ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

MATHICSE

Mathematics Institute of Computational Science and Engineering

School of Basic Sciences - Section of Mathematics

MATHICSE Technical Report
Nr. 15.2011
February 2011

Analysis of the finite element heterogeneous multiscale method for nonmonotone elliptic homogenization problems

Assyr Abdulle, Gilles Vilmart

http://mathicse.epfl.ch

Phone: +41 21 69 37648

Address:
EPFL - SB - MATHICSE (Bâtiment MA)
Station 8 - CH-1015 - Lausanne - Switzerland

Fax: +41 21 69 32545

Analysis of the finite element heterogeneous multiscale method for nonmonotone elliptic homogenization problems.

Assyr Abdulle and Gilles Vilmart

Abstract

A fully discrete analysis of the finite element heterogeneous multiscale method for a class of nonlinear elliptic homogenization problems of nonmonotone type is proposed. In contrast to previous results obtained for such problems in dimension $d \leq 2$ for the H^1 norm and for a semi-discrete formulation [W.E, P. Ming and P. Zhang, J. Amer. Math. Soc. 18 (2005), no. 1, 121–156], we obtain optimal convergence results for dimension $d \leq 3$ and for a fully discrete method, which takes into account the microscale discretization. In addition, our results are also valid for quadrilateral finite elements, optimal a-priori error estimates are obtained for the H^1 and L^2 norms, improved estimates are obtained for the resonance error and the Newton method used to compute the solution is shown to converge. Numerical experiments confirm the theoretical convergence rates and illustrate the behavior of the numerical method for various nonlinear problems.

Keywords: nonmonotone quasilinear elliptic problem, numerical quadrature, finite elements, multiple scales, micro macro errors, numerical homogenization.

AMS subject classification (2010): 65N30,65M60,74D10,74Q05.

1 Introduction

We consider a finite element method (FEM) for the numerical solution of a class of nonlinear nonmonotone multiscale problems of the form

$$-\nabla \cdot (a^\varepsilon(x, u_\varepsilon(x))\nabla u_\varepsilon(x)) \quad = \quad f(x) \quad \text{in } \Omega, \tag{1}$$

in a domain $\Omega \subset \mathbb{R}^d,\ d \leq 3$, with suitable boundary conditions and where $a^\varepsilon(x, u)$ is a $d \times d$ tensor. Diffusion phenomena in highly heterogeneous medium from a wide range of applications are modeled by the nonlinear equations (1), where ε represents a small scale in the problem. For example, the stationary form of the Richards problem [11], problems related to phase changes in materials [32], the modeling of the thermal conductivity of the Earth's crust [35], or the heat conduction in composite materials [30] can be modeled with the help of (1). Yet, often the multiscale nature of the medium, described in (1) through a nonlinear multiscale conductivity tensor $a^\varepsilon(x, u_\varepsilon(x))$, is not taken into account in the modeling due to the difficulty in solving numerically (1). Indeed, standard numerical methods such as the FEM or the finite difference method (FD) require a grid resolving the medium's finest scale which is often computationally too demanding. Upscaling of equation (1) is thus needed for an efficient numerical treatment. Rigorously described by the mathematical homogenization theory [12],[29], coarse graining (or homogenization) aims at averaging the finest scales of a multiscale equation and deriving a homogenized equation that captures the essential macroscopic features of the problem as $\varepsilon \to 0$. The mathematical homogenization of (1) has been developed in [9],[14], and [28], where it is shown that the homogenized equation is of the same quasilinear type as the original equation, with $a^\varepsilon(x, u_\varepsilon(x))$ replaced by a homogenized tensor $a^0(x, u_0(x))$ depending nonlinearly on a homogenized solution u_0 (the limit in a certain sense of u_ε as $\varepsilon \to 0$).

Several difficulties arise when trying to compute a numerical solution of the homogenized equation. Since the tensor is not proportional to the identity in general, the Kirschoff transformation (see for instance [34]) cannot be used to treat the non-linearity. Also, as it depends on the point x of the computational domain, the tensor $a^0(x, u_0(x))$ can only be computed at a finite number of points. This amounts to define adequate quadrature points and to define a modified bilinear form based on quadrature formulas (QF). Then, as $a^0(x, u_0(x))$ has to be computed numerically, only an approximation of the tensor can be obtained. The accuracy of the numerical tensors has to be taken into account when solving numerically the homogenization problems as both the existence and the convergence properties of a numerical approximation of u_0 depend on it. For numerical simulation, a linearization scheme relying on the modified bilinear form has to be constructed and shown to be convergent. Finally, a-priori convergence rates have to be derived to control the accuracy of a numerical solution.

While numerical methods for linear elliptic homogenization problems have been studied in many papers - see [3],[23],[25], and the references therein - the literature for the numerical homogenization of nonlinear nonmonotone elliptic problems is less abundant. Numerical methods based on the multiscale finite element method (MsFEM) [25] for nonlinear elliptic problems of the form $-\nabla \cdot (a^\varepsilon(x, u_\varepsilon(x), \nabla u_\varepsilon(x)) = f(x)$ have been studied in [26],[25], where a monotonicity assumption has been used to derive *convergence rates*. This assumption leads essentially to problems of the type $-\nabla \cdot (a^\varepsilon(x, \nabla u_\varepsilon(x)) = f(x)$. MsFEM has been studied for problem (1) in [17] and for the finite element heterogeneous multiscale method (FE-HMM) in [24]. The approaches in these papers rely on the two-grid discretization techniques introduced in [36]. There, the analysis of FEM for nonlinear partial differential equation (PDE) relies on the linearization of the equation at the exact solution and the study of its FEM discretization. However, for numerical homogenization, an additional term arises in the linearization due to the discrepancy between the exact bilinear form and the bilinear form based on numerical quadrature. We also note that the analysis in [36] is only valid for two-dimensional problems as it relies on bounds for discrete Green functions which are not available for three-dimensional problems [36, p.1760].

In this paper we analyze a numerical homogenization method based on the FE-HMM for problem (1). First results for the FE-HMM applied to (1) have been obtained in [24]. Our analysis is different and allows for significant generalizations and new results. It combines recent results [7] on optimal convergence rates for standard FEMs with numerical quadrature for nonlinear nonmonotone problems, with the analysis of the FE-HMM for linear problems. For the convenience of the reader we briefly put our results in perspective.

Our analysis is valid in dimension $d \leq 3$ in bounded convex polyhedral domains. The analysis in [24] relies on the use of a discrete Green functions G_H^z and the logarithmic bound $\sup_{z\in\overline{\Omega}} \|G_H^z\|_{W^{1,1}(\Omega)} \leq C|\log H|$ (see [24, equ.(5.16)]). Such an estimate to the best of our knowledge is not available in dimension $d = 3$ for arbitrary bounded convex polyhedral domains.

We propose a fully discrete analysis taking into account the H^1 and L^2 errors at both the microscopic and the macroscopic grid of the FE-HMM scheme. In contrast, the results in [24] were derived for a semi-discrete formulation of the FE-HMM and only for the H^1 and $W^{1,\infty}$ norms. The derivation of convergence rates in the L^2 norm does not follow from a standard duality argument (due to the nonlinearity of the problem and the use of numerical integration). Here we use the new estimates for FEM with numerical quadrature for indefinite linear elliptic problems obtained in [7].

