1} \otimes \cdots \otimes \Delta^{i_M}\big) \tag{4.24}$$

for $q \geq M,\ q \in \mathbb{N}$ indicating the level of accuracy.

Formula (4.24) can be equivalently written as

$$\mathcal{A}(q,M) = \sum_{q-M+1 \leq |i| \leq q} (-1)^{q-|i|} \binom{M-1}{q-|i|} \cdot (\mathcal{U}^{i_1} \otimes \cdots \otimes \mathcal{U}^{i_M}). \tag{4.25}$$

Therefore the grid only consists of points

$$\mathcal{H}(q,M) = \bigcup_{q-M+1 \leq |i| \leq q} \left(\vartheta^{i_1} \times \cdots \times \vartheta^{i_M}\right), \tag{4.26}$$

where $\vartheta^i = \{y_1^i, \ldots, y_{m_i}^i\}$ is the set of collocation points for $\mathcal{U}^i$.

The appropriate choice of abscissas is an open question for specific applications. We will introduce two of the probably mostly used ones.

Clenshaw-Curtis formulas

There are two questions we need to specify. The number of points m_i for each $i \in \mathbb{N}_+$ and the choice of the abscissas y_j^i. Clenshaw-Curtis formulas suggest us to use the extrema of the Chebyshev polynomials

$$\begin{aligned} y_j^i &= \cos\left(\frac{\pi(j-1)}{m_i - 1}\right), \quad j = 1, \ldots, m_i, \\ y_j^i &= 0, \quad m_i = 1. \end{aligned} \tag{4.27}$$

To answer the first question we set

$$m_i = \begin{cases} 1 & \text{if } i = 1 \\ 2^{i-1} + 1 & \text{if } i > 1. \end{cases}$$

This choice ensures us that the grid points are nested, i.e. $\mathcal{H}(q,M) \subset \mathcal{H}(q+1,M)$. We obtain the degree $m_i - 1$ of exactness, which is the same as for any other choice of m_i abscissas. The Chebyshev polynomials have, however, great interpolation properties and the Clenshaw-Curtis quadrature formulas have accuracy comparable to the Gaussian quadrature formulas.

Gaussian formulas

We know that the Gaussian formulas have the maximum level of exactness $2m_i - 1$. We will set m_i to be the same as in (4.27) although in general it will not supply us with nested formulas

$$\begin{aligned} y_j^i &= \cos\left(\frac{\pi(j-1)}{m_i - 1}\right), \quad j = 1, \ldots, m_i, \\ y_j^i &= 0, \quad m_i = 1. \end{aligned} \tag{4.28}$$

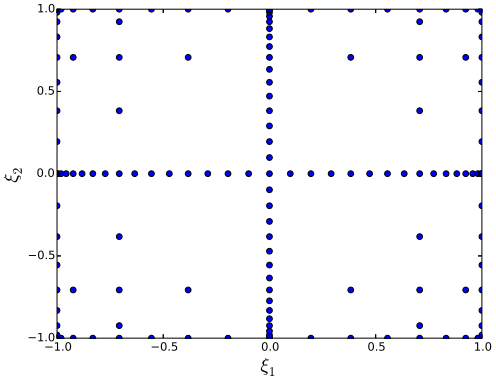


Figure 4.2: Clenshaw-Curtis $\mathcal{H}(5,2)$

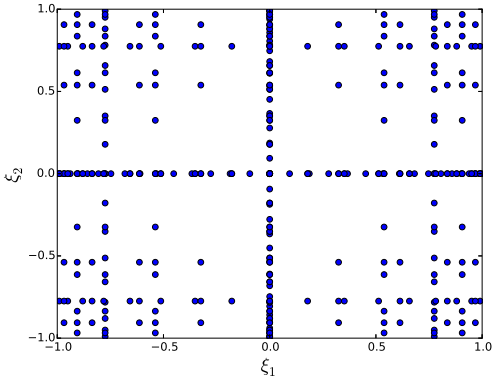


Figure 4.3: Gaussian $\mathcal{H}(5,2)$

As for the choice of abscissas we will proceed analogically to the full-tensor case, i.e. for each dimension $n = 1, \ldots, M$ taking m_{i_n} roots of the m_{i_n} degree orthogonal that is $\hat{\delta}_n$-orthogonal to all polynomial of a lower degree.

A Clenshaw-Curtis sparse grid of level accuracy $q = 5$ and dimension $M = 2$ can be seen in the Figure (4.2). The Figure (4.3) depicts the Gaussian sparse grid for uniformly distributed random variables ξ_1, ξ_2 over the interval $[-1; 1]$, i.e. $\hat{\delta}_1 = \hat{\delta}_2 = 1/2$, q, M are the same. As you can see, the Gaussian sparse grid contains a lot more points than The Clenshaw-Curtis one, even though m_i is defined the same for both of them. The reason is that the Clenshaw-Curtis formulas produce nested grids. The Gaussian formula compensates this disadvantage by being more accurate.

4.4 Convergence analysis

This section is dedicated to an analysis of the convergence rates for all of the proposed methods. Additional assumptions for some of them have to be stated, all of which will be meaningful for the applications computed in the last chapter.

4.4.1 Monte Carlo method

We are trying to estimate the rate of convergence for the error $E(\tilde{A}^\rho) - Q_L(\tilde{A}^\rho)$ in the norm $\|\cdot\|_{L^2(\Gamma,\nu^2_{pot}(\square))}$ which in this case becomes

$$\begin{aligned}\left\|E(\tilde{A}^\rho) - Q_L(\tilde{A}^\rho)\right\|^2_{L^2(\Gamma,\nu^2_{pot}(\square))} &= E\Big[\big(E(\tilde{A}^\rho) - Q_L(\tilde{A}^\rho), E(\tilde{A}^\rho) - Q_L(\tilde{A}^\rho)\big)_{\mathbb{R}^{m\times m}}\Big] \\ &\equiv E\Big[\big(E(\tilde{A}^\rho) - Q_L(\tilde{A}^\rho)\big)^2\Big].\end{aligned}$$

This quantity is also known as the mean squared error of the mean (expected value), for which it holds

$$E\Big[\big(E(\tilde{A}^\rho) - Q_L(\tilde{A}^\rho)\big)^2\Big] = \frac{\mathrm{var}(\tilde{A}^\rho)}{L}, \tag{4.29}$$

where $\mathrm{var}(\tilde{A}^\rho)$ is the variance of the random vector $\tilde{A}^\rho \in \mathbb{R}^{m\times m}$.

Therefore we can state that the Monte Carlo method converges in the natural norm with rate $L^{-1/2}$, where L is the number of 'collocation points'.

4.4.2 Collocation method

The collocation method is easily applicable for problems with an unbounded right-hand side and unbounded random variables as well. In what follows we are describing the required maximal growth of f and decay of δ at infinity.

We introduce an auxiliary weight $\sigma(y) = \prod_{n=1}^{M} \sigma_n(y_n) \leq 1$, where

$$\sigma(y_n) = \begin{cases} 1 & \text{if } \Gamma_n \text{ is bounded} \\ e^{-\alpha_n |y_n|} \text{ for some } \alpha_n > 0 & \text{if } \Gamma_n \text{ is unbounded} \end{cases}$$

and the functional space

$$C^0_\sigma(\Gamma; V) \equiv \left\{ v : \ \Gamma \to V, \, v \, \text{continuous in } y, \, \max_{y\in\Gamma} \|\sigma(y)v(y)\|_V < \infty \right\},$$

where V is a Banach space of functions defined in $\square$.

Assumptions

1. $a_{ij} \in C^0_\sigma(\Gamma; L^2(\square)), \quad a_{ij} \in C^0_{loc}(\Gamma; L^\infty(\square))$,
2. The joint probability density δ satisfies

$$\delta(y) \leq C_\delta e^{-\sum_{n=1}^{M} (\vartheta_n y_n)^2}, \quad y \in \Gamma \tag{4.30}$$

   where $C_\delta > 0$ and ϑ_n is strictly positive if Γ_n is unbounded and zero otherwise.

To explain the meaning of the preceding assumptions we shall say that α_n controls the growth of the right-hand side $A(x,\omega)\lambda$ whenever Γ is unbounded, while ϑ_n describes the rate of the decay of δ at infinity. Most trivially said, we want the growth of the right-hand side to be at most exponential and the joint density δ to decay as a Gaussian weight.

