TECHNISCHE
UNIVERSITÄT
WIEN
Vienna University of Technology

DIPLOMARBEIT

ERROR ESTIMATES AND OPTIMAL APPROACHES TO THE STOCHASTIC HOMOGENIZATION IN ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS

Ausgeführt am Institut für Analysis und Scientific Computing
der Technischen Universität Wien

unter der Anleitung von Prof. Dr. Clemens Heitzinger

durch

Caroline Geiersbach
Weyrgasse 7/16, 1030 Wien

10. Dezember 2015

Abstract

This thesis is focused on the numerical treatment of stochastic homogenization in elliptic partial differential equations. A major application, and one that we keep in mind in this work, can be found in materials science, where one is interested in obtaining an average conductivity for a composite containing a fine microscopic structure. Such problems are also called "multiscale" due to the differences in (length) scales that are of interest. Determining an average conductivity requires first resolving at least part of the fine structure, which we can use to estimate the conductivity on the macroscopic scale. A central issue in numerically solving a homogenization problem comes from the fact that in order to minimize the error, one would need to determine conductivity on the microscale for the entire body; this is however not possible given the sheer size of such systems. One needs to therefore content oneself with a sample of the material and use this information to compute an average conductivity. This can be accomplished by first solving the so-called "cell problem." The essential contribution in this thesis is the estimation of the error of the cell problem, which we express as a function of domain size, mesh fineness and number of samples. We will quantify the work needed to solve this problem and then present an optimal approach to solving the problem.

Zusammenfassung

Die vorliegende Diplomarbeit beschäftigt sich mit der stochastischen Homogenisierung von elliptischen partiellen Differentialgleichungen. Homogenisierung findet vor allem in den Materialwissenschaften Anwendung. Eine Fragestellung dieses Gebiets, auf welche wir uns in dieser Arbeit konzentrieren, ist die Berechnung der mittleren Leitfähigkeit eines Verbundwerkstoffs mit einer feinen mikroskopischen Struktur. Solche Probleme nennt man "Mehrskalenprobleme", da die involvierten Größenskalen sehr unterschiedlich sind. Um die makroskopische mittlere Leitfähigkeit zu berechnen, muss zuerst die heterogene mikroskopische Struktur aufgelöst werden. Von zentraler Bedeutung ist die Minimierung des numerischen Fehlers für das Homogenisierungsproblem. Um diesen zu minimieren, müsste man eigentlich das numerische Problem für das ganze Gebiet in der mikroskopischen Skala lösen. Jedoch ist der numerische Aufwand für dieses Problem wegen der bloßen Größe der zu untersuchenden Objekte nicht vertretbar. Stattdessen wird das Problem auf eine kleinere repräsentative Zelle reduziert, worauf eine Approximation der mittleren Leitfähighkeit berechnet wird. Dafür muss das sogennante "Zellproblem" gelöst werden. Das wesentliche Resultat dieser Arbeit sind Fehlerabschätzungen für das Zellproblem als Funktion der Gebietsgröße, Gitterweite und Anzahl der verwendeten Stichproben. Weiters wird der numerische Aufwand für die Lösung des Problems quantifiziert und eine optimale Herangehensweise präsentiert.

Contents

CHAPTER 1

Introduction

1.1 Introductory Examples

In this work, we will be concentrating on a body of problems that share the common property of possessing multiple scales. Such *multiscale* problems can be found in numerous physics and engineering applications, where an object has a geometrical size that is many orders of magnitude larger than the atoms and molecules it contains. Multiscale problems can also refer to different time scales: for example, the behavior of atoms occur on a time scale of femto-seconds (10^{-15} second) [10, p. 6], whereas the time interval of interest may be closer to one second. Most systems possessing multiple scales can be adequately approximated without taking its microscopic properties into account, but in some cases, the microscopic properties play a decisive role in their macroscopic behavior. For such problems, the goal is to describe the average behavior of the system, taking details from all length or time scales into account.

A large class of multiscale problems is focused on describing the average properties of heterogeneous materials, or materials that consist of different phases, such as composites or polycrystals. Processes for such materials are generally described by differential equations with coefficients that rapidly oscillate between the phases. Media possessing a microstructure, where the length scale of one or more phases is much smaller than the overall material, is a particular kind of heterogeneous material that presents its own challenges. An example of such a medium with a microstructure can be seen in Figure 1.1, which shows an example of a SEM micrograph of a composite. The behavior of such a medium at the macroscopic level is often quite different from that at the microscopic level, yet resolving the fine structure is critical to understanding the overall behavior of the medium. However, it is not feasible to resolve the microscopic behavior in its entirety; the behavior is too complex and contains too much information that is of

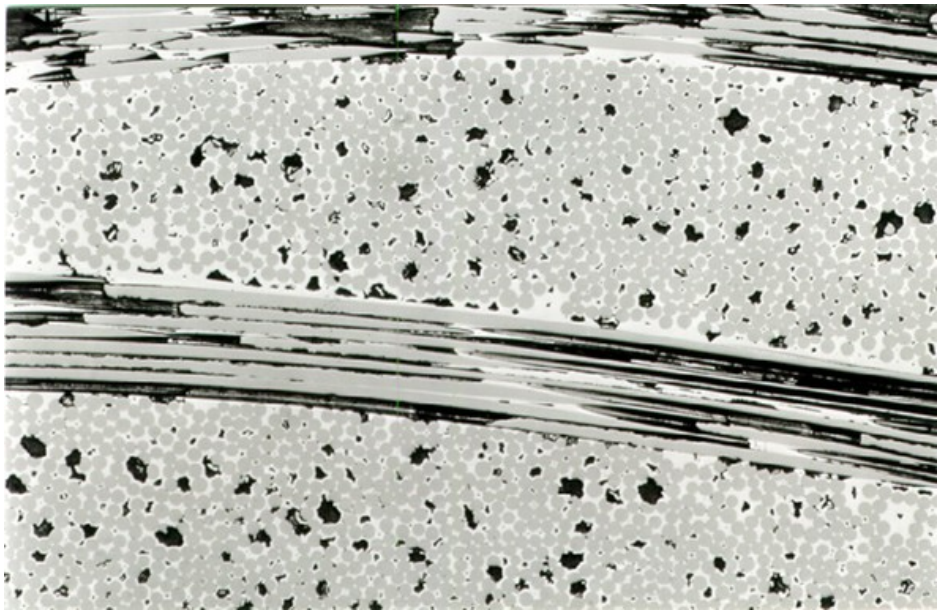


Figure 1.1: SEM micrograph showing detail of a ceramic matrix SiC/SiC composite, manufactured via a CVI-process. The micrograph reveals a structure possessing multiple scales. Used with permission by MT Aerospace AG, Augsburg via Wiki Commons.

no interest to the problem at hand. In these cases, it is often enough to analyze one or more samples of the microstructure in order to characterize the material as a whole. Homogenization is a mathematical tool that is necessary to predict properties of a composite if the microstructure is known. With the same tool, new materials with desired properties that optimize the composite's performance can be modeled.

The enormity of the problem in characterizing such materials becomes apparent when one considers the number of variables that demonstrably affect a material's behavior. Let us consider a simple example of a two-phase material that is comprised of grains or inclusions and a matrix surrounding the grains. Say we are interested in the elasticity of the material at the macroscopic level. Clearly, a number of variables could influence this behavior; a few variables include: the size of the inclusions compared to macroscopic scale, the distribution of inclusion size, the density of the inclusions, clustering of inclusions, the degree of difference between the properties of the matrix versus the inclusions, and the effect of the grains' orientation (in the anistropic case). When one considers a model for the macroscopic behavior under an applied force, more questions surface: how large must our microscale sample be in order to give us an understanding of how this force impacts the material as a whole? What sort of boundary conditions should we impose on the microstructure that will minimize the error in the macroscopic problem? If we are taking multiple samples of a material to characterize the average behavior, how many samples do we need?

Periodic problems were among the first to be well-understood given their simple

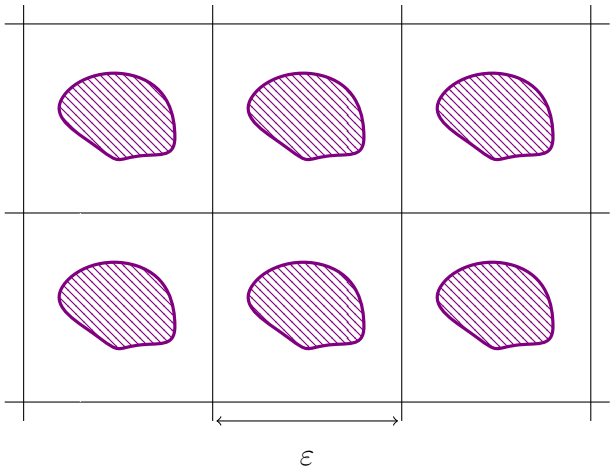


Figure 1.2: Cross-section of a periodic composite cut across the fibers.

structure; these are structures that have repeating cells as in Figure 1.2. Early contributors to this field were Bakhvalov [3] and Berdichevskii [5]. These models are rather unrealistic, however. Early, more general, results for the non-periodic case can be found in the works of Kozlov [19], Yurinskii [33], and Papanicolaou and Varadhan [27], who arrived at similar results, albeit by somewhat different methods.

The question of how large a sample must be is of particular importance. In applications, a sample that is "large enough" in the sense that it is statistically representative of the behavior of the entire body is known as a *representative volume element* (RVE). In an extracted rock core, it may contain information about the pore scale distribution, for instance. In order for it to be representative of the whole, it must contain a large number of composite microheterogeneities, but this of course leads to much larger domains, which can be computationally expensive. Figure 1.3 shows a simulation of a cross-section of a fibrous material and a sample in higher resolution. Clearly, there is a limit to how small a RVE can be for this material, and it depends not only on the fiber size, but also the density. There is not yet a unified approach to determine RVEs for an arbitrary material, but much progress has been made on this topic; see for example the papers by [21] or [26].

An important question in multiscale methods is how the boundary condition chosen for the microproblem affects the error of the solution. The boundary conditions used in the microproblem are in a certain sense artificial; they are needed due to the fact that the computational domains are truncated and localized. In [6], Bourgeat and Piatnitskii examined the use of periodization and other "cut-off" procedures to approximate a RVE for second-order elliptic operators in divergence form. They were able to approximate convergence rates when an additional mixing condition was assumed. Yue and E [32] studied the convergence rates numerically for the same class of problems. The authors tested several benchmark problems with periodic, Dirichlet, and Neumann boundary conditions; it was found that while all three boundary conditions perform reasonably well, periodic boundary conditions generally perform best.

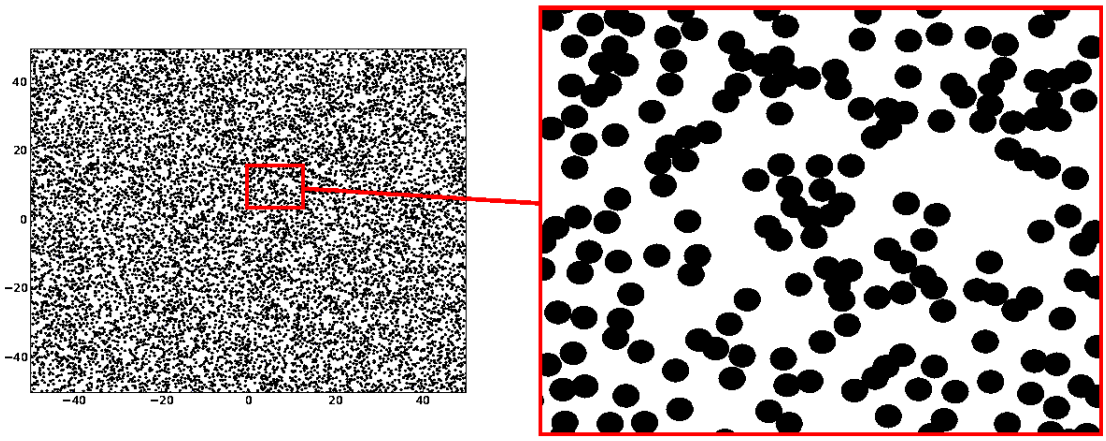


Figure 1.3: Cross-section of a fibrous material, zoomed in 10 times to reveal the microstructure. The size of the sample plays an important role in multiscale methods.

In the following pages, we will be focusing on characterizing and modeling materials with a specific subset of physical laws, namely those that can be described using elliptic partial differential equations. While we shall focus on systems containing two widely separated scales, methods described here can be extended in a natural way to deal with systems involving many scales. To begin, let us look at two major applications belonging to this class of problems.

Example 1.1. (2D Thermal Field in a Composite [4]) Modeling a thermal field in a composite amounts to solving the elliptic problem

$$\frac{\partial}{\partial x_1}\left(A_\varepsilon(x_1,x_2)\frac{\partial u}{\partial x_1}\right)+\frac{\partial}{\partial x_2}\left(A_\varepsilon(x_1,x_2)\frac{\partial u}{\partial x_2}\right)=f(x_1,x_2)\quad\forall(x_1,x_2)\in D, \tag{1.1a}$$

$$u=0\qquad\forall(x_1,x_2)\in\partial D \tag{1.1b}$$

where $u(x_1,x_2)$ quantifies the temperature at the point (x_1,x_2) in the 2D plane, $f(x_1,x_2)$ is a smooth function describing the density of heat sources in the composite, and A_ε is the conductivity coefficient of the composite, which rapidly oscillates on the microscale between two real values:

$$A_\varepsilon(x_1,x_2)=\begin{cases}A_1, & \text{if } (x_1,x_2) \text{ is in a fiber,}\\ A_2, & \text{if } (x_1,x_2) \text{ is in the matrix surrounding the fiber.}\end{cases}$$

The continuity conditions $[u]=0$ and $\left[A_\varepsilon\frac{\partial u}{\partial n}\right]=0$ are also imposed, where the notation $[w]$ means the difference between values of w on two sides of a surface. One is interested in obtaining an equation with an "averaged" $\bar{A}$ that represents conductivity for the macroscale.

Example 1.2. (Linear Elasticity Problem [30]) Consider a problem in elastostatics, where a body $D=D_1\cup D_2\subset\mathbb{R}^3$ is clamped on one side, denoted by ∂D_1,

and where a force $\boldsymbol{f}$ is applied to the other side, denoted by ∂D_2. If $\boldsymbol{u}$ denotes the displacement of the body, then the strain tensor $\boldsymbol{\varepsilon}$ can be expressed as

$$\boldsymbol{\varepsilon}_{ij} = \frac{1}{2}\left(\frac{\partial \boldsymbol{u}_i}{\partial x_j} + \frac{\partial \boldsymbol{u}_j}{\partial x_i}\right).$$

A generalization of Hooke's Law gives a linear relation between the stress tensor $\boldsymbol{\tau}$ and the strain tensor $\boldsymbol{\varepsilon}$ such that

$$\boldsymbol{\tau}_{ij} = \mathbf{C}_{ijkh} : \boldsymbol{\varepsilon}_{kh}.$$

Here, the Einstein summation convention is used, where repeated indices are summed, meaning $c_i x_i$ represents the sum $\sum_{i=1}^{n} c_i x_i$. Assume that $\mathbf{C}$ is symmetric, i.e. $\mathbf{C}_{ijkh} = \mathbf{C}_{jikh} = \mathbf{C}_{ijhk} = \mathbf{C}_{khij}$, and positive in the sense that

$$\mathbf{C}_{ijkh}\boldsymbol{\varepsilon}_{ij}\boldsymbol{\varepsilon}_{kh} \geq \alpha \boldsymbol{\varepsilon}_{ij}\boldsymbol{\varepsilon}_{ij} \qquad \alpha > 0 \quad \forall \boldsymbol{\varepsilon}_{ij}.$$

If $\boldsymbol{f}_i$ are the components of body forces and $\boldsymbol{F}_i$ the components of surface forces, then the governing equations are given by

$$\frac{\partial \boldsymbol{\tau}_{ij}}{\partial x_j} + \boldsymbol{f}_i = 0 \qquad \text{in } D, \tag{1.2a}$$

$$\boldsymbol{u}_i = 0 \qquad \text{on } \partial D_1, \tag{1.2b}$$

$$\boldsymbol{\tau}_{ij}\boldsymbol{n}_j = \boldsymbol{F}_i \qquad \text{on } \partial D_2 \tag{1.2c}$$

with $\boldsymbol{n}$ denoting the unit outer normal vector to ∂D. Under the assumption that the coefficients $\mathbf{C}_{ijkh}$ are periodic and rapidly oscillating, a homogenized stress-strain relation of the form

$$\bar{\boldsymbol{\tau}} = \mathbf{C}^h_{ijkh} : \bar{\boldsymbol{\varepsilon}}_{kh}$$

can be constructed [30]. This new constitutive relation is used to construct the "average" equation on the macroscale.

1.2 Overview of Homogenization

Up until now, we have used the terms "averaged equations" and "averaged coefficients" to gain an intuitive idea of homogenization. We will now be more precise. In Example 1.1, we defined a rapidly oscillating coefficient A_ε that took one of two values depending on the location. We also discussed the problem that generally speaking, composites have inclusions that are on a much smaller scale than the size of the body itself. The difference in scales has serious practical concerns; numerically, it is not possible to resolve the microscales and macroscales at the same time. A technique for mathematically handling these concerns is called *homogenization*. Homogenization aims at obtaining *homogenized equations* with coefficients that are not rapidly oscillating, but which still yield solutions that are close to solutions to the original equation. In the homogenized equations, the original coefficients are replaced by so-called *effective coefficients*. The advantage of these

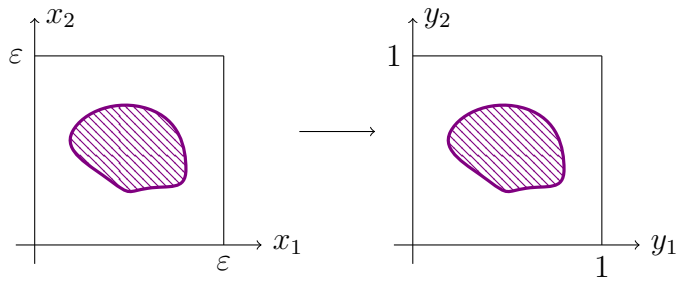


Figure 1.4: Transformation of the elementary cell.

equations is that they can more easily be solved numerically, unlike the equations with rapidly oscillating coefficients, where the mesh would need to be be very fine, at the very least at transitions between materials. The homogenized equation of Example 1.1 is

$$\frac{\partial}{\partial x_1}\left(\bar{A}(x_1,x_2)\frac{\partial u_0}{\partial x_1}\right)+\frac{\partial}{\partial x_2}\left(\bar{A}(x_1,x_2)\frac{\partial u_0}{\partial x_2}\right)=f(x_1,x_2) \quad \forall (x_1,x_2)\in D, \tag{1.3a}$$

$$u_0=0 \quad \forall (x_1,x_2)\in \partial D, \tag{1.3b}$$

where u_0 is the *homogenized solution*, which is close to the solution u of the original problem and $\bar{A}$ is the composite's effective coefficient of conductivity [4]. In the next chapter, we will derive this result in the stochastic case for a general number of dimensions d.

The central idea behind the technique is as follows: to obtain the homogenized equations, we introduce fast variables $y_i := x_i/\varepsilon$. This turns an elementary cell into one with unit length 1; see Figure 1.4. We make the assumption that our fast variables y_i are on a much smaller scale than the slow variables x_i; for this reason, homogenization is also referred to as the "method of multiple scales." We then seek a solution in the form of a formal power series in powers of the small parameter ε with coefficients depending on both x_i and y_i; in one dimension this ansatz looks like

$$u=\sum_{i=0}^{\infty}\varepsilon^i u_i(x,y). \tag{1.4}$$

This series is the substituted into the original equation and coefficients are equated that have the same powers of ε. This results in a system of decoupled equations of the same scale that can be combined to obtain the homogenized solution. In Section 2.1, we will see how this procedure works in the periodic case.

1.3 Outline of Thesis

Now we will summarize the contents of the remaining pages. In the next chapter, will derive the homogenized equation for Poisson's equation, providing the appropriate definitions and framework for the numerical work. The corrector equation

(cell problem) will be introduced, which is an essential equation in homogenization, and which will play an important role in our error analysis and numerical results. Additionally, we will introduce the effective coefficient matrix, which we obtain using the solution to the corrector equation.

In the third chapter, we will focus on the numerical solution to the corrector equation using the finite element method. We will see how sampling techniques can be used for the stochastic problem. The chapter closes with a brief survey of methods used in numerical homogenization.

In the fourth chapter, we will focus on new theoretical results obtained as part of this study. Error bounds will be derived for the numerical solution to the corrector equation and these bounds will be justified using both existing theory and numerical simulations. An optimal approach to solving our problem will be presented after quantifying work and solving an optimization problem.

Numerical results will be presented in Chapter 5. We will present a specific framework for testing and focus on a couple of benchmark equations. Solutions to both the corrector equation as well as the effective coefficient matrix will be calculated. Values for the optimization problem will be determined based on these simulations.

In Chapter 6, we will give a brief description of the implementation of our solver, which was for the most part built from scratch. Finally, opportunities for future research will be discussed in the conclusion.

CHAPTER 2

Stochastic Homogenization of Elliptic PDEs

2.1 Derivation of the Homogenized Equation in the Stochastic Case by Periodization

In this section, we will derive the homogenized equation for Poisson's equation

$$-\nabla \cdot (A_\varepsilon \nabla u_\varepsilon) = f \qquad \forall x \in D, \tag{2.1a}$$

$$u_\varepsilon = 0 \qquad \forall x \in \partial D. \tag{2.1b}$$

This problem describes several different types of physical settings defined on an open and bounded set $D \subset \mathbb{R}^d$. In diffusion, A_ε represents diffusivity or thermal conductivity and u_ε represents some quantity that diffuses, as in a concentration of a chemical solution or the temperature in a medium. Example 1.1 belongs to this class of problems. In electrostatics, A_ε represents permittivity and u_ε is the electric potential. Our focus will be on a multiscale problem in which the coefficient A_ε fluctuates rapidly on the microscale, such as in a composite or a mixture of several materials with different properties. The parameter $\varepsilon > 0$ is the ratio of the typical length scale associated with the domain D to the typical length scale associated with variations in conductivity [27]. To obtain the homogenized equation, we will combine techniques found in [27] and [28].

We will focus on the periodic and stochastic case. Let $(\Omega, \mathcal{F}, P)$ be a probability space where $\omega \in \Omega$ represents a single realization of a medium, $\mathcal{F}$ is an appropriate σ-algebra and P is a probability measure defined on $(\Omega, \mathcal{F})$. Each realization ω of a medium can be identified with a coefficient $A(x, \frac{x}{\varepsilon}, \omega) =: A_\varepsilon(x, \omega)$. We will assume that the coefficient A_ε takes the form $A(\frac{x}{\varepsilon}, \omega)$, i.e. fluctuations in the coefficient only take place on a microscale; the methods below can be extended to the more general case. A is furthermore a matrix-valued random field that is assumed to be strictly positive and bounded, meaning

$$\exists \alpha^+, \alpha^- > 0 : \quad \alpha^- |\xi|^2 \leq \xi^\top A(y, \omega) \xi \leq \alpha^+ |\xi|^2 \tag{2.2}$$

for all $y \in \mathbb{R}^d, \omega \in \Omega$, and $\xi \in \mathbb{R}^d$. We will also assume that A is 1-periodic in all d directions. Let us recall the definition of a periodic function.

Definition 2.1. A function $f : \mathbb{R}^d \to \mathbb{R}$ that satisfies

$$f(y + e_i) = f(y) \qquad \forall y \in \mathbb{R}^d, \quad \forall i \in \{1, \dots, d\}$$

where $e_i, i \in \{1, \dots, d\}$ denotes the standard basis of $\mathbb{R}^d$, is called a *1-periodic function.*

The problem (2.1) in the stochastic form is

$$-\nabla \cdot (A_\varepsilon(x, \omega)\nabla u_\varepsilon(x, \omega)) = f(x) \qquad \forall (x, \omega) \in D \times \Omega, \tag{2.3a}$$

$$u_\varepsilon(x, \omega) = 0 \qquad \forall (x, \omega) \in \partial D \times \Omega, \tag{2.3b}$$

where we assume, for simplicity, that f is deterministic. The goal is to find the homogenized equation of (2.3) for a *fixed* realization $\omega \in \Omega$.

We seek a solution in the form of a power series expansion in ε as in the ansatz (1.4). We use the idea inherent in multiscale methods: that in addition to our "slow" variable x, our solution u is dependent on another "fast" variable, which we introduce as $y := \frac{x}{\varepsilon}$. The ansatz (1.4) becomes

$$u_\varepsilon(x, \omega) = u_0\left(x, \frac{x}{\varepsilon}, \omega\right) + \varepsilon u_1\left(x, \frac{x}{\varepsilon}, \omega\right) + \varepsilon^2 u_2\left(x, \frac{x}{\varepsilon}, \omega\right) + \cdots . \tag{2.4}$$

We will treat x and y as independent variables, a main idea behind scale separation and justified in standard literature on the subject; see for example [28, Chapter 19].