We also improve the so-called modeling or resonance error (r_{MOD}) obtained in [24] for locally periodic tensor. In Theorem 3.7 we show the estimate $r_{MOD} \leq C(\delta + \varepsilon/\delta)$, whereas $r_{MOD} \leq C(\delta + \sqrt{\varepsilon/\delta})$ was obtained in [24, Thm. 5.5] (here ε is the size of the period and δ the length, in each spatial direction, of the sampling domains).

In [24, equ. 5.21] the difference between the weak form for the exact problem and the discretized problem based on numerical quadrature is estimated by $|A(u^H; u^H, w^H) - A_H(u^H; u^H, w^H)| \leq CH^\ell \|w^H\|_{H^1(\Omega)}$, [1] using results obtained for linear problems [20]. However, C depends (nonlinearly due to the nonlinearity of the tensor) on the broken norms of u^H in Sobolev spaces of the type $W^{\ell+1,p}(\Omega)$. Thus, a priori bounds (independent of H) are needed for these high-order broken norms of the solution u^H. This issue has not been discussed in [24] and we do not know how to derive such bounds for $\mathcal{P}^\ell$-simplicial FEs when $\ell > 2$ or for $\mathcal{Q}^\ell$-quadrilateral FEs when $\ell \geq 1$.[2] In contrast, our analysis is valid for arbitrary high-order simplicial or quadrilateral FEs. The above estimates involving u^H are never used as we rely on the projection of the exact homogenized solution when estimating the quadrature error. This allows to use the regularity assumed on the exact solution.

While a local uniqueness of the semi-discrete scheme was proved in [24, Lemma 5.3], we prove in Theorem 3.1 that *any sequence* of numerical solutions for the fully discrete scheme converges with optimal convergence rates (on sufficiently fine meshes) and give in Theorem 3.3 necessary conditions on the parameters of the problem for the numerical solution u^H to be unique. Our results also show that the Newton method, used in practice to compute a solution of the nonlinear discretized problem, converges (see Theorem 4.11).

A basic assumption in [24] is that the linearized operator of the original homogenized equation at the exact solution u_0 is an isomorphism (which is difficult to check in practice). Here, we only rely on structure assumptions of the original tensor (see (4), (5)). Of course in both our proof and in [24], appropriate smoothness of the oscillating and the homogenized tensors is required.

Finally, we present a post-processing procedure similar as for linear problems, to approximate numerically the oscillating solution u^ε in the energy norm $H^1(\Omega)$. As the FE-HMM yields an approximation u^H to the homogenized solution u_0 (itself an $\mathcal{O}(1)$ approximation of u_ε in the energy norm), such reconstruction procedures are needed to capture the oscillating solution u_ε in the H^1 norm.

Our paper is organized as follows. In Sect. 2 we introduce the homogenization problem for nonlinear nonmonotone problems and we describe the multiscale method. In Sect. 3 we state our main results. The analysis of the numerical method is given in Sect. 4. In Sect. 5 we present and analyze a reconstruction procedure to capture numerically the fine scales of the oscillatory problem. In Sect. 6 we first discuss an efficient implementation of the linearization scheme used for solving the nonlinear macroscopic equation and present various numerical experiments which confirm the sharpness of our a priori error bounds and illustrate the versatility of our method.

Notations. Let $\Omega \subset \mathbb{R}^d$ be open and denote by $W^{s,p}(\Omega)$ the standard Sobolev space. For $p =$

[1] Here $A(u^H; u^H, w^H) = \int_\Omega a^0(x, u^H)\nabla u^H \nabla w^H dx$ and A_H is a corresponding nonlinear form based on numerical quadrature.

[2] For simplicial $\mathcal{P}^2$ elements, an a priori bound can be obtained by combining the $W^{1,\infty}$ estimates in [24] with the inverse inequality. Unfortunately, these arguments cannot be used for $\mathcal{P}^\ell$, $\ell > 2$ or $\mathcal{Q}^\ell$.

2 we use the notation $H^s(\Omega)$ and $H^1_0(\Omega)$, and denote by $W^1_{per}(Y) = \{v \in H^1_{per}(Y);\ \int_Y v\,dx = 0\}$, where $H^s_{per}(Y)$ is defined as the closure of $\mathcal{C}^\infty_{per}(Y)$ (the subset of $\mathcal{C}^\infty(\mathbb{R}^d)$ of periodic functions in $Y = (0,1)^d$ with respect to the H^s norm. For a domain $D \subset \Omega$, $|D|$ denotes the measure of D. Given a $d \times d$ tensor a, we denote $\|a\|_F = \sqrt{\sum_{m,n} |a_{mn}|^2}$ its Frobenius norm.

2 Homogenization and multiscale method

Let Ω be a bounded convex polyhedral subset of $\mathbb{R}^d$, where $d \leq 3$. We consider the quasilinear elliptic problems (1), where for simplicity we take homogeneous Dirichlet boundary conditions, i.e., $u_\varepsilon(x) = 0$ on $\partial\Omega$. Associated to $\varepsilon > 0$, a sequence of positive real numbers going to zero, we consider a sequence of tensors $a^\varepsilon(\cdot, s) = (a^\varepsilon_{mn}(\cdot, s))_{1 \leq m,n \leq d}$ assumed to be uniformly elliptic and bounded, continuous on $\overline{\Omega} \times \mathbb{R}$ and uniformly Lipschitz continuous with respect to s. We further assume that $f \in H^{-1}(\Omega)$. Under the above assumptions, for all fixed $\varepsilon > 0$, the weak form of (1) has a unique solution $u_\varepsilon \in H^1_0(\Omega)$ (see for example [18, Theorem 11.6]),which satisfies the bound

$$\|u_\varepsilon\|_{H^1(\Omega)} \leq \lambda^{-1} \|f\|_{H^{-1}(\Omega)}. \tag{2}$$

Thus, standard compactness arguments implies the existence of a subsequence of $\{u_\varepsilon\}$ converging weakly in $H^1(\Omega)$. The aim of homogenization theory is to provide a limiting equation for u_0. The following result is shown in [14, Theorem 3.6] (see also [28]): there exists a subsequence of $\{a^\varepsilon(\cdot, s)\}$ (again indexed by ε) such that the corresponding sequence of solutions $\{u_\varepsilon\}$ converges weakly to u_0 in $H^1(\Omega)$, where u_0 is solution of the so-called homogenization problem

$$\begin{aligned} -\nabla \cdot \big(a^0(x, u_0(x)) \nabla u_0(x)\big) &= f(x) \ \text{ in } \Omega, \\ u_0(x) &= 0 \ \text{ on } \partial\Omega, \end{aligned} \tag{3}$$

and where the tensor $a^0(x, s)$, called the homogenized tensor, can be shown to be again uniformly elliptic, bounded and Lipschitz continuous with respect to s.[3] We summarize next our basic assumptions on the homogenized problem to analyze the proposed numerical method:

- the coefficients $a^0_{mn}(x, s)$ are continuous functions on $\overline{\Omega} \times \mathbb{R}$ which are uniformly Lipschitz continuous with respect to s, i.e., there exist $\Lambda_1 > 0$ such that
$$|a^0_{mn}(x, s_1) - a^0_{mn}(x, s_2)| \leq \Lambda_1 |s_1 - s_2|, \ \forall x \in \Omega, \forall s_1, s_2 \in \mathbb{R}, \forall\ 1 \leq m, n \leq d. \tag{4}$$