Lemma 4.3. *Under the assumptions stated above, the solution to the problem (4.14) satisfies* $v \in C^0_\sigma(\Gamma; \nu^2_{pot}(\square))$.

Proof follows directly from the definition of the problem (4.14). Another assumption concerns the regularity of the solution to (4.4). For that we are introducing new notation:

$$\Gamma^*_n = \prod_{\substack{j=1 \\ j\neq n}}^{M} \Gamma_j, \qquad y^*_n \ \text{ arbitrary element of } \Gamma^*_n.$$

3. We need the solution of the problem (4.4) to satisfy: For each $n = 1, \ldots, M$, there exists $\tau_n > 0$ such that the function $v(x, y_n, y_n^*)$ as a function of y_n, $v : \Gamma_n \to C^0(\Gamma_n^*; \nu_{pot}^2(\square))$ admits an analytic extension $v(x, z, y_n^*),\ z \in \mathbb{C}$, in the region of the complex plane

$$\Sigma(\Gamma_n; \tau_n) \equiv \{z \in \mathbb{C},\ dist(z, \Gamma_n) \leq \tau_n\}.$$

Moreover, $\forall z \in \sum(\Gamma_n; \tau_n)$,

$$\|v(z)\|_{C^0(\Gamma_n^*; \nu_{pot}^2(\square))} \leq \beta, \tag{4.31}$$

where β is a constant independent of n.

This assumption has to be verified for each application individually, with the distance τ_n having direct impact on the rate of convergence.

As explained before, the whole process consists of approximating the solution $v \in C^0(\Gamma; \nu_{pot}^2(\square))$ by finitely many function values, each of which is computed by the Fourier-Galerkin method described in Chapter 3. The fully discrete solution $v_{p,N}$ can be expressed as

$$v_{p,N} = \mathcal{A}(q, M)\pi_N v,$$

in case of sparse grid approximation or

$$v_{p,N} = \mathcal{I}_i \pi_N v,$$

in case of full tensor grids.

We are going to estimate the error $v - v_{p,N}$ in the natural norm $\|\cdot\|_{L^2_\delta(\Gamma; \nu_{pot}^2(\square))}$ as defined in (4.8), also referred to as $\|\cdot\|_{V_\delta}$. Controlling the error in this norm comes in handy also when wanting to control the error in the mean value of the solution:

$$\|E(v - v_{p,N})\|_{\nu_{pot}^2(\square)} \leq E(\|v - v_{p,N}\|_{\nu_{pot}^2(\square)}) \leq \|v - v_{p,N}\|_{V_\delta}. \tag{4.32}$$

We can derive

$$\|v - v_{p,N}\|_{V_\delta} \leq \|v - \pi_N v\|_{V_\delta} + \|\pi_N(v - \mathcal{A}(q, M)v)\|_{V_\delta}. \tag{4.33}$$

The first term can be estimated by the expression derived in Chapter 3 (3.11).

By virtue of the Céa lemma we know that

$$\|w - \pi_N w\|_{\nu_{pot}^2(\square)} \leq C_\pi \min_{z \in \nu_{pot,N}^2(\square)} \|w - z\|_{\nu_{pot}^2(\square)}, \quad \forall w \in \nu_{pot}^2(\square), \tag{4.34}$$

where the constant C_π is independent of the discretization parameter N.

Therefore the for the second term of (4.33) we can write

$$\|\pi_N(v - \mathcal{A}(q, M)v)\|_{V_\delta} \leq C_\pi \|v - \mathcal{A}(q, M)v\|_{V_\delta}.$$

These preceding derivation are valid for $\mathcal{I}_i$ instead of $\mathcal{A}(q, M)$ as well.

In the next sections we focus only on the estimation of $\|v - \mathcal{A}(q, M)v\|_{V_\delta}$ or $\|v - \mathcal{I}_i v\|_{V_\delta}$.

4.4.3 Full-tensor grid with Gaussian abscissas

Let $p = (p_1, \ldots, p_M) = (m_{i_1} - 1, \ldots, m_{i_M} - 1)$ denote the polynomial degree, i.e. m_{i_n} denote the number of Gaussian abscissas in each dimension. Then we can formulate the following theorem stating a sub-exponential rate of convergence with respect to the polynomial degree.

Theorem 4.1. *Under the assumptions (4.4.2), there exist positive constants r_n, $n = 1, \ldots, M$ and C, independent of N and p, such that*

$$\|v - \mathcal{I}_i v\|_{V_\delta} \leq C \sum_{n=1}^{M} \beta_n(p_n) \, \exp(-r_n \, p_n^{\theta_n}), \tag{4.35}$$

where

- *if Γ_n is bounded:*

$$\theta_n = \beta_n = 1,$$
$$r_n = \log\Big[\frac{2\tau_n}{|\Gamma_n|}\big(1 + \sqrt{1 + \frac{|\Gamma_n|^2}{4\tau_n^2}}\big)\Big],$$

- *Γ_n is unbounded:*

$$\theta_n = 1/2, \quad \beta_n = O(\sqrt{p_n}),$$
$$r_n = \tau_n \vartheta_n.$$

Proof. A detailed proof of this theorem can be found in [Babuška et al., 2007, p. 20].

□

What we are more interested in are the estimates with respect to the number of collocation points. For the sake of simplicity let's assume $m_{i_n} = \tilde{m}$, $\forall i$. Then the number of points in the full tensor grid is $\eta = \tilde{m}^M$.

We can simplify the estimation (4.35) to be isotropic in each dimension and obtain

$$err \equiv \|v - \mathcal{I}_i v\|_{V_\delta} \leq C \exp(-rp),$$

which expressed with respect to the number of collocation points becomes

$$err \leq C \exp(-r\eta^{1/M}).$$

Observe that for M large

$$\eta^{1/M} \approx 1 + \log(\eta)/M,$$

which causes the effective rate to be algebraic, rather than exponential

$$err \leq C\eta^{-r/M}.$$

For M large we can come up with worse convergence rates than in the Monte Carlo method. This phenomenon is referred to as the curse of dimensionality. Now we will look at what sparse grids have to offer.

4.4.4 Clenshaw-Curtis sparse grid

For the sake of simplicity we focus our study on bounded variables only. A similar result can be obtained for unbounded variables as well.

Firstly consider an approximation error in a 1-dimensional case.

Lemma 4.4. *Given a function $v \in C^0(\Gamma^1; \nu^2_{pot}(\square))$ which admits an analytic extension in the region of the complex plane $\Sigma(\Gamma^1; \tau) = \{z \in \mathbb{C}.\ dist(z, \Gamma^1) \leq \tau\}$ for some $\tau > 0$, i.e. satisfies the property from the assumption 3. for $M = 1$, there holds*

$$E_{m_i} \equiv \min_{w \in V_{m_i}} \|v - w\|_{C^0(\Gamma^1; \nu^2_{pot}(\square))} \leq \frac{2}{\varrho - 1} e^{-m_i log(\varrho)} \max_{z \in \Sigma(\Gamma^1; \tau)} \|v(z)\|_{\nu^2_{pot}(\square)}, \quad (4.36)$$

where $1 < \varrho = \frac{2\tau}{|\Gamma^1|} + \sqrt{1 + \frac{4\tau^2}{|\Gamma^1|^2}}$.

Proof. A detailed proof of this lemma is available in [Babuška et al., 2007, p. 17].

□

In the multidimensional case the size of the analyticity region will depend on n and will be denoted by τ_n. In what follows we set

$$\varrho \equiv \min_n \varrho_n.$$

Instead of estimating the error in the norm of $L^2_\delta(\Gamma; \nu^2_{pot}(\square))$, we will use the norm of $L^\infty(\Gamma; \nu^2_{pot}(\square))$ justified by

$$\|v\|_{L^2_\delta(\Gamma; \nu^2_{pot}(\square))} \leq \|v\|_{L^\infty(\Gamma; \nu^2_{pot}(\square))}, \quad \forall v \in L^\infty(\Gamma; \nu^2_{pot}(\square)).$$

For the operator $\mathcal{U}^i$ it holds

$$\left\|v - \mathcal{U}^i(v)\right\|_{L^\infty(\Gamma; \nu^2_{pot}(\square))} \leq E_{m_i - 1}(v) \cdot (1 + \Lambda_{m_i}), \quad (4.37)$$

where in the case of Clenshaw-Curtis formulas Λ_{m_i} can be bounded by

$$\Lambda_{m_i} \leq \frac{2}{\pi} \log(m_i - 1) + 1, \quad \text{for } m_i \geq 2.$$

By virtue of the Lemma 4.4 we have

$$E_{m_i}(v) \leq C \varrho^{-m_i},$$

C is independent of τ.

