SIAM/ASA J. Uncertainty Quantification
Vol. 4, pp. 1246–1262



# Optimal Approximation of the First-Order Corrector in Multiscale Stochastic Elliptic PDE*

Caroline Geiersbach†, Clemens Heitzinger‡, and Gerhard Tulzer†

**Abstract.** This work addresses the development of an optimal computational scheme for the approximation of the first-order corrector arising in the stochastic homogenization of linear elliptic PDEs in divergence form. Equations of this type describe, for example, diffusion phenomena in materials with a heterogeneous microstructure, but require enormous computational efforts in order to obtain reliable results. We derive an optimization problem for the needed computational work with a given error tolerance, then extract the governing parameters from numerical experiments, and finally solve the obtained optimization problem. The numerical approach investigated here is a stochastic sampling scheme for the probability space connected with a finite-element method for the discretization of the physical space.

**Key words.** numerical stochastic homogenization, stochastic elliptic PDE, optimization, multiscale problems

**AMS subject classifications.** 35B27, 60H25, 60H35

**DOI.** 10.1137/16M106011X

## 1. Introduction.

Stochastic multiscale differential equations describe the behavior of various important physical systems, including metamaterials, glass, fiber-reinforced composites, or granular media [25], where the processes of interest comprise diffusion, electrostatic problems, and heat transfer [17] or flow [1, 2]. However, an analytical solution to such problems does not exist in general cases, and even the calculation of reliable approximations of the solution requires an enormous computational effort, which is even increased if a weak computational scheme is chosen. It is therefore crucial to have a high-performance computational setup in order to reach manageable workloads.

In particular, when calculating the solution, the overall error comprises several sources. In order to obtain an optimal scheme, one needs to find a balance between the involved error components in order to maximize accuracy by minimizing the computational effort. In this study, we derive an optimization problem for this task, we apply the framework to our implementation of the stochastic homogenization problem, and we compute the desired parameters to obtain a minimal computational effort for a given error tolerance. To this end,

*Received by the editors February 5, 2016; accepted for publication (in revised form) August 8, 2016; published electronically October 19, 2016.

http://www.siam.org/journals/juq/4/M106011.html

**Funding:** This work was supported by FWF (Austrian Science Fund) START project Y660 *PDE Models for Nanotechnology*.

†Institute for Analysis and Scientific Computing, TU Vienna, Wiedner Hauptstrasse 8–10, A-1040 Vienna, Austria (caroline.geiersbach@tuwien.ac.at, gerhard.tulzer@tuwien.ac.at).

‡Institute for Analysis and Scientific Computing, TU Vienna, Wiedner Hauptstrasse 8–10, A-1040 Vienna, Austria (clemens.heitzinger@tuwien.ac.at), and School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ 85287-1804 (clemens.heitzinger@asu.edu).

all the input parameters governing the error function as well as the workload are derived from numerical experiments.

The framework of stochastic homogenization for linear elliptic PDE in divergence form goes back to the works by Kozlov [22], Yurinskii [28], and Papanicolaou and Varadhan [24], who independently showed the admissibility of homogenization under certain conditions and also provided an expression for the homogenized coefficient function. However, since the calculation of the latter requires the integration over an abstract probability space, this formula cannot be used for the numerical computation of the desired quantity. Overcoming this challenge has been addressed from different angles, but there are still various open questions, including the optimization of the computational scheme.

Bourgeat and Piatnitskii [5] used a combination of periodization and cut-off procedures to approximate the homogenized coefficient and furthermore provided error estimates in terms of the cut-off under additional mixing conditions. Costaouec considered the stochastic case as a small perturbation of a periodic case and derived error estimates depending on the perturbation scale [9]. Gloria and Otto derived optimal error estimates for the corrector and for the homogenized coefficients in the discrete setting [14, 15] and also presented a quantitative error estimate for the continuous case under further assumptions on the spectral gap of the diffusion coefficient [12, 13].

A rather common approach in the framework of stochastic homogenization is the use of representative volume elements (RVE) [20, 21, 23]. However, the main open question here is the appropriate choice of the RVE, which is considered here in connection with other crucial parameters, namely, the mesh size and the number of samples in the Monte Carlo scheme.

Homogenization results have also be derived for nonlinear problems. The existence of a homogenized equation and convergence rates for fully nonlinear elliptic problems have been theoretically investigated by Caffarelli, Souganidis, and Wang [7] and Caffarelli and Souganidis [6]. However, in this work, we restrict ourselves to linear problems.

This paper is organized as follows. The concept of stochastic homogenization for linear elliptic PDEs in divergence form together with theoretical results is presented for our purposes in section 2. We derive the optimization problem for the computational scheme in section 3. Numerical experiments including their results are described in section 4, which are then used to solve the optimization problem. The findings are discussed in section 5.

## 2. The stochastic multiscale problem.

The problem formulation for the stochastic homogenization method is mainly based on [22, 24, 28]. In the following, we will briefly recall the theoretical results that are necessary for our numerical investigations.

### 2.1. Formulation of the problem.

The stochastic multiscale problem giving rise to the application of a homogenization method is

$$-\nabla \cdot (A_\varepsilon(x,\omega)\nabla u_\varepsilon(x,\omega)) = f(x) \qquad \text{in } D, \tag{1a}$$

$$u_\varepsilon(x,\omega) = 0 \qquad \text{on } \partial D, \tag{1b}$$

where $A_\varepsilon(x,\omega) = A(\frac{x}{\varepsilon},\omega)$, $u_\varepsilon(x,\omega) = u(x,\frac{x}{\varepsilon},\omega)$, $f(x)$ is (for simplicity) a deterministic function, and $D \subset \mathbb{R}^d$ is a bounded domain. The simplification in the choice of $A$ means that fluctuations only take place on a microscale. Here, $(\Omega, \mathcal{F}, P)$ is a probability space where

$\omega \in \Omega$ represents a single realization of a medium, $\mathcal{F}$ is an appropriate $\sigma$-algebra, and $P$ is a probability measure defined on $(\Omega, \mathcal{F})$. Since the coefficient $A_\varepsilon$ is a random field, so is the solution $u_\varepsilon$. Using homogenization, one seeks to find a limiting problem that is independent of the fast variable $y := \frac{x}{\varepsilon}$ and that allows the calculation of the statistics of the random field solving problem (1). In order to obtain the limiting problem, we need the definitions of *stationarity* of a random field and *ergodicity* of a map.

Definition 1. *A random field is called* 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* $h \in \mathbb{R}^d$.

Definition 2. *Let* $T : \Omega \to \Omega$ *be a measure preserving transformation. We say that* $T$ *is* ergodic *with respect to the measure* $P$ *if for any* $F \in \mathcal{F}$ *with* $T(F) \subset F$ *we have either* $P(F) = 0$ *or* $P(F) = 1$.

We are now ready to state the theorem that lays the ground for the numerical investigations presented in this study. This result is based on the presentation in [24].