- the tensor $a^0(x, s)$ is uniformly elliptic and bounded, i.e., there exist $\lambda, \Lambda_0 > 0$ such that
$$\lambda \|\xi\|^2 \leq a^0(x, s)\xi \cdot \xi, \quad \|a^0(x, s)\xi\| \leq \Lambda_0 \|\xi\|, \quad \forall \xi \in \mathbb{R}^d, \forall s \in \mathbb{R}, \forall x \in \Omega. \tag{5}$$

Under these assumptions, the homogenized problem (3) has also a unique solution $u_0 \in H^1_0(\Omega)$. Let us further mention the following characterization of the homogenized tensor,

[3] With the Lipschitzness assumption on a^ε, problem (3) also has a unique solution. We emphasize that [14, Theorem 3.6] holds under weaker assumptions than assumed here. In particular equi-continuity with respect to s is assumed for a^ε instead of uniform Lipschitzness.

instrumental to derive the homogenization result. Let $\{a^{\varepsilon}(\cdot,s)\}$ be the subsequence of tensor considered above, then for all fixed real parameter s, the tensor $x \mapsto a^{0}(\cdot,s)$ is the homogenized tensor for the linear problem

$$-\nabla \cdot (a^{\varepsilon}(x,s)\nabla v_{\varepsilon}(x)) = f(x) \ \text{ in } \Omega, \quad v_{\varepsilon}(x) = 0 \ \text{ on } \partial\Omega. \tag{6}$$

If the homogenized tensor $a^{0}(x,s)$ is locally periodic, e.g., $a^{\varepsilon}(x,s) = a(x, x/\varepsilon, s)$ where $a(x,y,s)$ is Y periodic with respect to y, then weak convergence of u^{ε} to the solution of (3) holds for the whole sequence. The homogenized tensor can be characterized in the following way [10]:

$$a^{0}(x,s) = \int_{Y} a(x,y,s)(I + J^{T}_{\chi(x,y,s)})dy, \qquad \text{for } x \in \Omega, s \in \mathbb{R}, \tag{7}$$

where $J_{\chi(x,y,s)}$ is a $d \times d$ matrix with entries $J_{\chi(x,y,s)_{ij}} = (\partial \chi^{i})/(\partial y_{j})$ and $\chi^{i}(x,\cdot,s)$, $i = 1,\ldots,d$ are the unique solutions of the cell problems

$$\int_{Y} a(x,y,s)\nabla_{y}\chi^{i}(x,y,s) \cdot \nabla w(y)dy = -\int_{Y} a(x,y,s)\mathbf{e}_{i} \cdot \nabla w(y)dy, \quad \forall w \in W^{1}_{per}(Y), \tag{8}$$

where $\mathbf{e}_{i}$, $i = 1,\ldots,d$ is the canonical basis of $\mathbb{R}^{d}$.

2.1 Multiscale method

We define here the homogenization method based on the framework of the HMM [23]. The numerical method is based on a macroscopic FEM defined on QF and microscopic FEMs recovering the missing macroscopic tensor at the macroscopic quadrature points. While the macroscopic FEM will be nonlinear, the microscopic FEMs are modeled as linear problems.

2.1.1 Macro finite element space.

Let $\mathcal{T}_{H}$ be a triangulation of Ω in simplicial or quadrilateral elements K of diameter H_{K} and denote $H = \max_{K \in \mathcal{T}_{H}} H_{K}$. We assume that the family of triangulations $\{\mathcal{T}_{H}\}$ is conformal and shape regular and that it satisfies the inverse assumption

$$\frac{H}{H_{K}} \leq C \ \text{ for all } K \in \mathcal{T}_{H} \text{ and all } \mathcal{T}_{H}. \tag{9}$$

Notice that (9) is often assumed for the analysis of FEM for non-linear problems, see for instance [33, 27, 34, 36] in the context of one-scale problems and [24, 16] for multi-scale problems. In our analysis, the assumption (9) is used only in the proof of an L^{2} estimate (see Lemma 4.2 in Sect. 4.1) and for the uniqueness of the numerical solution (Sect. 4.3).

For each partition $\mathcal{T}_{H}$, we define a FE space

$$S^{\ell}_{0}(\Omega,\mathcal{T}_{H}) = \{v^{H} \in H^{1}_{0}(\Omega);\ v^{H}|_{K} \in \mathcal{R}^{\ell}(K),\ \forall K \in \mathcal{T}_{H}\}, \tag{10}$$

where $\mathcal{R}^{\ell}(K)$ is the space $\mathcal{P}^{\ell}(K)$ of polynomials on K of total degree at most ℓ if K is a simplicial FE, or the space $\mathcal{Q}^{\ell}(K)$ of polynomials on K of degree at most ℓ in each variables if K is a quadrilateral FE. We call $\mathcal{T}_{H}$ the macro partition, $K \in \mathcal{T}_{H}$ being a macro element, and $S^{\ell}_{0}(\Omega,\mathcal{T}_{H})$ is called the macro FE space. By macro partition, we mean that H is allowed to be much larger than ε and, in particular, $H < \varepsilon$ is not required for convergence.

2.1.2 Quadrature formula.

For each element K of the of the macro partition we consider a C^1-diffeomorphism F_K such that $K = F_K(\hat{K})$, where $\hat{K}$ is the reference element (of simplicial or quadrilateral type). For a given quadrature formula $\{\hat{x}_j, \hat{\omega}_j\}_{j=1}^J$ on $\hat{K}$, the quadrature weights and integration points on $K \in \mathcal{T}_H$ are then given by $\omega_{K_j} = \hat{\omega}_j |\det(\partial F_K)|$, $x_{K_j} = F_K(\hat{x}_j)$, $j = 1, \ldots, J$. We make the following assumptions on the quadrature formulas, which are standard assumptions also for linear elliptic problems [19, Sect. 29]:

(**Q1**) $\hat{\omega}_j > 0,\ j = 1, \ldots, J, \quad \sum_{j=1}^J \hat{\omega}_j |\nabla \hat{p}(\hat{x}_j)|^2 \geq \hat{\lambda} \|\nabla \hat{p}\|^2_{L^2(\hat{K})},\ \forall \hat{p}(\hat{x}) \in \mathcal{R}^\ell(\hat{K})$, where $\hat{\lambda} > 0$;

(**Q2**) $\int_{\hat{K}} \hat{p}(x)dx = \sum_{j=1}^J \hat{\omega}_j \hat{p}(\hat{x}_j),\ \forall \hat{p}(\hat{x}) \in \mathcal{R}^\sigma(\hat{K})$, where $\sigma = \max(2\ell-2, \ell)$ if $\hat{K}$ is a simplicial FE, or $\sigma = \max(2\ell - 1, \ell + 1)$ if $\hat{K}$ is a rectangular FE.