After gathering the preceding estimates we have

$$\left\|v - \mathcal{U}^i(v)\right\|_{L^\infty(\Gamma; \nu^2_{pot}(\square))} \leq C \log(m_i) \varrho^{-m_i} \leq C i \varrho^{-2^i}.$$

For the operator Δ in 1 dimension we finally obtain

$$\begin{aligned}\left\|(\Delta^i)(v)\right\|_{L^\infty(\Gamma;\nu^2_{pot}(\square))} &= \left\|(\mathcal{U}^i - \mathcal{U}^{i-1})(v)\right\|_{L^\infty(\Gamma;\nu^2_{pot}(\square))} \\ &\leq \left\|(I - \mathcal{U}^i)(v)\right\|_{L^\infty(\Gamma;\nu^2_{pot}(\square))} + \left\|(I - \mathcal{U}^{i-1})(v)\right\|_{L^\infty(\Gamma;\nu^2_{pot}(\square))} \\ &\leq Ei\varrho^{-2^{i-1}},\end{aligned} \tag{4.38}$$

where C, E depend on v but not on i.

Applying (4.38) for each dimension in the special linear combinations (4.25) we come to the final estimate.

Theorem 4.2. *For functions $v \in C^0(\Gamma; \nu^2_{pot}(\square))$ satisfying the Assumption 3. in (4.4.2) the Smolyak formula (4.25) based on Clenshaw-Curtis abscissas satisfies:*

$$\begin{gathered}\|(I - \mathcal{A}(q, M)(v)\|_{L^2_\delta(\Gamma;\nu^2_{pot}(\square))} \leq C(r_{min}, M)\eta^{-\mu_1}, \\ \mu_1 = \frac{r_{min}}{1 + \log(2M)}.\end{gathered} \tag{4.39}$$

Proof. A detailed proof is provided in [Nobile et al., 2008].

□

4.4.5 Gaussian sparse grid

For the Gaussian sparse grid interpolation we proceed in a similar manner.

Firstly we need to embed the auxiliary density $\hat{\delta}$ into our work by

$$\|v\|_{L^2_\delta(\Gamma;\nu^2_{pot}(\square))} \leq \left\|\frac{\delta}{\hat{\delta}}\right\|_{L^\infty(\Gamma)} \cdot \|v\|_{L^2_{\hat{\delta}}(\Gamma;\nu^2_{pot}(\square))} \quad \forall v \in C^0(\Gamma; \nu^2_{pot}(\square)).$$

Now we will derive some estimates for the 1-dimensional case, i.e. $M = 1$.

Lemma 4.5. *For every function $v \in C^0(\Gamma^1; \nu^2_{pot}(\square))$ it holds*

$$\left\|v - \mathcal{U}^i(v)\right\|_{L^2_{\hat{\delta}}(\Gamma^1;\nu^2_{pot}(\square))} \leq C_{\hat{\delta}} \inf_{w \in V_{m_i}} \|u - w\|_{L^\infty_{\hat{\delta}}(\Gamma^1;\nu^2_{pot}(\square))}.$$

Proof. The proof can be found in [Nobile et al., 2008].

□

Analogically to the previous procedure, by applying the Lemma 4.4 we obtain

$$\left\|v - \mathcal{U}^i(v)\right\|_{L^2_{\hat{\delta}}(\Gamma^1;\nu^2_{pot}(\square))} \leq \tilde{C}\varrho^{-2^i}.$$

For the estimation of the operator Δ^i we can derive

$$\begin{aligned}\left\|\Delta^i(v)\right\|_{L^2_{\hat{\delta}}(\Gamma^1;\nu^2_{pot}(\square))} &= \left\|(\mathcal{U}^i - \mathcal{U}^{i-1})(v)\right\|_{L^2_{\hat{\delta}}(\Gamma^1;\nu^2_{pot}(\square))} \\ &\leq \left\|(I_1 - \mathcal{U}^i(v))\right\|_{L^2_{\hat{\delta}}(\Gamma^1;\nu^2_{pot}(\square))} + \left\|(I_1 - \mathcal{U}^{i-1})(v)\right\|_{L^2_{\hat{\delta}}(\Gamma^1;\nu^2_{pot}(\square))} \\ &\leq \tilde{E}\varrho^{-2^{i-1}},\end{aligned}$$

where, as before, $\tilde{C}, \tilde{E}$ depend on v but not i.

By employing the estimates above we obtain the final estimate.

Theorem 4.3. *For functions $v \in C^0(\Gamma; \nu^2_{pot}(\square))$ satisfying the Assumption 3. from (4.4.2) the Smolyak formula (4.25) based on Gaussian abscissas satisfies*

$$\begin{aligned}\|(I - \mathcal{A}(q, M)(v)\|_{L^2_{\hat{\delta}}(\Gamma;\nu^2_{pot}(\square))} &\leq \sqrt{(\left\|\delta/\hat{\delta}\right\|_{L^\infty(\Gamma)})}C(r_{min}, M)\eta^{-\mu_2}, \\ \mu_2 &= \frac{r_{min}e\log(2)}{\varsigma + \log(M)},\end{aligned} \tag{4.40}$$

where $\varsigma \approx 2.1$ and the constant $C(r_{min}, M)$ tends to zero as $r_{min} \to \infty$.

Proof. A detailed proof is provided in [Nobile et al., 2008].

□

As we can see, sparse grid does not entirely solve the curse of dimensionality but helps to bring down its influence. The Figure 4.4 depicts the dependence of the number of collocation points on the number of random variables, where each grid has the maximum number of 5 points employed in each direction. In the last chapter Applications we will see that after crossing a certain number of random variables, Monte Carlo techniques might be the best choice.

4.5 Estimates For The Approximated Effective Matrix

In the previous section we managed to estimate the rate of convergence for the error $v - v_{p,N}$. However, what we are interested in is the error in the effective matrix $\tilde{A}^\rho_y$ or its mean value $E(\tilde{A}^\rho_y)$.

4.5.1 Deterministic case

In this part we are assuming the problem does not depend on any random variables, therefore $\tilde{A}^\rho_y = \tilde{A}^\rho$. The error of $\tilde{A}^\rho$ will be estimated in the norm $\|\cdot\|_{L^2(\square)}$. Note that the norm in $\nu^2_{pot}(\square)$ is the same as in $L^2(\square)$.

As we know, the exact solution $\tilde{A}^\rho$ is computed by

$$\tilde{A}^\rho\lambda = \langle A^\rho_{per}(v + \lambda)\rangle = \frac{1}{\square}\int_\square A^\rho_{per}(x)\big(v(x) + \lambda\big)\,\mathrm{d}x, \tag{4.41}$$

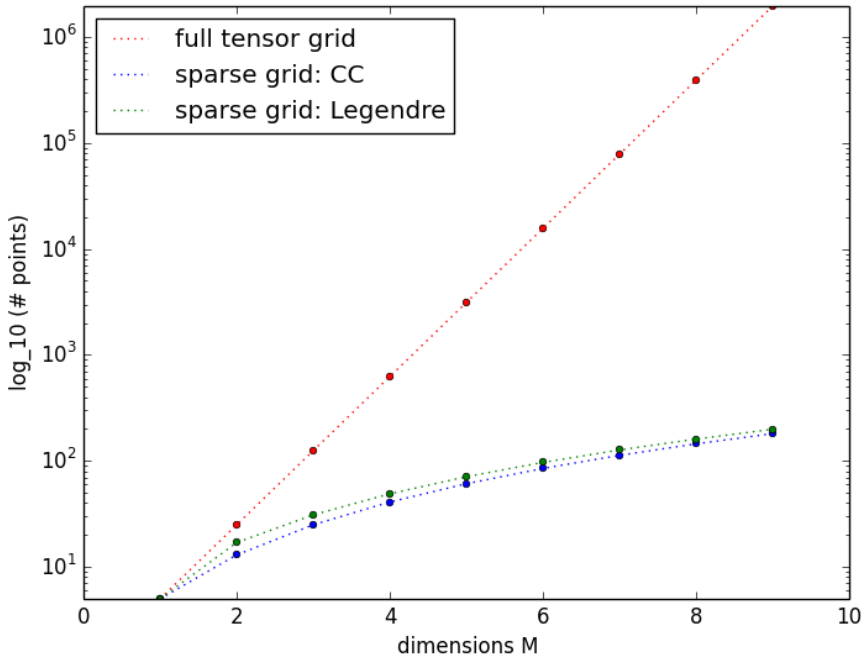


Figure 4.4: Dependence of the number of collocation points on the number of random variables, each grid has the maximum number of 5 points employed in each direction.