Theorem 3. *Let* $A_\varepsilon(x, \omega) := A_\varepsilon(T_x(\omega))$ *be a random field for an ergodic transformation* $T$ *and let* $A_\varepsilon$ *be strictly stationary and obey the conditions*

$$\exists\exists\, \underline{a}, \overline{a} > 0\colon \quad \forall\forall y, \xi \in \mathbb{R}\colon \quad \forall \omega \in \Omega\colon \quad \underline{a}|\xi|^2 \leq \xi^\top A(y, \omega)\xi \leq \overline{a}|\xi|^2. \tag{2}$$

*Then, as* $\varepsilon \to 0+$, *the solution to problem* (1) *converges to the solution of the deterministic problem*

$$-\nabla \cdot \left(\overline{A}\nabla u_0(x)\right) = f(x) \qquad \text{in } D, \tag{3a}$$

$$u_0(x) = 0 \qquad \text{on } \partial D, \tag{3b}$$

*in the sense that*

$$\lim_{\varepsilon \to 0} \mathbb{E}\left(\int_D \left(u_\varepsilon(x) - u_0(x)\right)^2 \, dx\right) = 0. \tag{4}$$

*Here,* $\overline{A}$ *is defined by*

$$\begin{aligned} \xi^\top \cdot \overline{A}\widetilde{\xi} &= \lim_{L \to \infty} \frac{1}{L^d} \left(\int_{[-L/2, L/2]^d} (\xi + \nabla \chi_\xi(y, \omega)) \cdot A(y, \omega)(\widetilde{\xi} + \nabla \chi_{\widetilde{\xi}}(y, \omega)) \, dy\right) \\ &= \mathbb{E}\left[(\xi + \nabla \chi_\xi) \cdot A(\widetilde{\xi} + \nabla \chi_{\widetilde{\xi}})\right] \end{aligned} \tag{5}$$

*for all* $\xi, \widetilde{\xi} \in \mathbb{R}^d$, *where* $\chi_\xi$ *is the* first-order corrector *solving the so-called auxiliary or cell problem* [19]

$$-\nabla \cdot A(\xi + \nabla \chi_\xi) = 0 \qquad \text{in } \mathbb{R}^d, \tag{6a}$$

$$\nabla \chi_\xi \text{ is stationary}, \tag{6b}$$

$$\int \nabla \chi_\xi(y, \omega) \, dy = 0 \qquad \forall \omega \in \Omega. \tag{6c}$$

■

This work addresses the approximation of the first-order corrector $\chi_\xi$ by numerically solving problem (6). The main objective is the optimization of the computational efficiency for a given approximation error tolerance. In order to do so, we need expressions for the necessary computational work as well as for the error introduced by discretization, cut-off, and sampling.

**2.2. Setting.** We will consider a domain that describes a matrix material containing circular hard-sphere inclusions at random positions, which means that the inclusions cannot overlap. The two materials shall differ in their physical properties modeled by the coefficient function $A$. In applications, there are many physical interpretations of this coefficient function, ranging from the meaning of a diffusion coefficient to electric permittivity to heat-transfer coefficient. In leading examples, $A$ is considered to be piecewise constant throughout $D$ and to have the form

$$A(y,\omega) = A_C \mathbb{1}_C(y,\omega) + A_M \mathbb{1}_M(y,\omega), \tag{7}$$

where $\mathbb{1}_C$ and $\mathbb{1}_M$ denote the indicator for the circular inclusion and for the matrix material, respectively, and $A_C$ and $A_M$ represent constant matrices. For a coefficient function of this type, the conditions stated in (2) are valid, and the theory applies.

The most straightforward choice for vectors $\xi$ in the auxiliary problem is the choice of unit vectors $e_i$, where we will denote the respective solution by $\chi_i$ with integer $i$. In particular, to obtain all the entries of $\overline{A}$, we use $\chi = (\chi_1, \dots, \chi_d)^\top$, so the problem here is a vectorial one. However, the equations for the different components are independent and can be computed one after another. Since the structure of the problems is the same, we will set $\chi := \chi_1$ and only perform the calculations for the first component. The whole procedure applies for $\chi_2, \dots, \chi_d$ in an analogous way.

**2.3. Computational scheme.** There are various approaches for the approximation of the integral in the probability space. Numerous publications employ basic, but powerful, sampling techniques such as Monte Carlo, quasi Monte Carlo [16], and multilevel Monte Carlo [3, 8] methods, which are favorable due to their simplicity, since algorithms for deterministic problems can be used. Furthermore, there are also spectral techniques based on polynomial chaos expansions [11, 18, 26], which require a new implementation of the solution algorithm for the stochastic case.

Here, the computational scheme is based on a Monte Carlo finite-element method combining approximations of the probability space and of the physical space. The overall approximation error therefore comprises both parts and will in fact include a further part related to artificial introduction of boundary conditions.

The integral over the probability space, i.e., the calculation of the expected value, is approximated employing a Monte Carlo method. The parameter governing this method is the number $N$ of trials evaluated in the sampling process.

The PDE defined on the physical domain is solved using a finite-element method. The typical parameter of importance for error estimation is the size $h$ of the mesh. However, since we are approximating a problem on the whole space, we also need to introduce a cut-off distance $L$ that defines a bounded domain on which we solve the auxiliary problem. These

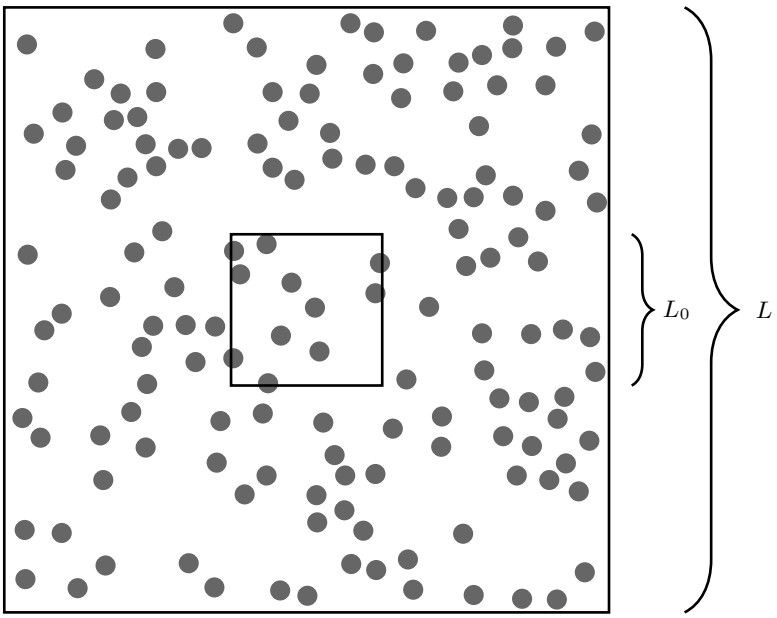



**Figure 1.** *Sketch of the investigated domain. The random circular inclusions are also shown. The convergence is studied only on the small square of length $L_0$ in the inside of the large domain. The parameters chosen here are $L := 12$, $L_0 := 3$, and $r := 0.2$.*

two parameters will be taken into account for the performance analysis of the implemented scheme.