2.1.3 Micro finite elements method.

For each macro element $K \in \mathcal{T}_H$ and each integration point $x_{K_j} \in K$, $j = 1, \ldots, J$, we define the sampling domains

$$K_{\delta_j} = x_{K_j} + \delta I, \quad I = (-1/2, 1/2)^d \quad (\delta \geq \varepsilon).$$

We consider a conformal and shape regular (micro) partition $\mathcal{T}_h$ of each sampling domain K_{δ_j} in simplicial or quadrilateral elements Q of diameter h_Q and denote $h = \max_{Q \in \mathcal{T}_H} h_Q$. Usually, the size of δ scales with ε, which implies that the complexity of the FEM presented below remains unchanged as $\varepsilon \to 0$. We then define a micro FE space

$$S^q(K_{\delta_j}, \mathcal{T}_h) = \{z^h \in W(K_{\delta_j});\ z^h|_Q, \in \mathcal{R}^q(Q),\ Q \in \mathcal{T}_h\}, \tag{11}$$

where $W(K_{\delta_j})$ is either the Sobolev space

$$W(K_{\delta_j}) = W^1_{per}(K_{\delta_j}) = \{z \in H^1_{per}(K_{\delta_j});\ \int_{K_{\delta_j}} z dx = 0\} \tag{12}$$

for a periodic coupling or

$$W(K_{\delta_j}) = H^1_0(K_{\delta_j}) \tag{13}$$

for a coupling through Dirichlet boundary conditions. The choice of the Sobolev space $W(K_{\delta_j})$ sets the coupling condition between macro and micro solvers. It has important consequences in the numerical accuracy of the method as will be discussed in Section 4.2. The micro FEM problems on each micro domain K_{δ_j} is defined as follows. Let $w^H \in S^\ell_0(\Omega, \mathcal{T}_H)$ and consider its linearization

$$w^H_{lin,j}(x) = w^H(x_{K_j}) + (x - x_{K_j}) \cdot \nabla w^H(x_{K_j}) \tag{14}$$

at the integration point x_{K_j}. For all real parameter s, we define a micro FE function $w^{h,s}_{K_j}$ such that $(w^{h,s}_{K_j} - w^H_{lin,j}) \in S^q(K_{\delta_j}, \mathcal{T}_h)$ and

$$\int_{K_{\delta_j}} a^\varepsilon(x, s) \nabla w^{h,s}_{K_j}(x) \cdot \nabla z^h(x) dx = 0 \quad \forall z^h \in S^q(K_{\delta_j}, \mathcal{T}_h). \tag{15}$$

2.1.4 Finite element heterogeneous multiscale method (FE-HMM).

We have now all the ingredients to define our multiscale method. The method, essentially similar to the method proposed in [24] reads as follows[4]. Find $u^H \in S_0^\ell(\Omega, \mathcal{T}_H)$ such that

$$B_H(u^H; u^H, w^H) = F_H(w^H), \qquad \forall w^H \in S_0^\ell(\Omega, \mathcal{T}_H), \tag{16}$$

where

$$B_H(u^H; v^H, w^H) := \sum_{K\in\mathcal{T}_H} \sum_{j=1}^{J} \frac{\omega_{K_j}}{|K_{\delta_j}|} \int_{K_{\delta_j}} a^\varepsilon(x, u^H(x_{K_j})) \nabla v_{K_j}^{h,u^H(x_{K_j})}(x) \cdot \nabla w_{K_j}^{h,u^H(x_{K_j})}(x) dx, \tag{17}$$

and the linear form F_H on $S_0^\ell(\Omega, \mathcal{T}_H)$ is an approximation of $F(w) = \int_\Omega f(x)w(x)dx$, obtained for example by using quadrature formulas. Here, $w_{K_j}^{h,u^H(x_{K_j})}$ denotes the solution of the micro problem (15) with parameter $s = u^H(x_{K_j})$ (and similarly for $v_{K_j}^{h,u^H(x_{K_j})}$).

Remark 2.1 *Provided that we use for F_H a QF satisfying (**Q2**), for $f \in W^{\ell,p}(\Omega)$ with $\ell > d/p$ and $1 \le p \le \infty$, we have[5] [20, Thm. 4]*

$$|F_H(w^H) - F(w^H)| \le CH^\ell \|w^H\|_{H^1(\Omega)}, \ \forall w^H \in S_0^\ell(\Omega, \mathcal{T}_H). \tag{18}$$

If in addition $f \in W^{\ell+1,p}(\Omega)$, then [20, Thm. 5]

$$|F_H(w^H) - F(w^H)| \le CH^{\ell+1} \Big(\sum_{K\in\mathcal{T}_H} \|w^H\|^2_{H^2(K)} \Big)^{1/2}, \ \forall w^H \in S_0^\ell(\Omega, \mathcal{T}_H). \tag{19}$$

The above constants C depend on $\|f\|_{W^{\ell,p}(\Omega)}$ and $\|f\|_{W^{\ell+1,p}(\Omega)}$ respectively, but they are independent of H.

If we assume a locally periodic tensor, i.e. $a^\varepsilon(x, s) = a(x, x/\varepsilon, s)$, Y-periodic with respect to the second variable $y \in Y = (0, 1)^d$, we shall consider the slightly modified bilinear form

$$\widetilde{B}_H(u^H; v^H, w^H) := \sum_{K\in\mathcal{T}_H} \sum_{j=1}^{J} \frac{\omega_{K_j}}{|K_{\delta_j}|} \int_{K_{\delta_j}} a(x_{K_j}, \frac{x}{\varepsilon}, u^H(x_{K_j})) \nabla v_{K_j}^{h,u^H(x_{K_j})}(x) \cdot \nabla w_{K_j}^{h,u^H(x_{K_j})}(x) dx, \tag{20}$$

where $w_{K_j}^{h,u^H(x_{K_j})}$ is the solution of the micro problem (15) with tensor $a(x_{K_j}, x/\varepsilon, u^H(x_{K_j}))$ (and similarly for $v_{K_j}^{h,u^H(x_{K_j})}$), where compared to (17), the tensor $a(x, y, s)$ is collocated in the slow variable x at the quadrature point x_{K_j}.

We shall discuss now the existence of a solution of (16). We first recall here a result for the analysis of the FE-HMM, shown in [1], [24] in the context of linear problems (see [3, Sect. 3.3.1] for details). The proof is similar in the nonlinear case and is thus omitted.

[4] In [24] (17) is based on exact micro functions v_{K_j}, w_{K_j} instead of the FE micro functions $v_{K_j}^{h,s}, w_{K_j}^{h,s}$ and the micro-problems are nonlinear (see [24, equs. (5.3)-(5.4)]).

[5] Notice that the assumption (**Q1**) is not needed for the quadrature formula in F_H.

Lemma 2.2 *Assume that* (**Q1**) *holds and that the tensor* a^{ε} *satisfies* (4),(5). *Then the bilinear form* $B_H(z^H;\cdot,\cdot)$, $z^H \in S_0^{\ell}(\Omega,\mathcal{T}_H)$ *is uniformly elliptic and bounded. Precisely, there exist two constants again denoted* $\lambda, \Lambda_0 > 0$ *such that*

$$\lambda \|v^H\|^2_{H^1(\Omega)} \leq B_H(z^H; v^H, v^H), \quad |B_H(z^H; v^H, w^H)| \leq \Lambda_0 \|v^H\|_{H^1(\Omega)} \|w^H\|_{H^1(\Omega)}, \tag{21}$$

for all $z^H, v^H, w^H \in S_0^{\ell}(\Omega,\mathcal{T}_H)$. *Similar formulas also hold for the modified bilinear form* $\widetilde{B}_H(z^H;\cdot,\cdot)$ *defined in* (20).

Notice at this stage that in Lemma 2.2 no structure assumption (as for example local periodicity) is required for the tensor a^{ε}.

Since the micro problems (15) are linear with a uniformly bounded and coercive tensor (5), their solutions $w_{K_j}^{h,s} \in S^q(K_{\delta_j},\mathcal{T}_h)$ are always uniquely defined, in particular there is no restriction on the mesh size h. The macro solution u^H of the FE-HMM is solution of the nonlinear system (16) and the existence of a solution u^H of (16) follows from a classical fixed point argument.