$v \in \nu^2_{pot}(\square)$ being a solution of the problem (3.5).

When solving this problem numerically, we come up with an approximate solution $v_N \in \nu^2_{pot,N}(\square)$ by the Fourier-Galerkin method. According to [Vondřejc et al., 2015, Lemma 39], we can manage to compute the integral in (4.41) exactly by means of the Fourier transformation. We obtain $\tilde{A}^{\rho,N}$ by

$$\tilde{A}^{\rho,N}\lambda = \langle A^\rho_{per}(v_N + \lambda)\rangle = \frac{1}{\square}\int_\square A^\rho_{per}(x)\big(v_N(x) + \lambda\big)\,\mathrm{d}x, \tag{4.42}$$

$v_N \in \nu^2_{pot,N}(\square)$ being the Fourier-Galerkin approximation of v.

Let's denote $\tilde{A}^{\rho,N}$ by $\pi_N\tilde{A}^\rho$.

Then we can estimate

$$\begin{aligned}\left\|\tilde{A}^\rho\lambda - \pi_N\tilde{A}^\rho\lambda\right\|_{L^2(\square)} &\le \frac{1}{\square}\int_\square \left\|A^\rho_{per}(v+\lambda) - A^\rho_{per}(v_N+\lambda)\right\|_{L^2(\square)}\,\mathrm{d}x \\ &\le \left\|A^\rho_{per}\right\|_{L^\infty(\square)}\,\|v - v_N\|_{L^2(\square)} \\ &\le Ch^s\left\|A^\rho_{per}\right\|_{W^{s,\infty}(\square)}\,\|v\|_{H^s(\square)},\end{aligned} \tag{4.43}$$

h defined as $h = \max_{i=1,\dots,m}\frac{\rho}{N_i}$.

4.5.2 Stochastic case

Now consider $\tilde{A}^\rho_y$ and $v(x,y)$ dependent on random variables $y \in \Gamma$.

Thanks to the Lemma 4.3 and Assumption 4.4.2 we know that $\tilde{A}^{\rho}_{y}$ defined by

$$\tilde{A}^{\rho}_{y}\lambda = \frac{1}{\square}\int_{\square} A^{\rho}_{per,y}(x)\big(v(x)+\lambda\big)\,\mathrm{d}x,$$

$v \in \nu^2_{pot}(\square)$ being a solution of the problem (3.5), belongs to the space $C^0(\Gamma, \mathbb{R}^{m\times m})$.

This justifies the use of interpolation operators $\mathcal{U}^i$. All of the estimates for the interpolation error in the random variables are valid for functions from the space $C^0(\Gamma, V)$, provided that the analyticity condition from Assumption 3. (4.4.2) is satisfied. Therefore we must require $\tilde{A}^{\rho}_{y}$ to satisfy this condition as well. By $\tilde{A}^{\rho}_{p,N}$ we denote the approximated $\tilde{A}^{\rho}$ in both random and spatial variables (similarly to v), i.e. either $\tilde{A}^{\rho}_{p,N} = \mathcal{A}(q,M)\pi_N\tilde{A}^{\rho}$ or $\tilde{A}^{\rho}_{p,N} = \mathcal{I}_i\pi_N\tilde{A}^{\rho}$. From now on we follow the theory derived above and obtain

$$\left\|\tilde{A}^{\rho} - \tilde{A}^{\rho}_{p,N}\right\|_{L^2_\delta(\Gamma,\mathbb{R}^{m\times m})} \leq \underbrace{\left\|\tilde{A}^{\rho} - \pi_N\tilde{A}^{\rho}\right\|_{L^2_\delta(\Gamma,\mathbb{R}^{m\times m})}}_{(\mathbb{I})} + \underbrace{\left\|\pi_N\tilde{A}^{\rho} - \tilde{A}^{\rho}_{p,N}\right\|_{L^2_\delta(\Gamma;\mathbb{R}^{m\times m})}}_{(\mathbb{III})}, \tag{4.44}$$

where for $(\mathbb{I})$ we can use (4.43)

$$(\mathbb{I}) \leq Ch^s\Big(\int_\Gamma \left\|A^{\rho}_{per,y}\right\|_{W^{s,\infty}(\square)} \left\|v(y)\right\|_{H^s(\square)}\delta(y)\,\mathrm{d}y\Big)^{1/2}$$

and the estimate of $(\mathbb{III})$ is just the same as in (4.35), (4.39) or (4.40).

As explained in (4.3), what we are most interested in is the quantity $E(\tilde{A}^{\rho})$. The collocation method suggests an easy way of how to obtain its approximation

$$E(\tilde{A}^{\rho}) \approx E(\tilde{A}^{\rho}_{p,N}) = \sum_{k\in\mathcal{K}} \tilde{A}^{\rho}_{p,N,y_k}\omega_k,$$

where $\{y_k\}_{k\in\mathcal{K}}$ is the grid of all collocation points and ω_k are correspondent tensor products of weights from the quadrature formulas, i.e. integrals of Lagrange polynomials.

To express the estimate of the error of $E(\tilde{A}^{\rho}) - E(\tilde{A}^{\rho}_{p,N})$ we derive

$$\left\|E(\tilde{A}^{\rho}) - E(\tilde{A}^{\rho}_{p,N})\right\|_{\mathbb{R}^{m\times m}} \leq E\big(\left\|\tilde{A}^{\rho} - \tilde{A}^{\rho}_{p,N}\right\|_{\mathbb{R}^{m\times m}}\big) \leq \left\|\tilde{A}^{\rho} - \tilde{A}^{\rho}_{p,N}\right\|_{L^2_\delta(\Gamma;\mathbb{R}^{m\times m})},$$

through which we come directly to the estimates (4.44).

5. Applications

This chapter shows some practical applications of the methods introduced in previous chapters. We are focusing on analyzing the impact of the collocation methods rather than the Fourier-Galerkin method, since there already were some publications in that direction.

For computing the deterministic part I used a Python solver [Vondřejc, 2016–2017]. The Monte-Carlo or collocation method was programmed by me and my sparse grid generator for arbitrary level q and dimension M is provided in Appendix, developed in Python as well.

5.1 Checkerboard problem

First problem we are dealing with concerns a 2-dimensional periodic medium with the matrix of material coefficients expressed as

$$A(x,\omega) = \sum_{k\in\mathbb{Z}^2} \mathbf{1}_{Q+k}(x)\, a_k(\omega) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{5.1}$$

where $Q = (0;1)^2$ and $\{a_k\}_{k\in\mathbb{Z}^2}$ is an independent identically distributed sequence of random variables, each of which is uniformly distributed over the interval $[1;10]$, i.e.

$$a_k(\omega) \sim \mathtt{U}\big([1;10]\big).$$

Figure (5.1) depicts a segment of our domain for a specific realization.

Lemma 5.1. *The matrix of material coefficients defined in (5.1) can be expressed in terms of a 2-dimensional ergodic dynamical system as*

$$A(x,\omega) = \mathcal{A}(T(x)\omega), \quad x \in \mathbb{R}^2, \omega \in \Omega.$$

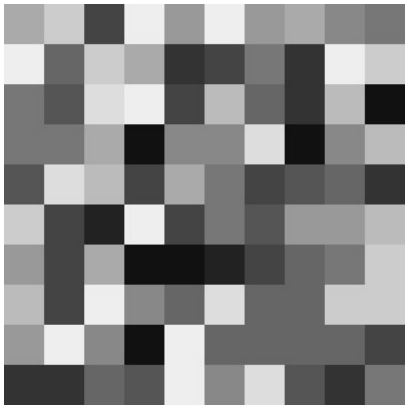


Figure 5.1: Realization of the checkerboard problem

Proof. The proof can be found in [Heida, 2017, p. 12].