It is also necessary to prescribe boundary conditions for this given domain, which will also introduce an error to the solution. In order to minimize this error, we consider periodic boundary conditions, which have been shown to lead to the lowest error [27]. Moreover, to further decrease the introduced error by the finite domain size, we will only investigate the solution on a part of the domain which can be considered far away from the boundary and hence its influence. We therefore consider a domain of length $L_0$, where $L_0 < L$, on which we study convergence. A sketch of this idea is shown in Figure 1.

**3. Optimization of the computational scheme.** The goal of this section is to find the optimal parameters for the computational scheme when calculating the first-order corrector. Optimization is understood here in the sense that for a given error tolerance, the computational work should be minimized.

As mentioned before, we combine a Monte Carlo sampling method for the probability space and a finite-element method for the physical space in this derivation of the error and work functions. We also formulate our main result, which is Proposition 5, in terms of these functions. However, the result is still valid for any other type of numerical approximation, if the expressions for the error and the work are adapted once according to the respective method.

**3.1. Error estimates.** We will now estimate the error involved in the calculation of the first-order corrector solving the auxiliary problem (6). As the homogenized coefficient depends

on $\nabla\chi$, this will be the quantity of interest in the estimation. Furthermore, the $L^2$-norm of the gradient also introduces a norm on the space used to evaluate one single realization of the auxiliary problem, namely, on

$$H := \left\{ u \in H^1_{\#}(D_{L_0}) \;\middle|\; \int_{D_{L_0}} u \,\mathrm{d}y = 0 \right\}, \tag{8}$$

where the hash indicates that the functions must be periodic on $D_{L_0}$. The overall error estimation will be performed on a smaller physical square $D_{L_0} := [-L_0/2, L_0/2)^d$ in order to attenuate the influence of the artificial boundary conditions (see Figure 1). The appropriate space to consider the error is then the Bochner space $L^2(\Omega; L^2(D_{L_0}))$, where $L^2(\Omega; X)$ is the space which consists of all measurable functions $u\colon \Omega \to X$ for which the norm

$$\|u\|_{L^2(\Omega;X)} := \left( \int_\Omega \|u(\cdot,\omega)\|_X^2 \,\mathrm{d}P(\omega) \right)^{1/2} = \mathbb{E}\left( \|u(\cdot,\omega)\|_X^2 \right)^{1/2} \tag{9}$$

is finite.

There are three sources of error arising during the numerical approximation of $\nabla\chi$:

(i) Discretization error: This is the error introduced by the approximation of the exact solution $\nabla\chi$ by a discretized solution $\nabla\chi_h$ calculated on the whole space.

(ii) Error due to cut-off and artificial boundary conditions: Since the auxiliary problem cannot be solved numerically on the whole space, a truncation of the domain is necessary. This includes the prescription of artificial boundary conditions that do not exist for a problem defined on the whole space. This approximation of $\nabla\chi_h$ will be denoted by $\nabla\chi_{L,h}$, where $L$ denotes the side length of the truncated domain.

(iii) Statistical error: The expected value of $\nabla\chi_{L,h}$ is approximated by a stochastic sampling method. The sample mean over $N$ independent realizations will be denoted by $\mu_{N,L,h}$, yielding

$$\mu_{N,L,h} = \frac{1}{N} \sum_{i=1}^{N} \nabla\chi_{L,h}^{(i)}, \tag{10}$$

where the superscript $(i)$ indicates specific realizations of $\nabla\chi_{L,h}$. 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.

The quantity we are then interested in is the overall error in the Bochner space.

Proposition 4. *Assume that the error $e_{\mathrm{FEM}}$ introduced by the discretization, the error $e_{\mathrm{BC}}$ introduced by cutting off the domain, and the error $e_{\mathrm{MC}}$ introduced by Monte Carlo sampling are given by*

$$e_{\mathrm{MC}} \leq \nu_0 N^{-\sigma}, \tag{11a}$$

$$e_{\mathrm{FEM}} \leq \nu_1 h^{\alpha}, \tag{11b}$$

$$e_{\mathrm{BC}} \leq \nu_2 L^{-\beta}. \tag{11c}$$

*Then the error bound*

$$\|\mu_{N,L,h} - \mathbb{E}[\nabla\chi]\|_{L^2(\Omega;L^2(D_{L_0}))} \leq \nu_0 N^{-\sigma} + \nu_1 h^{\alpha} + \nu_2 L^{-\beta} =: E(h,N,L) \tag{12}$$

*holds for the Monte Carlo finite element method estimator* $\mu_{N,L,h}$.

*Proof.* We have

$$\begin{aligned}
&\|\mu_{N,L,h} - \mathbb{E}[\nabla\chi]\|_{L^2(\Omega;L^2(D_{L_0}))} \\
&\quad\leq \|\mu_{N,L,h} - \mathbb{E}(\mu_{N,L,h})\|_{L^2(\Omega;L^2(D_{L_0}))} + \|\mathbb{E}(\mu_{N,L,h}) - \mathbb{E}(\nabla\chi_h)\|_{L^2(\Omega;L^2(D_{L_0}))} \\
&\qquad\qquad + \|\mathbb{E}(\nabla\chi_h) - \mathbb{E}(\nabla\chi)\|_{L^2(\Omega;L^2(D_{L_0}))} \\
&\quad\leq \underbrace{\|\mu_{N,L,h} - \mathbb{E}(\nabla\chi_{L,h})\|_{L^2(\Omega;L^2(D_{L_0}))}}_{=:e_{\mathrm{MC}}} + \underbrace{\|\nabla\chi_{L,h} - \nabla\chi_h\|_{L^2(\Omega;L^2(D_{L_0}))}}_{=:e_{\mathrm{BC}}} \\
&\qquad\qquad + \underbrace{\|\nabla\chi_h - \nabla\chi\|_{L^2(\Omega;L^2(D_{L_0}))}}_{=:e_{\mathrm{FEM}}},
\end{aligned} \tag{13}$$

by the triangle inequality, where we have used Jensen's inequality in the last step. Since $\mathbb{E}(\mu_{N,L,h})$ is an unbiased estimator, we have furthermore used $\mathbb{E}(\mu_{N,L,h}) = \mathbb{E}(\nabla\chi_{L,h})$. By using (11) for each of the components of the error, the assertion follows. ∎

To be able to make use of this proposition, we check the assumptions on the error estimates. The convergence of the statistical error in Monte Carlo sampling methods is well-known and therefore requires no further attention. The important part is the estimate in terms of the two other parameters. In particular, the question is whether these expressions are sufficient or whether a mixed term is necessary. We therefore calculated the error for many different parameters in a representative range for $h$ and $L$ and calculated the best fit for the estimate above as well as for an expression with an additional mixed term $\nu_3 h^{\rho} L^{-\tau}$. It was found that the error for the expression with the additional mixed term is marginally better (less than 1% difference). Given that the new expression comprises three additional parameters ($\nu_3$, $\rho$, and $\tau$), the error must be smaller, yet in view of the small size of the improvement it is safe to say that the assumptions used here are valid.