Theorem 2.3 *Assume that the bilinear form* $B_H(z^H;\cdot,\cdot)$, $z^H \in S_0^{\ell}(\Omega,\mathcal{T}_H)$, *defined in* (17) *is uniformly elliptic and bounded* (21), *that it depends continuously on* z^H, *and that* $f \in W^{\ell,p}(\Omega)$ *with* $\ell p > d$. *Then, for all* $H, h > 0$, *the nonlinear problem* (16) *possesses at least one solution* $u^H \in S_0^{\ell}(\Omega,\mathcal{T}_H)$. *A solution* u^H *is uniformly bounded in* $H_0^1(\Omega)$, *i.e.*

$$\|u^H\|_{H^1(\Omega)} \leq C \|f\|_{W^{\ell,p}(\Omega)}, \tag{22}$$

where C *is independent of* H.

The proof of Theorem 2.3 follows standard argument ([21], see also [15]). It relies on the Brouwer fixed point theorem applied to the nonlinear map $S_H : S_0^{\ell}(\Omega,\mathcal{T}_H) \to S_0^{\ell}(\Omega,\mathcal{T}_H)$ defined by

$$B_H(z^H; S_H z^H, w^H) = F_H(w^H), \ \forall w^H \in S_0^{\ell}(\Omega,\mathcal{T}_H). \tag{23}$$

In contrast, the proof of the uniqueness of a solution u^H is non trivial and is discussed in Theorem 3.3.

2.2 Reformulation of the FE-HMM

Similarly to the case of linear multiscale problems, the FE-HMM can be reformulated as a standard FEM with numerical integration applied to a modified macro problem. We emphasize that this reformulation is not used for the numerical implementation, but is very useful for the analysis of the method.

A straightforward computation shows that for all scalar s, the solution $w_{K_j}^{h,s}$ of the linear cell problem (15) is given by

$$w_{K_j}^{h,s}(x) = w_{lin,j}^H(x) + \sum_{i=1}^{d} \psi_{K_j}^{i,h,s}(x) \frac{\partial v_{lin,j}^H}{\partial x_i}, \qquad \text{for } x \in K_{\delta_j}, \tag{24}$$

where $\psi_{K_j}^{i,h,s}$, $i = 1, \ldots, d$ are the solutions of the following auxiliary problems. Let $\mathbf{e}_i$, $i = 1 \ldots d$ denote the canonical basis of $\mathbb{R}^d$. For each scalar s and for each $\mathbf{e}_i$ we consider the problem: find $\psi_{K_j}^{i,h,s} \in S^q(K_{\delta_j},\mathcal{T}_h)$ such that

$$\int_{K_{\delta_j}} a^{\varepsilon}(x,s) \nabla \psi_{K_j}^{i,h,s}(x) \cdot \nabla z^h(x) dx = -\int_{K_{\delta_j}} a^{\varepsilon}(x,s) \mathbf{e}_i \cdot \nabla z^h(x) dx, \ \forall z^h \in S^q(K_{\delta_j},\mathcal{T}_h), \tag{25}$$

where $S^q(K_{\delta_j}, \mathcal{T}_h)$ is defined in (11) with either periodic or Dirichlet boundary conditions.

We also consider for the analysis the following problems (26), (28), which are analogue to (15),(25), but without FEM discretization (i.e. with test functions in the space $W(K_{\delta_j})$ defined in (12) or (13)): find $w^s_{K_j}$ such that $(w^s_{K_j} - w^H_{lin,j}) \in W(K_{\delta_j})$ and

$$\int_{K_{\delta_j}} a^\varepsilon(x,s)\nabla w^s_{K_j}(x) \cdot \nabla z(x) dx = 0 \quad \forall z \in W(K_{\delta_j}). \tag{26}$$

Similarly to (24), it can be checked that the unique solution of problem (26) is given by

$$w^s_{K_j}(x) = w^H_{lin,j}(x) + \sum_{i=1}^d \psi^{i,s}_{K_j}(x)\frac{\partial v^H_{lin,j}}{\partial x_i}, \tag{27}$$

where for each scalar s and for each $\mathbf{e}_i$, $\psi^{i,s}_{K_j}$ are the solutions of the following problem: find $\psi^{i,s}_{K_j} \in W(K_{\delta_j})$ such that

$$\int_{K_{\delta_j}} a^\varepsilon(x,s)\nabla \psi^{i,s}_{K_j}(x) \cdot \nabla z(x) dx = -\int_{K_{\delta_j}} a^\varepsilon(x,s)\mathbf{e}_i \cdot \nabla z(x) dx, \ \forall z \in W(K_{\delta_j}). \tag{28}$$

Consider for all scalar s the two tensors

$$a^0_{K_j}(s) := \frac{1}{|K_{\delta_j}|}\int_{K_{\delta_j}} a^\varepsilon(x,s)\left(I + J^T_{\psi^{h,s}_{K_j}(x)}\right)dx, \tag{29}$$

$$\bar{a}^0_{K_j}(s) := \frac{1}{|K_{\delta_j}|}\int_{K_{\delta_j}} a^\varepsilon(x,s)\left(I + J^T_{\psi^{s}_{K_j}(x)}\right)dx, \tag{30}$$

where $J_{\psi^s_{K_j}(x)}$ and $J_{\psi^{h,s}_{K_j}(x)}$ are $d \times d$ matrices with entries $\left(J_{\psi^s_{K_j}(x)}\right)_{i\ell} = (\partial\psi^{i,s}_{K_j})/(\partial x_\ell)$ and $\left(J_{\psi^{h,s}_{K_j}(x)}\right)_{i\ell} = (\partial\psi^{i,h,s}_{K_j})/(\partial x_\ell)$, respectively. The Lemma 2.4 below has been proved in [6],[2] in the context of linear elliptic problems and is a straightforward consequence of (24),(27). It permits to interpret the FE-HMM (16)-(17) as a standard FEM applied with a modified tensor.

Lemma 2.4 *Assume that the tensors a^0, a^ε are continuous on $\overline{\Omega} \times \mathbb{R}$ and satisfy* (5). *For all $v^H, w^H \in S^\ell_0(\Omega, \mathcal{T}_H)$, all sampling domains K_{δ_j} centered at a quadrature node x_{K_j} of a macro element $K \in \mathcal{T}_H$ and all scalar s, the following identities hold:*

$$\begin{aligned}
\frac{1}{|K_{\delta_j}|}\int_{K_{\delta_j}} a^\varepsilon(x,s)\nabla v^{h,s}_{K_j} \cdot \nabla w^{h,s}_{K_j} dx &= a^0_{K_j}(s)\nabla v^H(x_{K_j}) \cdot \nabla w^H(x_{K_j}),\\
\frac{1}{|K_{\delta_j}|}\int_{K_{\delta_j}} a^\varepsilon(x,s)\nabla v^{s}_{K_j} \cdot \nabla w^{s}_{K_j} dx &= \overline{a}^0_{K_j}(s)\nabla v^H(x_{K_j}) \cdot \nabla w^H(x_{K_j}),
\end{aligned}$$

where $v^{h,s}_{K_j}$, $v^s_{K_j}$ are the solutions of (15), (26), *respectively, and the tensors $a^0_{K_j}(s)$, $\overline{a}^0_{K_j}(s)$ are defined in* (29), (30). *Similar formulas also hold for the terms in the right-hand side of* (20), *with $a^\varepsilon(x,s)$ replaced by $a(x_{K_j}, x/\varepsilon, s)$ in the above two identities and in* (26), (28), (15), (25), (29).