The chain rule yields

$$\nabla = \nabla_x + \frac{1}{\varepsilon}\nabla_y.$$

Defining the left hand side of (2.3a) as

$$\mathcal{A}_\varepsilon := -\nabla \cdot (A_\varepsilon \nabla)$$

we see

$$\mathcal{A}_\varepsilon = \frac{1}{\varepsilon^2}\mathcal{A}_0 + \frac{1}{\varepsilon}\mathcal{A}_1 + \mathcal{A}_2,$$

where

$$\begin{aligned}
\mathcal{A}_0 &:= -\nabla_y \cdot (A(y, \omega)\nabla_y), \\
\mathcal{A}_1 &:= -\nabla_y \cdot (A(y, \omega)\nabla_x) - \nabla_x \cdot (A(y, \omega)\nabla_y), \\
\mathcal{A}_2 &:= -\nabla_x \cdot (A(y, \omega)\nabla_x).
\end{aligned}$$

The stochastic problem (2.3) can therefore be written as

$$\left(\frac{1}{\varepsilon^2}\mathcal{A}_0 + \frac{1}{\varepsilon}\mathcal{A}_1 + \mathcal{A}_2\right) u_\varepsilon = f \qquad \forall (x, y, \omega) \in D \times \mathbb{T}^d \times \Omega, \tag{2.5a}$$

$$u_\varepsilon = 0 \qquad \forall (x, y, \omega) \in \partial D \times \mathbb{T}^d \times \Omega. \tag{2.5b}$$

Now, plugging the ansatz (2.4) into (2.5), we have

$$\left(\frac{1}{\varepsilon^2}\mathcal{A}_0 + \frac{1}{\varepsilon}\mathcal{A}_1 + \mathcal{A}_2\right)\left[u_0\left(x, \frac{x}{\varepsilon}, \omega\right) + \varepsilon u_1\left(x, \frac{x}{\varepsilon}, \omega\right) + \varepsilon^2 u_2\left(x, \frac{x}{\varepsilon}, \omega\right) + \cdots\right] = f$$
$$\implies \quad \frac{1}{\varepsilon^2}\left(\mathcal{A}_0 u_0\right) + \frac{1}{\varepsilon}(\mathcal{A}_0 u_1 + A_1 u_0) + (A_0 u_2 + A_1 u_1 + A_2 u_0) + \mathcal{O}(\varepsilon) = f.$$

Comparing coefficients for ε^{-2}, ε^{-1}, and ε^0, we have, respectively:

$$\mathcal{A}_0 u_0 = 0, \tag{2.6a}$$
$$\mathcal{A}_0 u_1 = -\mathcal{A}_1 u_0, \tag{2.6b}$$
$$\mathcal{A}_0 u_2 = -\mathcal{A}_1 u_1 - \mathcal{A}_2 u_0 + f. \tag{2.6c}$$

Naturally, this process can be extended further to obtain higher-order terms. In equation (2.6a), since $\mathcal{A}_0$ contains differentials solely dependent on y, we can set $u_0(x, y, \omega) := u(x)$. Later, we will see that it is indeed sensible to assume u's independence from ω, as it will be shown that u_ε converges weakly to a u that is the solution to a deterministic equation; see Section 2.2. The system (2.6) in expanded form is

$$-\nabla_y \cdot (A(y,\omega)\nabla_y u(x)) = 0, \tag{2.7a}$$
$$-\nabla_y \cdot (A(y,\omega)\nabla_y u_1(x,\omega)) = \nabla_y \cdot (A(y,\omega)\nabla_x u(x)) + \nabla_x \cdot (A(y,\omega)\nabla_y u(x)), \tag{2.7b}$$
$$\begin{aligned} -\nabla_y \cdot (A(y,\omega)\nabla_y u_2(x,\omega)) = {} & \nabla_y \cdot (A(y,\omega)\nabla_x u_1(x,\omega)) \\ & + \nabla_x \cdot (A(y,\omega)\nabla_y u_1(x,\omega)) \\ & + \nabla_x \cdot (A(y,\omega)\nabla_x u(x)) + f. \end{aligned} \tag{2.7c}$$

Note that this system can be solved sequentially. Solvability is guaranteed thanks to the Fredholm alternative, since we know that, given our assumptions on A, the elliptic equation

$$\mathcal{A}u = f, \qquad u \text{ is 1-periodic}$$

has a solution if and only if [28, p. 184]

$$\int_{\mathbb{T}^d} f(y)\,\mathrm{d}y = 0. \tag{2.8}$$

This solution is unique up to an additive constant. We will in the following fix this constant by using the condition that the solution should vanish over the unit torus, i.e.

$$\int_{\mathbb{T}^d} u(y)\,\mathrm{d}y = 0.$$

Note that the divergence of a matrix is defined by

$$(\nabla_z \cdot A)_i := \sum_j \partial_{z_j} A_{ij}$$

so that

$$\nabla_y \cdot (A\nabla_x u) = \nabla_y \cdot \begin{pmatrix} \sum_j A_{1j}\partial_{x_j} u \\ \vdots \\ \sum_j A_{dj}\partial_{x_j} u \end{pmatrix} = \sum_{i,j} \partial_{y_i}(A_{ij}\partial_{x_j} u)$$

and

$$(\nabla_y \cdot A^\top) \cdot \nabla_x u = \begin{pmatrix} \sum_i \partial_{y_i} A_{i1} \\ \vdots \\ \sum_i \partial_{y_d} A_{id} \end{pmatrix} \cdot \begin{pmatrix} \partial_{x_1} u \\ \vdots \\ \partial_{x_d} u \end{pmatrix} = \sum_{i,j} (\partial_{y_i} A_{ij})(\partial_{x_j} u)$$

are equal, given $u = u(x)$. Hence, (2.6b) reduces to

$$\begin{aligned} \mathcal{A}_0 u_1 &= \nabla_y \cdot (A(y,\omega)\nabla_x u) + \nabla_x \cdot (A(y,\omega)\nabla_y u) \\ &= \nabla_y \cdot (A(y,\omega)\nabla_x u) \\ &= (\nabla_y \cdot A^\top(y,\omega)) \cdot \nabla_x u, \end{aligned}$$

so that we obtain

$$\mathcal{A}_0 u_1 = (\nabla_y \cdot A^\top(y,\omega)) \cdot \nabla_x u, \qquad u_1(x,\cdot,\omega) \text{ is 1-periodic}, \qquad \int_{\mathbb{T}^d} u_1(x,y,\omega) = 0.$$

Generally, for a function $g \in C^1_\#(\mathbb{T}^d)$, i.e. a function that is continuously differentiable and periodic on the torus, the divergence theorem shows

$$\int_{\mathbb{T}^d} \nabla \cdot g(y)\,\mathrm{d}y = 0. \tag{2.9}$$

Thus, the solvability condition (2.8) is again fulfilled, since

$$\int_{\mathbb{T}^d} f(y)\,\mathrm{d}y = \int_{\mathbb{T}^d} (\nabla_y \cdot A^\top) \cdot \nabla_x u\,\mathrm{d}y = \nabla_x u \cdot \int_{\mathbb{T}^d} (\nabla_y \cdot A^\top)\,\mathrm{d}y = 0,$$

as A is periodic. We make the separation ansatz

$$u_1(x,y,\omega) = \chi(y,\omega) \cdot \nabla_x u(x), \qquad \chi : \mathbb{T}^d \times \Omega \to \mathbb{R}^d$$

and we get

$$\begin{aligned} & \mathcal{A}_0(\chi(y,\omega) \cdot \nabla_x u) = (\nabla_y \cdot A^\top(y,\omega)) \cdot \nabla_x u \\ \iff \quad & (\mathcal{A}_0 \chi(y,\omega)) \cdot \nabla_x u = (\nabla_y \cdot A^\top(y,\omega)) \cdot \nabla_x u. \end{aligned}$$

Thus, if χ solves the *cell problem* (or *corrector equation*)

$$\begin{aligned} -\nabla_y \cdot (A(y,\omega)\nabla_y \chi(y,\omega)) &= \nabla_y \cdot A^\top(y,\omega), \\ \chi \text{ is 1-periodic}, \qquad \int_{\mathbb{T}^d} \chi(y,\omega)\,\mathrm{d}y &= 0, \end{aligned} \tag{2.10}$$

then $u_1 = \chi(y) \cdot \nabla_x u(x)$ is a solution to (2.6b). The field χ is called the *first-order corrector*. Note that the cell problem has the form

$$-\nabla_y \cdot (A(y,\omega)\nabla_y \chi(y,\omega)) = h(y,\omega)$$

so that we can again use similar arguments to those used in the first step to establish existence and uniqueness: if $A(y,\omega)$ is strictly positive and uniformly bounded

in y and ω, then the cell problem in its weak form has a unique solution if and only if the integral of h disappears over the torus [27]. This is indeed the case, as the integral over the right hand side vanishes on the unit torus, so the problem is well-posed. The uniqueness of the solution χ is guaranteed due to the condition $\int_{\mathbb{T}^d} \chi \, \mathrm{d}y = 0$.

To solve (2.10), we assume that $\chi \in C^1_{\#}(\mathbb{T}^d)$. The cell problem can be written in the form of one equation per component of χ [28, p. 189] as

$$-\nabla_y \cdot (A(y,\omega)\nabla_y \chi_i) = \nabla_y \cdot (Ae_i), \qquad i \in \{1,\dots,d\}. \tag{2.11}$$

Here, e_i is the i^{th} unit vector in $\mathbb{R}^d$ and Ae_i is the i^{th} column of A. Using that $e_i = \nabla_y y_i$, we can write

$$-\nabla_y \cdot (A(y,\omega)(\nabla_y \chi_i + \nabla_y y_i)) = 0.$$

We multiply this equation by the test function $\varphi \in C^1_{\#}(\mathbb{T}^d)$ and partially integrate to get the weak form

$$\int_{\mathbb{T}^d} \left\langle \nabla_y \chi_i + \nabla_y y_i, A(y,\omega)\nabla_y \varphi \right\rangle \mathrm{d}y = 0,$$

where $\langle \cdot, \cdot \rangle$ represents the scalar product. Introducing the bilinear form

$$a(u,v) := \int_{\mathbb{T}^d} \left\langle \nabla_y v, A(y,\omega)\nabla_y u \right\rangle \mathrm{d}y,$$

we see that solving the cell problem amounts to finding $\chi_i \in C^1_{\#}(\mathbb{T}^d)$, $i \in \{1,\dots,d\}$ such that

$$a(\varphi, \chi_i + y_i) = 0 \qquad \forall \varphi \in C^1_{\#}(\mathbb{T}^d). \tag{2.12}$$

Remark 2.2. The equations (2.12) can be written more compactly as

$$\int_{\mathbb{T}^d} \nabla_y \chi \cdot A(y,\omega) \cdot \nabla_y \varphi \, \mathrm{d}y = -\int_{\mathbb{T}^d} A(y,\omega) \cdot \nabla_y \varphi \, \mathrm{d}y \qquad \forall \varphi \in C^1_{\#}(\mathbb{T}^d), \tag{2.13}$$

where

$$\nabla_y \chi = \begin{pmatrix} \nabla_{y_1}\chi_1 & \nabla_{y_2}\chi_1 \\ \nabla_{y_1}\chi_2 & \nabla_{y_2}\chi_2 \end{pmatrix}$$

in the two-dimensional case.

Now we shall turn to the third equation (2.6c). The solvability condition is that the integral over the right hand side should be equal to zero, i.e.

$$\int_{\mathbb{T}^d} \mathcal{A}_2 u_0 + \mathcal{A}_1 u_1 \, \mathrm{d}y = \int_{\mathbb{T}^d} f(x) \, \mathrm{d}y = f(x),$$

since we assumed that f is independent of y. We see that

$$\begin{aligned} \int_{\mathbb{T}^d} \mathcal{A}_2 u_0 \, \mathrm{d}y &= -\int_{\mathbb{T}^d} \nabla_x \cdot (A(y,\omega)\nabla_x u(x)) \, \mathrm{d}y \\ &= -\nabla_x \cdot \left(\int_{\mathbb{T}^d} A(y,\omega) \, \mathrm{d}y \, \nabla_x u(x) \right) \end{aligned}$$

and

$$\begin{aligned}\int_{\mathbb{T}^d} \mathcal{A}_1 u_1 \,\mathrm{d}y &= \int_{\mathbb{T}^d} -\nabla_y \cdot (A(y,\omega)\nabla_x u_1) - \nabla_x \cdot (A(y,\omega)\nabla_y u_1)\,\mathrm{d}y \\ &= -\int_{\mathbb{T}^d} \nabla_x \cdot \Big(A(y,\omega)\nabla_y(\chi(y,\omega)\cdot\nabla_x u(x))\Big)\,\mathrm{d}y \\ &= -\int_{\mathbb{T}^d} \nabla_x \cdot \Big(A(y,\omega)(\nabla_y \chi(y,\omega))^\top \nabla_x u(x)\Big)\,\mathrm{d}y.\end{aligned}$$

Combining these terms, our solvability condition becomes

$$-\nabla_x \cdot \left(\int_{\mathbb{T}^d} A(y,\omega) + A(y,\omega)(\nabla_y \chi(y,\omega))^\top \,\mathrm{d}y\, \nabla_x u\right) = f.$$

In the deterministic case, i.e. $A(y,\omega) = A(y)$ we would define the *effective coefficient* as

$$\bar{A} := \int_{\mathbb{T}^d} A(y) + A(y)(\nabla_y \chi)^\top \,\mathrm{d}y. \tag{2.14}$$

and for $0 < \varepsilon \ll 1$, the solution u_ε of (2.1) would be approximately given by the solution u to the homogenized equation (see [28, p. 185]):

$$-\nabla \cdot (\bar{A}\nabla u) = f \qquad \forall x \in D, \tag{2.15a}$$

$$u = 0 \qquad \forall x \in \partial D. \tag{2.15b}$$

We will see how this result is generalized for the stochastic case in the next section.

Remark 2.3. The effective coefficient can also be expressed in the form

$$\bar{A}_{ij} = a(\chi_j + y_j, \chi_i + y_i), \quad i,j \in \{1,\dots,d\}.$$

This is due to the fact that for $\chi_i \in C^1_{\#}(\mathbb{T}^d)$, $a(\chi_i, \chi_j + y_j) = 0$ for all $i,j \in \{1,\dots,d\}$. The expression is used to justify the symmetry of $\bar{A}$ when A is symmetric; see [28, p. 189].

2.2 Theoretical Results for the Stochastic Homogenized Problem

In the previous section, we derived the homogenized equation for a fixed realization ω in the periodic case. Now we wish to summarize some of the main findings in the papers by Kozlov [19], Yurinskii [33], and Papanicolaou and Varadhan [27].

Let $D_L := \left[-\frac{L}{2}, \frac{L}{2}\right)^d \subset \mathbb{R}^d$ be a cube with length L centered at the origin. First, let us recall the definition of an ergodic transformation.

Definition 2.4. Let $(\Omega, \mathcal{F}, P)$ be a probability space and $T : \Omega \to \Omega$ be a measure-preserving transformation. We say that T is *ergodic* with respect to the measure P if any $F \in \mathcal{F}$ with $T(F) \subset F$ implies $P(F) = 0$ or $P(F) = 1$.

Now we recall the definition of a homogeneous random field.

Definition 2.5. [19] Let the probability measure P be invariant with respect to the translation group $T_x : \Omega \to \Omega$, which is given by

$$(T_x\omega)(y) = \omega(y - x) \qquad \forall x, y \in \mathbb{R}^d.$$

Assume that T_x is ergodic. A *homogeneous random field* $v : \mathbb{R}^d \times \Omega \to \mathbb{R}^n$ satisfies

$$v(x + y, \omega) = v(x, T_y\omega), \qquad \forall x, y \in \mathbb{R}^d.$$

For a homogeneous random field, Birkhoff's ergodic theorem [19] states that the spatial average exists, in other words

$$\langle v \rangle := \lim_{L\to\infty} \frac{1}{L^d} \int_{D_L} v(x)\,\mathrm{d}x.$$

In the ergodic case, it holds that $\langle v \rangle = \mathbb{E}v$. In other words, the result of averaging over all realizations of an ensemble is equivalent to averaging over the volume for one realization in the infinite-volume limit [31, p. 29].

The main result by Kozlov can be summarized as follows.

Theorem 2.6 (Kozlov [19])**.** *Assume that A is uniformly elliptic and bounded, i.e.* (2.2) *holds, where A is a $d \times d$ matrix whose elements form homogeneous random fields. Then the following statements hold.*

1. *There exists a unique homogeneous random field $\psi = (\psi_i^j(y))$, $i, j \in \{1, \dots, d\}$, which furnishes the minimum of the variational problem*

$$\langle \psi_i^j(y) A_{jk} \psi_i^k(y) \rangle \to \min!,$$

$$\operatorname{curl}(\psi_i) = \frac{\partial \psi_i^j}{\partial y_k} - \frac{\partial \psi_i^k}{\partial y_j} = 0, \quad \langle \psi_j^i(y) \rangle = \delta_j^i, \quad \mathbb{E}[\psi_j^i \psi_j^i] < \infty.$$

2. *The effective coefficient $\bar{A}$ is given by*

$$\bar{A}_{ij} = \langle \psi_i^k(y) A_{km}(y) \psi_j^m(y) \rangle = \langle A_{ik} \psi_j^k \rangle. \tag{2.16}$$

3. *The weak solution $u_\varepsilon \in H_0^1(D)$ to* (2.3) *almost surely converges weakly in $H^1(D)$ to $u \in H_0^1(D)$, the solution of the deterministic elliptic problem*

$$-\nabla \cdot (\bar{A} \nabla u) = f(x) \qquad \forall x \in D, \tag{2.17a}$$

$$u(x) = 0 \qquad \forall x \in \partial D. \tag{2.17b}$$

In other words, when ε is small, the stochastic equation can be approximated by a deterministic equation with constant effective coefficient $\bar{A}$. Until Kozlov's paper, homogenized equations had only been constructed in the case of periodic and

almost periodic microstructures; the paper gave homogenized equations in the general case.

It is worth mentioning that another author, Yurinskii [33], independently gave the same results under the additional condition of strong mixing. Since this condition is more restrictive, however, Kozlov's result is considered more general and is therefore more frequently cited in the literature. The following theorem from Yurinskii's paper detailed the ergodic properties of the cell problem.

Theorem 2.7 (Yurinskii [33])**.** *Given the boundary value problem*

$$-\nabla \cdot (A(\nabla \chi + b)) = 0 \qquad \forall y \in D_L, \tag{2.18a}$$

$$\chi = 0 \qquad \forall y \in \partial D_L, \tag{2.18b}$$

where the components $A_{ij}(y,\omega)$ satisfying (2.2) *are homogeneous random fields that are measurable for all $y \in \mathbb{R}^d$ and $\omega \in \Omega$, assume that A is subject to a condition of strong mixing in the form*

$$\sup_{|\xi_i| \le 1, i \in \{1,2\}} |\mathbb{E}\xi_1\xi_2 - \mathbb{E}\xi_1\mathbb{E}\xi_2| \le \varphi(d(\bar{D}, \bar{\tilde{D}})) \tag{2.19}$$

where $d(\bar{D}, \bar{\tilde{D}})$ is the distance between the closures of the bounded domains D and $\tilde{D}$. Let $\chi_b = \chi_{b,L}$ be a solution to (2.18)*. Then the limit*

$$\langle Ab, b' \rangle = \lim_{L\to\infty} \mathbb{E} L^{-d} \int_{D_L} \left\langle A(\nabla \chi_b + b), b' \right\rangle \mathrm{d}y \tag{2.20}$$

exists for any $b, b' \in \mathbb{R}^d$.

Remark 2.8. Equation (2.18) corresponds to the corrector equation with Dirichlet boundary conditions instead of periodic boundary conditions. Equation (2.20) gives an explicit form for the effective coefficient $\bar{A}$ if we set $b = e_i$ and $b' = e_j$, the unit vectors on $\mathbb{R}^d$.

A companion work to the papers by Yurinksii and Kozlov can be found in the paper by Papanicolaou and Varadhan [27]. The authors created an analytical framework in which the stochastic homogenization result can be understood. We will summarize some of the more important elements here. Instead of working with homogeneous random fields, Papanicolaou and Varadhan made stationarity, ergodicity and uniform ellipticity assumptions on the tensor A. We recall the definition of strict stationarity.

Definition 2.9. A is referred to as *strictly stationary* if the joint distribution of $A(y_1,\omega), \ldots, A(y_n,\omega)$ is the same as that of $A(y_1 + h,\omega), \ldots, A(y_n + h,\omega)$ for all $y_i \in \mathbb{R}^d, i \in \{1,\ldots,n\}$ and for all $h \in \mathbb{R}^d$.

The processes $A(y,\omega)$ are assumed to be stochastically continuous in the sense that

$$\lim_{|\xi|\to 0} \mathbb{P}\{|A(y+\xi,\omega) - A(y,\omega)| > \delta\} = 0 \qquad \forall \delta > 0, \qquad \forall y \in \mathbb{R}^d.$$

Tilde notation is introduced to associate a function with its translates that form the stationary process; in particular,

$$f(x,\omega) = (\tau_x \tilde{f})(\omega) = f(\tau_{-x}\omega).$$

For the periodic case, the solving of the cell problem amounts to solving the elliptic equation on the torus, for which we have the Poincaré inequality. In the non-periodic case, another strategy must be employed. As in the papers by Kozlov and Yurinskii, Papanicolaou and Varadhan consider a modified cell problem

$$-\nabla \cdot A(I + \nabla \chi_T) + T^{-1}\chi_T = 0. \tag{2.21}$$

with a zero-order term and $T^{-1} > 0$, which is meant to circumvent the lack of coercivity of the elliptic operator in probability [17]. We give a main result (Theorem 2) from their paper, which is essentially a restatement of Theorem 2.6.

Theorem 2.10 (Papanicolaou and Varadhan [27])**.** *Let $\mathcal{H} := L^2(\Omega, \mathcal{F}, P)$ and let $\mathcal{H}^1 := \bigcap_{i=1}^d C_0^\infty(D_i)$. There exist uniquely defined functions $\tilde{\psi}_i^k(\omega) \in \mathcal{H}$ such that*

$$\sum_{i,j=1}^{d} \mathbb{E}\left[\tilde{A}_{ij}(\delta_{jk} + \tilde{\psi}_j^k) D_i \tilde{\varphi}\right] = 0 \qquad \forall \tilde{\varphi} \in \mathcal{H}^1, \quad k \in \{1,\dots,d\}, \tag{2.22}$$

and

$$\mathbb{E}[\tilde{\psi}_i^k] = 0. \tag{2.23}$$

Furthermore, the coefficients $\bar{A}_{ij}$ in the homogenized (modified) equation are given by

$$\bar{A}_{ij} = \mathbb{E}\left[\sum_{k=1}^{d} \tilde{A}_{ik}(\delta_{kj} + \tilde{\psi}_k^j)\right], \quad i,j \in \{1,\dots,d\}. \tag{2.24}$$

There exist uniquely defined processes $\chi^k(x,\omega)$ that are not stationary, such that $\chi^k(0,\omega) = 0$ and

$$\frac{\partial \chi^k(x,\omega)}{\partial x_i} = \psi_i^k(x,\omega) = \tilde{\psi}_i^k(\tau_{-x}\omega)$$

so that their gradients are stationary.

In particular, this theorem agrees with existing theory for periodic problems; in the periodic case there exist functions $\tilde{\chi}_k \in \mathcal{H}^1$ that satisfy [27, p. 848]

$$\int_{\mathbb{T}^d} \sum_{i,j=1}^{d} \tilde{A}_{ij}(\omega)\left(\delta_{jk} + \frac{\partial \tilde{\chi}_k(\omega)}{\partial \omega_j}\right)\frac{\partial \tilde{\varphi}(\omega)}{\partial \omega_i}\,\mathrm{d}\omega = 0 \qquad \forall \tilde{\varphi} \in \mathcal{H}^1.$$

This corresponds to the usual cell problem in the deterministic case.

CHAPTER 3

Numerical Methods

3.1 Finite Element Solution to First-Order Corrector

We have seen that in order to calculate the effective coefficient matrix, we first need to solve the corrector equation (2.10). We will discuss how to approximate the solution to this equation using the finite element method for $d = 2$ (the two-dimensional setting). To begin with, assume that we are interested in solving a single realization of the cell problem (2.10), meaning that the matrix $A(y, \omega)$ is fixed. In this section, we will simplify the notation by setting $A(y, \omega) = A(y)$ and similarly for the solution χ. We will denote with

$$S^{1,1}(\mathcal{T}) = \left\{ u \in H^1(\mathbb{T}^d) \,\middle|\, u|_K \in \mathcal{P}_1(K) \quad \forall K \in \mathcal{T} \right\},$$

a finite element space associated with a shape-regular triangulation $\mathcal{T}$, where the set $\mathcal{P}_1(K)$ is the space of linear polynomials on the triangle $K \in \mathcal{T}$. The space

$$S^{1,1}_{\#}(\mathcal{T}) = S^{1,1}(\mathcal{T}) \cap H^1_{\#}(\mathbb{T}^d)$$

is the restriction of $S^{1,1}(\mathcal{T})$ to a space with periodic boundary conditions. With basis functions $\mathcal{B} = \{\varphi_i \,|\, i = 1, \ldots, \#\mathcal{N}\}$, where $\#\mathcal{N}$ denotes the number of nodes, the discrete solution χ_h can be written for each component i as a linear combination of these basis functions: $\chi_h^i = \sum_{j=1}^{\#\mathcal{N}} u_j \varphi_j$.

3.1.1 Solution for a Diagonal A

In Chapter 2, we derived the weak formulation to the cell problem. We found that the weak formulation (2.12) is that of finding a $\chi_i \in C^1_{\#}(\mathbb{T}^d)$ for each component i such that

$$a(\varphi, \chi_i + y_i) = 0 \qquad \forall \varphi \in C^1_{\#}(\mathbb{T}^d),$$

where

$$a(u,v) = \int_{\mathbb{T}^d} \langle \nabla_y v, A(y)\nabla_y u \rangle \, \mathrm{d}y.$$

For a 2×2 diagonal matrix

$$A = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix},$$

observe that

$$\begin{aligned} a(\varphi, \chi_1 + y_1) &= \int_{\mathbb{T}^d} \left\langle \begin{pmatrix} \partial_{y_1}\chi_1 + 1 \\ \partial_{y_2}\chi_1 \end{pmatrix}, \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix} \begin{pmatrix} \partial_{y_1}\varphi \\ \partial_{y_2}\varphi \end{pmatrix} \right\rangle \mathrm{d}y \\ &= \int_{\mathbb{T}^d} \left\langle \begin{pmatrix} \partial_{y_1}\chi_1 + 1 \\ \partial_{y_2}\chi_1 \end{pmatrix}, \begin{pmatrix} A_{11}\partial_{y_1}\varphi \\ A_{22}\partial_{y_2}\varphi \end{pmatrix} \right\rangle \mathrm{d}y \\ &= \int_{\mathbb{T}^d} (\partial_{y_1}\chi_1 + 1)A_{11}\partial_{y_1}\varphi + \partial_{y_2}\chi_1 A_{22}\partial_{y_2}\varphi \, \mathrm{d}y = 0 \end{aligned}$$

$$\iff a_1(\chi_1, \varphi) := \int_{\mathbb{T}^d} \partial_{y_1}\chi_1 A_{11}\partial_{y_1}\varphi + \partial_{y_2}\chi_1 A_{22}\partial_{y_2}\varphi \, \mathrm{d}y = -\int_{\mathbb{T}^d} A_{11}\partial_{y_1}\varphi \, \mathrm{d}y =: b_1(\varphi).$$

An analogous calculation shows that

$$a(\varphi, \chi_2 + y_2) = \int_{\mathbb{T}^d} \partial_{y_1}\chi_2 A_{11}\partial_{y_1}\varphi + (\partial_{y_2}\chi_2 + 1)A_{22}\partial_{y_2}\varphi \, \mathrm{d}y = 0$$

$$\iff a_2(\chi_2, \varphi) := \int_{\mathbb{T}^d} \partial_{y_1}\chi_2 A_{11}\partial_{y_1}\varphi + \partial_{y_2}\chi_2 A_{22}\partial_{y_2}\varphi \, \mathrm{d}y = -\int_{\mathbb{T}^d} A_{22}\partial_{y_2}\varphi \, \mathrm{d}y =: b_2(\varphi).$$

Therefore, using the basis functions $\varphi_i, i \in \{1, \dots, \#\mathcal{N}\}$, we can define components of the stiffness matrices via

$$B^1_{ij} := a_1(\varphi_j, \varphi_i), \qquad B^2_{ij} := a_2(\varphi_j, \varphi_i)$$

and components of the load vectors via

$$l^1_i := b_1(\varphi_i), \qquad l^2_i := b_2(\varphi_i).$$

The algorithm for calculating the approximation χ_h in the diagonal case can be summarized as follows:

1. Assemble the two stiffness matrices B^1, B^2 and the two load vectors l^1 and l^2.
2. Apply the periodic boundary conditions: For each master node m and corresponding slave node s on the boundary, add the contribution of the slave to the master, i.e.

$$l^1_m = l^1_m + l^1_s \tag{3.1}$$
$$l^2_m = l^2_m + l^2_s, \tag{3.2}$$

and then set $l^i_s = 0$, $i \in \{1, 2\}$. There is a corner case where a master has more than one slave.

3. Set $\tilde{\chi}_h^1 = (B^1)^{-1} l^1$ and $\tilde{\chi}_h^2 = (B^2)^{-1} l^2$.

4. Apply the uniqueness condition by calculating, using quadrature,

$$c := \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \int_{\mathbb{T}^2} \tilde{\chi}_h \, \mathrm{d}y.$$

Then the unique numerical solution satisfying the condition $\int_{\mathbb{T}^d} \chi \, \mathrm{d}y = 0$ is given by

$$\begin{pmatrix} \chi_h^1 \\ \chi_h^2 \end{pmatrix} = \begin{pmatrix} \tilde{\chi}_h^1 \\ \tilde{\chi}_h^2 \end{pmatrix} - \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}.$$

Remark 3.1. If we only need χ to calculate the effective coefficient $\bar{A}$, then step four can be omitted, since the gradient of the constant is zero.

3.1.2 Solution for a Non-Diagonal A

For completeness, we will write down the equations that need to be solved in the non-diagonal case. For a 2×2 matrix

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

observe that

$$\begin{aligned} a(\varphi, \chi_1 + y_1) &= \int_{\mathbb{T}^d} \left\langle \begin{pmatrix} \partial_{y_1}\chi_1 + 1 \\ \partial_{y_2}\chi_1 \end{pmatrix}, \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} \partial_{y_1}\varphi \\ \partial_{y_2}\varphi \end{pmatrix} \right\rangle \mathrm{d}y \\ &= \int_{\mathbb{T}^d} \left\langle \begin{pmatrix} \partial_{y_1}\chi_1 + 1 \\ \partial_{y_2}\chi_1 \end{pmatrix}, \begin{pmatrix} A_{11}\partial_{y_1}\varphi + A_{12}\partial_{y_2}\varphi \\ A_{21}\partial_{y_1}\varphi + A_{22}\partial_{y_2}\varphi \end{pmatrix} \right\rangle \mathrm{d}y \\ &= \int_{\mathbb{T}^d} (\partial_{y_1}\chi_1 + 1)(A_{11}\partial_{y_1}\varphi + A_{12}\partial_{y_2}\varphi) + \partial_{y_2}\chi_1(A_{21}\partial_{y_1}\varphi + A_{22}\partial_{y_2}\varphi) \, \mathrm{d}y = 0 \end{aligned}$$

$$\begin{aligned} \iff a_1(\chi_1, \varphi) &:= \int_{\mathbb{T}^d} \partial_{y_1}\chi_1(A_{11}\partial_{y_1}\varphi + A_{12}\partial_{y_2}\varphi) + \partial_{y_2}\chi_1(A_{21}\partial_{y_1}\varphi + A_{22}\partial_{y_2}\varphi) \, \mathrm{d}y \\ &= -\int_{\mathbb{T}^d} A_{11}\partial_{y_1}\varphi + A_{12}\partial_{y_2}\varphi \, \mathrm{d}y =: b_1(\varphi). \end{aligned}$$

An analogous calculation shows that

$$a(\varphi, \chi_2 + y_2) = \int_{\mathbb{T}^d} \partial_{y_1}\chi_2(A_{11}\partial_{y_1}\varphi + A_{12}\partial_{y_2}\varphi) + (\partial_{y_2}\chi_2 + 1)(A_{21}\partial_{y_1}\varphi + A_{22}\partial_{y_2}\varphi) \, \mathrm{d}y = 0$$

$$\begin{aligned}\iff a_2(\chi_2,\varphi) &:= \int_{\mathbb{T}^d} \partial_{y_1}\chi_2(A_{11}\partial_{y_1}\varphi + A_{12}\partial_{y_2}\varphi) + \partial_{y_2}\chi_2(A_{21}\partial_{y_1}\varphi + A_{22}\partial_{y_2}\varphi)\,\mathrm{d}y \\ &= -\int_{\mathbb{T}^d} A_{21}\partial_{y_1}\varphi + A_{22}\partial_{y_2}\varphi\,\mathrm{d}y =: b_2(\varphi).\end{aligned}$$

As before, using the basis functions $\varphi_i, i \in \{1,\dots,\#\mathcal{N}\}$, we can define components of the stiffness matrices via

$$B^1_{ij} := a_1(\varphi_j,\varphi_i), \qquad B^2_{ij} := a_2(\varphi_j,\varphi_i)$$

and components of the load vectors via

$$l^1_i := b_1(\varphi_i) \qquad l^2_i := b_2(\varphi_i)$$

and solve for χ_h using the algorithm from the previous section.