□

Remark. The domain Ω is not specified because in many problems it is not trivial to find it. A detailed description of building the space Ω and the dynamical system $T(x),\ x \in \mathbb{R}^2$ is available in [Alexanderian et al., 2012, p. 8, 9, 10, 11, 12]. In computations we can omit the domain of random events and work with random variables only, if the number of random variables is finite (as was suggested in Chapter 4).

Since for any choice of x and ω the material coefficients are bounded by 1 from bellow and 10 from above, we are able to write $\mathcal{A} \in L^\infty(\Omega)$.

We shall comment on the fact that $A(\cdot, \omega)$ does not belong to $C^0(\square)$ which is a necessary condition for the operator Q_N from Chapter 3 to be defined. In this case $A(\cdot, \omega)$ is piecewise smooth (even constant) so we need to ensure the picked discretization points x_N^k from (3.7) are inside of each square. In all of the simulations we proceeded there were 9 points inside of each square. We cannot directly expect the convergence results derived for smooth material coefficients, but in [Vondřejc, 2013, p. 115] there are some nice approximation properties proved by means of a mollifier. Since the focus of our research is the impact of the techniques in the stochastic domain, we will not discuss this issue any further.

5.1.1 Convergence by enlargement of domain

The strategy of obtaining the homogenized matrix A^0 is explained at the beginning of Chapter 4. Shortly said, we need to compute the quantity $E(\tilde{A}^\rho)$ for $\rho \to \infty$. In our case the domain S_ρ will be characterized by N, meaning S_ρ is a segment of $N \times N$ squares surrounding $[0, 0]$. We have to realize that enlarging the domain brings an exponential growth in the number of random variables, causing a major difficulty.

We know that for $A(x)$ symmetric the homogenized matrix is symmetric as well. For each N the matrix $E(\tilde{A}^\rho)$ becomes

$$E(\tilde{A}^\rho) = \begin{pmatrix} a & \varepsilon \\ \varepsilon & b \end{pmatrix} \tag{5.2}$$

with $b \approx a$.

The Figure 5.2 and the Table 5.1 depict the convergence of a with respect to the size of the domain N, where for each case we used a Monte Carlo method with approximately 60 000 simulations.

The condition of $\tilde{A}_y^\rho \in C^0(\Gamma; \mathbb{R}^{2\times 2})$ is fulfilled and thus the use of the collocation methods is justified. Now, let us discuss in more detail the performance of a sparse grid compared to the Monte Carlo method. For these results I used the Clenshaw-Curtis formulas.

N	a
3	4.8929
4	4.8916
5	4.8311
6	4.8476
7	4.8158
11	4.8026
15	4.8029
20	4.8071
30	4.8047

Table 5.1: Convergence of a w.r.t the size of the domain

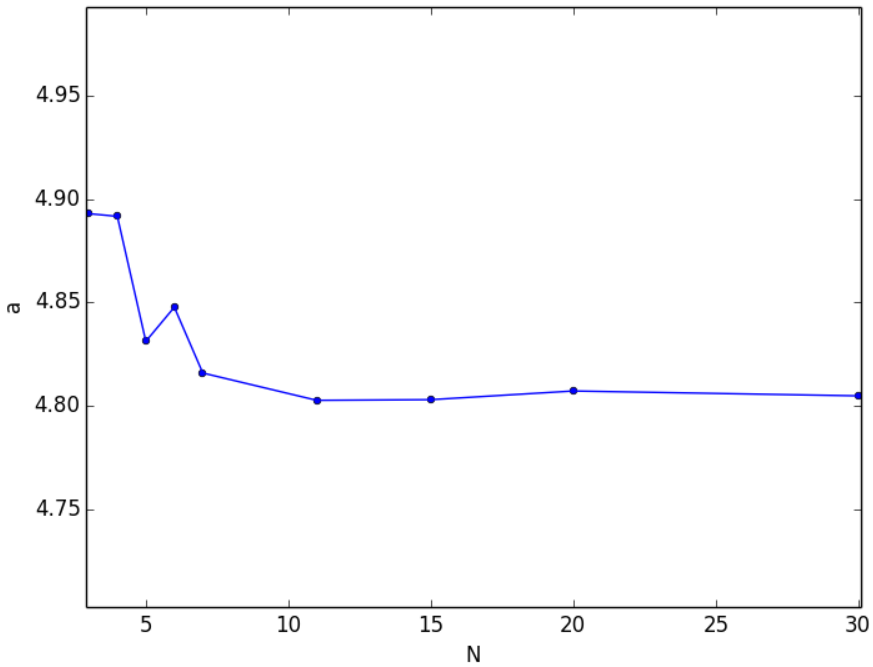


Figure 5.2: Convergence of a w.r.t the size of the domain

q of sparse grid	number of colloc. points	error of sparse grid	error of MC
2	19	0.0771	0.5568
3	181	0.0062	0.0247
4	1177	0.0020	0.0164

Table 5.2: Comparison of Clenshaw-Curtis sparse grid and Monte Carlo approach for the case of 9 random variables

q of sparse grid	number of colloc. points	error of sparse grid	error of MC
2	99	0.1020	0.1045
3	4901	0.0192	0.0114
4	161897	0.0026	0.0015

Table 5.3: Comparison of Clenshaw-Curtis sparse grid and Monte Carlo approach for the case of 49 random variables

5.1.2 $N = 3$

Since there is no well-known analytic solution for this problem that we could compare our results to, I computed a mean value by 100 000 Monte Carlo simulations and declared it as a reference solution.

The reference solution:

$$A_{3,ref} = \begin{pmatrix} 4.89578569165 & 0.000195380006229 \\ 0.000195380006229 & 4.89511273871 \end{pmatrix}$$

From the central limit theorem we know that the correct $\tilde{A}^\rho$ is normally distributed having $A_{3,ref}$ as mean, which partly justifies comparing new results to $A_{3,ref}$.

As can be seen in Table (5.2), sparse grid outperforms the Monte Carlo approach.

5.1.3 $N = 7$

Now let's have a look at the case of having 49 random variables. We followed a similar procedure as before, ran 1 350 000 Monte Carlo simulations and provided a reference solution $A_{7,ref}$

$$A_{7,ref} = \begin{pmatrix} 4.815820754415 & -2.3003151235000000 \cdot 10^{-6} \\ -2.3003151235000000 \cdot 10^{-6} & 4.81605258891 \end{pmatrix}.$$

The performance of sparse grid versus Monte Carlo can be seen in Table 5.3

q of sparse grid	number of colloc. points	error of sparse grid	error of MC
3	64969	0.0024	0.00464

Table 5.4: Comparison of Clenshaw-Curtis sparse grid and Monte Carlo approach for the case of 36 random variables

When looking at the results we see similar performance for both approaches. An important element to observe is the rapid increase in the number of collocation points when stepping up the level q. This brings a huge disadvantage for sparse grids when the number of random variables is moderately high. Therefore in the case $N = 7$ we can affirm that Monte Carlo would be a smarter choice.

5.1.4 $N = 6$

For completeness I examined the case $N = 6$, where the reference solution computed by 400 000 simulations of Monte Carlo is given by

$$A_{6,ref} = \begin{pmatrix} 4.84767532e+00 & -2.8551600 \cdot 10^{-4} \\ -2.8551600 \cdot 10^{-4} & 4.84753644e+00 \end{pmatrix}.$$

Based on the results in Table 5.3 and 5.5 we can see that for this particular application it would be smarter to use sparse grid for the case $N \leq 6$. For higher number of random variables the sparse grid collocation method faces the curse of dimensionality and from the methods we consider there is not a better choice other than the Monte Carlo. This result meets our expectations.