**3.2. Computational work.** In order to model the computational work, the expression for the work will depend on the mesh size $h$, the size of the domain $L$, and the number $N$ of realiziations used in the Monte Carlo simulation.

Calculating one realization of the auxiliary problem consists of several steps that all scale differently in $h$ and $L$. The procedure can be split into $n$ steps that all scale in the form

$$W_i \propto h^{-\xi_i} L^{\gamma_i}. \tag{14}$$

Altogether, the computational work for a whole Monte Carlo simulation comprising $N$ samples is therefore given by

$$W(h,N,L) = N \sum_{i=1}^{n} \mu_i h^{-\xi_i} L^{\gamma_i}. \tag{15}$$

In order to minimize this expression under the constraint $E(h,N,L) \leq \varepsilon$ for a given error tolerance $\varepsilon$, we employ Lagrange multipliers and the Karush–Kuhn–Tucker (KKT) conditions which yield necessary conditions for an optimal solution to our problem.

Proposition 5. *Parameters $(h,N,L)$ that are in the interior of the domain of admissible parameters and that minimize the computational work for a prescribed error tolerance $\varepsilon$ necessarily solve the system of equations*

$$\sum_{i=1}^{n} \left( \xi_i - \frac{\nu_1 \alpha h^\alpha}{\sigma(\varepsilon - \nu_1 h^\alpha - \nu_2 L^{-\beta})} \right) \mu_i h^{-\xi_i} L^{\gamma_i} = 0, \tag{16a}$$

$$\sum_{i=1}^{n} \left( \gamma_i - \frac{\beta \nu_2 L^{-\beta}}{\sigma(\varepsilon - \nu_1 h^\alpha - \nu_2 L^{-\beta})} \right) \mu_i h^{-\xi_i} L^{\gamma_i} = 0, \tag{16b}$$

$$N^\sigma - \frac{\nu_0}{\varepsilon - \nu_1 h^\alpha - \nu_2 L^{-\beta}} = 0. \tag{16c}$$

*Proof.* The Lagrange function associated to the optimization problem is given by

$$\mathcal{L}(h,N,L,s) := W(h,N,L) + s \cdot (E(h,N,L) - \varepsilon), \tag{17}$$

where $W(h,N,L)$ and $E(h,N,L)$ are given by (15) and (12), respectively. The necessary conditions for a minimum are obtained by calculating the partial derivatives of $\mathcal{L}$ with respect to $N$, $L$, $h$, and $s$, which yields

$$0 = \frac{\partial \mathcal{L}(h,N,L,s)}{\partial N} = \sum_{k=1}^{n} \mu_k L^{\gamma_k} h^{-\xi_k} - s\sigma \frac{\nu_0}{N^{\sigma+1}}, \tag{18a}$$

$$0 = \frac{\partial \mathcal{L}(h,N,L,s)}{\partial L} = N \sum_{k=1}^{n} \gamma_k \mu_k L^{\gamma_k - 1} h^{-\xi_k} - s\nu_2 \beta L^{-\beta-1}, \tag{18b}$$

$$0 = \frac{\partial \mathcal{L}(h,N,L,s)}{\partial h} = -N \sum_{k=1}^{n} \xi_k \mu_k L^{\gamma_k} h^{-\xi_k - 1} + s\nu_1 \alpha h^{\alpha-1}, \tag{18c}$$

$$0 = \frac{\partial \mathcal{L}(h,N,L,s)}{\partial s} = \frac{\nu_0}{N^\sigma} + \nu_1 h^\alpha + \nu_2 L^{-\beta} - \varepsilon. \tag{18d}$$

Equation (18a) translates to

$$s = \frac{N^{\sigma+1}}{\nu_0 \sigma} \sum_{k=1}^{n} \mu_k L^{\gamma_k} h^{-\xi_k}, \tag{19}$$

while we obtain

$$N^\sigma - \frac{\nu_0}{\varepsilon - \nu_1 h^\alpha - \nu_2 L^{-\beta}} = 0 \tag{20}$$

from (18d). Finally, inserting the expression for $s$ into (18b) and (18c) yields (16a) and (16b), which completes the proof. ∎

*Remark* 6. The coefficients as well as the exponents in (15) generally depend on the precise algorithm used and even on its implementation and the hardware. Therefore their exact numbers will be determined by numerical experiments in section 4.

*Remark* 7. After having computed these candidates for the minimum, it must be checked that the parameters actually yield a minimum. This can be achieved by calculating the determinant of the Hessian matrix of the Lagrange function $\mathcal{L}$. If this quantity is negative for the parameters, this point is a minimum. The determinant of this $4 \times 4$ matrix is a large expression, however, so it is not recorded here.

*Remark* 8. The analysis of the KKT conditions shows that the minimal work is obtained for $E(h, N, L) = \varepsilon$ (and not for $E(h, N, L) < \varepsilon$), i.e., one will not obtain a lower error than the prescribed one after minimization.

**4. Results.** For carrying out the numerical experiments, we used an Intel Core i7-3770 3.4GHz CPU with 32 GB of RAM running Linux.

Here, we used the coefficient function stated in (7) with

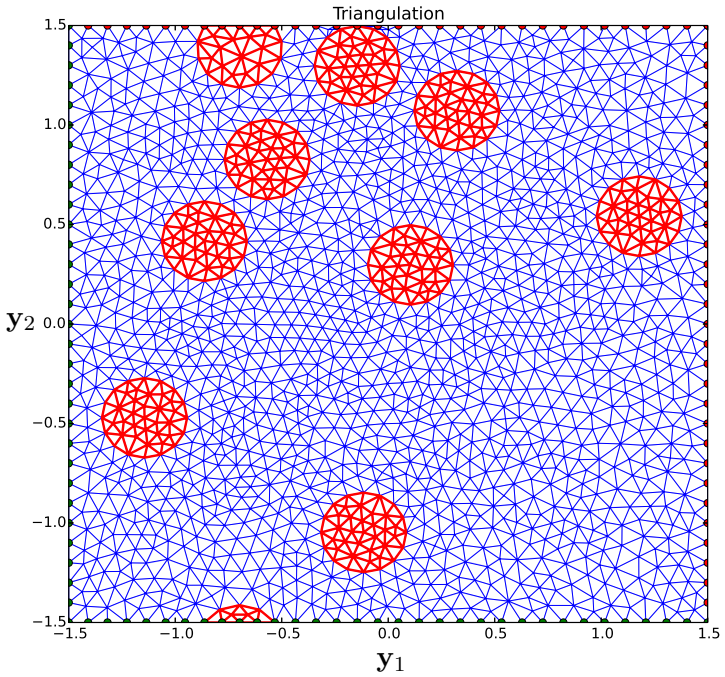
$$A_C := \begin{pmatrix} 20 & 0 \\ 0 & 10 \end{pmatrix}, \qquad A_M := \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}.$$

The density of the inclusions is set to one per unit square, where each inclusion is placed randomly in the domain such that there is no overlapping. The radius of the circular inclusions was chosen to be 0.2, which leads to a coverage of approximately 13%. For the subdomain, we always used a length of $L_0 := 3$. As already mentioned, we used periodic boundary conditions and employed a periodic continuation of circles at the boundary. A typical realization on $D_{L_0}$ together with its solution are shown in Figures 2(a) and 2(b), respectively.