3.2 Monte Carlo Finite Element Method

In this section, we will describe Monte Carlo FEM as applied to elliptic PDEs in two dimensions. Let $(\Omega,\mathcal{F},P)$ be a probability space and let $D\subset\mathbb{R}^2$ be a bounded domain. We assume that $\mathcal{F}$ is an appropriate σ-algebra. First, consider the generic problem

$$-\nabla\cdot(A(x,\omega)\nabla u(x,\omega)) = f(x,\omega) \qquad \forall (x,\omega)\in D\times\Omega, \tag{3.3a}$$
$$u(x,\omega) = g(x,\omega) \qquad \forall (x,\omega)\in \partial D\times\Omega. \tag{3.3b}$$

We will summarize the method as presented in [23, Chapter 9]. The PDE (3.3) is stochastic in the sense that $u(x)$ is a random field with realizations $u(x,\omega)$. The main idea behind Monte Carlo FEM is that for individual realizations of $A(\cdot,\omega)$ and $f(\cdot,\omega)$, Galerkin FEM can be used to approximate individual realizations of the solution, which are unique provided some regularity of the coefficients A. Monte Carlo sampling is then used to approximate $\mathbb{E}[u(x)]$ and $\mathbb{V}[u(x)]$. The advantage of this method is that one can use a deterministic solver for each realization and then aggregate the results.

More formally, given a finite element space $V_h\subset H^1(D)$ associated with the sequence of shape-regular triangulations $\mathcal{T}_h$, the random field $u_h(x)$ with realizations $u_h(\cdot,\omega)\in V_h$ satisfies the weak form

$$\int_D A(x,\omega)\nabla u_h(x,\omega)\cdot\nabla v(x)\,\mathrm{d}x = \int_D f(x)v(x)\,\mathrm{d}x \qquad \forall v\in V_h. \tag{3.4}$$

In practice, we only have approximations for A or else just exact samples at discrete points. Thus, we have the weak form, using the approximation $\tilde{A}$ for A,

$$\int_D \tilde{A}(x,\omega)\nabla \tilde{u}_h(x,\omega)\cdot\nabla v(x)\,\mathrm{d}x = \int_D f(x)v(x)\,\mathrm{d}x \qquad \forall v\in V_h. \tag{3.5}$$

For N i.i.d. realizations $\tilde{A}_r := \tilde{A}(\cdot, \omega_r)$, $r \in \{1, \dots, N\}$ of the diffusion coefficient, we generate i.i.d. samples $\tilde{u}_{h,r}(x) := \tilde{u}_h(x, \omega_r)$ of the finite element solution $\tilde{u}_h(x)$.

To estimate $\mathbb{E}[\tilde{u}]$, we use $\mathbb{E}[\tilde{u}_h]$, where $\tilde{u}_h$ is the FEM approximation of $\tilde{u}$ using a mesh fineness h, restricted to the domain D. We estimate $\mathbb{E}[u_h]$ using the sample mean

$$\mu_{N,h}(x) := \frac{1}{N} \sum_{i=1}^{N} \tilde{u}_{h,i}(x),$$

where $\tilde{u}_{h,i} = \tilde{u}_h(x, \omega^{(i)})$ is the i^{th} realization of the solution. The variance $\mathbb{V}[\tilde{u}_h]$ can be estimated by

$$\sigma^2_{N,h}(x) := \frac{1}{N-1} \sum_{i=1}^{N} \left(\tilde{u}_{h,i}(x) - \mu_{N,h}(x) \right)^2 = \frac{1}{N-1} \left(\sum_{i=1}^{N} \tilde{u}_{h,i}(x)^2 - N \mu_{N,h}(x)^2 \right).$$

Now we will apply this method to the cell problem, dropping the inconvenient tilde notation and solving on the domain $D_L := [-L/2, L/2)^2 \subset \mathbb{R}^2$. Given a finite element space $V_h \subset H^1_{\#}(D_L)$, we generate i.i.d. realizations $A_r := A(\cdot, \omega_r)$ for $r \in \{1, \dots, N\}$ and solve

$$a_r(\varphi, \chi_{i,r} + y_i) = 0 \qquad \forall \varphi \in H^1_{\#}(D_L), \tag{3.6}$$

where

$$a_r(u, v) := \int_{D_L} \left\langle \nabla_y v, A_r(y) \nabla_y u \right\rangle \mathrm{d}y.$$

With that, we obtain i.i.d. samples $\chi_{i,r}(x) := \chi_i(x, \omega_r)$ of the finite element solution for each $i \in \{1, 2\}$. The expected value $\mathbb{E}[\chi_i]$ and variance $\mathbb{V}[\chi_i]$ can be approximated as before.

3.3 Numerical Homogenization

Now that we have seen how the cell problem can be solved and the effective coefficient can be determined, numerically, let us see how this can be utilized in practice to get a solution to the problem (2.1). The technique for approximating this solution is called numerical homogenization, an important tool used in a wide variety of scientific and engineering simulations. When disparities between spatial and/or temporal scales are limited, then traditional numerical techniques can be employed. When scale separation is more pronounced, then multiscale techniques become necessary: some resolution of the details on the microscopic scale is needed to understand behavior at the macroscopic level. Even with supercomputers and the ability to compute processes in parallel, the sheer size of the computations involved in multiscale problems is prohibitive; an enormous amount of computer memory and CPU time is needed. Numerical homogenization is designed to numerically capture the small-scale effect on larger scales without the need to completely resolve the microscale.

An important concept in numerical homogenization is the representative volume element (RVE), which was already mentioned in the introduction. Often, input information about properties of a material is not available over the entire domain. These RVEs contain essential information about the heterogeneities and can be seen as a representative sample of the medium. Ostoja-Starzewski [26] distinguishes between the RVE found in continuum solid mechanics and the closely-related statistical volume element (SVE) found in stochastic solid mechanics. He argues that RVEs have their place in the (unrealistic) case of a unit cell in a periodic structure (as in Figure 1.2), or on an infinite set of microscale inclusions possessing statistically homogeneous and ergodic properties. He introduces a "mesoscale" on which the averaging takes place; this is some subset $D_L \subset D$ of the original domain. As the mesoscale grows, he shows that the SVE tends to become the RVE. We already discussed conditions under which a spatial average is equal to the statistical average, i.e. $\mathbb{E}v = \lim_{L\to\infty} L^{-d} \int_{D_L} v(x)\,\mathrm{d}x$. In practice, one does not actually solve the cell problem on the entire domain. One rather approximates the expectation on a finite domain using a finite number of sampling points [26, p. 118].

We will focus on numerical homogenization as applied to the deterministic problem (2.1). There are different approaches in numerical homogenization. In the Multiscale Finite Element Method (MsFEM), localized basis functions are used to capture the microscale's effect on the macroscale. These basis functions are computed on a RVE and are used to construct an average picture over the macroscopic domain. An introduction on MsFEM can be found in, for example, [11].

We will take a closer look at the Heterogeneous Multiscale Method (HMM) as described in [9]. This method has two main components: (1) the macroscopic scheme for macroscopic variables on a macroscopic grid; and (2) the estimation of missing macroscopic data from the microscale model. In the following, let $\mathcal{T}_H$ be a triangulation on the domain D with mesh fineness $H \gg \varepsilon$. Let us consider the (more realistic) case where the coefficient A varies on both the microscale and the macroscale. If we already knew the effective coefficient $\bar{A}$ for discrete points on a certain element of $\mathcal{T}_h$, we could evaluate the quadratic form

$$\int_D \nabla u(x) \cdot \bar{A}(x) \nabla u(x)\,\mathrm{d}x$$

by numerical quadrature: for any element $U \in V_H$, the finite element space on the macroscale,

$$\bar{A}_H(U,U) \approx \sum_{K\in\mathcal{T}_H} |K| \sum_{x_l \in K} w_l (\nabla U \cdot \bar{A} \nabla U)(x_l)$$

where $|K|$ signifies the area of the element K, x_l are quadrature points and w_l the corresponding quadrature weights. Since we do not know the value of $\bar{A}$, however, we approximate it on a smaller domain $D_L(x_l) := [x_l - L/2, x_l + L/2)^d$ by solving the cell problem at each quadrature point:

$$-\nabla \cdot (A(y) \nabla_y \chi_{L,l}(y)) = \nabla_y \cdot A^\top(y), \quad \chi \text{ is 1-periodic}, \quad \int_{D_L(x_l)} \chi(y)\,\mathrm{d}y = 0.$$

The solution to each of these cell problems gives us local estimates for A. To reduce the effect of the periodic boundary conditions, it is advised that a subdomain $D_{L_0}(x_l) \subset D_L(x_l)$ is used for the integration, so that our estimate for $\bar{A}$ takes the form

$$\bar{A}_H(x_l) = \fint_{D_{L_0}(x_l)} A(y)(I + \nabla_y \chi_{L,l}) \, \mathrm{d}y.$$

Thus, a considerable amount of computational work is saved by solving the local cell problems.

Theoretical error bounds exist in the periodic case for the effective coefficient. If we define

$$e(\mathrm{HMM}) := \max_{x_l \in K, K \in T_H} \|\bar{A}(x_l) - \bar{A}_h(x_l)\|,$$

then we have [9, p. 125]

$$e(\mathrm{HMM}) = \begin{cases} C\varepsilon & \text{if } D_L(x_l) = x_l + \varepsilon D_1(0), \\ C\left(\frac{\varepsilon}{L} + L\right) & \text{otherwise.} \end{cases}$$

Gloria [13] calls the error $\mathcal{O}(\varepsilon/L)$ the "resonance error." In that paper, a zero-order term is added to the equation, which is inspired by the modified corrector equation (2.21) used in the papers by Kozlov, Yurinskii, and Papanicolaou and Varadhan. This term has the effect of dramatically reducing the effect of spurious boundary conditions away from the boundary layer.

Note that we only discussed the deterministic, general case; these methods can be easily extended to the stochastic case where A does not vary on the macroscale, i.e. $A = A(y, \omega)$. For the random homogenization problem, error bounds for $\mathbb{E}e(\mathrm{HMM})$ exist for $d = 1$, taking the form $\mathcal{O}(\varepsilon/L)^\kappa$ with a κ arbitrarily close to $6/25$, or $d = 3$, taking the form $\mathcal{O}(\varepsilon/L)^{1/2}$. Error bounds for $d = 2$ do not yet exist; see [9, p. 126].

CHAPTER 4

Error Analysis and an Optimal Computation Scheme

In this chapter, we will focus on estimating the error for $d = 2$ in numerically computing the cell problem. Using these estimates, we will develop a scheme to compute the cell problem as efficiently as possible. Before we proceed, let us establish notation that will be used in the following section. As before, we will denote with $D_L := [-L/2, L/2)^2 \subset D \subset \mathbb{R}^2$ a square of length L.

We will now refine our definition of the cell problem, which was introduced in Chapter 2. The first-order corrector is defined on a torus of length L.

Definition 4.1. Let A be a stationary ergodic random field. The *first-order corrector* (or *corrector*) χ_L is a L-periodic function defined by the elliptic PDE

$$-\nabla_y \cdot (A(y,\omega)\nabla_y \chi_L(y,\omega)) = \nabla_y \cdot A^\top(y,\omega), \qquad \int_{D_L} \chi_L(y,\omega)\,\mathrm{d}y = 0. \tag{4.1}$$

The motivation behind this definition is a technique called "windowing." A major source of error is due to the periodic boundary conditions that we impose, which are artificial. To minimize the effects of spurious boundary conditions, one imposes boundary conditions far from the domain of interest and then uses a subdomain to approximate $\bar{A}$, as in the following definition.

Definition 4.2. Using the first-order corrector χ_L as defined above, we will denote with $\bar{A}_{L,L_0}$ for $L_0 < L$ the *effective coefficient matrix*, or "approximation by periodization"

$$\bar{A}_{L,L_0} = \frac{1}{L_0^d} \int_{D_{L_0}} A(y,\omega)(I + (\nabla \chi_L)^\top)\,\mathrm{d}y. \tag{4.2}$$

Even though quantifying error in the first-order corrector is the first step towards the quantification of the homogenization error, very few error estimates exist for

the first-order corrector. Abdulle [1, p. 452] estimated the error of the homogenized solution in (deterministic) numerical homogenization, and for that needed an estimate for the numerical error in the corrector. The case of discrete elliptic equations has been developed to a much greater extent, where optimal estimates for the corrector already exist (as well as approximations of the homogenized coefficients); see [15]. Gloria and Otto [16], [17] produced quantitative estimates for the periodic approximation of the corrector equation for stochastic homogenization of linear elliptic equations in divergence form. In that paper, the effective coefficients satisfied a spectral gap estimate in probability and results were obtained for dimensions $d > 2$.

4.1 Error in Numerical Calculation of the Corrector

We will now estimate the error involved in calculating the first-order corrector (4.1) using Monte Carlo FEM as outlined in Section 3.2. The error of the gradient of the first-order corrector and the effective coefficient are closely related, as we see in the formula for the effective coefficient (4.2). Thus, the first step in quantifying the error of the approximation to $\bar{A}$ is estimating the error incurred by numerically approximating $\nabla\chi$. To this end, we will seek to estimate the L^2 error of the gradient of each element of χ on the domain $D_1 = [-1/2, 1/2)^2$. This norm is sensible not only because we are interested in the effect of the error in χ on $\bar{A}$, but also since this defines a norm in the space

$$H = \Big\{ u \in H^1_{\#}(\mathbb{T}^d) \,\Big|\, \int_{\mathbb{T}^d} u \,\mathrm{d}y = 0 \Big\},$$

which is exactly the space we use to solve the cell problem. This space also comes equipped with a norm $\|\cdot\|_H$ that is defined by $\|u\|_H := \|\nabla u\|_{L^2(\mathbb{T}^d)}$ for $u \in H$ [28, p. 24].

The domain D_1 on which the error is calculated is chosen for ease of notation, but these results can naturally be extended to other domains. In the following, the norm $\|\cdot\|_{L^p(\Omega;X)}$ is defined by

$$\|u\|_{L^p(\Omega;X)} = \begin{cases} \left[\int_\Omega \|u(\cdot,\omega)\|_X^p \,\mathrm{dP}(\omega)\right]^{1/p} = \mathbb{E}[\|u(\cdot,\omega)\|_X^p]^{1/p} < \infty & 1 \le p < \infty, \\ \operatorname{ess\,sup}_{\omega\in\Omega} \|u(\cdot,\omega)\|_X & p = \infty. \end{cases}$$

For the remainder of this chapter, we will assume that we are calculating the error for one element of $\chi = (\chi^1, \chi^2)^\top$ only, meaning we are looking for either the error in χ^1 or χ^2. Thus we will simply use the notation χ for one or the other element.

We will now establish some notation to help quantify this error. χ will be the exact solution to the cell problem, i.e. the χ that satisfies (2.20). χ_h will be the discrete solution that is (theoretically) solved on $\mathbb{R}^2$ but for an *approximation* of the matrix A as discussed in Section 3.2. $\chi_{L,h}$ will be the discrete solution with periodic

boundary conditions on the domain D_L, and $\mu_{N,L,h}$ be the sample mean of N independent realizations of $\nabla\chi_{L,h,i} = \nabla\chi_{L,h}(x,\omega^{(i)})$, in other words

$$\mu_{N,L,h} = \frac{1}{N}\sum_{i=1}^{N} \nabla\chi_{L,h,i}.$$

Note that the notation $\nabla\chi_{L,h}$, as the gradient of a discrete solution $\chi_{L,h}$, is to be understood in a piecewise sense, i.e. on each element of a finite element solution. There are three sources of error present in the approximation of $\nabla\chi$:

- Discretization error: This is the error of $\mathbb{E}[\nabla\chi]$ by $\mathbb{E}[\nabla\chi_h]$, which is due to spatial discretization. This includes the error incurred by using approximate data as described in Section 3.2, which scales like the error of the spatial discretization as long as A does not vary on a length scale that cannot be captured by spatial discretization [23, p. 379].
- Error due to boundary conditions: The domain is artificially trimmed to a finite domain and boundary conditions are chosen that are not, a priori, given. The expected value of the spatial discretization $\mathbb{E}[\nabla\chi_h]$ is therefore approximated by $\mathbb{E}[\nabla\chi_{L,h}]$.
- Statistical error: The expected value $\mathbb{E}[\nabla\chi_{L,h}]$ is approximated by a sample mean $\mu_{N,L,h}$ of N realizations of $\nabla\chi_{L,h}$.

We will now state the main result.

Theorem 4.3. *Assume that for almost all $\omega \in \Omega$, realizations $A(\cdot,\omega)$ of the coefficient satisfy* (2.2)*. Assume that there exist constants α, β, ν_0, ν_1 and $\nu_2 > 0$ independent of L, h and N such that*

$$\|\nabla\chi_h - \nabla\chi\|_{L^2(\Omega;L^2(D_1))} \leq \nu_1 h^\alpha, \tag{4.3a}$$

$$\|\nabla\chi_{L,h} - \nabla\chi_h\|_{L^2(\Omega;L^2(D_1))} \leq \nu_2 L^{-\beta}, \tag{4.3b}$$

$$\mathbb{E}\big[\|\mathbb{E}[\nabla\chi_{L,h}] - \mu_{N,L,h}\|^2_{L^2(D_1)}\big] \leq \frac{\nu_0}{N} \tag{4.3c}$$

hold. Then the error of the MCFEM estimator $\mu_{N,L,h}$ satisfies

$$\|\mu_{N,L,h} - \mathbb{E}[\nabla\chi]\|_{L^2(\Omega;L^2(D_1))} = \mathcal{O}(N^{-1/2}) + \mathcal{O}(h^\alpha) + \mathcal{O}(L^{-\beta}). \tag{4.4}$$

Proof. Since $\|\cdot\|_{L^2(\Omega;L^2(D_1))}$ is a norm, we can use the triangle inequality twice to obtain

$$\begin{aligned}\|\mathbb{E}[\nabla\chi] - \mu_{N,L,h}\|_{L^2(\Omega;L^2(D_1))} \leq{}& \underbrace{\|\mathbb{E}[\nabla\chi] - \mathbb{E}[\nabla\chi_h]\|_{L^2(\Omega;L^2(D_1))}}_{:=e_{\mathrm{FEM}}} + \\ &+ \underbrace{\|\mathbb{E}[\nabla\chi_h] - \mathbb{E}[\nabla\chi_{L,h}]\|_{L^2(\Omega;L^2(D_1))}}_{:=e_{\mathrm{BC}}} + \\ &+ \underbrace{\|\mathbb{E}[\nabla\chi_{L,h}] - \mu_{N,L,h}\|_{L^2(\Omega;L^2(D_1))}}_{:=e_{\mathrm{MC}}}.\end{aligned} \tag{4.5}$$

For the first two errors, we use assumptions (4.3a) and (4.3b) to get

$$\begin{aligned} e_{\text{FEM}} + e_{\text{BC}} &= \mathbb{E}\left[\|\nabla\chi - \nabla\chi_h\|_{L^2(D_1)} + \|\nabla\chi_h - \nabla\chi_{L,h}\|_{L^2(D_1)}\right] \\ &\leq \nu_1 h^\alpha + \nu_2 L^{-\beta}. \end{aligned}$$

For the third error, the inequality (4.3c) means that $\mathbb{E}[e_{\text{MC}}^2] \leq \frac{\nu_0}{N}$, which implies [23, p. 385] that for the random variable e_{MC} and any $\epsilon > 0$,

$$\mathbb{P}(e_{\text{MC}} \geq N^{-1/2+\epsilon}) \leq \nu_0^2 N^{-2\epsilon}.$$

This means that the statistical error is $\mathcal{O}(N^{-1/2})$; see [23, p. 386]. □

In the proof, we already showed the motivation behind the bound for the statistical error (4.8c). The constant ν_0 can be explicitly determined by the expression

$$\nu_0 := \frac{K_p}{\alpha^-}\|f\|_{L^2(D_1)},$$

where K_p is the Poincaré constant and f is the right-hand side of the partial differential equation [23, p. 385]. The error from the boundary, e_{BC}, will be shown to have the form $\mathcal{O}(L^{-\beta})$ later in numerical tests. The theoretical basis for this convergence behavior comes from the idea of windowing as discussed at the beginning of this chapter. The discretization error (4.3a) has some more solid basis in the theory, so we will highlight it here.

Let us recall that deterministic bounds of the form $\|\nabla\chi_h - \nabla\chi\|_{L^2(D_1)} \leq \nu_1 h^\alpha$ have their basis in finite element theory. The standard result is formulated from [29, pp. 95-96] in Theorem 4.4, where we recall the definition of the seminorm

$$|u|_{H^k(D)} := \sqrt{\sum_{|\boldsymbol{\alpha}|=k} \int_D (D^{\boldsymbol{\alpha}} u)^2 \,\mathrm{d}x}.$$

Theorem 4.4. *Let $u \in V$ be the exact solution to the generic elliptic problem*

$$\textit{find } u \in V: \quad a(u,v) = f(v) \qquad \forall v \in V,$$

where V is a (Hilbert) subspace of $H^1(\Omega)$, $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is a continuous and coercive bilinear form, and $f(\cdot) : V \to \mathbb{R}$. Let u_h denote the discrete solution to

$$\textit{find } u \in V: \quad a(u_h,v_h) = f(v_h) \qquad \forall v_h \in V_h,$$

where V_h is a finite-dimensional subspace of V with basis functions of order p. Then, if $u \in H^{p+1}(D)$, there exists a constant C_1 independent of h and u such that

$$\|u - u_h\|_{H^1(D)} \leq \frac{\beta}{\alpha} C_1 h^p |u|_{H^{p+1}(D)}, \tag{4.6}$$

where β and α refer to the boundedness and coercivity constants, respectively, of the bilinear form $a(\cdot,\cdot)$. Moreover, there exists a constant $\tilde{C}_1$ such that

$$\|u - u_h\|_{L^2(D)} \leq \tilde{C}_1 h^{p+1} |u|_{H^{p+1}(D)}. \tag{4.7}$$

Theorem 4.4 can be used to now justify the error bound (4.3a) in the stochastic case. We assume some regularity for χ, in this case that there exists a constant $C_2 > 0$ such that $\chi \in L^2(\Omega, H^{p+1}(D_1))$ and

$$|\chi|_{L^2(\Omega,H^{p+1}(D_1))} := \mathbb{E}[|\chi|^2_{H^{p+1}}]^{1/2} \leq C_2\|f\|_{L^2(D_1)} \qquad \forall f \in L^2(D_1).$$

For each $\chi(\cdot,\omega) \in V$ and corresponding finite element approximation $\chi_h(\cdot,\omega) \in V_h$, it follows from (4.6) that

$$\|\nabla\chi(\cdot,\omega)-\nabla\chi_h(\cdot,\omega)\|^2_{L^2(D_1)} \leq \|\chi(\cdot,\omega)-\chi_h(\cdot,\omega)\|^2_{H^1(D_1)} \leq \left(\frac{\beta}{\alpha}C_1h^p|\chi(\cdot,\omega)|_{H^{p+1}(D_1)}\right)^2.$$

Taking the expectation on both sides of the inequality yields

$$\mathbb{E}\left[\|\nabla\chi - \nabla\chi_h\|^2_{L^2(D_1)}\right] \leq \left(\frac{\beta}{\alpha}C_1h^p\right)^2 \mathbb{E}\left[|\chi|^2_{H^{p+1}(D_1)}\right] \leq \left(\frac{\beta}{\alpha}C_1C_2h^p\|f\|_{L^2(D_1)}\right)^2.$$

Finally, taking square roots on both sides yields an inequality of the form (4.3a).

Remark 4.5. In other words, we should expect the error $\|\nabla\chi_h - \nabla\chi\|_{L^2(\Omega;L^2(D_1))}$ to scale like $\mathcal{O}(h)$ if we are using linear basis functions. Later, to determine an efficient solving strategy, we will need to provide an explicit bound for the constant ν_1 given in the theorem. Methods for explicitly computing the bounds for the finite element method can be found in, for example, [2], [22], or [24]. However, for our numerical simulations, we will be using bounds that are generated numerically, since they also depend on the implementation.

4.1.1 Alternate Error Estimate

Now we will present an alternative estimate that involves using our knowledge of the variance of $\nabla\chi_{L,h}$ to estimate the error.

Theorem 4.6. *Using the notation introduced at the beginning of this section, assume that there exist constants α, β, ν_0, ν_1, and $\nu_2 > 0$ independent of L, h and N such that*

$$\|\nabla\chi_h - \nabla\chi\|^2_{L^2(\Omega;L^2(D_1))} \leq \frac{\nu_1}{4}h^\alpha, \tag{4.8a}$$

$$\|\nabla\chi_{L,h} - \nabla\chi_h\|^2_{(L^2(\Omega;L^2(D_1))} \leq \frac{\nu_2}{4}L^{-\beta}, \tag{4.8b}$$

$$\|\mathbb{V}\left[\nabla\chi_{L,h}\right]\|_{L^1(D_1)} \leq \frac{\nu_0}{2} \tag{4.8c}$$

hold. Then the error of the MCFEM estimator $\mu_{N,L,h}$ satisfies

$$||\mu_{N,L,h} - \mathbb{E}[\nabla\chi]||^2_{L^2(\Omega;L^2(D_1))} \leq \frac{\nu_0}{N} + \nu_1h^\alpha + \nu_2L^{-\beta}. \tag{4.9}$$

Proof. We have, using the triangle inequality,

$$\begin{aligned}
&\|\mu_{N,L,h} - \mathbb{E}[\nabla\chi]\|^2_{L^2(\Omega;L^2(D_1))} \\
&\quad\leq \mathbb{E}\left[\left(\|\mu_{N,L,h} - \mathbb{E}(\mu_{N,L,h})\|_{L^2(D_1)} + \|\mathbb{E}(\mu_{N,L,h}) - \mathbb{E}(\nabla\chi)\|_{L^2(D_1)}\right)^2\right] \qquad (4.10)\\
&\quad\leq 2\mathbb{E}\left[\|\mu_{N,L,h} - \mathbb{E}(\mu_{N,L,h})\|^2_{L^2(D_1)}\right] + 2\mathbb{E}\left[\|\mathbb{E}(\mu_{N,L,h}) - \mathbb{E}(\nabla\chi)\|^2_{L^2(D_1)}\right] = F_1 + F_2.
\end{aligned}$$

We used the simple inequality $(a+b)^2 \leq 2(a^2+b^2)$ to obtain the last line. Now,

$$\begin{aligned}
F_1 &= 2\int_{D_1} \mathbb{E}\left(\mu_{N,L,h} - \mathbb{E}(\mu_{N,L,h})\right)^2 \, \mathrm{d}y = 2\int_{D_1} \mathbb{V}\left(\mu_{N,L,h}\right) \, \mathrm{d}y \\
&= 2\|\mathbb{V}\left(\mu_{N,L,h}\right)\|_{L^1(D_1)} = 2\|\mathbb{V}\left(\frac{1}{N}\sum_{i=1}^{N}\nabla\chi^{(i)}_{L,h}\right)\|_{L^1(D_1)} \qquad (4.11)\\
&\leq \frac{2}{N^2}\sum_{i=1}^{N}\|\mathbb{V}\left(\nabla\chi^{(i)}_{L,h}\right)\|_{L^1(D_1)} \leq \frac{2}{N}\|\mathbb{V}\left(\nabla\chi_{L,h}\right)\|_{L^1(D_1)} \leq \frac{\nu_0}{N}.
\end{aligned}$$

Using $\mathbb{E}\left(\mu_{N,L,h}\right) = \mathbb{E}\left(\nabla\chi_{L,h}\right)$ since $\mu_{N,L,h}$ is an unbiased estimator,

$$\begin{aligned}
F_2 &\leq 2\mathbb{E}\left(\|\mathbb{E}\left(\nabla\chi_{L,h}\right) - \mathbb{E}\left(\nabla\chi_h\right)\|_{L^2(D_1)} + \|\mathbb{E}\left(\nabla\chi_h\right) - \mathbb{E}\left(\nabla\chi\right)\|_{L^2(D_1)}\right)^2 \\
&\leq 4\mathbb{E}\left(\|\nabla\chi_{L,h} - \nabla\chi_h\|^2_{L^2(D_1)}\right) + 4\mathbb{E}\left(\|\nabla\chi_h - \nabla\chi\|^2_{L^2(D_1)}\right) \qquad (4.12)\\
&\leq \nu_2 L^{-\beta} + \nu_1 h^{\alpha}.
\end{aligned}$$

Combining the estimates in equations (4.11) and (4.12) yields the desired result.