5.2 Autocorrelation function problem

The second problem we present is a highly frequent problem inspired by engineering applications. It concerns a material coefficient matrix of the form

$$A(x, \omega) = \begin{pmatrix} a(x, \omega) & 0 \\ 0 & a(x, \omega) \end{pmatrix}, \tag{5.3}$$

where the random field $g(x, \omega) = \log\big(a(x, \omega) - a_{min}\big)$ is characterized by its autocorrelation function $cov[g] : \overline{Q \times Q} \to \mathbb{R}$.

From the theory of stochastic processes we know that under 2 assumptions

- $cov[g]$ is a continuous function
- $g \in L^2(Q, \Omega)$,

there is a well-defined Karhunen-Loève expansion (KL expansion) as proved in [Alexanderian, 2015, p. 7,8].

By choosing the first M terms in the expansion correspondent to the M highest eigenvalues we approximate the infinite KL expansion by a finite number of terms, i.e. we end up having M uncorrelated random variables with zero mean value and unit variance.

In our application we look at the case $Q = [-0.5; 0.5]^2$ and

$$a_{min} = 1$$
$$cov[g](x, y) = \exp(\frac{-(x^2 + y^2)}{0.1}), \quad x, y \in Q$$

which makes $a(x, \omega)$ to be of the form

$$a(x, \omega) = 1 + 10 \exp(\sum_{n=1}^{M} b_n(x) Y_n(\omega)), \tag{5.4}$$

where Y_n are independent random variables normally distributed with zero mean value and unit variance.

In our application we assume both the material coefficient matrix $A(x, \omega)$ and the autocorrelation function $cov[g]$ to be $[-0.5; 0.5]^2$-periodic. We have to realize that for $A(x, \omega)$ defined by (5.3) and (5.4) it is not possible to express it in terms of an ergodic dynamical system, which ensures the existence of a homogenized matrix in a stochastic setting as was described in Theorem 2.3.

However, if $E(\tilde{A}^\rho)$ was to converge to a constant matrix, it would have to be $E(\tilde{A}^{0.5})$, since anytime $\rho = k0.5,\ k \in \mathbb{N}$ we get $E(\tilde{A}^\rho) = E(\tilde{A}^{0.5})$ thanks to the periodicity of A and $cov[g]$.

We also will not be able to satisfy the condition $\mathcal{A} \in L^\infty(\Omega)$, since the value of a can be arbitrarily large. However, the needed rate of growth of a in $y \in \Gamma$ and decay of δ (4.4.2) are preserved. The assumptions on the smoothness of A in both the spatial and random variables are satisfied.

5.2.1 $M = 8$

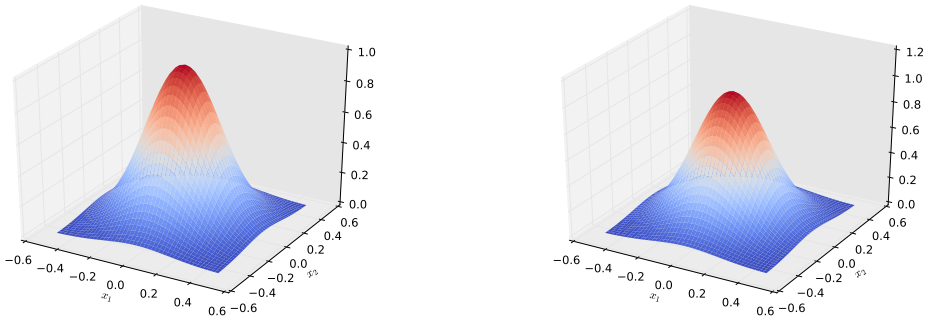
In this section we focus on the problem of approximating the KL expansion by 8 terms. Figure 5.3 depicts a comparison of the autocorrelation function before and after approximation by 8 terms.

The Fourier-Galerkin method provides fast convergence properties for A smooth and it suffices to use 25 (5×5) discretization points. We will compare 3 methods, Monte Carlo method, full tensor grid method and Gaussian sparse grid method. Since the considered random variables are normally distributed, for creating of the full tensor and the sparse grid we used the Hermite polynomials.

The sought matrix $E(\tilde{A}^{0.5})$ will be of the same form as the matrix in (5.2). We will focus on the convergence of the quantity a.

Results can be seen in Table 5.5 and Figures 5.4, 5.5. Note that the values in the Table 5.5 are just to see the differences clearly. The real value is obtained by

$$a = 11.01 + r \cdot 10^{-3}, \tag{5.5}$$



(a) before approximation (b) approximated by 8 terms

Figure 5.3: Autocorrelation function before and after approximation by 8 terms of KL expansion

q of sparse grid	number of colloc. points	full tensor grid r	sparse grid r	Monte Carlo r
2	145		5.4412	6.8149
	256	5.0605		4.7965
3	849		5.4449	4.6931
4	3937		5.4449	6.31785
	6561	5.4465		3.2162
	65536	5.4449		4.9850
	100000			5.1766

Table 5.5: Comparison of 3 different methods for solving the autocor. problem with 8 random variables where the sought quantity $a = 11.01 + r \cdot 10^{-3}$

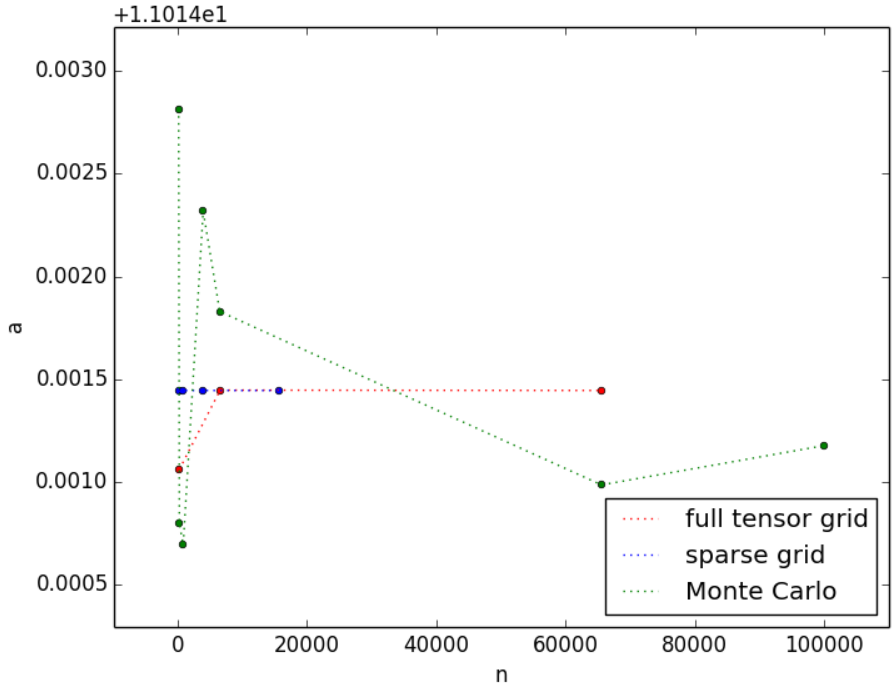


Figure 5.4: Comparison of convergence of all 3 methods for the autocor. problem with 8 random variables

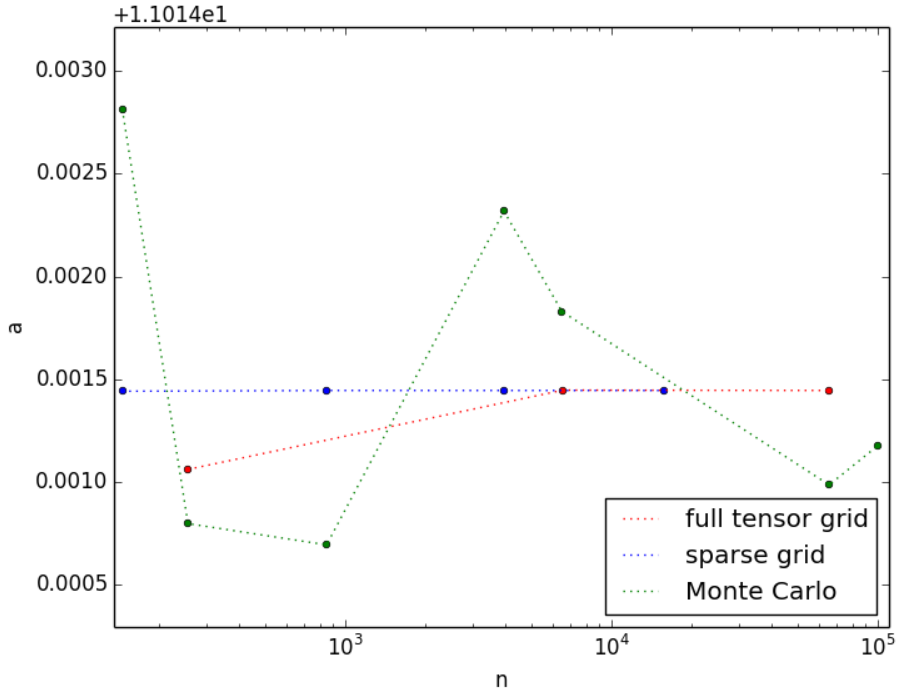


Figure 5.5: Comparison of convergence of all 3 methods for the autocor. problem with 8 random variables with logarithmic scale in the x-axis