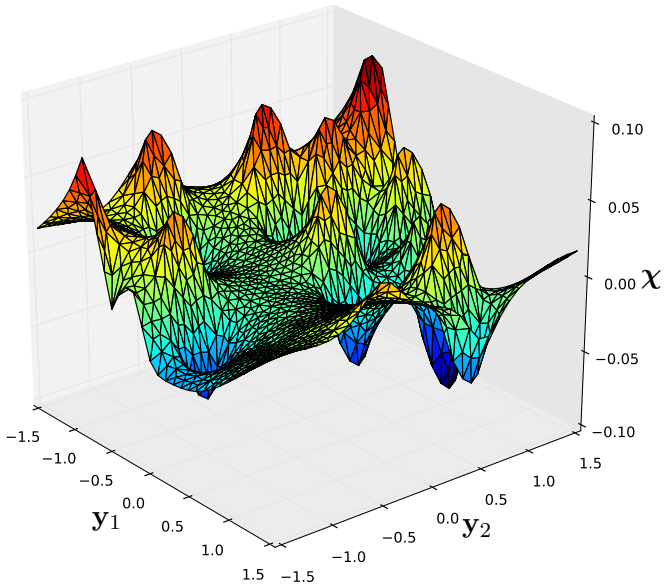
**4.1. Error function.** Before we can address the optimization problem, we need to calculate the parameters involved in the error function as well as in the function describing the computational work. To this end, we performed Monte Carlo simulations for different mesh sizes and cut-off lengths and determined the required coefficients by a regression after a logarithmic conversion. Since exact solutions of the auxiliary problem are unknown, we computed reference solutions to which we could compare the results on smaller domains and on coarser meshes, yielding the convergence rates. The obtained errors compared to the reference solutions are shown in Figure 3. The extracted parameters for the error function are given in Table 1. Although theoretical results exist on the convergence rates with respect to the mesh size and the number of realizations, we also derived these coefficients from the numerical experiments, since correct optimization results can only be obtained based on the actual coefficients, because they provide sharper estimates than theoretical bounds do.

**4.2. Work function.** Solving the auxiliary problem consists of several computational steps. Measurements of the computational work show that each of them scales differently in terms of $h$ and $L$. In order to obtain a proper expression for the work function, we consider each step by itself and add the terms. In particular, we consider four steps:

1. the mesh generation,
2. the assembly of the stiffness matrix for the finite-element method,




(a) Mesh showing a typical realization for the auxiliary problem on the subdomain $D_{L_0}$.




(b) Auxiliary problem solution to the realization shown in Figure 2(a).

**Figure 2.** *Mesh for a typical realization and solution to the auxiliary problem based on this realization.*

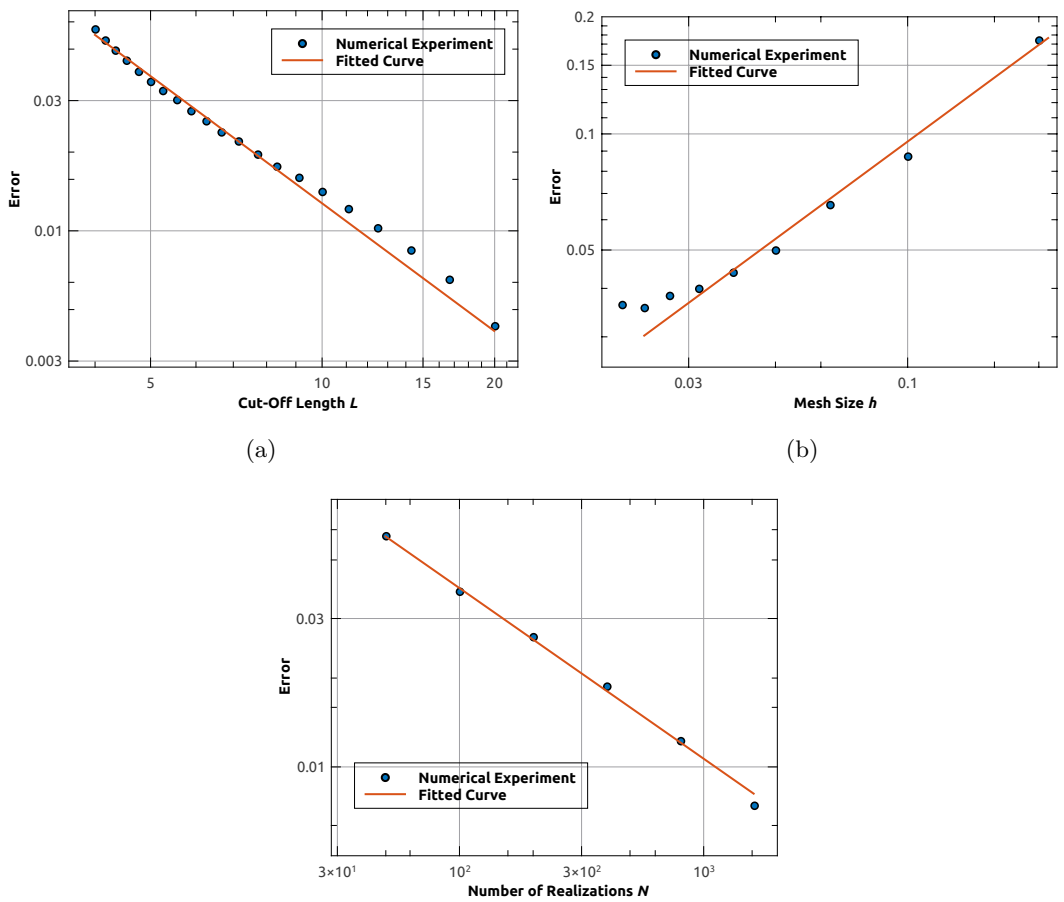



**Figure 3.** *Error in terms of* (a) *cut-off length* $L$, (b) *mesh size* $h$, *and* (c) *number* $N$ *of samples compared to the reference solution. The convergence rates and coefficients obtained from a least-squares fit are given in Table* 1.