□

4.2 Optimal Monte Carlo Method

In any homogenization problem, and especially in the case of stochastic homogenization, it would be useful to construct a method that will yield the lowest error with the least amount of computational effort. In the following, we will present an optimal approach to computing the solution to the cell problem (4.1).

As we have already discussed, there are three main sources of error in this problem: discretization error, error due to the boundary, and statistical error. The error takes the general form

$$E(N,L,h) = \frac{\nu_0}{\sqrt{N}} + \nu_1 h^{\alpha} + \nu_2 L^{-\beta},$$

where the ν_i, α, and β are positive constants. N, L and h refer to the number of Monte Carlo simulations, the length of the sample domain and the mesh fineness, respectively. It is noted that in order for N, L and h to be physically meaningful, these are all positive numbers.

The general expression for computational work is given by

$$W(N,L,h) = N\sum_{k=1}^{n}\mu_k L^{\gamma_k}h^{-\xi_k}, \tag{4.13}$$

where n indicates the number of differently scaled steps in the computation and μ_k, γ_k, and ξ_k are also positive constants. To minimize the objective function $W(N,L,h)$ given the inequality constraint $E(N,L,h) \leq \varepsilon$ for a given error tolerance ε, we employ Karush-Kuhn-Tucker (KKT) conditions. To summarize, we wish to minimize the Lagrange function

$$\mathcal{L}(N,L,h,s) = W(N,L,h) + s(E(N,L,h) - \varepsilon). \tag{4.14}$$

Proposition 4.7. *Assuming an error of the form*

$$E(N,L,h) = \frac{\nu_0}{\sqrt{N}} + \nu_1 h^{\alpha} + \nu_2 L^{-\beta},$$

the optimal choice of N, L, and h to achieve minimal error $E(N,L,h) \leq \varepsilon$ is equivalent to solving the system of equations

$$\sum_{k=1}^{n}\left(\nu_1\alpha\gamma_k h^{\alpha} - \nu_2\beta\xi_k L^{-\beta}\right)\mu_k L^{\gamma_k}h^{-\xi_k} = 0, \tag{4.15a}$$

$$\sqrt{N} - \frac{\nu_0}{\varepsilon - \nu_1 h^{\alpha} - \nu_2 L^{-\beta}} = 0, \tag{4.15b}$$

$$\sum_{k=1}^{n}\left(\gamma_k - \frac{2\nu_2\beta L^{-\beta}}{\varepsilon - \nu_1 h^{\alpha} - \nu_2 L^{-\beta}}\right)\mu_k L^{\gamma_k}h^{-\xi_k} = 0. \tag{4.15c}$$

Proof. The Lagrange function is given by

$$\mathcal{L}(N,L,h,s) = N\sum_{k=1}^{n}\mu_k L^{\gamma_k}h^{-\xi_k} + s\left(\frac{\nu_0}{\sqrt{N}} + \nu_1 h^{\alpha} + \nu_2 L^{-\beta} - \varepsilon\right).$$

The necessary conditions for a minimum translate to the system of equations

$$\frac{\partial\mathcal{L}}{\partial N} = \sum_{k=1}^{n}\mu_k L^{\gamma_k}h^{-\xi_k} - s\frac{\nu_0}{2N^{3/2}} = 0, \tag{4.16a}$$

$$\frac{\partial\mathcal{L}}{\partial L} = N\sum_{k=1}^{n}\gamma_k\mu_k L^{\gamma_k-1}h^{-\xi_k} - s\nu_2\beta L^{-\beta-1} = 0, \tag{4.16b}$$

$$\frac{\partial\mathcal{L}}{\partial h} = -N\sum_{k=1}^{n}\xi_k\mu_k L^{\gamma_k}h^{-\xi_k-1} + s\nu_1\alpha h^{\alpha-1} = 0, \tag{4.16c}$$

$$\frac{\partial\mathcal{L}}{\partial s} = \frac{\nu_0}{\sqrt{N}} + \nu_1 h^{\alpha} + \nu_2 L^{-\beta} - \varepsilon = 0. \tag{4.16d}$$

Clearly, condition (4.16a) implies

$$s = \frac{2N^{3/2}}{\nu_0}\sum_{k=1}^{n}\mu_k L^{\gamma_k}h^{-\xi_k}.$$

Inserting s into (4.16b) yields

$$N\sum_{k=1}^{n}\gamma_k\mu_k L^{\gamma_k-1}h^{-\xi_k}-\left(\frac{2N^{3/2}}{\nu_0}\sum_{k=1}^{n}\mu_k L^{\gamma_k}h^{-\xi_k}\right)\nu_2\beta L^{-\beta-1}=0$$

$$\iff \qquad \sum_{k=1}^{n}\left(\gamma_k-\frac{2\sqrt{N}\nu_2\beta L^{-\beta}}{\nu_0}\right)\mu_k L^{\gamma_k-1}h^{-\xi_k}=0 \tag{4.17}$$

$$\iff \qquad \sum_{k=1}^{n}\mu_k L^{\gamma_k-1}h^{-\xi_k}=\frac{\nu_0}{2\sqrt{N}\nu_2\beta L^{-\beta}}\sum_{k=1}^{n}\gamma_k\mu_k L^{\gamma_k-1}h^{-\xi_k}. \tag{4.18}$$

Similarly, inserting s into (4.16c) yields

$$-N\sum_{k=1}^{n}\xi_k\mu_k L^{\gamma_k}h^{-\xi_k-1}+\left(\sum_{k=1}^{n}\frac{2N^{3/2}}{\nu_0}\mu_k L^{\gamma_k}h^{-\xi_k}\right)\nu_1\alpha h^{\alpha-1}=0$$

$$\iff \qquad \sum_{k=1}^{n}\left(-\xi_k+\frac{2\sqrt{N}\nu_1\alpha h^{\alpha}}{\nu_0}\right)\mu_k L^{\gamma_k}h^{-\xi_k-1}=0$$

$$\iff \qquad \sum_{k=1}^{n}\left(-\xi_k+\frac{2\sqrt{N}\nu_1\alpha h^{\alpha}}{\nu_0}\right)\mu_k L^{\gamma_k-1}h^{-\xi_k}=0$$

$$\iff \qquad \sum_{k=1}^{n}\mu_k L^{\gamma_k-1}h^{-\xi_k}=\frac{\nu_0}{2\sqrt{N}\nu_1\alpha h^{\alpha}}\sum_{k=1}^{n}\xi_k\mu_k L^{\gamma_k-1}h^{-\xi_k}. \tag{4.19}$$

Combining (4.18) and (4.19), we see that

$$\frac{1}{\nu_2\beta L^{-\beta}}\sum_{k=1}^{n}\gamma_k\mu_k L^{\gamma_k-1}h^{-\xi_k}=\frac{1}{\nu_1\alpha h^{\alpha}}\sum_{k=1}^{n}\xi_k\mu_k L^{\gamma_k-1}h^{-\xi_k}. \tag{4.20}$$

Clearly, this is equivalent to (4.15a). Next, a trivial calculation shows that (4.16d) implies

$$\sqrt{N}=\frac{\nu_0}{\varepsilon-\nu_1h^{\alpha}-\nu_2L^{-\beta}},$$

which is equivalent to (4.15b). We substitute the expression for $\sqrt{N}$ into (4.17) to get

$$\sum_{k=1}^{n}\left(\gamma_k-\frac{2\nu_2\beta L^{-\beta}}{\varepsilon-\nu_1h^{\alpha}-\nu_2L^{-\beta}}\right)\mu_k L^{\gamma_k-1}h^{-\xi_k}=0,$$

which yields (4.15c). □

In the special case $\gamma_k=\xi_k$ for each step of the work, which we will later be able to justify numerically, we can simplify even further:

Corollary 4.8. *Under the additional assumption that $\gamma_k = \xi_k$ for every $k \in \{1, \dots, n\}$, the optimal choice of N, L, and h to achieve minimal error $E(N, L, h) \leq \varepsilon$ is equivalent to solving the equation*

$$\sum_{k=1}^{n} \left(\gamma_k - \frac{2\nu_2 \beta L^{-\beta}}{\varepsilon - \nu_1 h^{\alpha} - \nu_2 L^{-\beta}} \right) \mu_k L^{\gamma_k - 1} \left(\frac{\nu_2 \beta}{\nu_1 \alpha L^{\beta}} \right)^{-\xi_k/\alpha} = 0.$$

The resulting h is then given by

$$h = \left(\frac{\nu_2 \beta}{\nu_1 \alpha L^{\beta}} \right)^{1/\alpha}$$

and N is given by

$$N = \left(\frac{\nu_0}{\varepsilon - \nu_1 h^{\alpha} - \nu_2 L^{-\beta}} \right)^2.$$

Proof. Since $\gamma_k = \xi_k$, we follow from (4.20) that

$$h = \left(\frac{\nu_2 \beta}{\nu_1 \alpha L^{\beta}} \right)^{1/\alpha}.$$

We can substitute this into the expressions for h and N into (4.17) to get the equation

$$\sum_{k=1}^{n} \left(\gamma_k - \frac{2\nu_2 \beta L^{-\beta}}{\varepsilon - \nu_1 h^{\alpha} - \nu_2 L^{-\beta}} \right) \mu_k L^{\gamma_k - 1} \left(\frac{\nu_2 \beta}{\nu_1 \alpha L^{\beta}} \right)^{-\xi_k/\alpha} = 0,$$

which we can solve numerically for L. □

Remark 4.9. The setup of the solution procedure is such that solutions can be computed in parallel. If we have m cores available for computing, then we should adjust the work function to

$$W = \frac{N}{m} \sum_{k=1}^{n} \mu_k L^{\gamma_k} h^{-\xi_k}.$$

4.2.1 Optimal Method Using the Alternate Error Estimate

In this section, we will state the optimal method using the alternate error estimate. The proofs for these methods are nearly identical, so they will be omitted.

Proposition 4.10. *Assuming an error of the form*

$$E(N, L, h) = \frac{\nu_0}{N} + \nu_1 h^{\alpha} + \nu_2 L^{-\beta},$$

the optimal choice of N, L, and h to achieve minimal error $E(N, L, h) \leq \varepsilon$ is equivalent to solving the system of equations

$$\sum_{k=1}^{n} \left(\nu_1 \alpha \gamma_k h^{\alpha} - \nu_2 \beta \xi_k L^{-\beta}\right) \mu_k L^{\gamma_k - 1} h^{-\xi_k} = 0, \tag{4.21a}$$

$$N - \frac{\nu_0}{\varepsilon - \nu_1 h^{\alpha} - \nu_2 L^{-\beta}} = 0, \tag{4.21b}$$

$$\sum_{k=1}^{n} \left(\gamma_k - \frac{\nu_2 \beta L^{-\beta}}{\varepsilon - \nu_1 h^{\alpha} - \nu_2 L^{-\beta}}\right) \mu_k L^{\gamma_k - 1} h^{-\xi_k} = 0. \tag{4.21c}$$

Corollary 4.11. *Under the additional assumption that $\gamma_k = \xi_k$, the optimal combination of N, L, and h to achieve minimal error $E(N, L, h) \leq \varepsilon$ is equivalent to solving the equation*

$$\sum_{k=1}^{n} \left(\gamma_k - \frac{\nu_2 \beta L^{-\beta}}{\varepsilon - \nu_1 h^{\alpha} - \nu_2 L^{-\beta}}\right) \mu_k L^{\gamma_k - 1} \left(\frac{\nu_2 \beta}{\nu_1 \alpha L^{\beta}}\right)^{-\xi_k/\alpha} = 0$$

for a given maximal error ε for L. The resulting h is then given by

$$h = \left(\frac{\nu_2 \beta}{\nu_1 \alpha L^{\beta}}\right)^{1/\alpha}$$

and N is given by

$$N = \frac{\nu_0}{\varepsilon - \nu_1 h^{\alpha} - \nu_2 L^{-\beta}}.$$

The alternate error estimate is based on the mean square error, which was hoped to provide a sharper estimate. In our case, in addition to the statistical error, we have two other errors (due to the boundary conditions and the discretization) that require us, twice, to use the rather coarse estimate $(a + b)^2 \leq 2(a^2 + b^2)$. Thus, the first error estimate in Theorem 4.3 is preferred, as we do not need to make estimates of this kind. The alternate error estimate is stated here for completeness. Additionally, the alternate optimal method shows that it is still possible to determine an optimal method if the statistical error scales differently from $\mathcal{O}(N^{-1/2})$. This would be of use should an optimization problem be solved using, for instance, error estimates obtained through quasi-Monte Carlo sampling instead of Monte Carlo sampling, the former of which generally converges faster.

CHAPTER 5

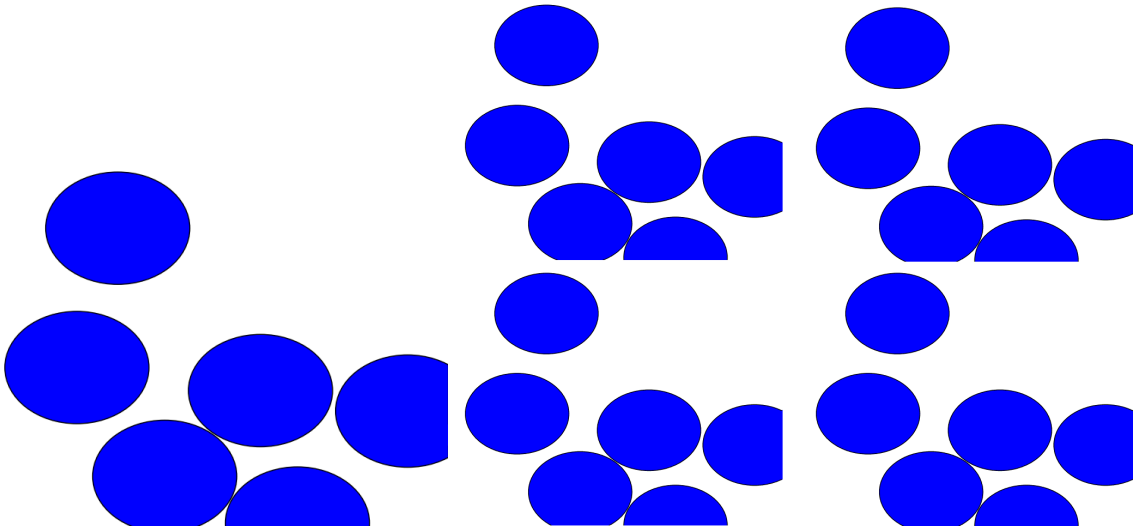
Numerical Results

In this chapter, we will discuss the results of the numerical tests. First, we will discuss the basic setup and introduce notation. Then, we will see how the error of the solution to the cell problem scales as a function of domain length, mesh fineness, and number of realizations. In the next section, given a sample problem, we will determine the optimal strategy for solving the cell problem given certain error bounds. In the final section, we will present the results of tests done on the effective diffusion tensor $\bar{A}$.

5.1 General Setup of Numerical Tests

We will now discuss the general setup of the numerical tests that are used for the remainder of this chapter. Our focus will be on the numerical solution to the cell problem as well as the related effective coefficient matrix. In all tests, we will use a domain $D_L = [-L/2, L/2)^2 \subset \mathbb{R}^2$ in the two-dimensional plane. To remain consistent with the theory presented in previous chapters, we will assume that A is periodic, and hence the first-order corrector χ from the cell problem. Our goal is to simulate a hard-sphere system, i.e. a composite that has circular inclusions that cannot overlap. We will assume that the reference domain has a fixed number of circles. Such systems have many applications, including liquids, glasses, fiber-reinforced composites, particulate composites, packed beds and granular media [31].

For the generation of an set of hard spheres, we use a random sequential addition (RSA) process. In this process, inclusions are placed randomly, irreversibly and sequentially such that none are overlapping. If an inclusion overlaps another inclusion in the process, another attempt is made until a nonoverlapping placement can be made. Of course, this method is subject to a limit at which no further inclusions can be placed; for RSA, the saturation limit in $\mathbb{R}^2$ for circles with identical



(a) Circles randomly generated using a RSA process without periodization on the torus $\mathbb{T}^2$.

(b) A naive periodization of the unit cell from (a), where the cell is repeated on a 2×2 grid. New shapes are introduced.

Figure 5.1: Incorrect periodization of the unit cell.

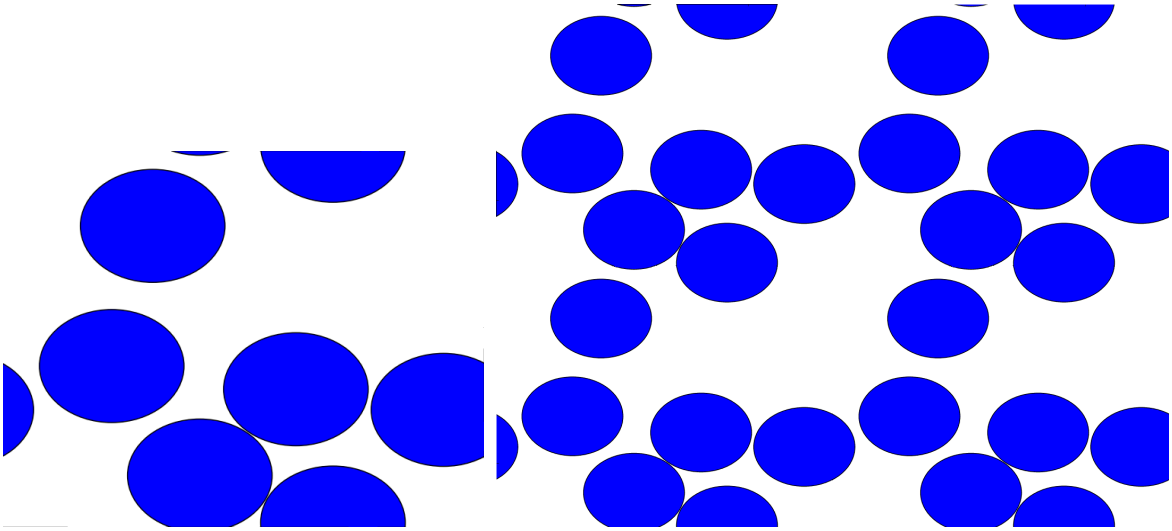
radii is approximately 55 percent coverage [31, p. 87].

The naive approach to periodize the domain D_L is to simply generate inclusions according to the RSA process and then solve the corrector equation on that domain with periodic boundary conditions. This introduces new shapes on the torus, however, so that the process is no longer stationary; see Figure 5.1. The correct way to periodize is to construct the coefficient field A on the *torus*, meaning that inclusions that would be cut on one side are periodically extended to the opposite side; see Figure 5.2. This results in inclusions that are not cut on the boundary, meaning that the statistics of the periodized coefficient field are translation invariant, just like the original matrix A. This process also produces an ergodic ensemble [31, p. 148]. The coefficient field A created in this manner is a stationary and ergodic; in particular, the theory introduced in Section 2.2 applies. This idea is not limited to a configuration with hard spheres; Gloria [14] describes this periodization applied to a homogeneous material with spherical inclusions that are distributed according to a Poisson point process.

All tests in this chapter began with the generation of a domain D_L of length L with a fixed number of inclusions, which were periodically extended as in Figure 5.2. For such a configuration, a new mesh needed to be created that was aligned with the circles; see Figure 5.3. This ensemble then determined the structure of the function A for a given test. Each of these matrices had the form

$$A(y,\omega) = A^{(c)}\mathbb{1}^{(c)}(y,\omega) + A^{(m)}\mathbb{1}^{(m)}(y,\omega),$$

where the function $\mathbb{1}^{(i)}(y,\omega)$ represents the indicator function for the circular inclusions ($i = c$) and the surrounding matrix ($i = m$), respectively. $A^{(c)}$ and $A^{(m)}$ gave the values for the coefficient A in the circles and surrounding matrix. Since



(a) Circles randomly generated using a RSA process with periodization on the torus $\mathbb{T}^2$.

(b) Correct periodization of the unit cell from (a), where the cell is repeated on a 2×2 grid.

Figure 5.2: A correct periodization of the unit cell yields a stationary and ergodic coefficient field.

the mesh was aligned with the circles, each element of the mesh had a constant value for A. In the numerical tests, we will consider two different functions for A: the function

$$A_1(y,\omega) = \begin{pmatrix} 20 & 0 \\ 0 & 10 \end{pmatrix} \mathbb{1}^{(c)}(y,\omega) + \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \mathbb{1}^{(m)}(y,\omega),$$

which will be used later as the example for the optimization problem, and the high-contrast function

$$A_2(y,\omega) = \begin{pmatrix} 200 & 0 \\ 0 & 100 \end{pmatrix} \mathbb{1}^{(c)}(y,\omega) + \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \mathbb{1}^{(m)}(y,\omega).$$

Once a mesh was generated for a given configuration, the cell problem (4.1) could be solved on each element for this randomly generated realization. The solutions for the first and second component of $\chi = (\chi^1, \chi^2)^\top$ using the mesh in Figure 5.3 can be seen in Figure 5.4. The outlines of the circles are clearly visible.

In the following section, we will focus on the convergence of the solution to the cell problem. We recall the following definitions, where we use χ^i as the reference solution (best discrete approximation) to the i^{th} component of the exact solution and χ^i_h as a less exact discrete solution. For a given domain D_L, $\mathcal{T}_h$ will denote the triangulation and K an element of the triangulation. The error in the L^2 norm is

$$\|\chi^i - \chi^i_h\|^2_{L^2(D_L)} = \int_{D_L} |\chi^i - \chi^i_h|^2 \,\mathrm{d}y = \sum_{K \in \mathcal{T}_h} \int_K |\chi^i - \chi^i_h|^2 \,\mathrm{d}y.$$

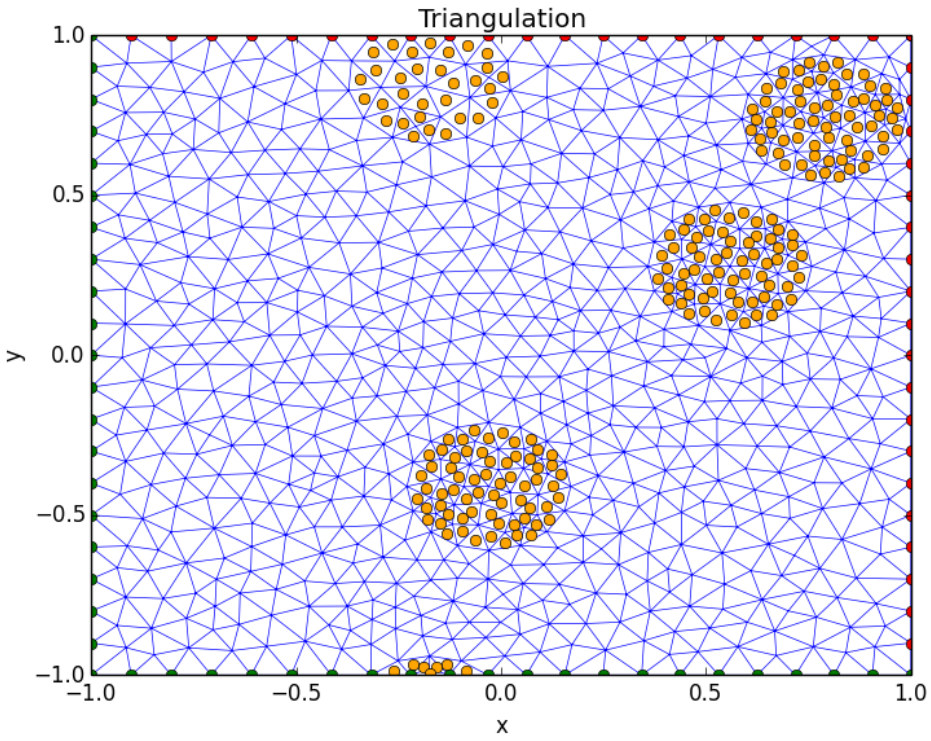
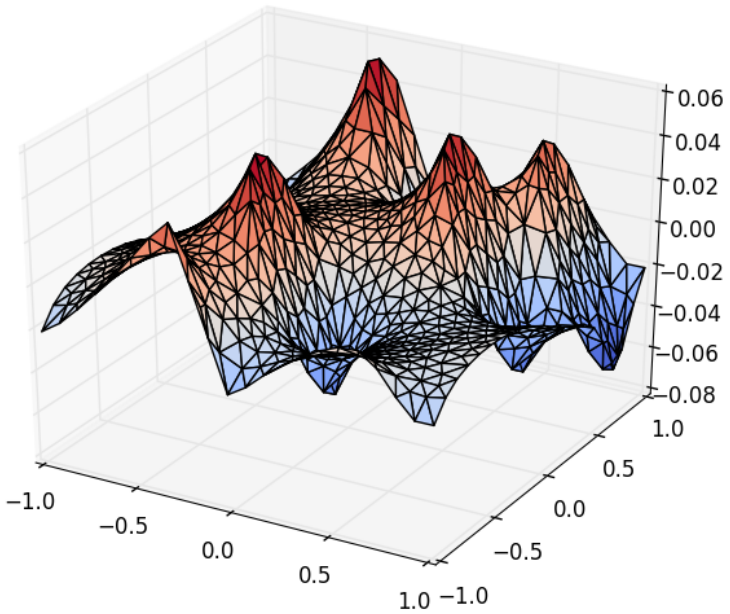
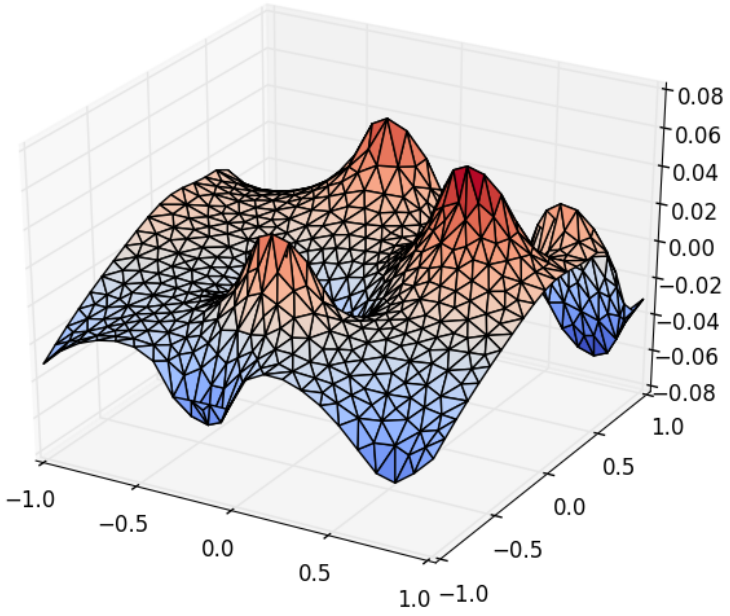


Figure 5.3: A sample mesh created from the random placement of four circles on a domain of length 2. The orange points represent the centroids of elements that belong to circles.



(a) χ_h^1



(b) χ_h^2

Figure 5.4: The discrete solution χ_h, solved on the mesh generated in Figure 5.3.

The error in the H^1 norm is

$$\begin{aligned}\|\chi^i - \chi_h^i\|_{H^1(D_L)}^2 &= \int_{D_L} |\chi^i - \chi_h^i|^2 \,\mathrm{d}y + \int_{D_L} |\nabla\chi^i - \nabla\chi_h^i|^2 \,\mathrm{d}y \\ &= \sum_{K\in\mathcal{T}_h} \int_K |\chi^i - \chi_h^i|^2 \,\mathrm{d}y + \int_K |\nabla\chi^i - \nabla\chi_h^i|^2 \,\mathrm{d}y.\end{aligned}$$

Finally, the error in the H^1 seminorm is

$$|\chi^i - \chi_h^i|_{H^1(D_L)}^2 = \int_{D_L} |\nabla\chi^i - \nabla\chi_h^i|^2 \,\mathrm{d}y = \int_K |\nabla\chi^i - \nabla\chi_h^i|^2 \,\mathrm{d}y.$$

5.2 Numerical Solution to Cell Problem

In this section, we will examine some convergence results for the solution to the cell problem. Throughout, an obstacle in measuring error will be the fact that we do not know the exact solution χ for a given realization of A. It will be possible to approximate this solution, however, in order to obtain convergence rates.

5.2.1 Error of Solution as a Function of Domain Length

In this section, we will see how the error in the numerical solution to the cell problem scales is affected by the domain length. For the reference solution, we used $D_{25} = [-25/2, 25/2)^2$; this is the domain on which the reference solution χ_h is was calculated. A fixed number of circles ($n = 25^2 = 625$) was chosen for the tests. Since each circle had radius $r = 0.2$, this corresponded to approximately 13 percent coverage by circles. A mesh with fixed fineness $h = 0.05$ was chosen.