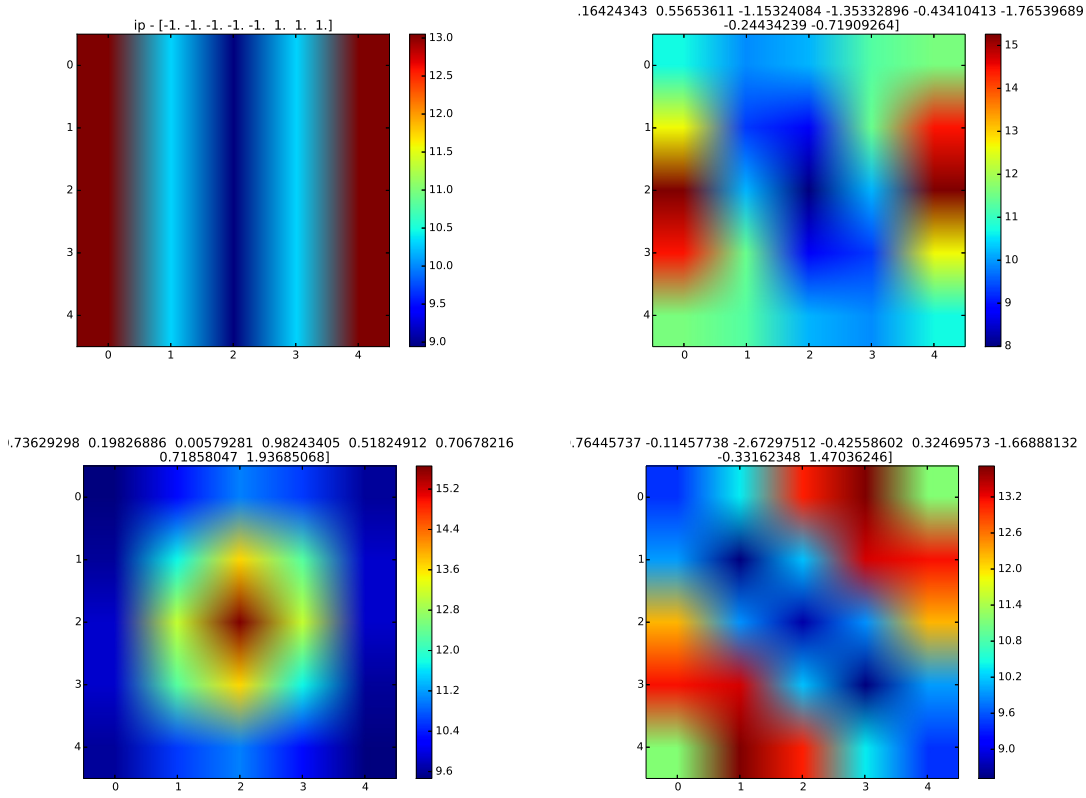


Figure 5.6: The realizations of $a(x, \omega)$ for collocation points specified at the top

where r is the value from the Table 5.5, i.e. the range is $[11.0132162; 11.0168149]$.

As we can observe, the sparse grid method was able with even the lowest number of collocation points quite accurately approximate the searched quantity. The tensor grid reached the same value with 65536 collocation points, i.e. having 4 points in each dimension. As for the Monte Carlo approach, the method converges very slowly and even after 100 000 simulations we would not be able to conclude a result.

We also include pictures illustrating the realizations of $a(x, \omega)$ for particular choices of the collocation points $y(\omega) \in \Gamma$ specified in the top of each picture.

Conclusion

We formulated the theory that stands behind some of the results of homogenization. The two most limiting assumptions are the periodicity in the deterministic case and the ergodicity in the stochastic case, which as could have been seen in the last chapter, are not always satisfied. We analyzed the Fourier-Galerkin method and the collocation method based on full tensor grids and sparse grids. The applications were focused on the comparison of all of the collocation methods to the Monte Carlo method, as a representative of the nowadays most widely used approach.

The collocation method offers several advantages:

- It decouples the problem into a number of deterministic problems, each of which can be computed by the Fourier-Galerkin method, and thus offers a trivial way of parallelization.
- It is able to easily deal with unbounded random variables assuming a rate of decay (Gaussian or exponential ones).
- It treats efficiently the case with nonindependent random variables by introducing an auxiliary density.

When computing the applications from Chapter 5, we highly benefited from the first advantage. Since the collocation method decouples the problem into a number of deterministic problems, we could use a solver for the deterministic part as a black box and apply it separately to each of these problems. In this thesis, I used a Python Fourier-Galerkin solver [Vondřejc, 2016–2017].

From the computed results we can see that the collocation method, in particular the sparse grid method, is an efficient tool when dealing with a problem that does not depend on a too large number of random variables, since the method suffers from the curse of dimensionality. If the problem involved too many random variables, we could have seen that the Monte Carlo outperformed the collocation method. If trying to identify the critical number of random variables, we would have to refer the reader to [Nobile et al., 2008].

Bibliography

A. Alexanderian. A brief note on the Karhunen-Loève expansion. North Carolina State University, `https://arxiv.org/abs/1509.07526`, September 2015.

A. Alexanderian, M. Rathinam, and R. Rostamian. Homogenization, symmetry, and periodization in diffusive random media. *Acta Mathematica Scientia*, 32 (1):129–154, 2012.

A. Anantharaman, R. Costauec, C. Le Bris, F. Legoll, and F. Thomines. Introduction to numerical stochastic homogenization and the related computational challanges: some recent developments. In W Bao and Q. Du, editors, *Multiscale modeling and analysis for materials simulation*, pages 198–268. World Scientific, Singapore, 2011.

I. Babuška, F. Nobile, and R. Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 45(3):1005–1034, 2007.

A. Bourgeat and A. Piatnitski. Approximations of effective coefficients in stochastic homogenization. *Ann. I. H. Poincaré*, 40(3):153–165, 2004.

D. Cioranescu and P. Donato. *An Introduction to Homogenization*. Oxford Univeristy Press, 1999. ISBN 0 19 856554 2.

K. Dajani and S. Dirksin. A simple introduction to ergodic theory. University of Utrecht, Lecture notes in Ergodic Theory, December 2008.

N. Dunford and J. T. Schwartz. *Linear Operators*. Wiley-Interscience, New York, 1988. ISBN 978-0471608486.

L. C. Evans. *Partial differential equations*. Graduate Studies in Mathematics. American Mathematical Society, 2010. ISBN 978-0821849743.

M. Heida. Stochastic homogenization of rate-independent systems and applications. *Continuum Mechanics and Thermodynamics*, 29:853–894, 2017.

V. V. Jikov, S. M. Kozlov, and O. A. Oleinik. *Homogenization of Differential Operators and Integral Functionals*. Druhé opravené vydání. Springer-Verlag Berlin Heidelberg, 1994. ISBN 978-3-642-84661-8.

F. Nobile, R. Tempone, and C. G. Webster. A sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 46(5):2309–2345, 2008.

J. Vondřejc. FFT-based method for homogenization of periodic media: Theory and applications. *Doctoral thesis. Department of Mechanics, Faculty of Civil Engineering, Czech Technical Univeristy*, page 143, 2013.

J. Vondřejc. FFTHomPy. `https://github.com/vondrejc/FFTHomPy`, 2016–2017.

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

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

List of Figures

List of Tables

Appendix

We present a collection of functions developed in Python for creating sparse grids $\mathcal{H}(q, M)$ (recall 4.26) and corresponding weights for an arbitrary level q and arbitrary dimension M. The level, dimension and the used orthogonal polynomials are specified as input parameters.