As a consequence of the above lemma the form B_H in (17) can be rewritten as

$$B_H(u^H; v^H, w^H) = \sum_{K\in\mathcal{T}_H}\sum_{j=1}^J \omega_{K_j} a^0_{K_j}(u^H(x_{K_j}))\nabla v^H(x_{K_j}) \cdot \nabla w^H(x_{K_j}).$$

3 Main results

We shall state in this section the main results of this paper. Given a solution u^H of (16) the aim is to estimates the errors $\|u_0 - u^H\|_{H^1(\Omega)}$ and $\|u_0 - u^H\|_{L^2(\Omega)}$, where u_0 is the unique solution of the homogenized problem (3) and to prove the uniqueness of a numerical solution u^H.

For the analysis of the FE-HMM, we shall consider quantity

$$r_{HMM} := \sup_{K \in \mathcal{T}_H, x_{K_j} \in K, s \in \mathbb{R}} \|a^0(x_{K_j}, s) - a^0_{K_j}(s)\|_F, \tag{31}$$

where a^0 is the homogenized tensor in (3) and $a^0_{K_j}$ is the tensor defined in (29).

Recall from Theorem 2.3 that the FE-HMM solution in (16) exists for all $H > 0$ and all $h > 0$. The first results give optimal H^1 and L^2 error estimates, as functions of the macro mesh size H, for the FE-HMM without specific structure assumption on the nature of the small scales (e.g. as periodicity or stationarity for random problems).

Theorem 3.1 *Consider u_0 the solution of problem* (3)*. Let $\ell \geq 1$. Let $\mu = 0$ or 1. Assume (**Q1**), (**Q2**),* (9), (18)*, and*

$$\begin{aligned} & u_0 \in H^{\ell+1}(\Omega) \cap W^{1,\infty}(\Omega), & \\ & a^0_{mn} \in W^{\ell+\mu,\infty}(\Omega \times \mathbb{R}), & \forall m, n = 1 \ldots d. \end{aligned}$$

In addition to (4), (5)*, assume that $\partial_u a^0_{mn} \in W^{1,\infty}(\Omega \times \mathbb{R})$, and that the coefficients $a^0_{mn}(x, s)$ are twice differentiable with respect to s, with the first and second order derivatives continuous and bounded on $\overline{\Omega} \times \mathbb{R}$, for all $m, n = 1 \ldots d$.*

Then, there exist $r_0 > 0$ and $H_0 > 0$ such that, provided

$$H \leq H_0 \quad \textit{and} \quad r_{HMM} \leq r_0, \tag{32}$$

any solution u^H of (16) *satisfies*

$$\|u_0 - u^H\|_{H^1(\Omega)} \leq C(H^\ell + r_{HMM}) \quad \textit{if } \mu = 0, 1, \tag{33}$$

$$\|u_0 - u^H\|_{L^2(\Omega)} \leq C(H^{\ell+1} + r_{HMM}) \quad \textit{if } \mu = 1 \textit{ and } (19) \textit{ holds.} \tag{34}$$

Here, the constants C are independent of H, h, r_{HMM}.

The proof of Theorem 3.1 is postponed to Sect. 4.1. In contrast to previous results [24, Thm 5.4], Theorem 3.1 is also valid for $d = 3$ and arbitrary high order simplicial and quadrilateral macro FEs.

Remark 3.2 *We emphasize that the constants H_0 and r_0 in Theorem 3.1 are independent of H, h, ε, δ. This makes possible a fully discrete error analysis, where the micro FE discretization errors are also taken into account by essentially re-using results obtained for linear problems (see Section 4.2). In contrast the a priori estimates of [24, Thm 5.4] rely, besides a smallness assumption on H and e_{HMM} (the semi-discrete version of r_{HMM}) [24, Lemma 3.3], on a kind of continuity of e_{HMM} (see [24, (5.13)]) which is proved to hold under some assumptions (see [24, Lemma 5.9]). This latter results seems to be non trivial to generalize in order to take also into account the micro FEM errors.*

For the uniqueness result, we shall consider the quantity

$$r'_{HMM} := \sup_{K\in\mathcal{T}_H, x_{K_j}\in K, s\in\mathbb{R}} \left\| \frac{d}{ds}\left(a^0(x_{K_j},s) - a^0_{K_j}(s)\right)\right\|_F . \tag{35}$$

For r'_{HMM} to be well defined and for the subsequent analysis, we need the assumption

$$s\in\mathbb{R}\mapsto a^\varepsilon(\cdot,s)\in (W^{1,\infty}(\Omega))^{d\times d} \text{ is of class } C^2 \text{ and } |\partial_u^k a^\varepsilon(\cdot,s)|_{W^{1,\infty}(\Omega)}\leq C\varepsilon^{-1},\ k\leq 2, \tag{36}$$

where C is independent of s and ε.

Theorem 3.3 *Assume that the hypothesis of Theorem 3.1 and* (36) *hold. Then, there exists* H_0, R_0, *such that if*

$$r_{HMM}\leq H\leq H_0 \textit{ and } r'_{HMM}\leq R_0,$$

then the solution u^H *of* (16) *is unique.*

If the oscillating coefficients are smooth and locally periodic coefficients (see (**H1**) and (**H2**) below), then the assumptions for the uniqueness result can be stated solely in terms of the size of the macro and micro meshes.

Corollary 3.4 *In addition to the hypothesis of Theorem 3.3, assume (***H1***) and (***H2***) as defined below. Assume* $W(K_{\delta_j}) = W^1_{per}(K_{\delta_j})$ *(periodic coupling conditions),* $\delta/\varepsilon\in\mathbb{N}$ *and that* (20) *is used for the solution* u^H *of* (16)*. Then, there exists a positive constant* H_0 *such that for all*

$$(h/\varepsilon)^{2q}\leq H\leq H_0,$$

the solution u^H *of* (16) *is unique.*

Remark 3.5 *In Corollary 3.4, if we use the form* (17) *for the solution* u^H *of* (16)*, to obtain the uniqueness of of* u^H*, we need to assume in addition that* δ *is small enough (*$\varepsilon\leq\delta\leq CH$*).*

We next describe our fully discrete a priori error estimates. For that, let us split r_{HMM} into

$$\begin{aligned} r_{HMM} \quad\leq\quad & \underbrace{\sup_{K\in\mathcal{T}_H, x_{K_j}\in K, s\in\mathbb{R}} \|a^0(x_{K_j},s)-\bar{a}^0_{K_j}(s))\|_F}_{r_{MOD}} \\ + \quad & \underbrace{\sup_{K\in\mathcal{T}_H, x_{K_j}\in K, s\in\mathbb{R}} \|\bar{a}^0_{K_j}(s)-a^0_{K_j}(s))\|_F,}_{r_{MIC}} \end{aligned} \tag{37}$$

where $\bar{a}^0_{K_j}$ is the tensor defined in (30). Here r_{MIC} stands for the micro error (error due to the micro FEM) and r_{MOD} for the modeling or resonance error. The first result gives explicit convergence rates in terms of the micro discretization. Some additional regularity and growth condition of the small scale tensor a^ε is needed in order to have appropriate regularity of the cell function $\psi^{i,s}_{K_j}$ defined in (28) and involved in the definition of $\bar{a}^0_{K_j}$. We note that if $a^\varepsilon_{ij}|_K\in W^{1,\infty}(K)\ \forall K\in\mathcal{T}_H$ and $|a^\varepsilon_{ij}|_{W^{1,\infty}(K)}\leq C\varepsilon^{-1}$, then classical H^2 regularity results ([31, Chap. 2.6]) imply that $|\psi^{i,s}_{K_j}|_{H^2(K_{\delta_j})}\leq C\varepsilon^{-1}\sqrt{|K_{\delta_j}|}$ when Dirichlet boundary conditions (13) are used. If a^ε_{ij} is locally periodic, we can also use periodic boundary conditions (12) and similar bounds for $\psi^{i,s}_{K_j}$ in terms of ε can be obtained, provided that we collocate the slow variables in each sampling domain. In that case, higher regularity for $\psi^{i,s}_{K_j}$ can be shown, provided $a^\varepsilon(\cdot, s)$ is smooth enough (see e.g., [13, Chap. 3]). As it is more convenient to state the regularity conditions directly for the function $\psi^{i,s}_{K_j}$, we assume