**Table 1**

*Numerical values for the parameters governing the error function as given in* (11). *All these constants have been derived from numerical experiments addressing the convergence of the solution.*

| Parameter | Numerical value |
|---|---|
| $\nu_0$ | 0.574 |
| $\nu_1$ | 0.645 |
| $\nu_2$ | 0.539 |
| $\alpha$ | 0.827 |
| $\beta$ | 1.630 |
| $\sigma$ | 0.577 |

**Table 2**

*List of steps necessary to solve the auxiliary problem together with the coefficients $\mu_i$, $\xi_i$, and $\gamma_i$ as given in (15). A more detailed description of all steps is given in the text.*

| Step | Description | $\mu_i$ | $\xi_i$ | $\gamma_i$ |
|---|---|---|---|---|
| 1 | Generate mesh | $5.09 \cdot 10^{-5}$ | 2.00 | 2.22 |
| 2 | Assemble stiffness matrix | $2.09 \cdot 10^{-6}$ | 2.13 | 2.54 |
| 3 | Solve the system | $4.97 \cdot 10^{-7}$ | 2.82 | 2.98 |
| 4 | Apply uniqueness condition | $8.26 \cdot 10^{-8}$ | 2.79 | 3.40 |

3. the solution of the system of equations, and
4. the application of the uniqueness conditions as described in the definition of the function space $H$ in (8).

The mesh is generated using the *GMSH* package [10], where the mesh was aligned with the circles such that each element had a constant value for $A$. The remaining steps have been performed using Julia [4]. The assembly of the stiffness matrix also includes the application of the periodic boundary conditions. Step 3 consists of applying the backslash operator implemented in Julia. The application of the uniqueness condition is necessary since the solution obtained is unique only up to an additive constant. However, this last step is not necessary if one only calculates the homogenized coefficient $\overline{A}$, since the constant disappears over the gradient.

The parameters related to the computational work function are presented in Table 2. Here, we used mesh sizes between 0.02 and 0.1 and cut-off distances between 6 and 22 to obtain the scaling of the different steps. The resulting work-function values for each step are shown in Figure 4. Since the axes show logarithmic scaling, the surfaces describing the data points are planes.

**4.3. Optimal method.** With these coefficients, we prescribe an error tolerance and solve (16). The resulting optimal values for the three quantities depending on the error tolerance are shown in Figure 5. A straightforward, but long, calculation shows that the determinant of the Hessian matrix is negative at all points considered, which confirms that we actually found a minimum of the Lagrange function within the region of admissible values for $L$, $N$, and $h$.

As an example, we consider the optimal values for two prescribed error tolerances, which are stated in Table 3 for further discussion. These values are all admissible in the sense that they lie in the interior the region of admissible values (a minimum requirement being that $L$, $N$, and $h$ are all positive and that $h$ is smaller than the diameter of the domain), and therefore no further restrictions on them are required to obtain a suitable optimum. This confirms that the optimization method works as intended.

For the first case, the mesh is very fine and 239 Monte Carlo samples are used. The computational work is approximately 1151 core hours on the computer we used and is parallelized trivially.

The second example shows the results for a larger error tolerance. Although the mesh is still relatively fine (the radius of one circle is 0.2 such that its shape is resolved very well) and the domain size is still around 10, the core hours required are down to approximately 3 in this case, which is due to the much lower number of samples of 84.

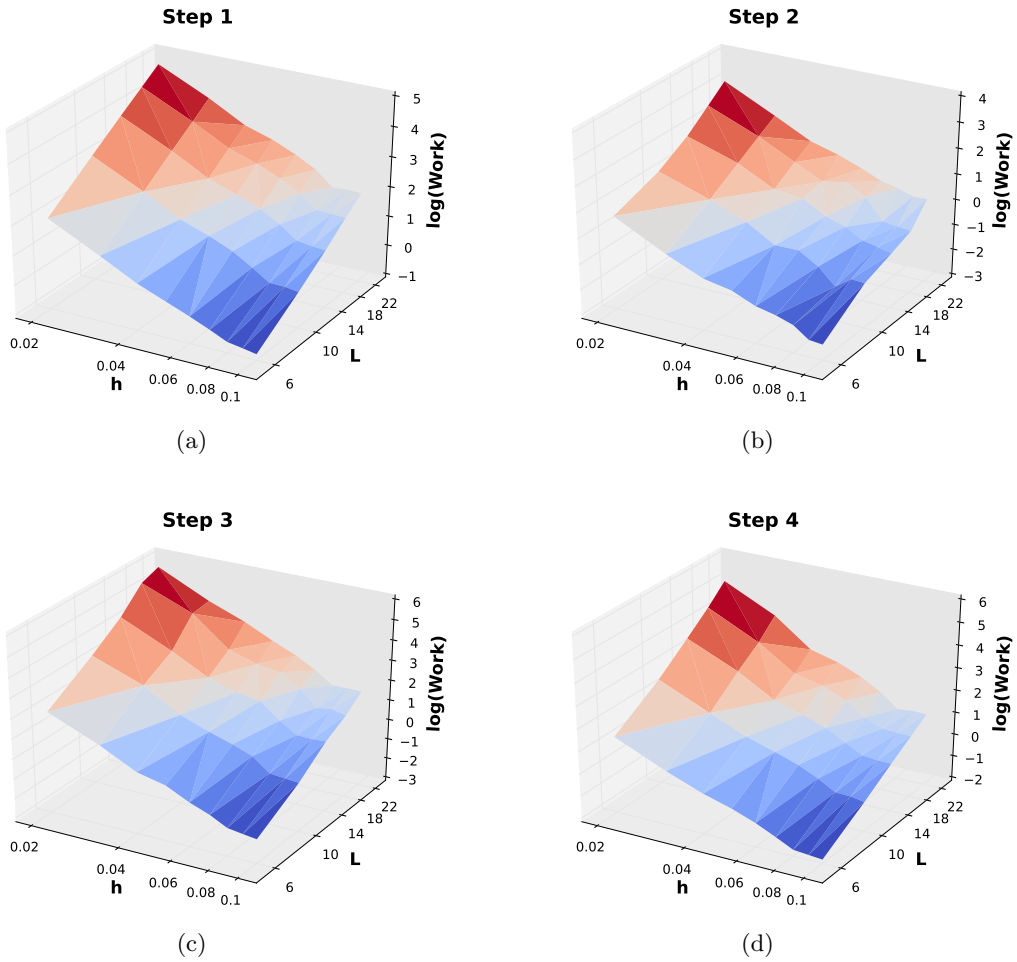



**Figure 4.** *Computational work necessary for one realization depending on the mesh size $h$ and the cut-off length $L$. Each diagram shows one step. Note the logarithmic scaling of the axes. The parameters found are shown in Table 2.*

As can be seen from these two examples, the computational work highly increases with decreasing error tolerance. This dependence is shown in Figure 6. A fit in the log-log scale reveals that the relation between $W$ and $\varepsilon$ is

$$W \propto \varepsilon^{-5.66}. \tag{21}$$

This large number for the exponent again emphasizes the necessity of a well-balanced setup that leads to an optimal simulation scheme.