For each of the samples, a solution χ_h was calculated on the reference domain. Then, for each of the lengths $L \in \{4.0, 6.0, 8.0, \ldots, 24.0\}$, a solution $\chi_{L,h}$ was calculated on the subdomain $D_L \subset D_{25}$. The error between the solution on the subdomain and the solution on the reference domain was computed in the L^2 and H^1 norms and H^1 seminorm as defined in the previous section. To enable a fair comparison and to observe the decreasing influence of the boundary conditions with increasing L, this error was computed on the subdomain $D_3 = [-3/2, 3/2)^2$.

For each length L, to obtain an approximate error in the norm $\|\cdot\|_{L^2(\Omega;X)}$, the square root of the empirical mean of N samples of the error in $\|\chi_h - \chi_{L,h}\|_X^2$ was computed. Then, linear regression was used to fit a relation of the form

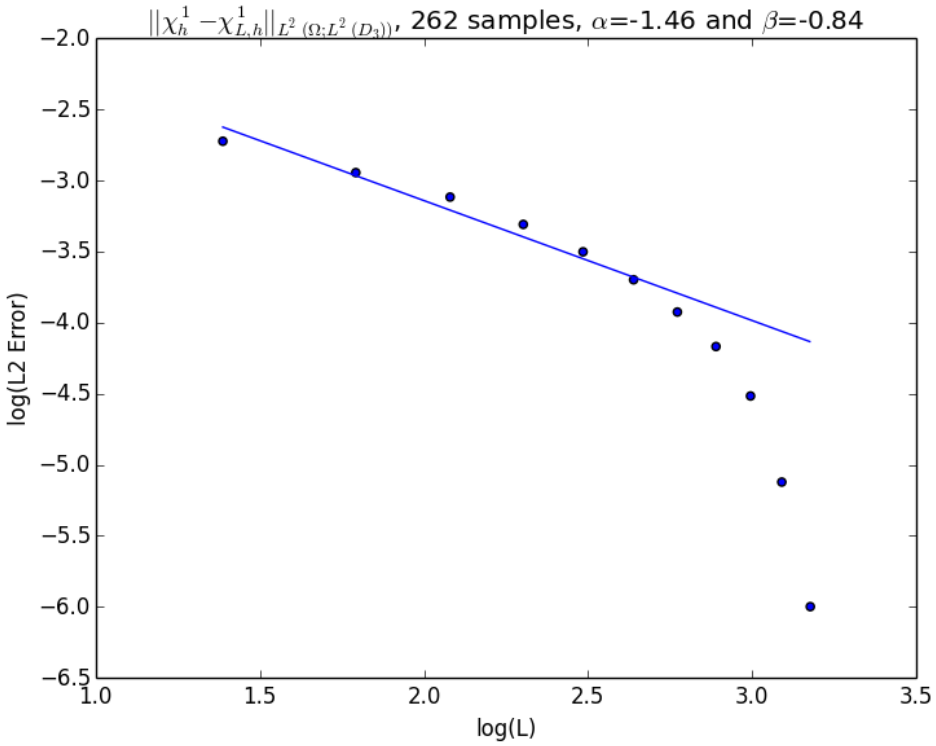
$$\log(\text{error}) = \beta \log(L) + \alpha,$$

which corresponds to an error function of the form

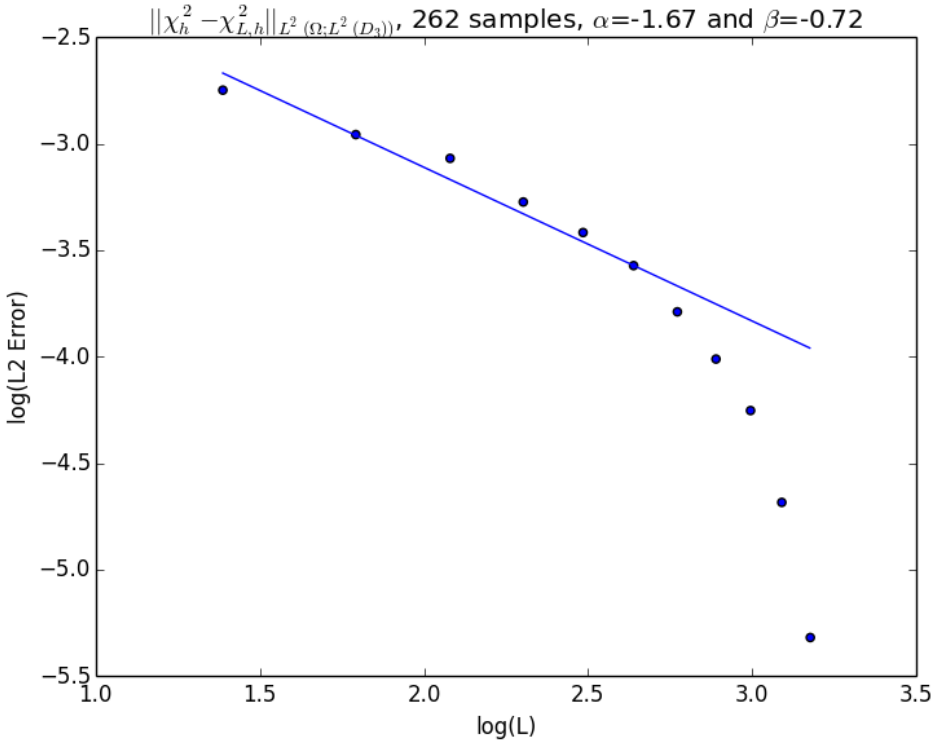
$$\text{error} = L^\beta \cdot e^\alpha.$$

In the first numerical example, we will look at convergence of the solution χ when $A = A_1$. A total of 262 samples were collected for this test. In Figure 5.5, we see the average L^2 error plotted as a function of length L (log-log scale) for each component of $\chi = (\chi_1, \chi_2)^\top$. Convergence is observed with respect to L. In Figures 5.6 and 5.7, convergence is observed in the H^1 norm and H^1 seminorm, respectively. In each plot, we observe an artificial "hyper-linear" convergence in the solution as described by Gloria in [14]; this comes from the fact that we are using a large domain as the reference solution. For that reason, when using linear regression, the last few points were thrown out since they skew the convergence. We see that not only do our results correspond to the theoretical error bound (4.8b), but we are able to estimate the constants in these bounds using numerical results.

In Figures 5.8, 5.9, and 5.10, we see the corresponding results for 191 samples when using the matrix $A = A_2$.

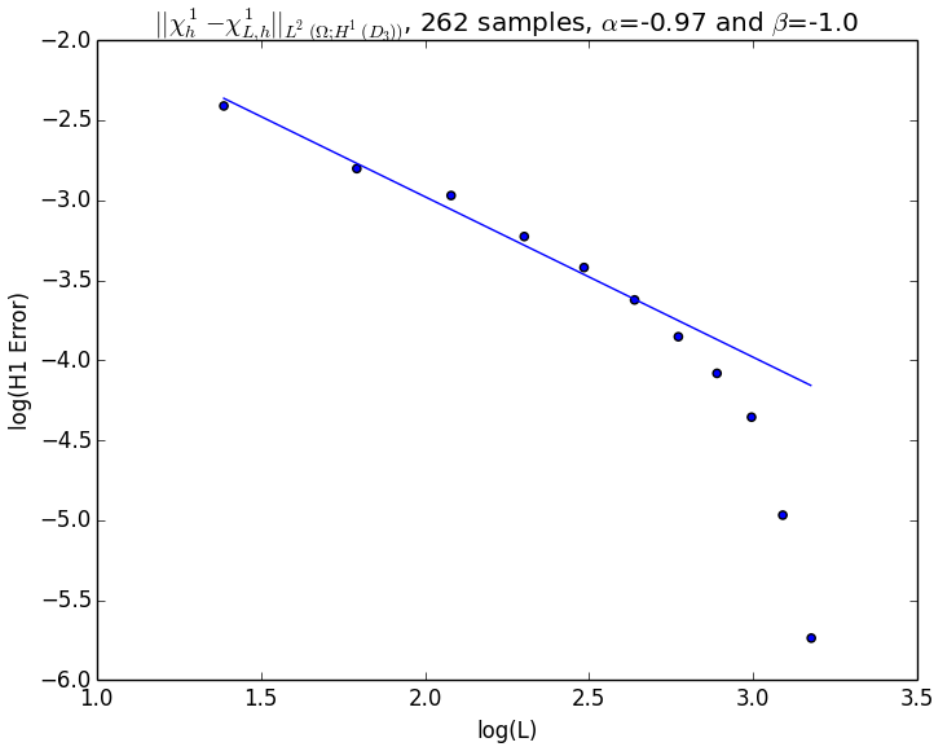


(a) χ^1

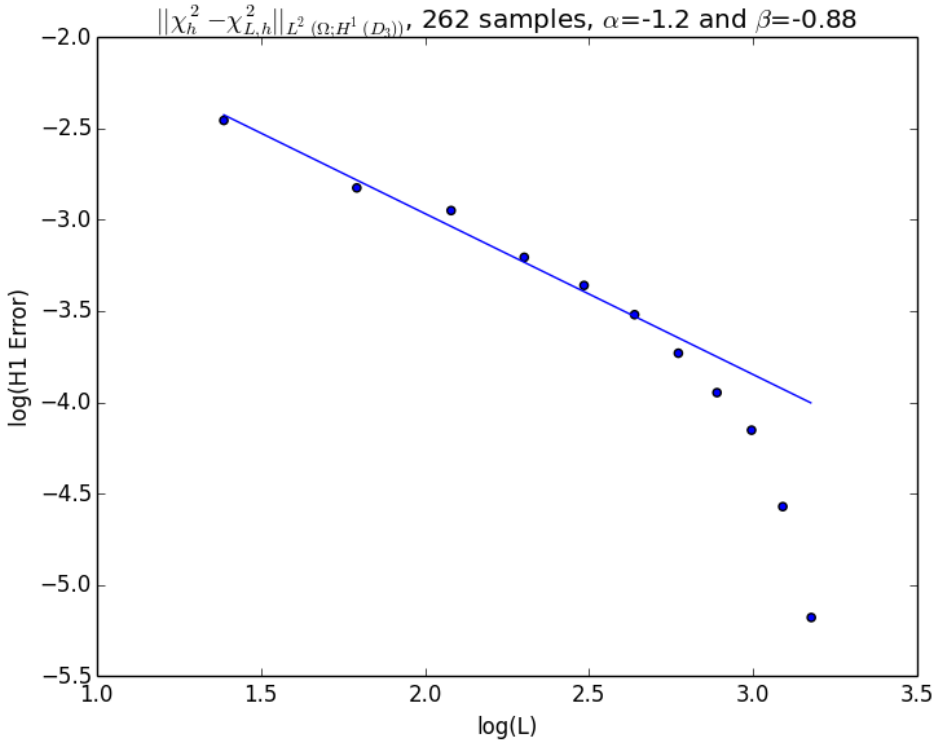


(b) χ^2

Figure 5.5: Convergence of solution χ in the L^2 norm with low-contrast example A_1.

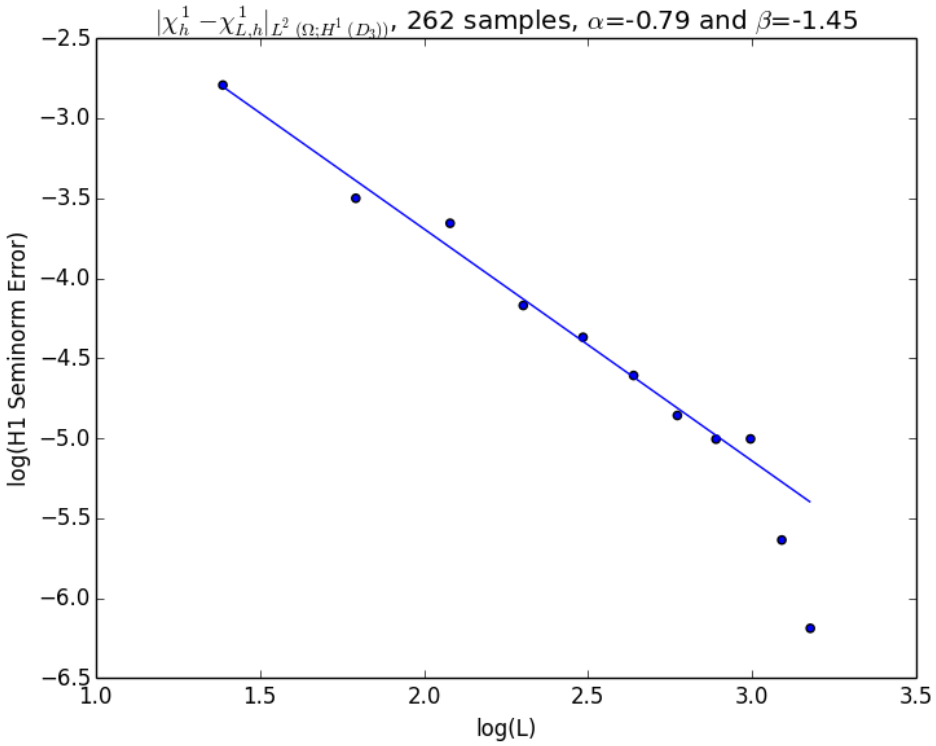


(a) χ^1

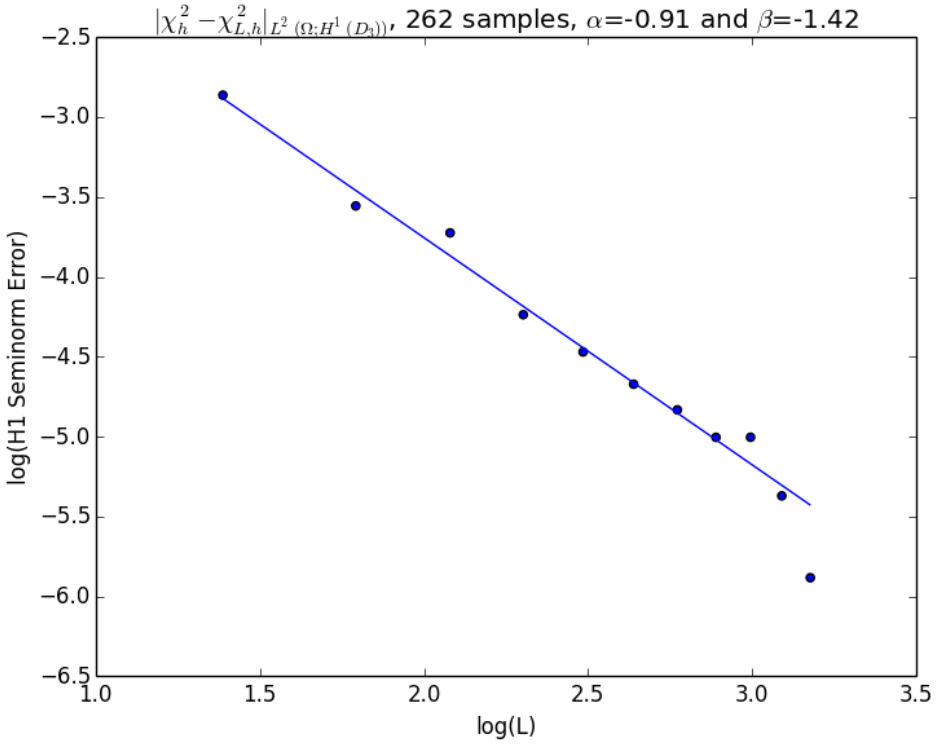


(b) χ^2

Figure 5.6: Convergence of solution χ in the H^1 norm with low-contrast example A_1.

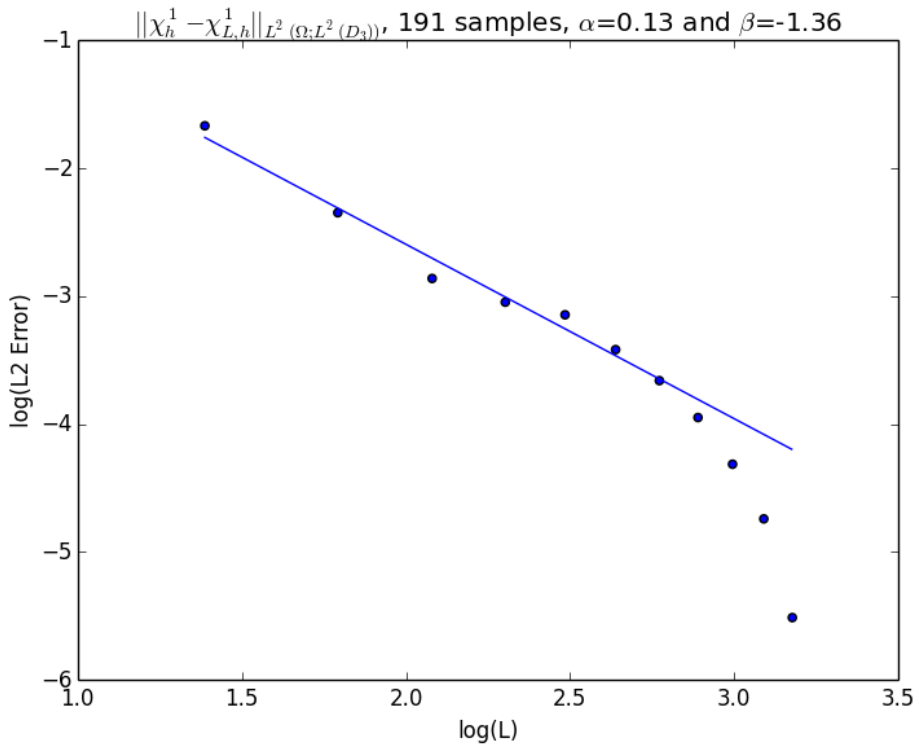


(a) χ^1

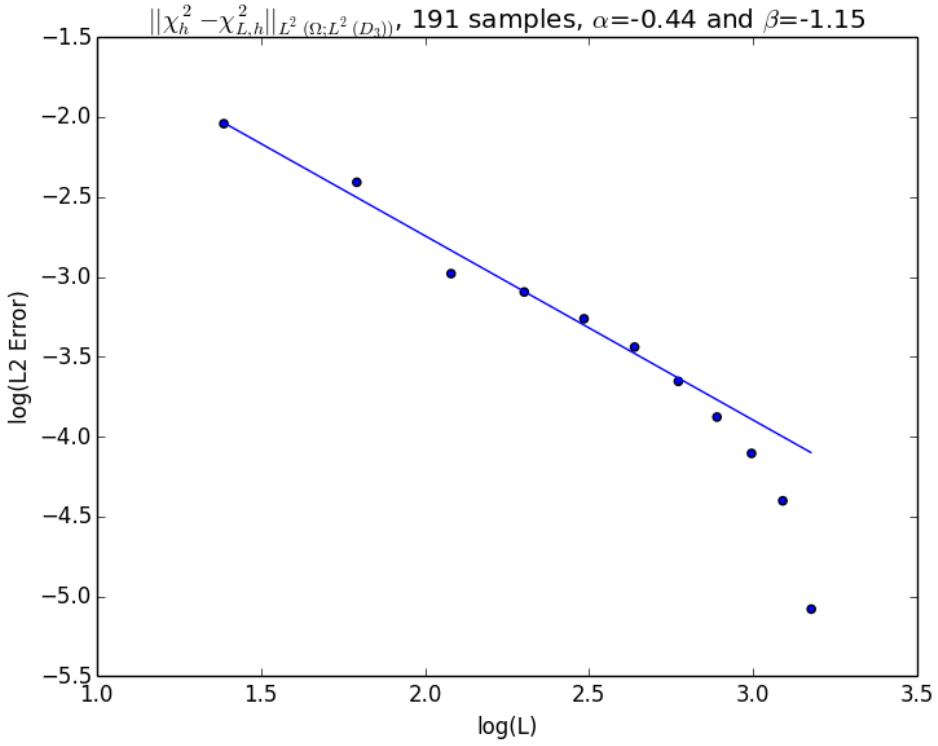


(b) χ^2

Figure 5.7: Convergence of solution χ in the H^1 seminorm with low-contrast example A_1.

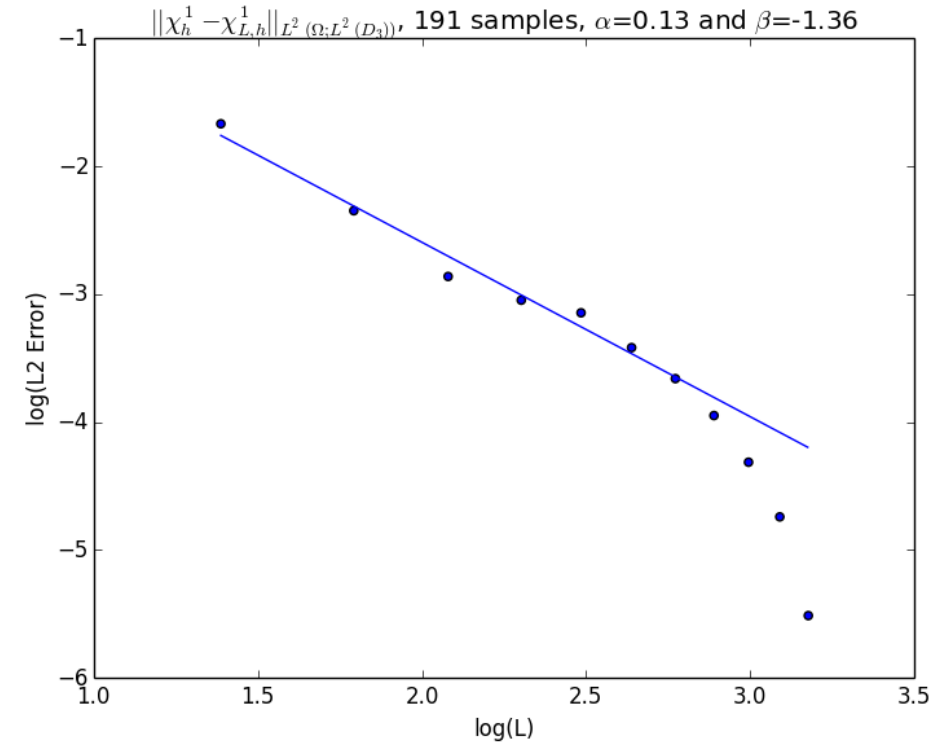


(a) χ^1

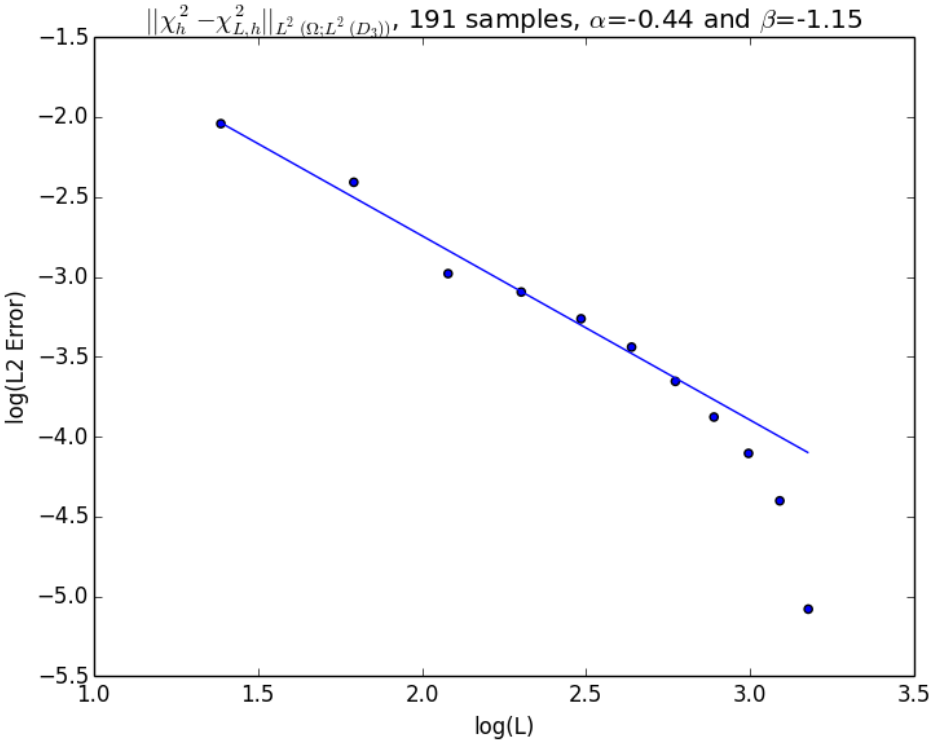


(b) χ^2

Figure 5.8: Convergence of solution χ in the L^2 norm with high-contrast example A_2.

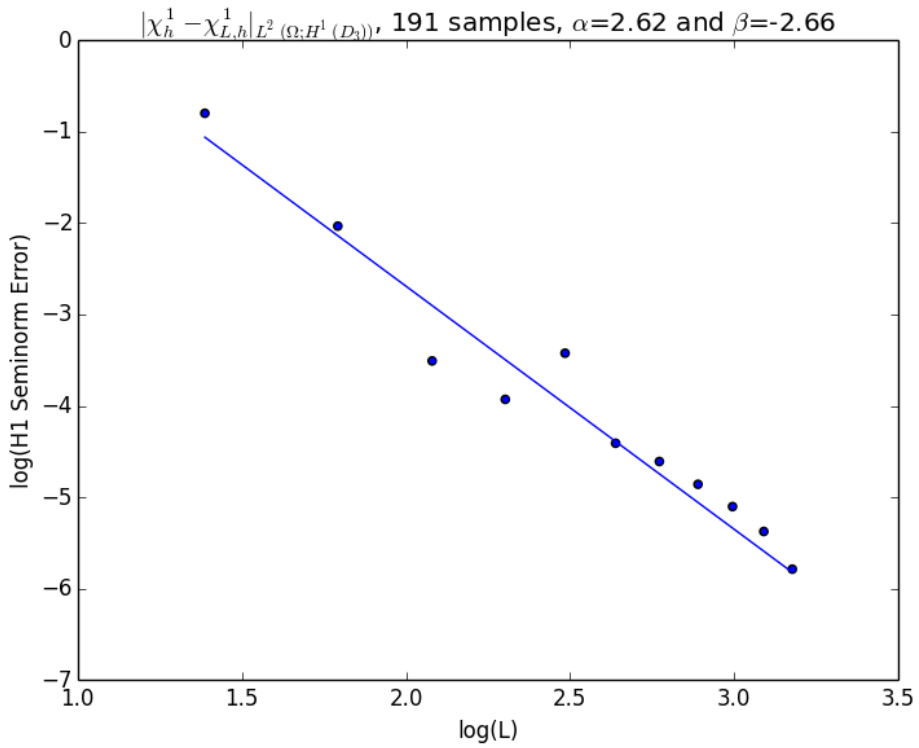


(a) χ^1

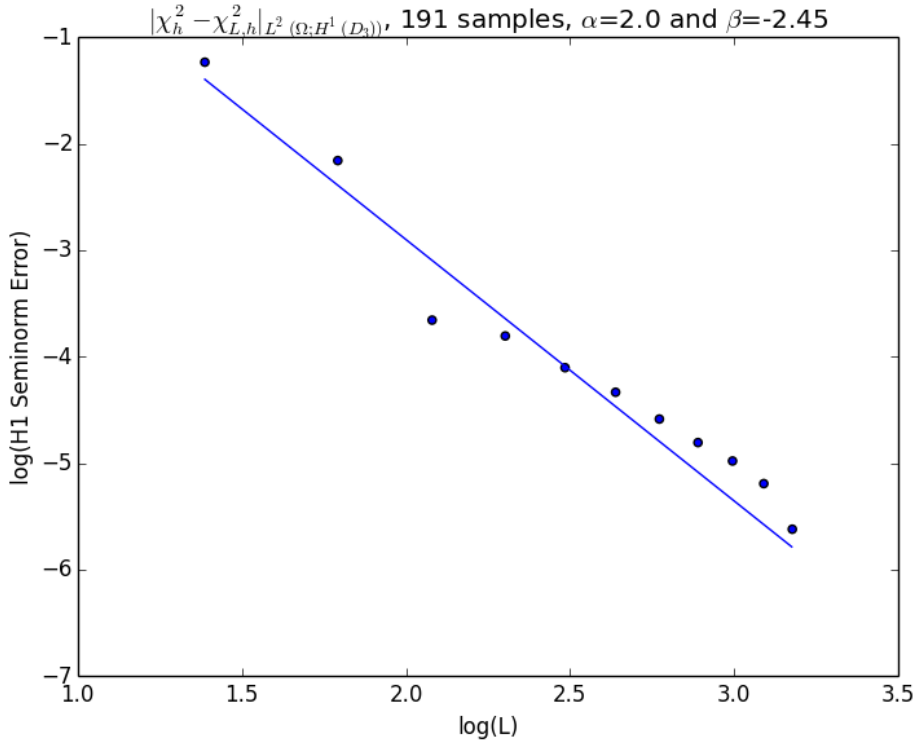


(b) χ^2

Figure 5.9: Convergence of solution χ in the H^1 norm with high-contrast example A_2.