(**H1**) Given $q \in \mathbb{N}$, the cell functions $\psi^{i,s}_{K_j}$ defined in (28) satisfy

$$|\psi^{i,s}_{K_j}|_{H^{q+1}(K_{\delta_j})} \leq C\varepsilon^{-q}\sqrt{|K_{\delta_j}|},$$

with C independent of ε, the quadrature point x_{K_j}, the domain K_{δ_j}, and the parameter s for all $i = 1 \dots d$. The same assumption also holds with the tensor a^ε replaced by $(a^\varepsilon)^T$ in (28).

Theorem 3.6 *In addition to the assumptions of Theorem 3.1, assume (***H1***). Then,*

$$\begin{aligned}\|u_0 - u^H\|_{H^1(\Omega)} &\leq C(H^\ell + \left(\frac{h}{\varepsilon}\right)^{2q} + r_{MOD}),\\ \|u_0 - u^H\|_{L^2(\Omega)} &\leq C(H^{\ell+1} + \left(\frac{h}{\varepsilon}\right)^{2q} + r_{MOD})\end{aligned}$$

where we also assume (19) *for the above L^2 estimate. Here, C is independent of $H, h, \varepsilon, \delta$.*

To estimate further the modeling error r_{MOD} defined in (37), we need more structure assumptions on a^ε. Here we assume local periodicity as encoded in the following assumption.

(**H2**) for all $m, n = 1, \dots, d$, we assume $a^\varepsilon_{mn}(x, s) = a_{mn}(x, x/\varepsilon, s)$, where $a_{mn}(x, y, s)$ is y-periodic in Y, and the map $(x, s) \mapsto a_{mn}(x, \cdot, s)$ is Lipschitz continuous and bounded from $\overline{\Omega} \times \mathbb{R}$ into $W^{1,\infty}_{per}(Y)$.

Theorem 3.7 *In addition to the assumptions of Theorem 3.1 assume (***H1***) and (***H2***). Then, for $\mu = 0$ or 1,*

$$\|u_0 - u^H\|_{H^{1-\mu}(\Omega)} \leq \begin{cases} C(H^{\ell+\mu} + (\frac{h}{\varepsilon})^{2q} + \delta), & \text{if } W(K_{\delta_j}) = W^1_{per}(K_{\delta_j}) \text{ and } \frac{\delta}{\varepsilon} \in \mathbb{N},\\ C(H^{\ell+\mu} + (\frac{h}{\varepsilon})^{2q}), & \text{if } W(K_{\delta_j}) = W^1_{per}(K_{\delta_j}) \text{ and } \frac{\delta}{\varepsilon} \in \mathbb{N}, \text{ and the tensor is collocated at } x_{K_j} \text{ (i.e. (20) is used)}\\ C(H^{\ell+\mu} + (\frac{h}{\varepsilon})^{2q} + \delta + \frac{\varepsilon}{\delta}), & \text{if } W(K_{\delta_j}) = H^1_0(K_{\delta_j}) \ (\delta > \varepsilon), \end{cases} \tag{38}$$

where for $\mu = 1$ we also assume (19) *and we use the notation $H^0(\Omega) = L^2(\Omega)$. The constants C are independent of $H, h, \varepsilon, \delta$.*

4 Proof of the main results

We first show the a priori convergence rates at the level of the macro error (Sect. 4.1) before estimating the micro and modeling errors (Sect. 4.2). These estimates are useful to prove the uniqueness of the solution (Sect. 4.3).

4.1 Explicit convergence rates for the macro error

In this section, we give the proofs of Theorem 3.1. Consider for $z^H, v^H, w^H \in S_0^\ell(\Omega, \mathcal{T}_H)$,

$$A_H(z^H; v^H, w^H) := \sum_{K\in\mathcal{T}_H} \sum_{j=1}^{J} \omega_{K,j} a^0(x_{K_j}, z^H(x_{K_j}))\nabla v^H(x_{K_j}) \cdot \nabla w^H(x_{K_j}), \tag{39}$$

where $a^0(x, s)$ is the tensor defined in (3) and let $\widetilde{u}_0^H$ be a solution of

$$A_H(\widetilde{u}_0^H; \widetilde{u}_0^H, w^H) = F_H(w^H), \qquad \forall w^H \in S_0^\ell(\Omega, \mathcal{T}_H). \tag{40}$$

The problem (40) is a standard FEM with numerical quadrature for the problem (3). Convergence rates for this nonlinear problem are not trivial to derive and have recently been obtained in [7]. For the proof of Theorem 3.1, we first need to generalize several results of [7]. For that, consider

$$Q_H(z^H) := \sup_{w^H \in S_0^\ell(\Omega,\mathcal{T}_H)} \frac{|A_H(z^H, z^H, w^H) - F_H(w^H)|}{\|w^H\|_{H^1(\Omega)}}, \quad \forall z^H \in S_0^\ell(\Omega, \mathcal{T}_H). \tag{41}$$

We observe that $Q_H(\widetilde{u}_0^H) = 0$. The three lemmas below have been obtained in [7] for the special case $z^H = \widetilde{u}_0^H$. Allowing for an arbitrary function $z^H \in S_0^\ell(\Omega, \mathcal{T}_H)$ leads to introducing the additional term $Q_H(z^H)$. The proofs of these more general results remain, however, nearly identical to [7] and are therefore omitted.

Lemma 4.1 *If the hypotheses of Theorem 3.1 are satisfied, then*

$$\|u_0 - z^H\|_{H^1(\Omega)} \le C(H^\ell + Q_H(z^H) + \|u_0 - z^H\|_{L^2(\Omega)}), \tag{42}$$

for all $z^H \in S_0^\ell(\Omega, \mathcal{T}_H)$, where C is independent of H and z^H.

Proof. Follows the lines of the proof of [7, Lemma 4.1]. □

Lemma 4.2 *Assume the hypotheses of Theorem 3.1 are satisfied with $\mu = 0$ or 1. Then, for all $z^H \in S_0^\ell(\Omega, \mathcal{T}_H)$,*

$$\|u_0 - z^H\|_{L^2(\Omega)} \le C(H^{\ell+\mu} + Q_H(z^H) + \|u_0 - z^H\|^2_{H^1(\Omega)}), \tag{43}$$

where C is independent of H and z^H.

Proof. Follows the lines of the proof of [7, Lemma 4.3]. □

Lemma 4.3 *Assume the hypotheses of Theorem 3.1 are satisfied. Consider a sequence $\{z^H\}$ bounded in $H^1(\Omega)$ as $H \to 0$ and satisfying $Q_H(z^H) \to 0$ for $H \to 0$. Then,*

$$\|u_0 - z^H\|_{L^2(\Omega)} \to 0 \quad \text{for } H \to 0.$$

Proof. Follows the lines of the proof of [7, Theorem 2.6]. □

We next notice that $Q_H(z^H)$ can be bounded in terms of r_{HMM} defined in (31).