**4.4. Discussion.** First, it turns out that only very few realizations of the system are required for the estimation of the computational work, especially compared to the large number of samples necessary in the actual calculation of the effective coefficient. Furthermore, one can—at least in parts—include the realizations calculated during the optimization step in the homogenization procedure such that almost no computational time is spent in vain. What is

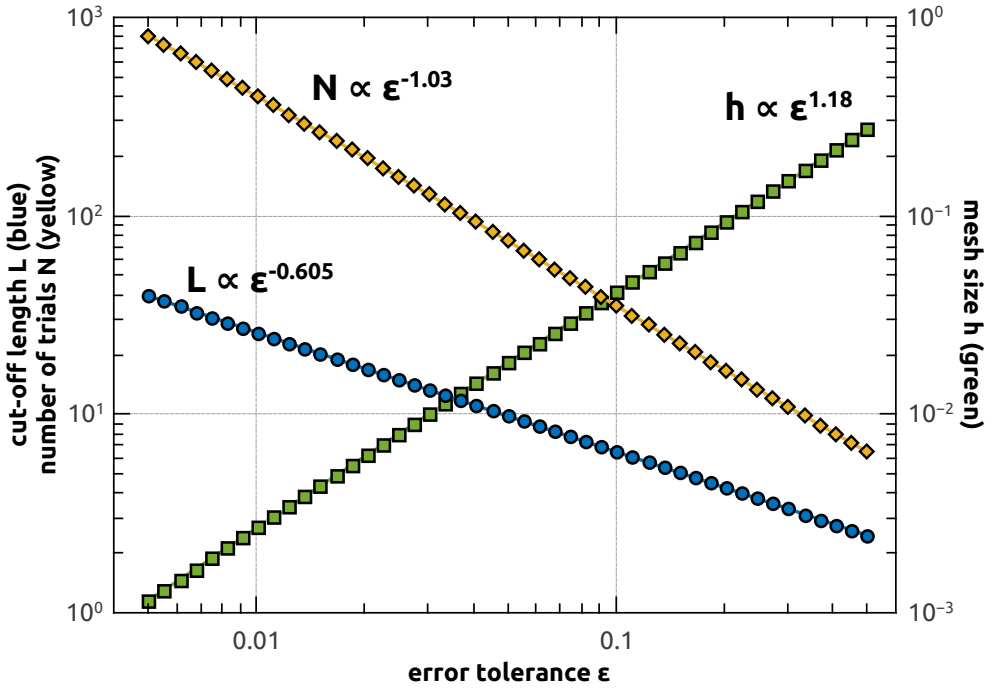



**Figure 5.** *Dependence of $L$ (blue), $N$ (yellow, both on left y-axis), and $h$ (green, on right y-axis) on the error tolerance $\varepsilon$.*

**Table 3**

*Example for optimal values for a prescribed error tolerance and resulting computational work. The optimal solution found is in the interior of the admissible values, so that no further restrictions are necessary.*

| $\varepsilon$ | $L$ | $h$ | $N$ | $W$ [h] |
|---|---|---|---|---|
| $1.7 \cdot 10^{-2}$ | 19.0 | $4.9 \cdot 10^{-3}$ | 239 | 1151 |
| $4.5 \cdot 10^{-2}$ | 10.4 | $1.6 \cdot 10^{-2}$ | 84 | 3 |

even more advantageous is the fact that the parameters related to the computational work need to be estimated only once if one calculates several effective coefficients, e.g., for different coefficient functions. This reduces the computational cost for the optimization procedure even more.

The data gathered for determining the governing coefficients shows linear behavior on the log-log scale, which justifies the use of power laws for the involved quantities.

However, after the calculation of the parameters giving a minimal work function value, one in general needs to check if the value found is indeed a minimum. First, it has to be admissible. Then, of course, the minimal property can be confirmed by calculating the determinant of the Hessian matrix of $\mathcal{L}(h, N, L, s)$.

The choice of a suitable error tolerance is a delicate task. Choosing a value too small such as $\varepsilon = 5.5 \cdot 10^{-3}$ leads to parameters that require enormous computational work (around $10^6$ hours $\approx 114$ years) and are therefore out of reach. Choosing a too large value, however, yields

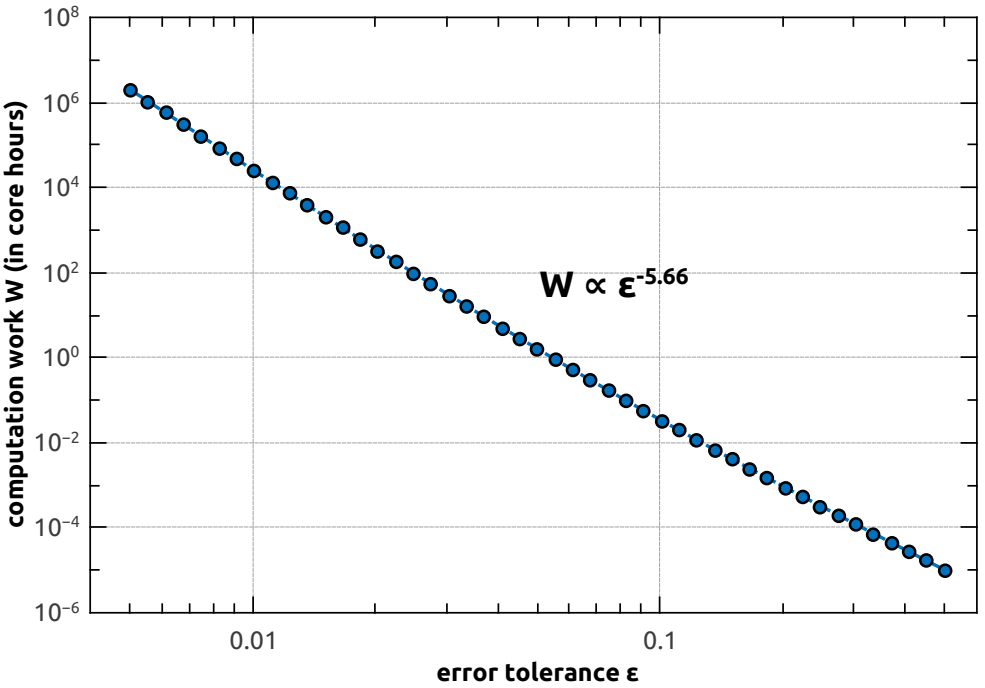



**Figure 6.** *Dependence of the computational work using optimal parameter values on the prescribed error tolerance $\varepsilon$. The slope here is $-5.66$. It is obvious that the solutions for the lowest error tolerances are way out of reach since they require too much computational effort.*

parameters that will not give meaningful results. For example, the mesh size always needs to be small enough such that the inclusions are properly resolved, the cut-off distance needs to be at least a bit larger than the inner domain on which the convergence is studied, and at least a few tens of realizations will be needed. These minimal requirements are necessary to obtain meaningful results. Of course, meaningful bounds for the error tolerance will always depend on the application.