(a) χ^1



(b) χ^2

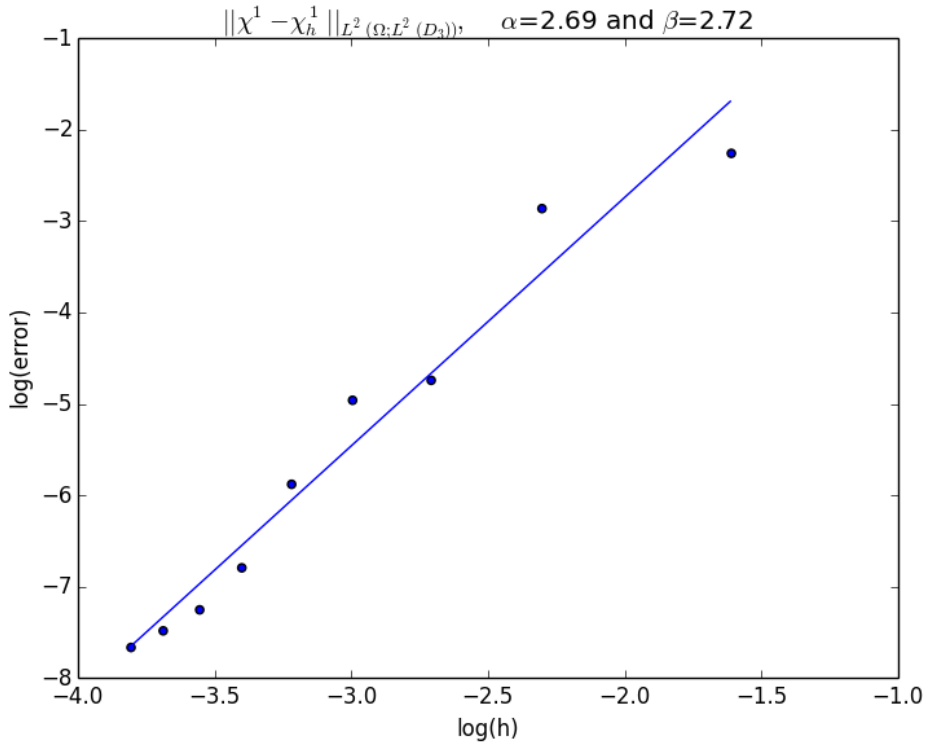
Figure 5.10: Convergence of solution χ in the H^1 seminorm with high-contrast example A_2.

5.2.2 Error as a Function of Mesh Fineness

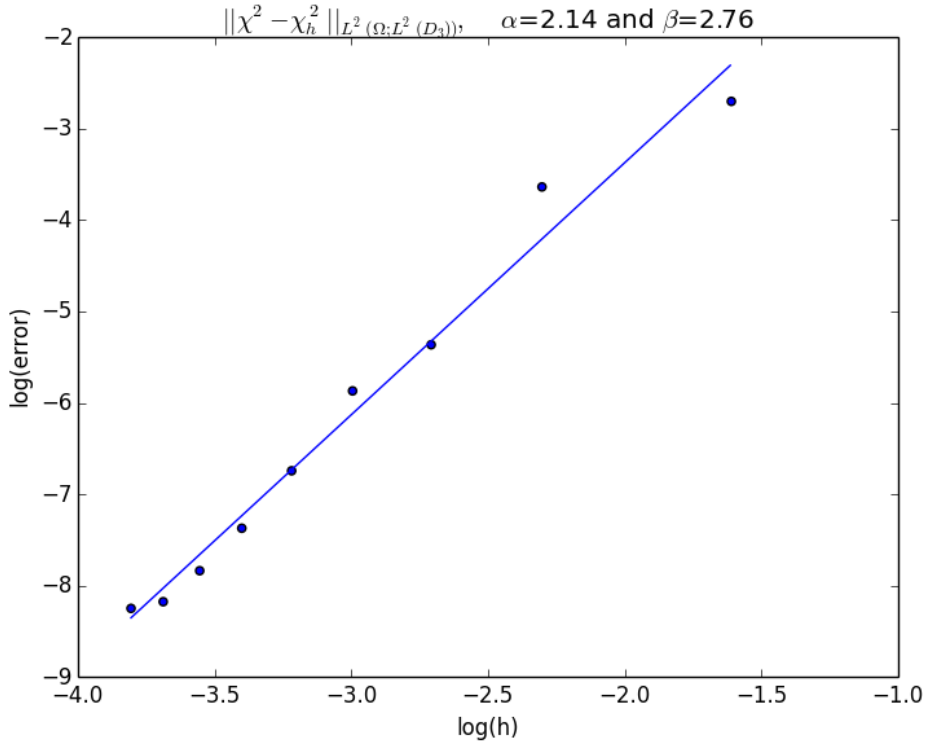
We use a standard method for approximating the error to the discrete solution. Samples were run on the domain $D_3 = [-1.5, 1.5)^2$. For each sample, a set of $n = 3^2 = 9$ circles were randomly placed in the domain according to an RSA process, using a mesh fineness of $h = 0.02$ as the reference solution. Solutions were then calculated on coarser grids with mesh fineness $\tilde{h} \in \{\frac{1}{5}, \frac{1}{10}, \frac{1}{15}, \ldots, \frac{1}{45}\}$. Linear interpolation was then used to calculate each solution onto a structured mesh $\tilde{\mathcal{T}}_h$ with equally spaced nodes $(x, y) = (-1.5 + hi, -1.5 + hj), i, j \in \{0, \ldots, \frac{3}{h}\}$. Then it was easily possible to approximate the error in each norm. The process of interpolating onto a structured mesh was extremely time-intensive, so only nine samples were collected for this test. The small sample size may partially explain why the error in the L^2 norm, see Figure 5.11, scales better than expected; observed error for χ^1 and χ^2 is proportional to $h^{2.72}$ and $h^{2.76}$, respectively. We would expect error to scale more like h^2 since this corresponds to the finite element estimate (4.7). The results in the H^1 norm and H^1 seminorm (Figures 5.12 and 5.13) are on the other hand less than the expected convergence rate of h^1. The main reason why there is a discrepancy between theoretical bounds and observed bounds, however, is likely due to the fact that we use a mesh that is more or less uniform, meaning each element has approximately the same diameter h. To obtain the theoretical convergence rate, it might be necessary to use an adaptive mesh to account for the rapid variation between the inclusions and surrounding domain.

5.2.3 Error as a Function of Number of Samples

Results of the convergence test for the statistical error are shown in Figure 5.14. We ran 2,000 samples of the solution to the cell problem solved on the domain D_6 with a coarse mesh of $h = 0.1$. The value $\mathbb{E}[\nabla \chi_{L,h}]$ was approximated by the sample mean of the gradient using 2,000 solutions; $\mu_{N,h}$ represents the sample mean using $N <$ 2,000 solutions. For this test, a structured mesh was chosen and all values were linearly interpolated onto this mesh. Then, the H^1 seminorm error for χ^1 on the domain D_3 was calculated. Results show that convergence is even better than the expected $N^{-1/2}$ with statistical error proportional to $N^{-0.59}$.

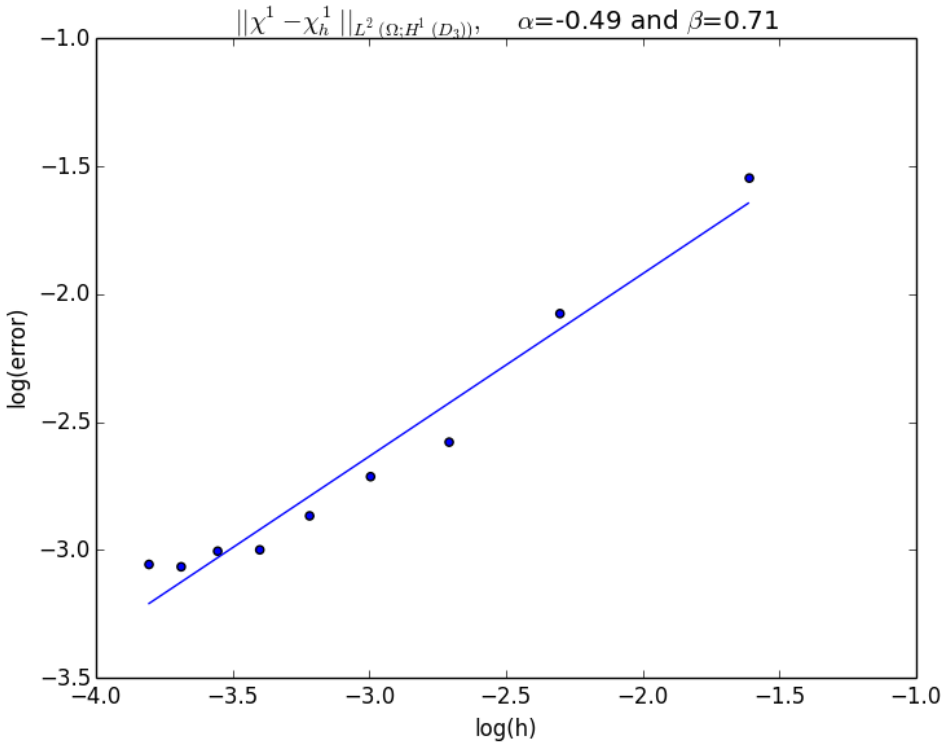


(a) χ^1

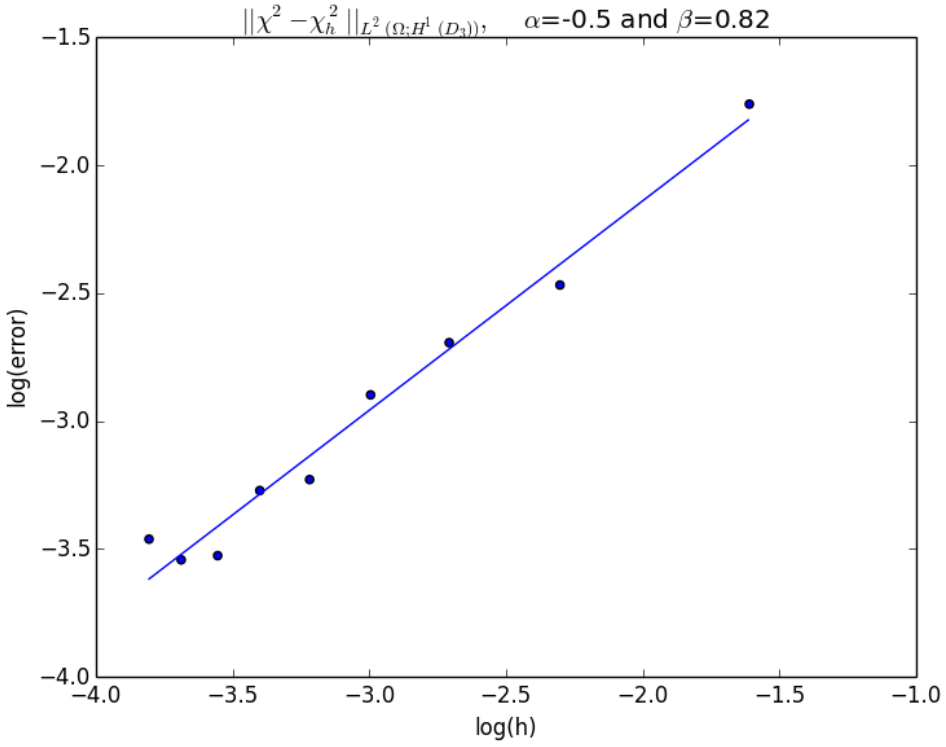


(b) χ^2

Figure 5.11: L^2 error as a function of h on the domain $[-1.5, 1.5)^2$.

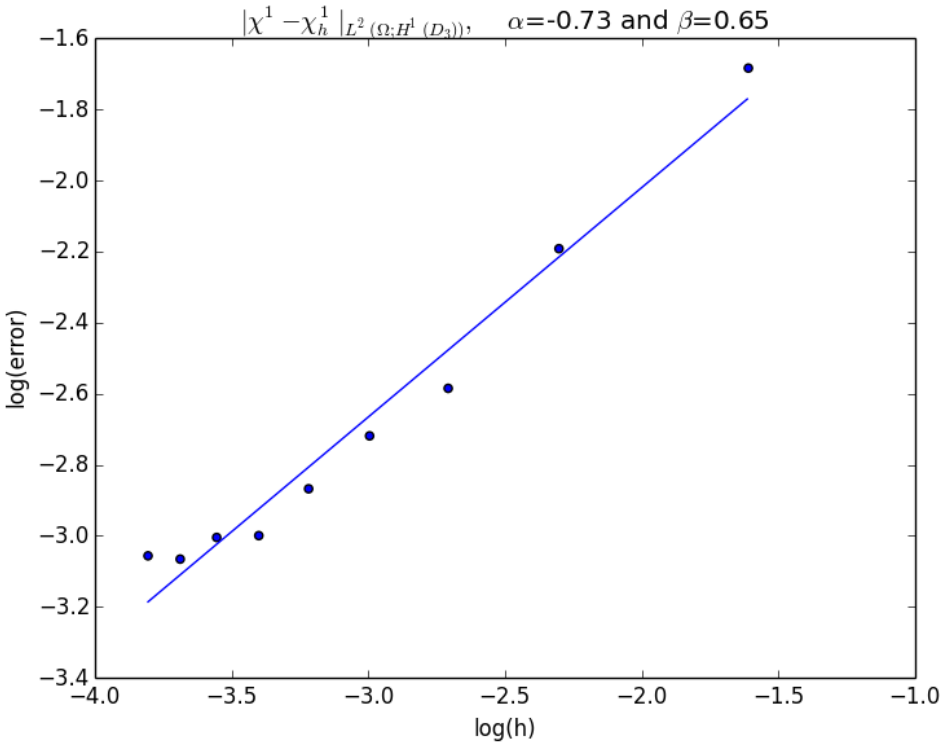


(a) χ^1

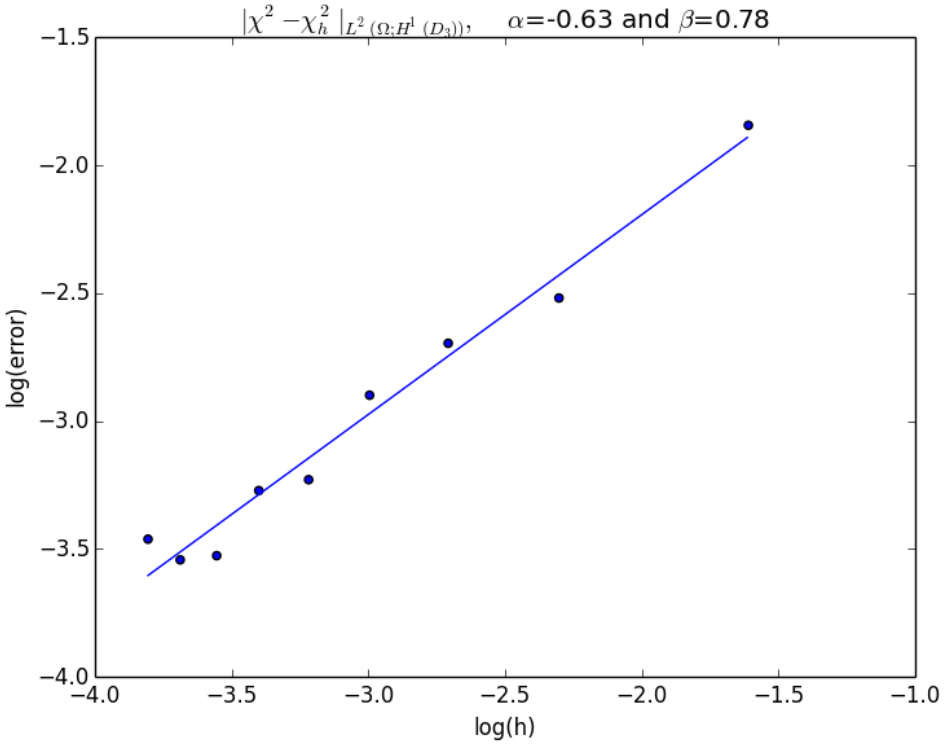


(b) χ^2

Figure 5.12: H^1 error as a function of h on the domain $[-1.5, 1.5)^2$.



(a) χ^1



(b) χ^2

Figure 5.13: H^1 seminorm error as a function of h on the domain $[-1.5, 1.5)^2$.

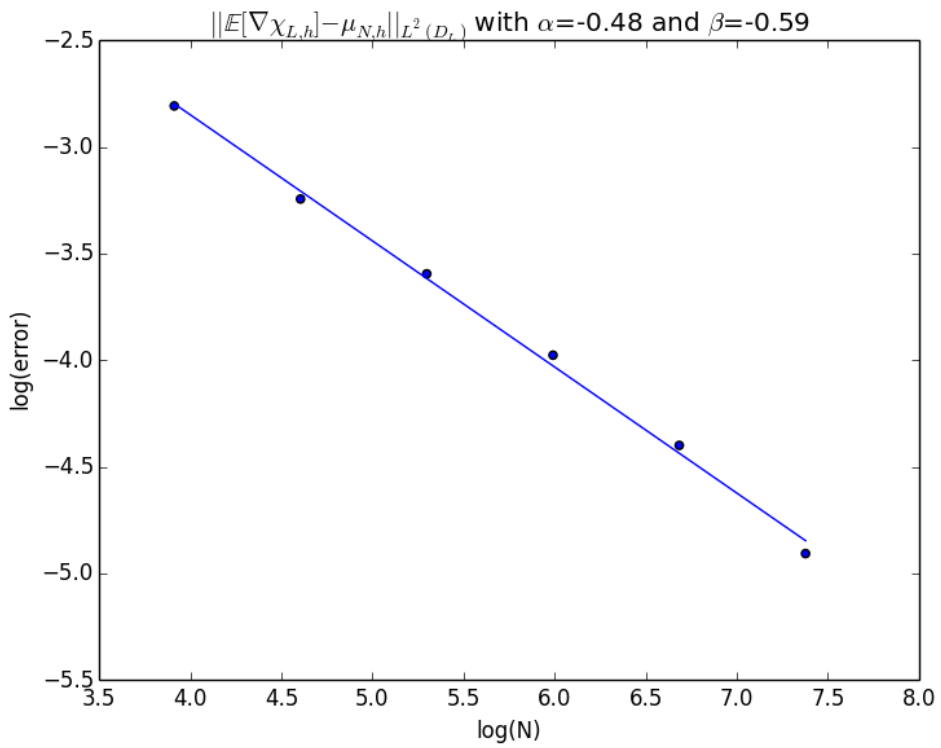


Figure 5.14: χ^1 error in the H^1 seminorm as a function of number of samples N.

5.3 Computation Time

In order to find an optimal method for solving a given problem, we need to minimize equation (4.14). A necessary ingredient for solving this problem is determining the coefficients μ_k, γ_k, ξ_k in the work function (4.13); it was also necessary to determine which parts of the solver scaled differently in time.

For fixed L and varying h, we make the ansatz

$$t = \mu_k h^{-\xi_k}$$

for each step. Discrete steps were determined by testing how certain functionality scaled as a function of mesh fineness h and domain length L. Those that scaled differently were split into groups. In order of function calls, times t_i were measured for the groups

1. Mesh generation;
2. Stiffness matrix M_i and load vector l_i assembly (one call for each component);
3. Solving of the linear system $M_i^{-1} l_i$ for $i \in \{1, 2\}$;
4. Application of the uniqueness condition $\int \chi_i \, \mathrm{d}y = 0$.

L	μ_1	ξ_1	μ_2	ξ_2	μ_3	ξ_3	μ_4	ξ_4
10.0	-3.877	-1.896	-5.295	-1.771	-5.814	-2.412	-5.302	-2.093
15.0	-2.928	-1.892	-5.129	-2.010	-4.808	-2.473	-5.261	-2.433
20.0	-2.342	-1.888	-4.709	-2.081	-4.578	-2.706	-4.988	-2.607

Table 5.1: Computation time $t = \mu_i h^{\xi_i}$ for fixed length L and different values of h for the four steps in the work function.

h	μ_1	γ_1	μ_2	γ_2	μ_3	γ_3	μ_4	γ_4
0.02	-1.639	2.285	-4.930	2.898	-4.096	3.415	-4.571	3.315
0.03	-2.326	2.229	-4.108	2.233	-5.545	3.521	-5.254	3.195
0.04	-2.784	2.154	-5.387	2.404	-4.878	2.944	-5.009	2.745
0.05	-3.528	2.318	-4.427	1.982	-5.914	3.187	-5.268	2.685
0.06	-3.302	2.111	-4.596	1.904	-5.747	2.962	-5.130	2.489
0.07	-4.111	2.288	-6.554	2.492	-5.645	2.749	-5.408	2.434
0.08	-4.375	2.215	-5.654	2.067	-6.775	2.971	-4.945	2.113
0.09	-4.561	2.274	-5.468	1.969	-6.110	2.673	-5.640	2.336
0.1	-4.555	2.268	-4.648	1.604	-7.200	3.085	-5.680	2.346

Table 5.2: Computation time $t = \mu_i L^{\gamma_i}$ for fixed length h and different values of L for the four steps in the work function.

In Table 5.1, for different fixed domain lengths L, we observe how the computation time scaled with respect to h for the four steps of our solver. Columns μ_1 and ξ_1 correspond to mesh generation, where we see that the scaling ξ_1 is about what we would expect, given that our mesh generator advertises linear assembly time (as a function of *nodes*). Assembly (columns μ_2 and ξ_2) appears to also be ideal; that will be confirmed in Chapter 6. The matrix division and the uniqueness condition scaling is given in the columns μ_3, ξ_3, μ_4 and ξ_4. Convergence plots for $L = 20$ for the four steps are shown in Figure 5.15.

Similarly, we looked at values for fixed h and varying L, where we made the ansatz

$$t = \mu_k L^{\gamma_k}.$$

Convergence for the four steps can be seen for different fixed values of h in Table 5.2. Plots for the four steps of the solver for fixed $h = 0.02$ are shown in Figure 5.16.

Recall that we made the assumption that $\gamma_k = \mu_k$ in order to simplify the expression for the optimization problem. This assumption led to the easier, decoupled systems of equations presented in Corollary 4.8 and Corollary 4.11. We now wish justify this assumption used in the next section's results. In Figure 5.17, we see the plots of work as a function of L/h for the four steps. The points do not fall entirely on the line obtained through linear regression, but some amount of variability in computation time is to be expected for our problem. In particular, the parameter h is the *input* parameter for the mesh generator. Depending on the ensemble, though, the generator may need to add elements to the mesh so that circles that are very close to each other are resolved.

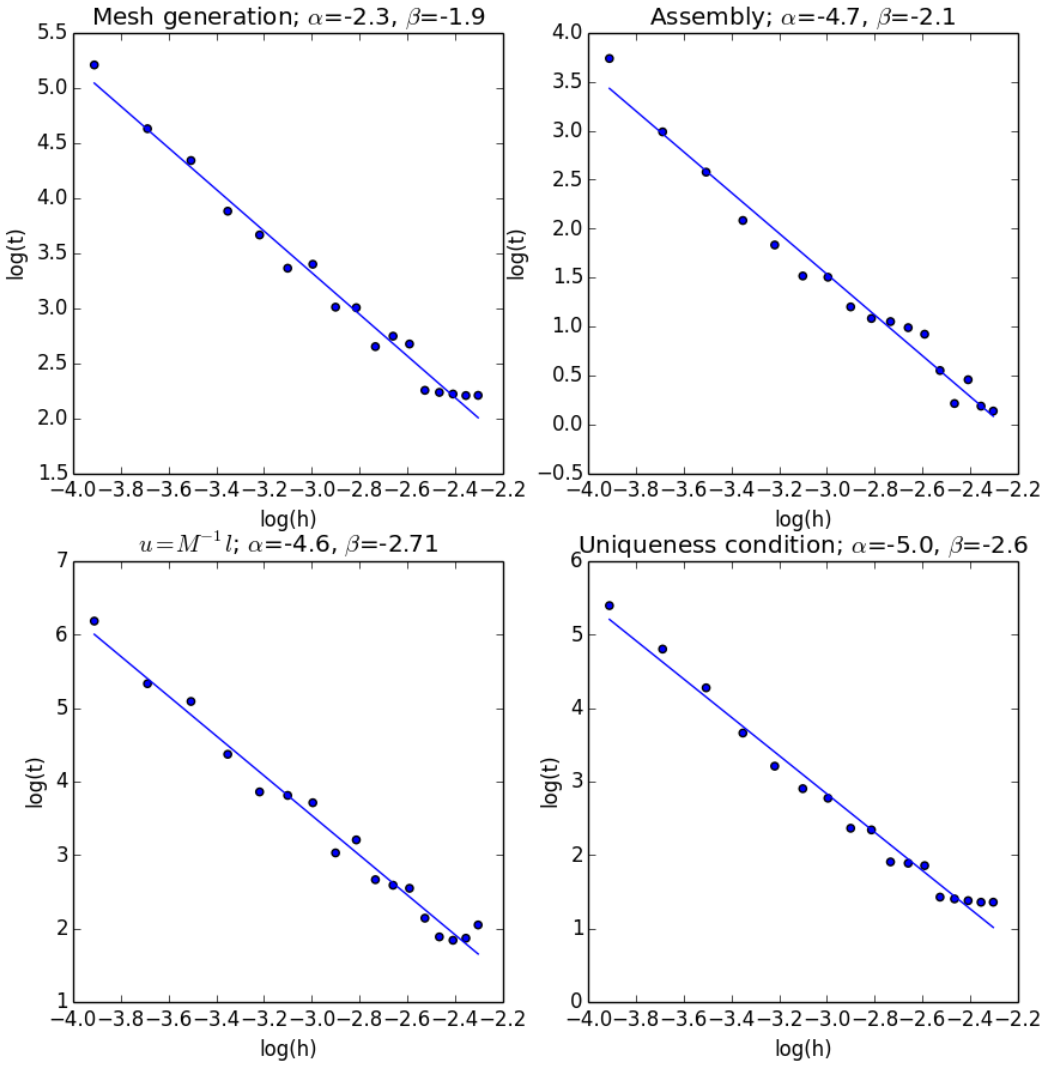


Figure 5.15: Work (computation time) as a function of mesh fineness h for a fixed domain length ($L = 20$) for different components of solver.

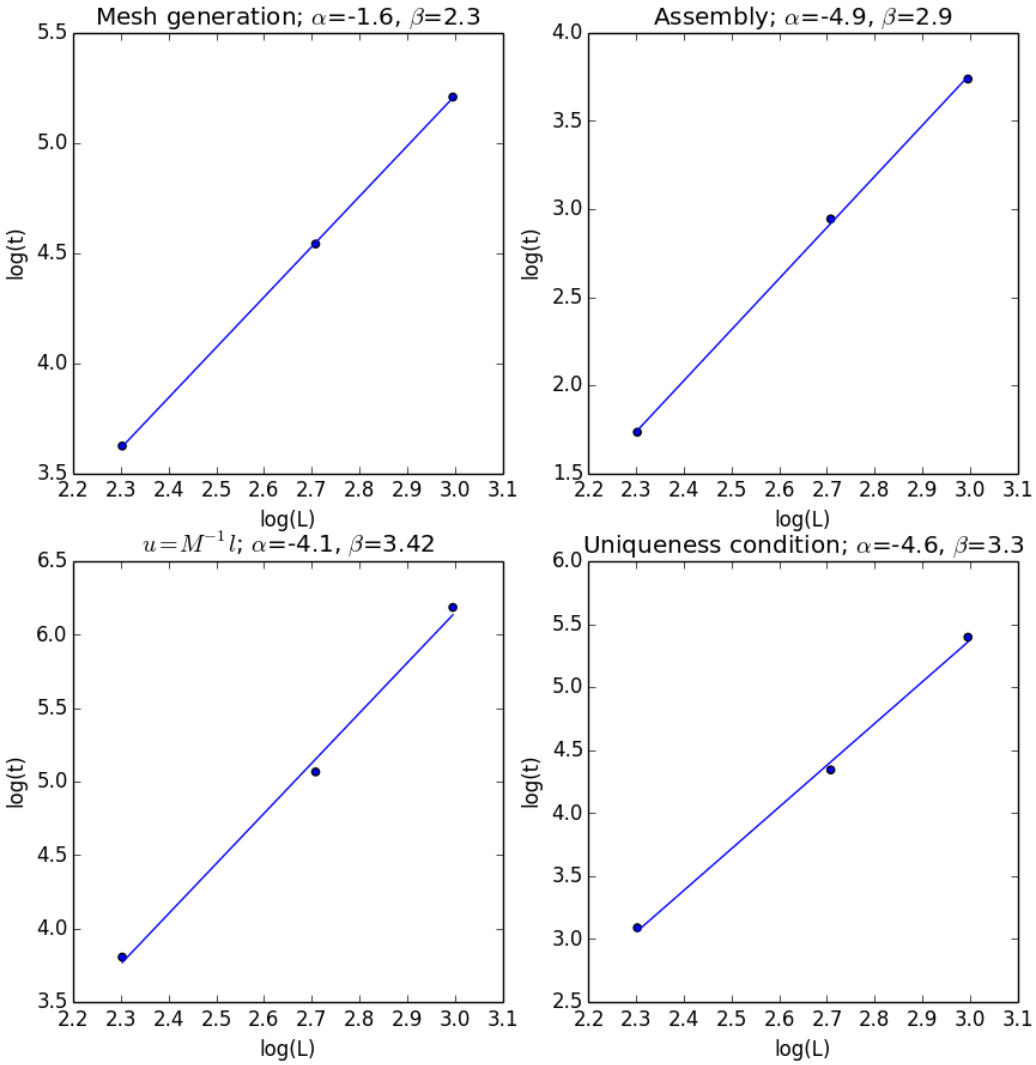


Figure 5.16: Work (computation time) as a function of L for a fixed mesh fineness ($h = 0.02$) for different components of solver.