We are giving more details on the main function *generator(level, dim, rule)*.

`def generator(level, dim, rule):`

INPUT:
`level (int)`: The accuracy level q in the Smolyak formula (4.24).
`dim (int)`: Dimension M in the Smolyak formula (4.24).
`rule (string)`: Quadrature rule, 3 options: "Clenshaw-Curtis", "Legendre", "Hermite".

OUTPUT:
`colloc points(array)`: 2-D array of coordinates of collocation points, shape $|\mathcal{K}| \times M$, $|\mathcal{K}|$ denoting the number of collocation points.
`weights(array)`: 1-D array of weights, length $|\mathcal{K}|$.

The generator recognizes only three types of quadrature rules but adding a new one is an easy extension of the code provided below. The computation of the collocation points and the weights follows the formula (4.25).

```
import itertools
import numpy as np
import math
import scipy.special

def generator(level, dim, rule):
    """Computes the collocation points and weights and returns
    an array of arrays with coordinates of collocation points in rows
    and array of corresponding weights.

    Args:
        level (int): The accuracy level q in the Smolyak formula.
        dim (int): Dimension M in the Smolyak formula.
        rule (string): Quadrature rule.

    Returns:
        colloc_points(array): 2-D array of coordinates
            of collocation points, i.e. array
            of dim-dimensional arrays.
        weights(array): 1-D array of weights.

    """

    multiindices = compute_multiindices(level, dim)
```

```
colloc_points = []
weights = []
tmp_colloc_points = []
tmp_weights = []

if (rule == 'Legendre'):
    for multiindex in multiindices:
    # multiindex stores combination of orders
        tmp_coords = []
        tmp_weighties = []

        for index in multiindex:
            [points, weighties] = np.polynomial.legendre.leggauss(
            m(index))
            tmp_coords.append(points)
            tmp_weighties.append(weighties)

        tmp_tensorcoords = itertools.product(*tmp_coords)
        tmp_tensorweights = itertools.product(*tmp_weighties)

        for c in tmp_tensorcoords:
            tmp_colloc_points.append(c)

        summ = sum(multiindex)
        tmp = scipy.special.binom(dim - 1, summ - level - 1)
        tmp *= (-1)**(level - summ + dim)

        for b in tmp_tensorweights:
            tmp_weights.append([np.prod(b) * tmp])

elif (rule == 'Hermite'):
    for multiindex in multiindices:
        tmp_coords = []
        tmp_weighties = []

        for index in multiindex:
            [points, weighties] = np.polynomial.hermite_e.hermegauss(
            m(index))
            tmp_coords.append(points)
            tmp_weighties.append(weighties)

        tmp_tensorcoords = itertools.product(*tmp_coords)
        tmp_tensorweights = itertools.product(*tmp_weighties)

        for c in tmp_tensorcoords:
            tmp_colloc_points.append(c)

        summ = sum(multiindex)
        tmp = scipy.special.binom(dim - 1, summ - level - 1)
        tmp *= (-1)**(level - summ + dim)

        for b in tmp_tensorweights:
            tmp_weights.append([np.prod(b) * tmp])


elif (rule == 'Clenshaw-Curtis'):
    for multiindex in multiindices:
    # multiindex stores combination of orders
```

```
            tmp_list_ind = []
            # tmp_list_ind stores lists of indices of points

            for index in multiindex:
                tmp_list_ind.append(range(1, m(index) + 1))

            tmp_tensor_ind = itertools.product(*tmp_list_ind)
            # tmp_tensor_ind - the tensor product of indices from tmp_list_ind

            summ = sum(multiindex)
            tmp = scipy.special.binom(dim - 1, summ - level - 1)
            tmp *= (-1)**(level - summ + dim)

            for comb in tmp_tensor_ind:
                coords = []
                weight = [1.]
                i = 1
                for j in comb:
                    coords.append(round(coord_CC(j,multiindex[i - 1]),11))
                    weight[0] = weight[0] * weight_CC(j, multiindex[i - 1])
                    i += 1
                weight[0] *= tmp
                tmp_colloc_points.append(coords)
                tmp_weights.append(weight)

    else:
        raise Exception("Quadrature rule not recognised!!")

##
##  removing nodes that occur more than once
##

    a = np.array(tmp_colloc_points)
    b = np.array(tmp_weights)
    c = np.hstack((a, b))

    b = c[np.lexsort(np.fliplr(c).T)]

    colloc_points = [b[0][:dim]]
    weights = [b[0][dim]]

    for i in range(1, len(b)):
        if (is_equal_vector(b[i][:dim], colloc_points[-1])):
            weights[-1] += b[i][dim]
        else:
            colloc_points.append(b[i][:dim])
            weights.append(b[i][dim])

    colloc_points = np.array(colloc_points)
    weights = np.array(weights)
    return colloc_points, weights
```

The rest of the functions are minor codes called from within the function *generator(level, dim, rule)*. We will not provide more details, however, Docstrings are available in the codes below.

```
def compute_multiindices(level, dim):
    """Computes all multiindices i = (i_1,...i_dim), s.t.
    i_j >= 1; level - dim + 1 <= sum_{j=1}^{dim}(i_j - 1) <= level.

    Args:
        level(int): Determines the values of the multiindex.
        dim(int): Determines the size of the multiindex.

    Returns:
        alpha(array): List of all multiindices satisfying formula above.

    """

    comb_number1 = int(scipy.special.binom(level + dim, dim))
    alpha = [[0] * dim for i in range(comb_number1 + 1)]


    if (level > 0):
        for i in range (1, dim + 1):
            alpha[i][i - 1] = 1

    if (level > 1):
        r = dim
        mat = [[0] * level for i in range(dim)]
        for i in range(0, dim):
            mat[i][0] = 1
        for k in range(1, level):
            L = r
            for i in range (0, dim):
                summ = 0
                for m in range (i, dim):
                    summ += mat[m][k - 1]
                mat[i][k] = summ
            for j in range(0, dim):
                for m in range(L - mat[j][k] + 1, L + 1):
                    r = r + 1
                    for i in range(0, dim):
                        alpha[r][i] = alpha[m][i]
                    alpha[r][j] = alpha[r][j] + 1

    comb_number2 = int(scipy.special.binom(level, dim) - 1)

    for i in range(comb_number2 + 1, comb_number1):
        for j in range(dim):
            alpha[i][j] += 1

    return alpha[comb_number2 + 1: comb_number1][:]
```

```
def weight_CC(ind, i):
    """Computes weight of ind-th point of i-th level of accuracyby
    Clenshaw-Curtis rule.

    Args:
        ind(int): Index of the point.
        i(int): Level of accuracy.

    Returns:
```

```
        float: Weight.

    """
    if (i == 1):
        return 2.
    else:
        nm = m(i)
        if ((ind == 1) or (ind == nm)):
            return (1 / (nm * (nm - 2.)))
        else:
            summ = 0
            for k in range (1, (nm - 3) / 2 + 1):
                summ += (1. / (4. * k * k - 1.))        \
                         * math.cos((2. * math.pi * k   \
                         * (ind - 1.)) / (nm - 1.))
            return (2. / (nm - 1.)) * (1. - ((math.cos(math.pi * (
                    ind - 1.))) / (nm * (nm - 2.))) - 2. * summ)
```

```
def coord_CC(ind, i):
    """Computes coordinate for Clenshaw-Curtis rule.

    Args:
        ind(int): Index of the point.
        i(int): Level of accuracy.

    Returns:
        float: Coordinate.

    """
    if (i == 1):
        return 0.
    else:
        return -math.cos((math.pi * (ind - 1)) / (m(i) - 1))
```

```
def m(i):
    "Computes the number of points for level i in 1-D."
    if (i <= 0):
        return 0
    if (i == 1):
        return 1
    else:
        return 2**(i - 1) + 1
```

```
def is_equal_vector(a, b):
    for i in range(len(a)):
        if (not is_equal(a[i], b[i])):
            return False
    return True
```

```
def is_equal(a, b):
    eps = 1e-12
    if ((a > (b - eps)) and (a < (b + eps))):
        return True
    else:
        return False
```