The chosen stochastic process for the generation of the random inclusions can easily be generalized in several ways. First, the size of the inclusions could also be treated as a random process, which leads to variable quota for the two materials. Second, we prohibited overlapping of the circular inclusions, which is a suitable choice for many applications. However, for certain materials, allowing overlapping might be a better model, when the inclusions are free to form more complicated shapes.

**5. Conclusions.** In this work, we derived an optimal method for the numerical stochastic homogenization of elliptic problems, when a prescribed error tolerance is to be met. The optimization is based on finding the optimal variables, namely, the mesh size $h$, the cut-off distance $L$, and the number $N$ of samples in the Monte Carlo part. Several coefficients governing the work function and the error function were determined using numerical experiments, and afterward the optimization problem was solved for different error tolerances. The computational work for the optimization step is small compared to the calculation of the effective

coefficient, and the values obtained during the optimization step can even be reused in the latter part of the computation. The resulting optimal values for the variables are reasonable and as a consequence admissible without any further constraints on the optimization problem.

Calculations and the examples shown here confirmed that the necessary computational work strongly depends on the error tolerance and increases with a large exponent for a decreasing prescribed error tolerance. This fact again emphasizes the need for a properly designed and optimal computational scheme.

It goes without saying that the numerical results of the optimization problems also depend on the algorithms, implementations, and even hardware used. The coefficients in the error and the work functions should therefore be determined on the hardware used for the actual calculations in order to obtain optimal results.

Finally, the optimization algorithm presented in this work can be used for many similar computational problems. To do so, only the definitions for the error and work functions must be modified.

## REFERENCES


[1] B. Amaziane, A. Bourgeat, and J. Koebbe, *Numerical simulation and homogenization of two-phase flow in heterogeneous porous media*, Transp. Porous Media, 6 (1991), pp. 519–547.

[2] T. Arbogast, *Analysis of a two-scale, locally conservative subgrid upscaling for elliptic problems*, SIAM J. Numer. Anal., 42 (2004), pp. 576–598.

[3] A. Barth, C. Schwab, and N. Zollinger, *Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients*, Numer. Math., 119 (2011), pp. 123–161.

[4] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, *Julia: A Fresh Approach to Numerical Computing*, preprint, arXiv:1411.1607, 2014.

[5] A. Bourgeat and A. Piatnitski, *Approximations of effective coefficients in stochastic homogenization*, Ann. Inst. Henri Poincaré Probab. Stat., 40 (2004), pp. 153–165.

[6] L. A. Caffarelli and P. E. Souganidis, *Rates of convergence for the homogenization of fully nonlinear uniformly elliptic PDE in random media*, Invent. Math., 180 (2010), pp. 301–360.

[7] L. A. Caffarelli, P. E. Souganidis, and L. Wang, *Homogenization of fully nonlinear, uniformly elliptic and parabolic partial differential equations in stationary ergodic media*, Comm. Pure Appl. Math., 58 (2005), pp. 319–361.

[8] K. Cliffe, M. Giles, R. Scheichl, and A. L. Teckentrup, *Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients*, Comput. Vis. Sci., 14 (2011), pp. 3–15.

[9] R. Costaouec, *Asymptotic expansion of the homogenized matrix in two weakly stochastic homogenization settings*, Appl. Math. Res. Express AMRX, 2012 (2012), pp. 76–104.

[10] C. Geuzaine and J.-F. Remacle, *Gmsh: A 3-D finite element mesh generator with built-in pre-and post-processing facilities*, Internat. J. Numer. Methods Engrg., 79 (2009), pp. 1309–1331.

[11] R. G. Ghanem and P. D. Spanos, *Stochastic Finite Elements: A Spectral Approach*, Courier, 2003.

[12] A. Gloria and F. Otto, *Quantitative Results on the Corrector Equation in Stochastic Homogenization*, preprint, arXiv:1409.0801, 2014.

[13] A. Gloria and F. Otto, *Quantitative estimates on the periodic approximation of the corrector in stochastic homogenization*, ESAIM Proc. Surveys, 48 (2015), pp. 80–97.

[14] A. Gloria, F. Otto, et al., *An optimal variance estimate in stochastic homogenization of discrete elliptic equations*, Ann. Probab., 39 (2011), pp. 779–856.

[15] A. Gloria, F. Otto, et al., *An optimal error estimate in stochastic homogenization of discrete elliptic equations*, Ann. Appl. Probab., 22 (2012), pp. 1–28.

[16] I. G. Graham, F. Y. Kuo, D. Nuyens, R. Scheichl, and I. H. Sloan, *Quasi-Monte Carlo methods for elliptic PDEs with random coefficients and applications*, J. Comput. Phys., 230 (2011), pp. 3668–3694.

[17] T. Y. Hou and X.-H. Wu, *A multiscale finite element method for elliptic problems in composite materials and porous media*, J. Comput. Phys., 134 (1997), pp. 169–189.
[18] M. Jardak and R. Ghanem, *Spectral stochastic homogenization of divergence-type PDEs*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 429–447.
[19] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer, New York, 2012.
[20] T. Kanit, S. Forest, I. Galliet, V. Mounoury, and D. Jeulin, *Determination of the size of the representative volume element for random composites: Statistical and numerical approach*, Internat. J. Solids Structures, 40 (2003), pp. 3647–3679.
[21] Z. Khisaeva and M. Ostoja-Starzewski, *On the size of RVE in finite elasticity of random composites*, J. Elasticity, 85 (2006), pp. 153–173.
[22] S. M. Kozlov, *Averaging of random operators*, Mat. Sb., 151 (1979), pp. 188–202.
[23] M. Ostoja-Starzewski, X. Du, Z. Khisaeva, and W. Li, *Comparisons of the size of the representative volume element in elastic, plastic, thermoelastic, and permeable random microstructures*, Int. J. Multiscale Comput. Engrg., 5 (2007).
[24] G. C. Papanicolaou and Varadhan, *Boundary value problems with rapidly oscillating random coefficients*, Random Fields, 1 (1979), pp. 835–873.
[25] S. Torquato, *Random Heterogeneous Materials: Microstructure and Macroscopic Properties*, Vol. 16, Springer, New York, 2013.
[26] X. Xu and L. Graham-Brady, *A stochastic computational method for evaluation of global and local behavior of random elastic media*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 4362–4385.
[27] X. Yue and E. Weinan, *The local microscale problem in the multiscale modeling of strongly heterogeneous media: Effects of boundary conditions and cell size*, J. Comput. Phys., 222 (2007), pp. 556–572.
[28] V. Yurinskii, *Averaging an elliptic boundary-value problem with random coefficients*, Sib. Math. J., 21 (1980), pp. 470–482.