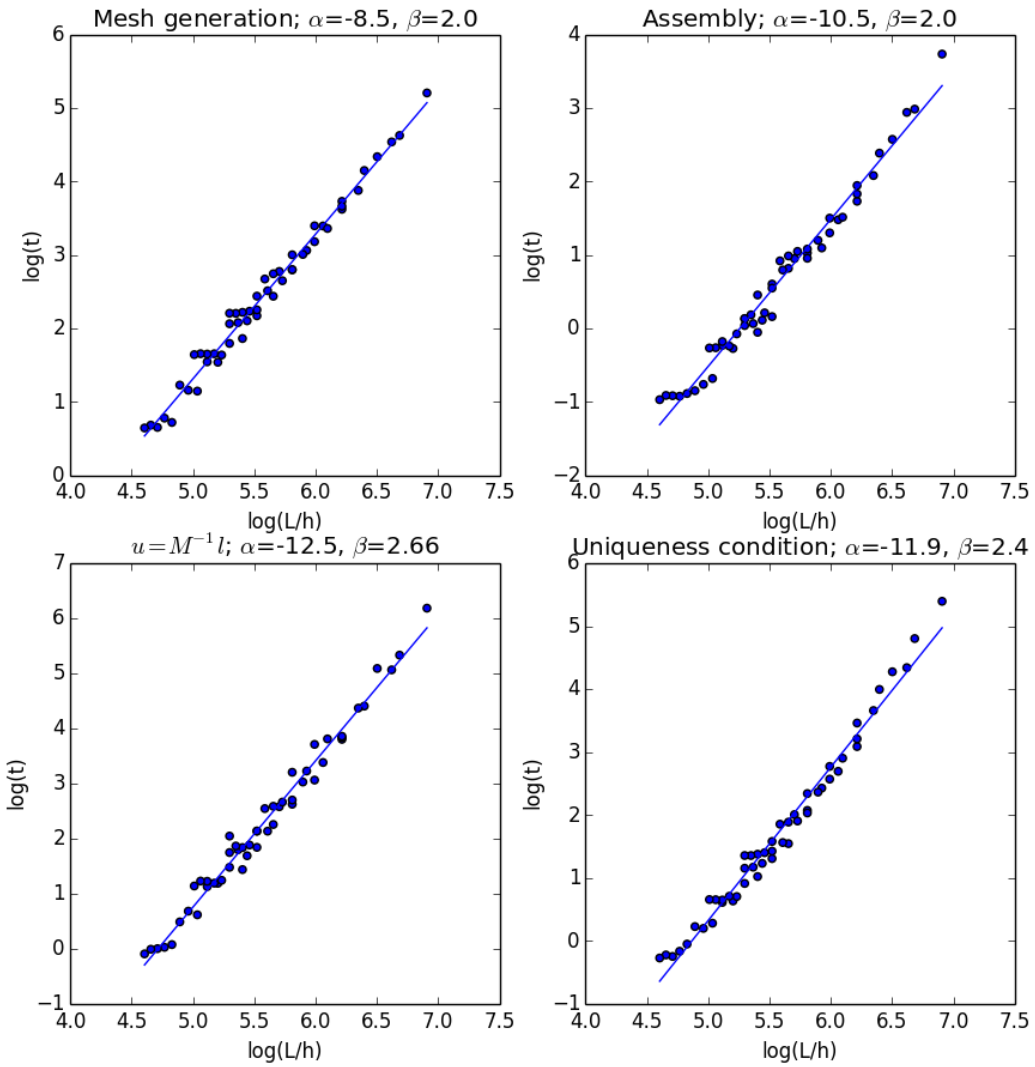


Figure 5.17: Work (computation time) as a function of L/h for different components of solver.

5.4 Optimal Method

We will now produce an optimal method for solving the cell problem for the specific example

$$A_1(y,\omega) = \begin{pmatrix} 20 & 0 \\ 0 & 10 \end{pmatrix} \mathbb{1}^{(c)}(y,\omega) + \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \mathbb{1}^{(m)}(y,\omega)$$

from Section 5.2.1. We can use the values that we obtained in Sections 5.2 and 5.3 to obtain an optimal method for this problem. First, we obtained the convergence coefficients for χ^1 from Figures 5.7, 5.13, and 5.14. We used the results from Figure 5.17 to make the assumption $\gamma_k = \xi_k$ for each step of the work. This means that we could use Corollary 4.8 to determine an optimal method. Note that we used this assumption for simplicity and that better results would be expected should we have solved the coupled system (4.15).

Using the `findroot` function in Mathematica [25], optimal combinations of L, h, and N were computed that minimize the Lagrange function for different errors; see Table 5.3. For error values $\varepsilon \geq 2 \cdot 10^{-2}$, reasonable values for our parameters were obtained, although for the lower error values, the values given for L, h, and N are not very realistic in terms of computation time and memory. Additionally, plots of optimal values as a function of error tolerance ε are shown for L, h, and N in Figures 5.18, 5.19, and 5.20, respectively. These values are combined on one plot to compare scaling; see Figure 5.21. We see that the number of trials needed is the most sensitive parameter with respect to error; the cut-off length is the least sensitive. Finally, it was also possible to plot the work function (4.13) as a function of error tolerance; see Figure 5.22.

To summarize, given a problem with a fixed distribution and corresponding matrix field A, we first obtained a-priori estimates to measure error in the cell problem as a function of domain cut-off length, mesh fineness and number of samples. We measured work for the solver and then minimized the Lagrange function (4.14) in order to obtain optimal combinations for our parameters given a certain error tolerance. In this way, we can proceed with our calculations in a systematic and efficient way. This algorithm could be extended to a wide array of stochastic homogenization problems using similar error estimates.

ε	L	h	N
$7 \cdot 10^{-3}$	41.36	0.00053	112 624
$1 \cdot 10^{-2}$	33.39	0.00092	54 748
$2 \cdot 10^{-2}$	22.05	0.00268	13 336
$4 \cdot 10^{-2}$	14.61	0.00773	3 178
$6 \cdot 10^{-2}$	11.49	0.01430	1 356

Table 5.3: Various optimal values for L, h, and N given an error tolerance ε.

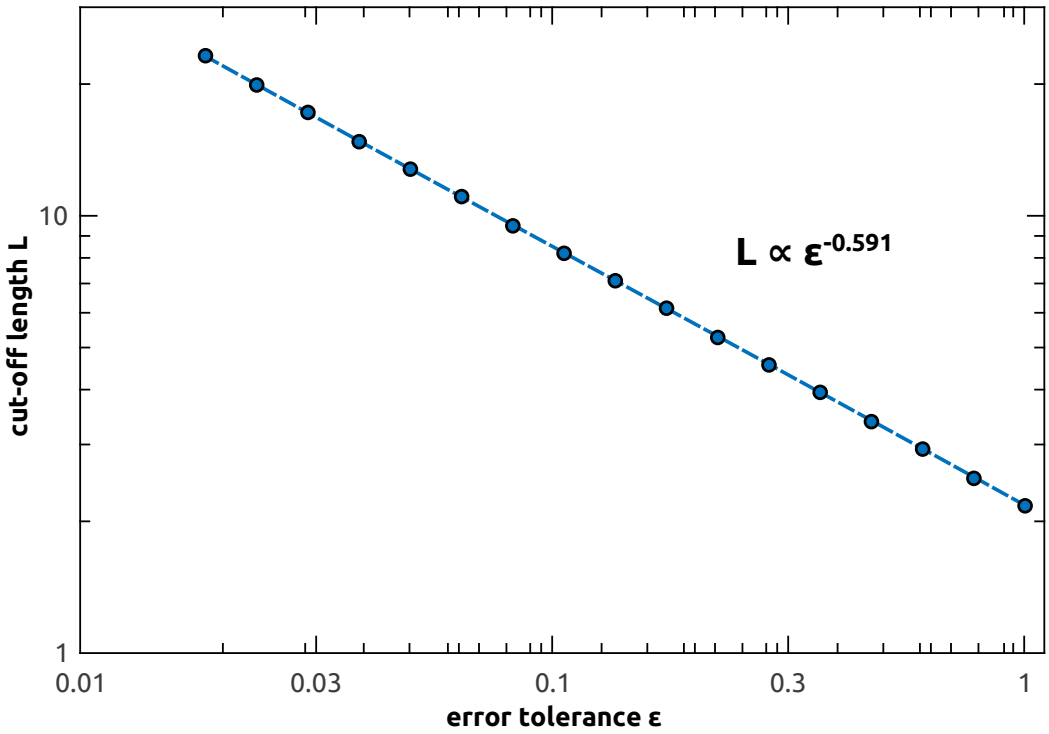


Figure 5.18: Optimal values for cut-off length L as a function of error.

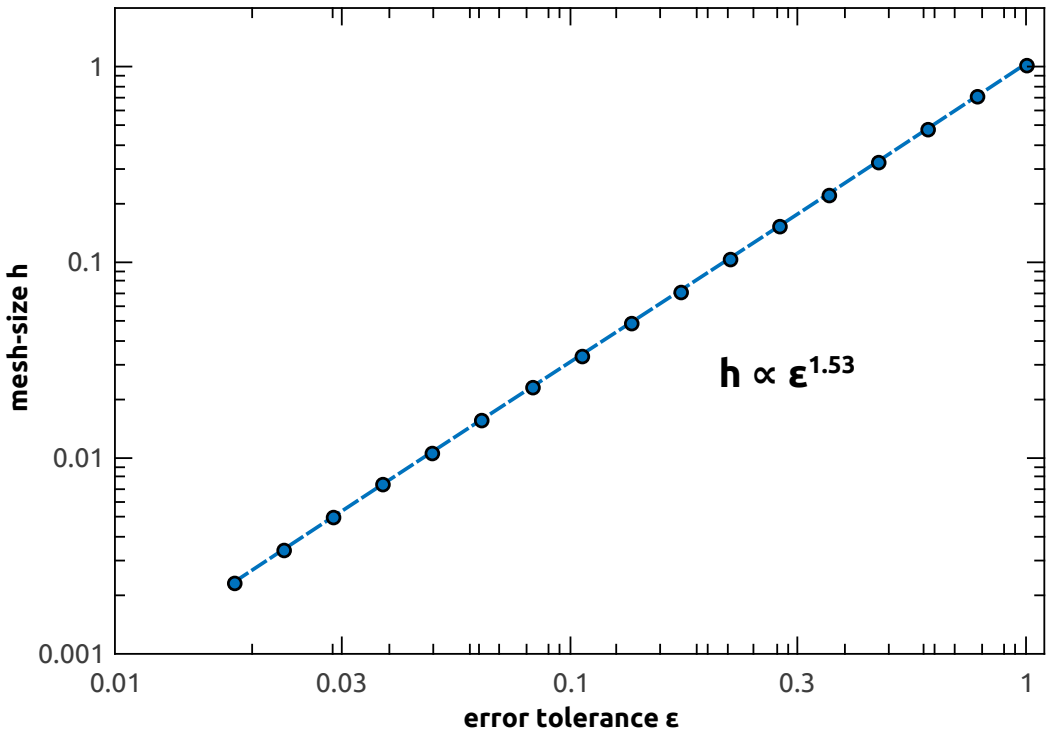


Figure 5.19: Optimal values for the mesh fineness h as a function of error.

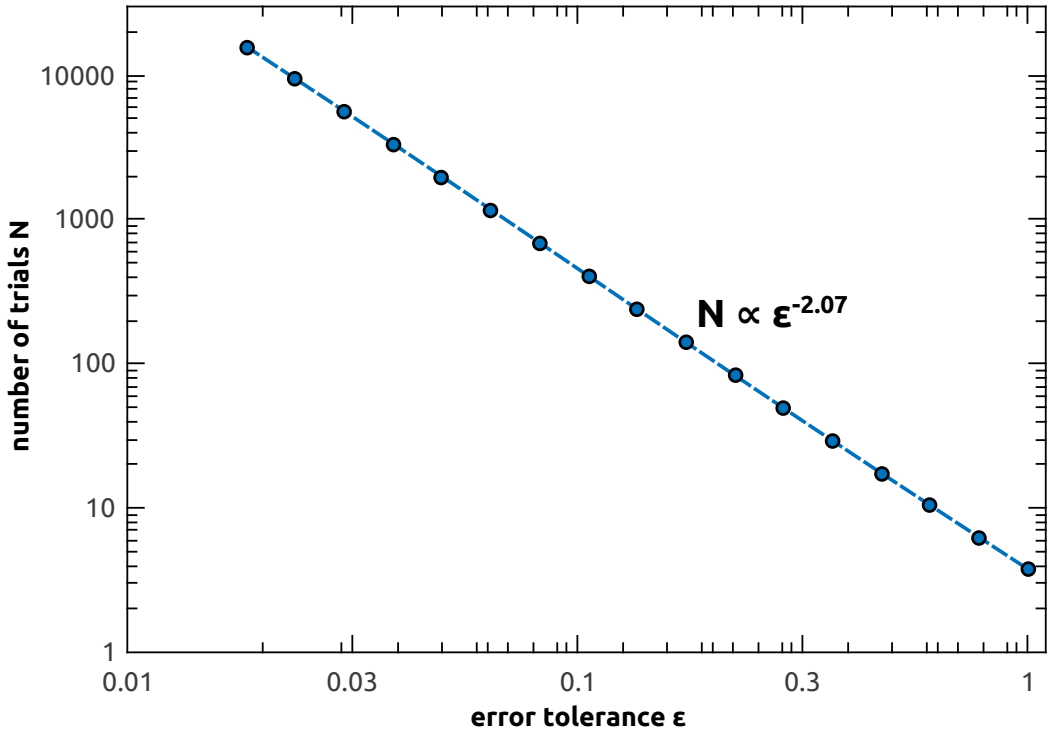


Figure 5.20: Optimal values for the number of samples N as a function of error.

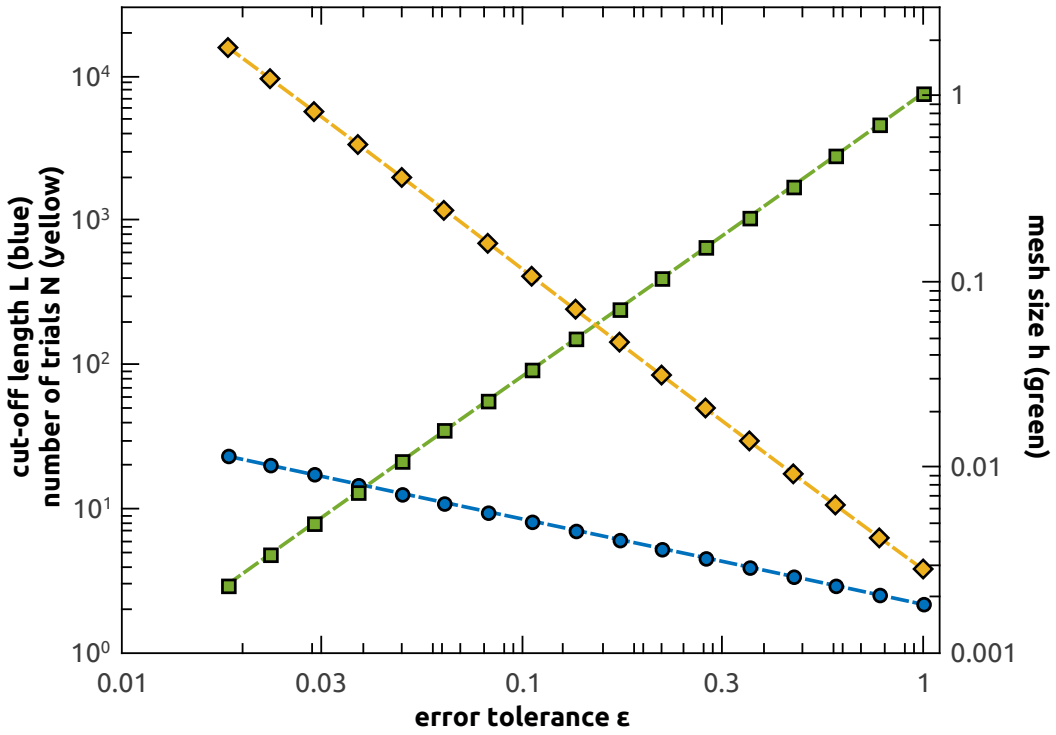


Figure 5.21: Combined plot of optimal parameters L, N, and h plotted as a function of error tolerance ε. For the left-hand y axis, optimal values for L are shown in blue and values for N are shown in yellow. For the right-hand y axis, optimal values for h are shown in green.

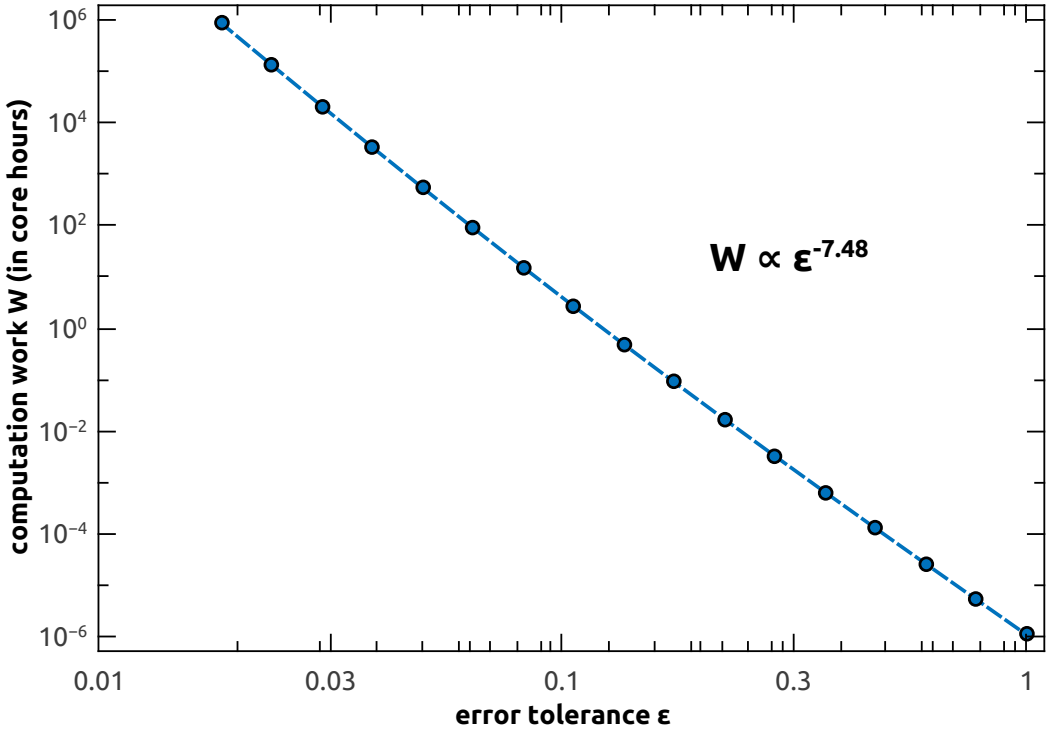


Figure 5.22: Computational work as a function of error.

5.5 Calculation of the Effective Coefficient

In this section, we would like to briefly discuss a few tests that were run on the effective coefficient. In Figure 5.23, we show a convergence plot for the effective coefficient matrix $\bar{A}$ as a function of domain length L. For this test, 50 samples were collected, where the reference solution $\bar{A}_{\text{ref}}$ was calculated on a domain $D_{25} = [-25/2, 25/2)^2$ as before, using the same mesh fineness, lengths and number of circles as described in Section 5.2.1. For each sample, the error $\|\bar{A}_{\text{ref}} - \bar{A}\|_F$ was calculated, where the Frobenius norm $\|\cdot\|_F$ is defined by

$$\|B\|_F = \left(\sum_{i,j} |B_{ij}|^2 \right)^{1/2}.$$

We see that the convergence exhibits the same hyper-linear convergence of the solution χ with $\|\bar{A}_{\text{ref}} - \bar{A}\|_F$ proportional to $L^{-0.53}$. As we mentioned in Section 3.3, convergence rates for the stochastic problem do not yet exist for $d = 2$, so it is uncertain whether our results coincide with the theory. It should be noted that for the periodic, *deterministic* case, numerical tests found that convergence rates for the effective coefficient matrix were closer to L^{-1} [32].

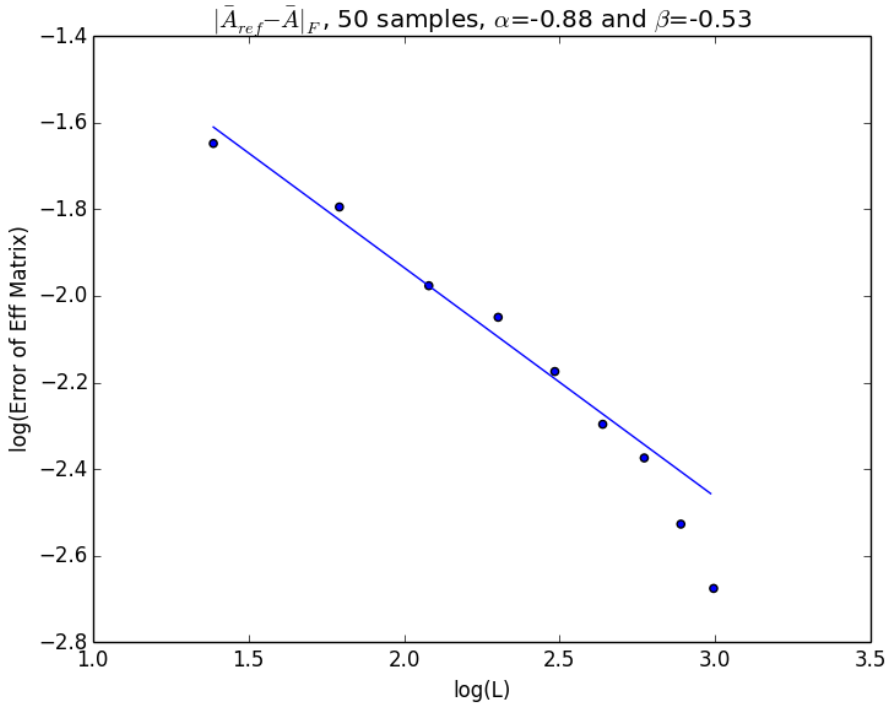


Figure 5.23: $\bar{A}$ error as a function of length.

Up until now, convergence rates for a fixed number of circles in a given sample were examined. We wish to see what occurs to the effective coefficient matrix when more circles are added to the domain. We will observe quantities for 10 percent and 30 percent coverage by circles; see Figure 5.24 for an example of realizations with these coverage percentages. We will observe $\bar{A}$ values for both the low-contrast example A_1 as well as the high-contrast A_2. To generate these values, a fixed mesh size of $h = 0.03$ was chosen and the cell problem was solved on a domain of length $L = 5$. The solution was then used to calculate the effective coefficient matrix on the same domain. A total of 5,000 samples were generated for each example, where it is noted that a couple of results needed to be discarded due to bad meshes.

Table 5.4 shows the mean and standard deviation of the 5,000 simulations for the low-contrast example A_1. Both the mean values and standard deviations are listed for each matrix element and each coverage percentage (10 and 30 percent). Both the mean values and the standard deviations are larger for 30 percent coverage. Table 5.5 shows the same test for the high-contrast example A_2. We see that the sample mean values shift as expected; in particular, the sample mean on the diagonal elements increases since the diagonal values of A in a circle are higher. Additionally, the standard deviations increase for the high-contrast example.

In Figures 5.25 and Figures 5.26, we see the distribution of values from Table 5.4 for each component of the effective coefficient matrix. The shift in the mean as well as the increase in the standard deviation is clearly visible. While coverage per-

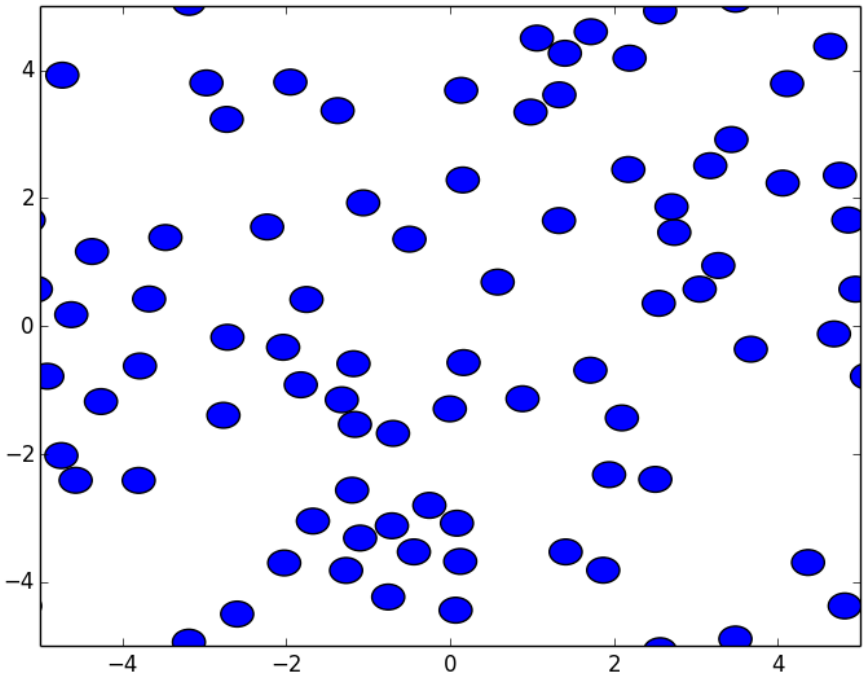
Matrix entry	μ_{10}	σ_{10}	μ_{30}	σ_{30}
$\bar{A}_{11}$	3.310	0.003	6.017	0.009
$\bar{A}_{12}$	0.000	0.002	0.000	0.007
$\bar{A}_{21}$	0.000	0.002	0.000	0.007
$\bar{A}_{22}$	1.672	0.002	3.073	0.007

Table 5.4: Four elements of $\bar{A}$ for low-contrast example A_1. Mean values σ_i and standard deviations μ_i are listed for coverage percentages $i = 10$, $i = 30$.

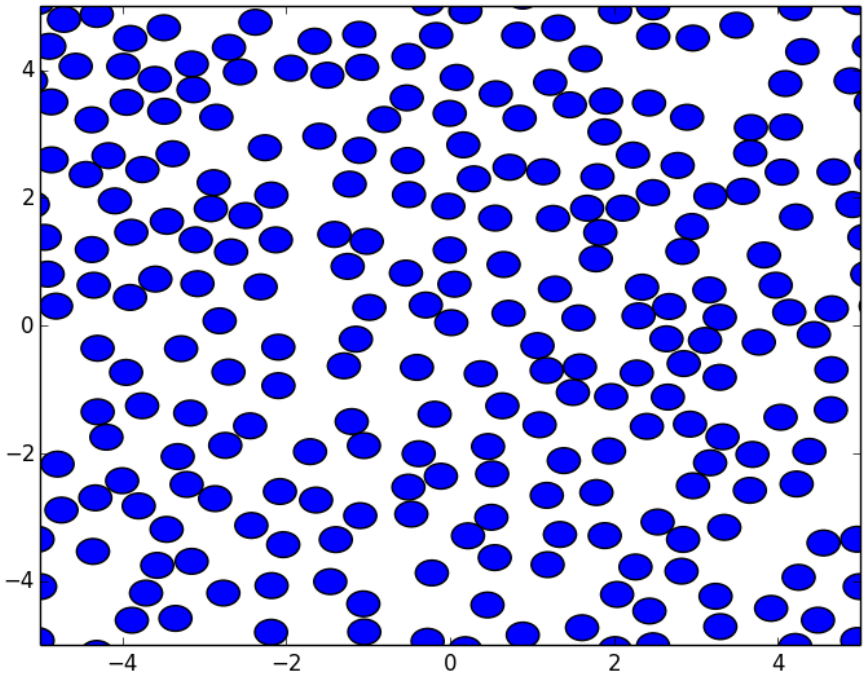
Matrix entry	μ_{10}	σ_{10}	μ_{30}	σ_{30}
$\bar{A}_{11}$	15.354	0.012	42.207	0.020
$\bar{A}_{12}$	0.000	0.014	0.000	0.023
$\bar{A}_{21}$	0.000	0.005	0.002	0.135
$\bar{A}_{22}$	7.704	0.007	21.217	0.120

Table 5.5: Four elements of $\bar{A}$ for high-contrast example A_2.

centage is only one parameter that can be adjusted in our model, this experiment demonstrates the need for our a-priori estimates when constructing an optimal approach.



(a) 10 percent coverage



(b) 30 percent coverage

Figure 5.24: Example of domain $D_{10} = [-5,5)^2$ with different coverage percentages.

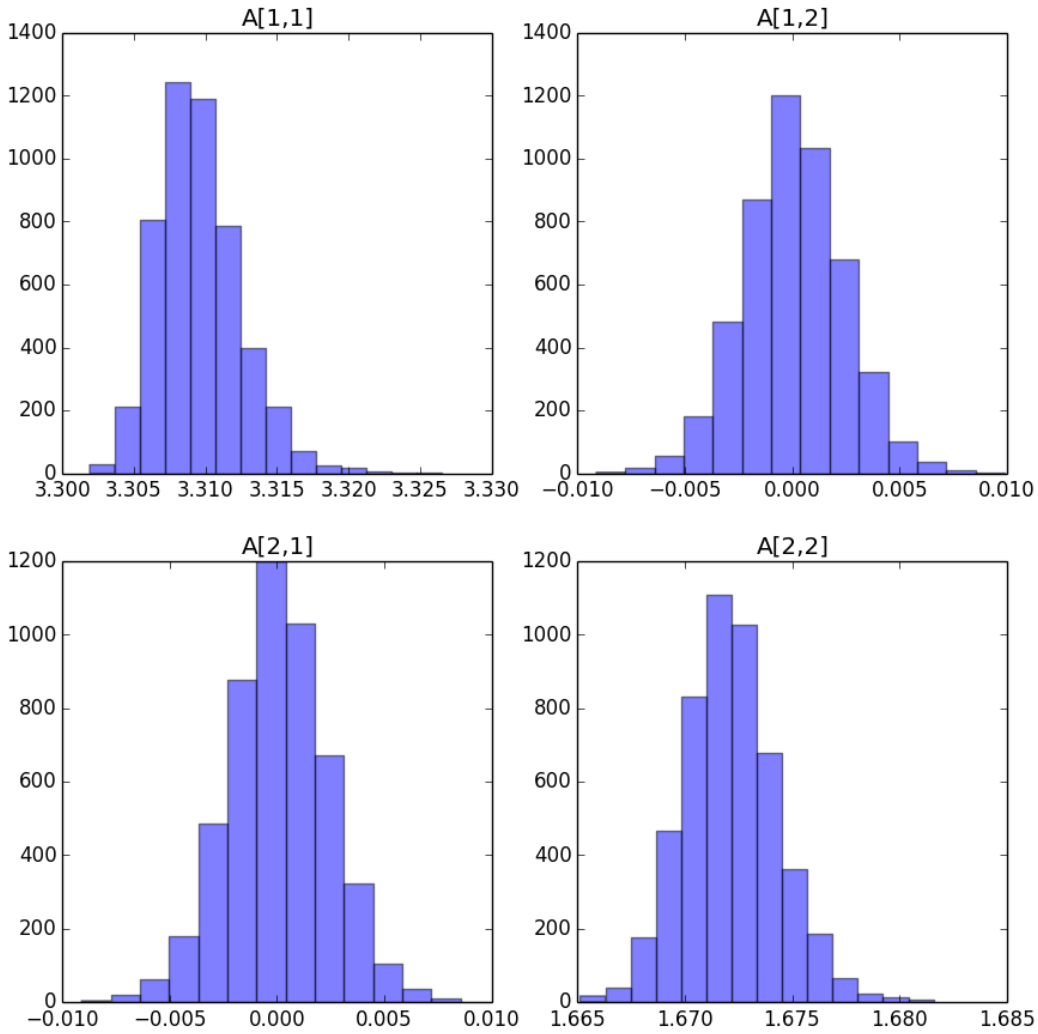


Figure 5.25: Low-contrast example A_1, 10 percent coverage. Results from $5\,000$ simulations on a domain of length $L = 5$ and mesh fineness $h = 0.03$.

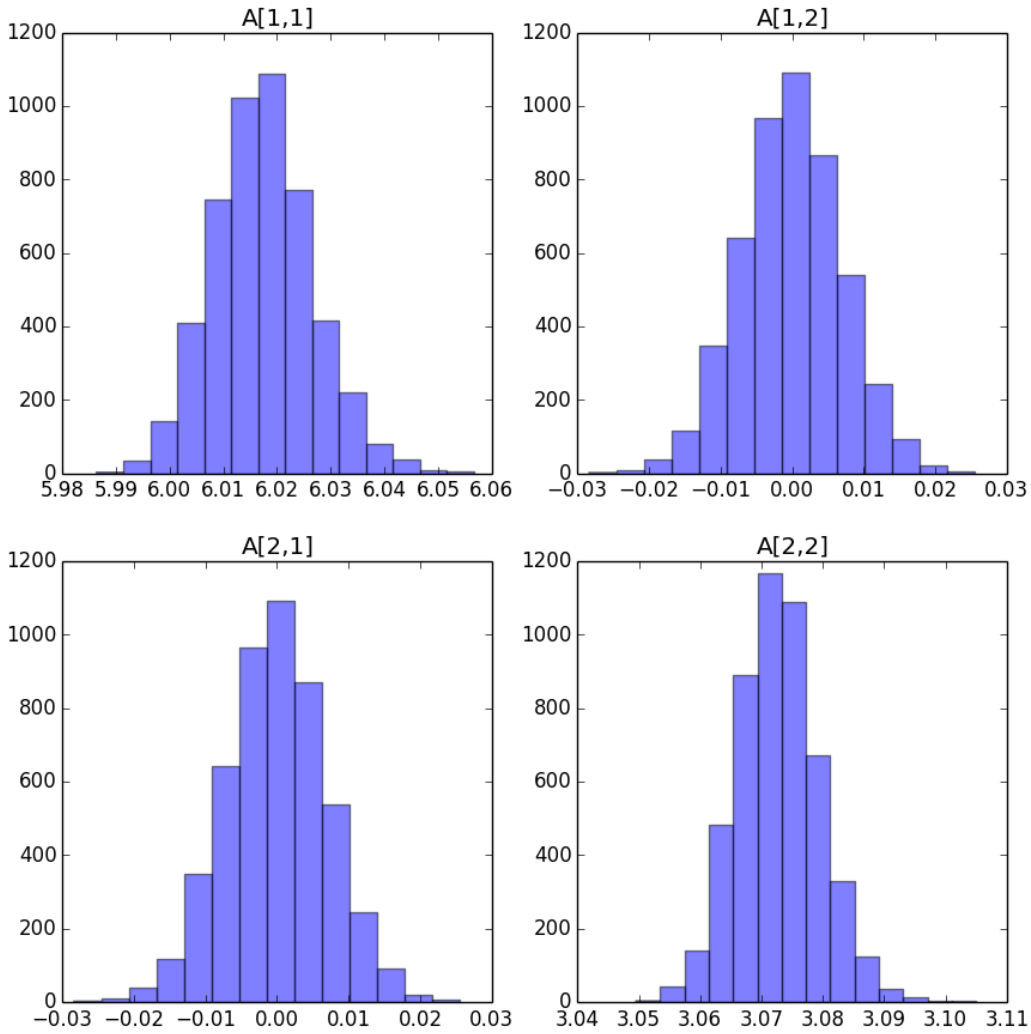


Figure 5.26: Low-contrast example A_1, 30 percent coverage. Results from $5\,000$ simulations on a domain of length $L = 5$ and mesh fineness $h = 0.03$.

CHAPTER 6

Implementation

6.1 Software

In this chapter, we will outline some of the implementation work that went into producing the results from the previous chapter. After testing FreeFem++ [18] and FEniCS [20] as possible solvers for the periodic problem, it was found that there was no existing software available for the unique makeup of this problem (namely, a periodic geometry, allowing for any number of circular inclusions with varying size, with periodic boundary conditions). Early on, efforts were therefore focused on creating a solver that would work for this specific problem. GMSH [12] was chosen to create the finite element mesh for a given geometry. Julia [7] was chosen as the programming language for the heart of the solver, as it has a reputation for being easy and fast; parallel computing, which is necessary for Monte Carlo simulations, is also especially convenient in this language. Since Julia is relatively new, however, the disadvantage was that a PDE package did not already exist. Part of this thesis was therefore focused on the practical questions of how one should implement an efficient FEM solver in Julia.

6.2 Description of Code

Here, we will briefly discuss what code was necessary to carry out the numerical tests in Chapter 5. The code can be divided into several functional groups:

1. Geometry,
2. Mesh generation,
3. Solver,

4. Tests and auxiliary functions.

Additionally, several new types were created to help make the code more readable. A `Circle` data type stored the midpoint of each circle, the radius, the shape (since circles overlapping an edge needed to be defined differently), and location (inner, south, east, north, and west). A `Mesh` data type stored all information about the mesh: nodes, edges, elements, edges, and their corresponding physical properties (such as whether they belonged to a boundary or circle). For the boundary, we distinguished between `master` and `slave` nodes. Data types mirroring those in GMSH were used to make the creation of the input files for GMSH convenient.

6.2.1 Geometry

For any given numerical test, it was always necessary to define the dimensions of the domain as well as the circles and their placement. Often, it was also necessary to define the geometry of a subdomain (that included only some of the circles from the original domain). The essential functions for these tasks included:

- Generation of circles: Given a desired number of circles and domain length L, generate the list of midpoints and radii that were created according to an RSA process. For each circle, random points $x, y \sim U[-L/2, L/2)$ (uniformly distributed on the interval $[-L/2, L/2)$) were created. The radius of circles was, for most tests, fixed. A test was run to see if the circle overlapped any others (including overlapping on the torus); if it did, a new set of points was created and the test was repeated until an appropriate placement could be found. A limit was set so that the function would throw an error if no placement was found after a certain number of tries. For each circle, the shape and location were also defined to make mesh generation easier.
- Periodization: Given a list of circles and a length L, add circles on the torus for each circle overlapping a side.
- Subdomain creation: Given a list of circles defined on D_L and a subdomain length $0 < L_0 < L$, determine the subset of circles that have a midpoint in the domain D_{L_0}. Shapes that were not intersecting the original domain were redefined to account for their new shape and location.

The creation of the subdomain has one unavoidable problem: when the subdomain intersected with a circle that was originally entirely contained in the full domain, and the subdomain was required to have a periodic structure, then it was possible that this circle intersected with another one on the torus. Thus, some circles had to be discarded that would otherwise be included on the subdomain.

6.2.2 Mesh Generation

Building the mesh used for the finite element solver involved:

- Input file creation: An input file for GMSH was written that included all pertinent details about the geometry (placement of circles, handling of circles that overlapped the boundaries, and the definition of periodic boundary conditions).
- Reading of output file and conversion into the `Mesh` data structure: This included a call to GMSH and subsequent reading of the output file, storing all information in the `Mesh` data type.

For some tests, errors for different solutions on subdomains of the same geometry needed to be calculated. For instance, the tests in Section 5.2.1 observed the error on the domain D_{L_0} for different subdomains $D_{L_1}, D_{L_2}, \ldots, D_{L_n}$ of D_L such that $D_{L_0} \subset D_{L_1} \subset \ldots \subset D_{L_n} \subset D_L$. To allow for the efficient calculation of the error, a special mesh was created for each subdomain that contained exactly the same points on D_{L_0}. An example of this idea is shown in Figure 6.1.

6.2.3 Solver

The functionality of the solver included:

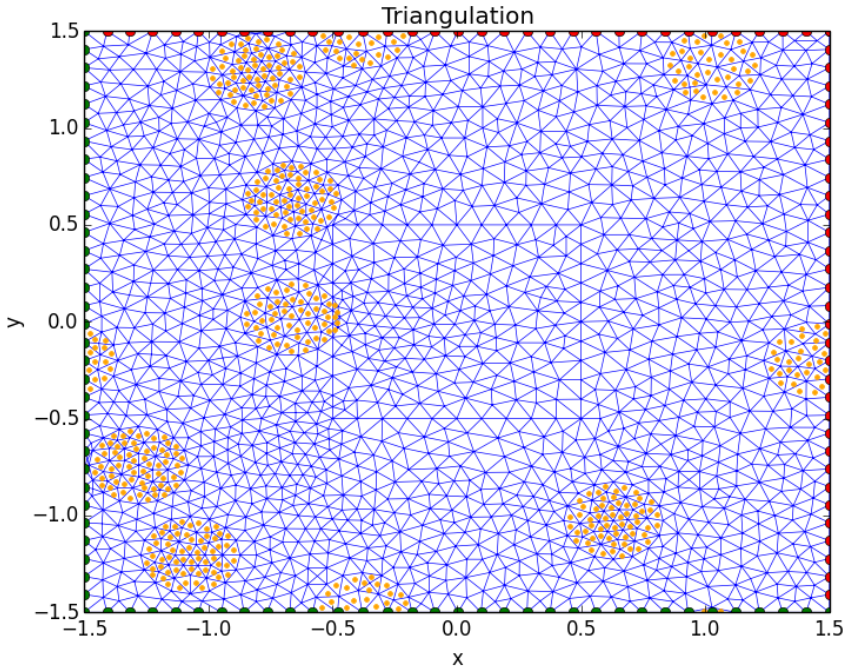
- Stiffness matrix and load vector assembly: for a problem of the form

$$\begin{aligned} -\nabla \cdot (A\nabla u) &= f \qquad \forall x \in D, \\ u &= g \qquad \forall x \in \partial D, \end{aligned}$$

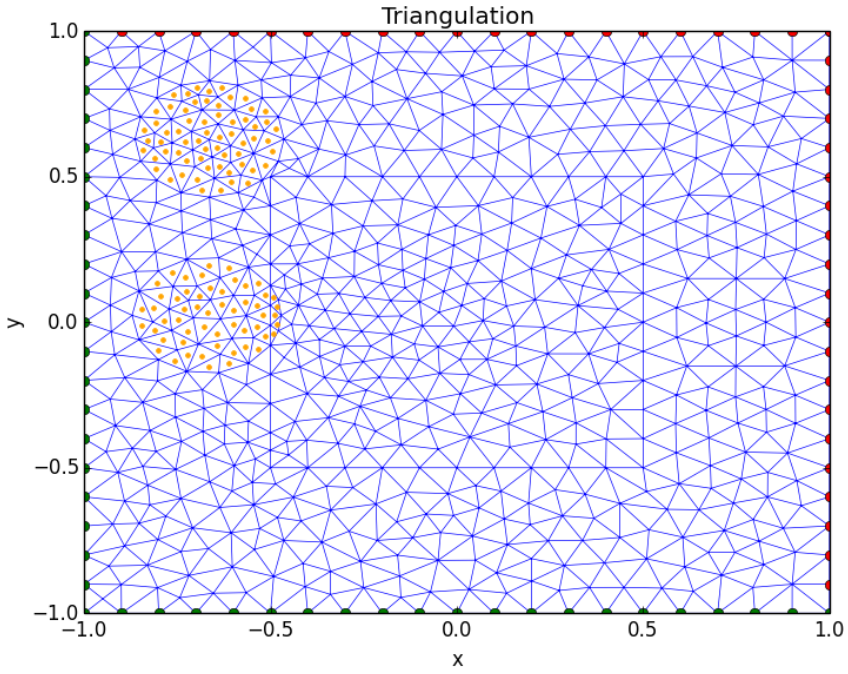
  assemble the stiffness matrix M and load vector l such that the numerical approximation $u_h \approx u$ is given by the relation $Mu_h = l$. This function had the option to assemble M and l given periodic boundary conditions (instead of Dirichlet, as displayed here). The stiffness matrix was assembled using an efficient algorithm presented in [8, pp. 15-16].
- Solving of system: Compute $u_h = M^{-1}l$ (here, using the backslash operator as one would use in MATLAB); also, where relevant, the uniqueness condition was applied.
- Computation of effective coefficient matrix: Once the numerical solution $\chi_{L,h}$ to the cell problem has been calculated, the effective coefficient matrix $\bar{A}_{L,L_0}$ as presented in (4.2) can be approximated by the matrix $\bar{A}_h$, where

$$\begin{aligned} \bar{A}_{L,L_0} &= \frac{1}{L_0^d} \int_{D_{L_0}} A(y)(I + (\nabla_y \chi_L)^\top)\,\mathrm{d}y \\ &\approx \sum_{K \in \mathcal{T}_h \cap D_{L_0}} \int_K \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} \partial_{y_1}\chi^1_{L,h} + 1 & \partial_{y_1}\chi^2_{L,h} \\ \partial_{y_2}\chi^1_{L,h} & \partial_{y_2}\chi^2_{L,h} + 1 \end{pmatrix} \mathrm{d}y := \bar{A}_h. \end{aligned}$$

  Integration over each element in K is calculated using quadrature.



(a) Mesh of fineness $h = 0.1$ on original domain $D_L = [-1.5, 1.5)^2$ with fixed $D_{L_0} = [-0.5, 0.5)^2$ and periodic circles added on the torus.



(b) Mesh of fineness $h = 0.1$ on subdomain $D_{L_1} = [-1, 1)^2$ with fixed $D_{L_0} = [-0.5, 0.5)^2$.

Figure 6.1: Two different meshes for the same geometry sharing a common inner structure on the square D_{L_0}. The midpoints of circles are marked with orange dots. Boundary nodes are marked by their category: slaves are in green and masters are in red.

6.2.4 Tests and Auxiliary Functions

There were a number of other functions used in the numerical tests. All tests that were presented in the numerical results first needed to be programmed. These tests included methods to calculate convergence, functions to perform parallel calculations for Monte Carlo simulations, and tests for computation time. Additionally, a number of auxiliary functions needed to be implemented, including the calculation of the L^2 and H^1 norms as well as the H^1 seminorm; functions that performed numerical quadrature; and functions for interpolation.

6.3 Performance

Figure 6.2 compares the performance of the assembly function (the assembly of the stiffness matrix M and load vector l) with the performance of the same in MATLAB for different meshes of fineness h. The performance of our solver compares quite favorably to that of MATLAB. Future work will involve further optimization of the Julia code. The first naive implementation of the solver was highly inefficient; several things were done that sped up calculations considerably.

- Julia's built-in `det` function was found to be a very inefficient way of calculating the area of a mesh element; this was replaced with a direct calculation based on the lengths of the sides of each triangle.
- For-loops were avoided where possible; operations were vectorized instead, considerably speeding up calculations. This meant that nested types were to be avoided: the `Mesh` data type was originally comprised of arrays of other types (`Nodes`, `Edges`, and `Elements`), which could not be vectorized. The `Mesh` type was therefore modified to include simple arrays that could be vectorized.

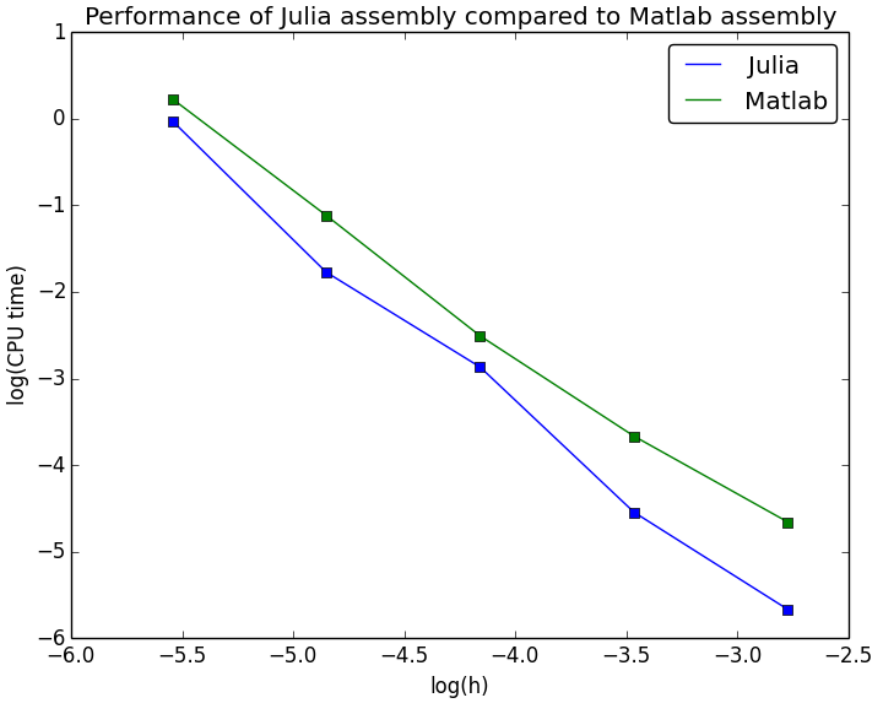


Figure 6.2: Performance of FEM assembly implemented in Julia compared to assembly in MATLAB.

CHAPTER 7

Conclusion

In this thesis, we focused on the stochastic homogenization of elliptic partial differential equations. We discussed the main theoretical results and applications as well as numerical methods used to compute approximations to two-dimensional solutions.

The main contribution of this work was Chapter 4, where error bounds for the solution to the cell problem were estimated. In this chapter, an optimal computational scheme was also presented. The theoretical error bounds agreed well with numerical results presented in Chapter 5. Additionally, convergence of the solution to the cell problem as well as the effective coefficient matrix was observed numerically. We ran tests to approximate the missing constants in the optimization problem and found an optimal method for stochastic homogenization.

Finally, the numerical results that were presented here were carried out using original Julia code that was summarized in Chapter 6 and that will be provided to the Julia programming community.

7.1 Future Work

We close with a few words about next steps in our research. In the development of the optimal approach, we made the assumption that the work and error of χ scaled at the same rate for both components; the error bounds and optimal approach in Chapter 4 were created using one component of χ. A natural next step would be to determine the optimal approach using both error and work values for χ^1 and χ^2.

While we focused on tests for diagonal A, these tests could easily be run for the non-diagonal case as described in Section 3.1.2. We focused entirely on the case of hard spheres (where overlapping is not allowed), but an area of future research

would be the overlapping case described by a Poisson point process. The theory for this case is much more developed; theoretical results already exist by Gloria and Otto [16] but numerical results are still missing. Part of what makes the numerical setup here so difficult, however, is the construction of a proper mesh. The geometry of a domain with overlapping circles introduces new shapes that must be handled individually, so modeling this is quite difficult. Some of those difficulties could potentially be overcome if we used quasi-Monte Carlo points for the circles. For certain setups with "small" circles, these points could be used to approximate ensembles derived from a Poisson Point Process.

The existing code could be pretty easily tweaked to allow for related shapes (like ellipses); also, tests could be easily run for circles with varying radii. Some work would be involved in extending the implementation to the three-dimensional case. The creation of the mesh and the setup of the solver (to ensure periodic conditions on all sides) would require adjustments. The use of an adaptive mesh might be beneficial in the two-dimensional case.

Gloria [13] has shown that the addition of a zero-order term in the deterministic problem can substantially reduce the error; this also reduces the hyper-linear convergence that we were seeing in the convergence plots in L. The question would be to what extent this strategy works in the stochastic case.

Finally, while out of scope for this thesis, it would be beneficial to compare the results of our simulations to measured outcomes in materials science.

List of Figures

Bibliography

[1] A. Abdulle. *On A Priori Error Analysis of Fully Discrete Heterogeneous Multiscale FEM.* Multiscale Model. Simul. Vol. 4, No. 2, pp. 447-459 (2005).

[2] M. Ainsworth and R. Rankin. *Guaranteed Computable Bounds on Quantities of Interest in Finite Element Computations.* Int. J. Numer. Meth. Engng 89: 1605-1634 (2012).

[3] N. Bakhvalov. *Average Characteristics of Bodies with Periodic Structure.* Dokl. Akad. Nauk SSSR, 218, No. 5, 1046-1048 (1974).

[4] N. Bakhvalov and G. Panasenko. *Homogenisation: Averaging Processes in Periodic Media.* Kluwer Academic Publishers (1989).

[5] V. Berdichevskii. *Spatial Averaging of Periodic Structures.* Dokl. Akad. Nauk SSSR, 222, No. 3, 565-567 (1975).

[6] A. Bourgeat and A. Piatnitski. *Approximations of Effective Coefficients in Stochastic Homogenization.* Annales de l'IHP Probabilités et Statistiques, Vol. 40, pp. 153-165 (2004).

[7] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. *Julia: A fresh approach to numerical computing.* http://arxiv.org/abs/1411.1607 (2014).

[8] F. Cuvelier, C. Japhet, and G. Scarella. *An Efficient Way to Perform the Assembly of Finite Element Matrices in Matlab and Octave.* ISSN 0249-6399. Research Report No 8305 (2013).

[9] W. E, P. Ming, and P. Zhang. *Analysis of the Heterogeneous Multiscale Method for Elliptic Homogenization Problems.* Journal of the American Mathematical Society. Vol 18, No 1, pp. 121-156 (2004).

[10] W. E, B. Engquist, X. Li, W. Ren, and E. Vanden-Eijnden. *The Heterogeneous Multiscale Method: A Review.* Commun. Comput. Phys. 2: 367-450 (2007).

[11] Y. Efendiev and T. Y. Hou. *Multiscale Finite Element Methods: Theory and Applications.* Springer (2009).

[12] C. Geuzaine and J. F. Remacle. *Gmsh: a Three-Dimensional Finite Element Mesh Generator with Built-in Pre- and Post-Processing Facilities.* Interna-

tional Journal for Numerical Methods in Engineering 79(11), pp. 1309-1331 (2009).

[13] A. Gloria. *Reduction of the Resonance Error - Part 1: Approximation of Homogenized Coefficients.* Mathematical Models and Methods in Applied Sciences, 21 (8), pp. 1601-1630 (2011).

[14] A. Gloria. *Numerical Homogenization: Survey, New Results, and Perspectives.* ESAIM: Proceedings, Vol. 37, pp. 50-116 (2012).

[15] A. Gloria, S. Neukamm, and F. Otto. *Quantification of Ergodicity in Stochastic Homogenization: Optimal Bounds via Spectral Gap on Glaubner Dynamics.* Invent. Math. (2014). DOI 10.1007/s00222-014-0518-z.

[16] A. Gloria and F. Otto. *Quantitative Estimates on the Periodic Approximation of the Corrector in Stochastic Homogenization.* ESAIM: Proceedings, Vol. 48, pp. 80-97 (2015).

[17] A. Gloria and F. Otto. *Quantitative Results on the Corrector Equation in Stochastic Homogenization.* arXiv:1409.0801v1 [math.AP] (2014).

[18] F. Hecht. *New Development in FreeFem++.* J. Numer. Math. 20, No. 3-4, pp. 251-265 (2012).

[19] S. M. Kozlov. *Averaging of Random Operators.* Math. USSR Sbornik Vol. 37, No. 2 (1980).

[20] A. Logg, K.-A. Mardal, G. N. Wells et al. *Automated Solution of Differential Equations by the Finite Element Method.* Springer. [doi:10.1007/978-3-642-23099-8] (2012).

[21] T. Kanit, S. Forest, I. Galliet, V. Mounoury, D. Jeulin. *Determination of the Size of the Representative Volume Element for Random Composites: Statistical and Numerical Approach.* International Journal of Solids and Structures 40 pp. 3647-3679 (2003).

[22] F. Kikuchi and X. Liu. *Estimation of Interpolation Error Constants for the P_0 and P_1 Triangular Finite Elements.* Comput. Methods Appl. Mech. Engrg. 196 (2007) 3750-3768.

[23] G. J. Lord, C .E. Powell, and T. Shardlow. *An Introduction to Computational Stochastic PDEs.* Cambridge University Press (2014).

[24] S. Mao and Z. Shi. *Explicit Error Estimates for Mixed and Nonconforming Finite Elements.* Journal of Computational Mathematics, Vol. 27, No. 4, pp. 425-440 (2009).

[25] Wolfram Research, Inc., Mathematica, Version 10.3, Champaign, IL (2015).

[26] M. Ostoja-Starzewski. *Material Spatial Randomness: From Statistical to Representative Volume Element.* Probabilistic Engineering Mechanics 21 pp. 112-132 (2006).

[27] G. C. Papanicolaou and S.R.S. Varadhan. *Boundary Value Problems with Rapidly Oscillating Random Coefficients*. Colloq. Math. Soc. Janos Bolyai, pp. 835-873 (1981).

[28] G. Pavliotis and A. Stuart. *Multiscale Methods: Averaging and Homogenization*. Springer (2008).

[29] A. Quarteroni. *Numerical Models for Differential Problems*. Springer (2009).

[30] E. Sanchez-Palencia. *Non-Homogeneous Media and Vibration Theory*. Springer (1980).

[31] S. Torquato. *Random Heterogenous Materials: Microstructure and Macroscopic Proporties*. Springer (2002).

[32] X. Yue and W. E. *The Local Microscale Problem in the Multiscale Modeling of Strongly Heterogeneous Media: Effects of Boundary Conditions and Cell Size*. Journal of Computational Physics 222 (2): pp. 556-572 (2007). DOI 10.1016/j.jcp.2006.07.034.

[33] V. Yurinskii. *Averaging an Elliptic Boundary-Value Problem with Random Coefficients*. Siberian Mathematical Journal, Vol. 21, No. 3, pp. 209-223 (1980).