Lemma 4.4 *Assume that the tensors a^0, a^ε are continuous on $\overline{\Omega} \times \mathbb{R}$ and satisfy* (5). *Then*

$$Q_H(z^H) \leq C r_{HMM} \|z^H\|_{H^1(\Omega)} \quad \forall z^H \in S_0^\ell(\Omega, \mathcal{T}_H), \tag{44}$$

where the constant C is independent of H, h, δ.

Proof. Using Lemma 2.4 and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&|A_H(z^H; z^H, w^H) - B_H(z^H; z^H, w^H)| \\
&= \left| \sum_{K \in \mathcal{T}_H} \sum_{j=1}^{J} \omega_{K,j} (a^0(x_{K_j}, z^H(x_{K_j})) - a^0_{K_j}(z^H(x_{K_j}))) \nabla z^H(x_{K_j}) \cdot \nabla w^H(x_{K_j}) \right| \\
&\leq C \sup_{K \in \mathcal{T}_H, x_{K_j} \in K, s \in \mathbb{R}} \|a^0(x_{K_j}, s)) - a^0_{K_j}(s))\|_F \|z^H\|_{H^1(\Omega)} \|w^H\|_{H^1(\Omega)}
\end{aligned}$$

where we used the estimate

$$\sum_{K \in \mathcal{T}_H} \sum_{j=1}^{J} \omega_{K,j} \|v^H(x_{K_j})\|^2 \leq C \|v^H\|^2_{L^2(\Omega)},$$

with $v^H = z^H$ and $v^H = w^H$, which holds for all piecewise continuous polynomials with respect to the partition $\mathcal{T}_H$, with C independent of H (but depending on the maximum degree of v^H). This concludes the proof. □

Corollary 4.5 *Consider u^H a solution of* (16). *Then $Q_H(u^H) \leq C r_{HMM}$, where $Q_H(u^H)$ is defined in* (41) *and the constant C is independent of H, h, δ.*

Proof. Follows from Lemma 4.4 and the a priori bound (22) on u^H. □

Proof of Theorem 3.1. We apply Lemmas 4.1, 4.2, 4.3 with $z^H = u^H$, the solution of (16). Let $\mu = 0$. This yields, for all H small enough

$$\begin{aligned}
\|u^H - u_0\|_{H^1(\Omega)} &\leq C(H^\ell + r_{HMM} + \|u^H - u_0\|_{L^2(\Omega)}), &(45)\\
\|u^H - u_0\|_{L^2(\Omega)} &\leq C(H^\ell + r_{HMM} + \|u^H - u_0\|^2_{H^1(\Omega)}), &(46)\\
\|u^H - u_0\|_{L^2(\Omega)} &\to 0 \text{ for } (H, r_{HMM}) \to 0, &(47)
\end{aligned}$$

where we recall that $Q_H(u^H) \leq C r_{HMM}$. Substituting (46) into (45), we obtain an inequality of the form

$$\|u^H - u_0\|_{H^1(\Omega)} \leq C(H^\ell + r_{HMM} + \|u^H - u_0\|^2_{H^1(\Omega)})$$

or equivalently

$$(1 - C\|u^H - u_0\|_{H^1(\Omega)}) \|u^H - u_0\|_{H^1(\Omega)} \leq C(H^\ell + r_{HMM}). \tag{48}$$

Using (45) and (47), we have $\|u^H - u_0\|_{H^1(\Omega)} \to 0$ if $(H, r_{HMM}) \to 0$. Thus, there exists H_0 and r_0 such that if $H \leq H_0$ and $r_{HMM} \leq r_0$, then $1 - C\|u^H - u_0\|_{H^1(\Omega)} \geq \nu > 0$ in (48), independently of the choice of the particular solution u^H. This concludes the proof of (33) for H and r_{HMM} small enough. For $\mu = 1$ inequality (46) can be replaced by

$$\|u^H - u_0\|_{L^2(\Omega)} \leq C(H^{\ell+1} + r_{HMM} + \|u^H - u_0\|^2_{H^1(\Omega)}).$$

This inequality together with the H^1 estimate (33) yields (34). □

4.2 Explicit convergence rates for the micro and modeling error

In this section we give the proof of Theorems 3.6 and 3.7. For that, we need to quantify r_{HMM} defined in (31) and involved in Theorem 3.1. In view of the decomposition (37) we shall further estimate r_{MIC} and r_{MOD}. We emphasize that the results in this section can be derived mutatis mutandis from the results for linear elliptic problems (i.e. when the tensor $a(x,s)$ is independent of s).

The following estimate of the micro error r_{MIC} was first presented in [1] for linear elliptic problems, generalized to high order in [3, Lemma 10],[2, Corollary 10] (see also [4]), and to non-symmetric tensors in [22]. We provide here a short proof which will be further useful in the proof of Lemma 4.12.

Lemma 4.6 *Assume that the tensors $a^0(x,s)$, $a^\varepsilon(x,s)$ are continuous on $\overline{\Omega}\times\mathbb{R}$ and uniformly elliptic and bounded* (5) *with respect to s. Assume (***H1***). Then*

$$r_{MIC} \leq C\Big(\frac{h}{\varepsilon}\Big)^{2q},$$

where C is independent of H, h, δ, ε.

Proof. From (29),(30) we deduce

$$\begin{aligned}(\bar{a}^0_{K_j}(s) - a^0_{K_j}(s))_{mn} &= \frac{1}{|K_{\delta_j}|}\int_{K_{\delta_j}} a^\varepsilon(x,s)\left(\nabla\psi^{n,s}_{K_j}(x) - \nabla\psi^{n,h,s}_{K_j}(x)\right)\cdot \mathbf{e}_m dx \\ &= \frac{-1}{|K_{\delta_j}|}\int_{K_{\delta_j}} a^\varepsilon(x,s)\left(\nabla\psi^{n,s}_{K_j}(x) - \nabla\psi^{n,h,s}_{K_j}(x)\right)\cdot \nabla\overline{\psi}^{m,s}_{K_j}(x) dx\end{aligned}$$

where $\overline{\psi}^{n,i}_{K_j}$, $i = 1,\ldots,d$ denote the solutions of (28) with $a^\varepsilon(x,s)$ replaced by $a^\varepsilon(x,s)^T$. Using (25), (28), the above identity remains valid with $\overline{\psi}^{m,s}_{K_j}(x)$ replaced by $\overline{\psi}^{m,s}_{K_j}(x) - z^h$ for all $z^h \in S^q(K_{\delta_j},\mathcal{T}_h)$. We take $z^h = \overline{\psi}^{m,h,s}_{K_j}$ (the solutions of (25) with $a^\varepsilon(x,s)$ replaced by $a^\varepsilon(x,s)^T$), and we obtain

$$(\bar{a}^0_{K_j}(s) - a^0_{K_j}(s))_{mn} = \frac{-1}{|K_{\delta_j}|}\int_{K_{\delta_j}} a^\varepsilon(x,s)\left(\nabla\psi^{n,s}_{K_j} - \nabla\psi^{n,h,s}_{K_j}\right)\cdot (\nabla\overline{\psi}^{m,s}_{K_j} - \nabla\overline{\psi}^{m,h,s}_{K_j})dx \quad (49)$$

The Cauchy-Schwarz inequality then yields

$$|(\bar{a}^0_{K_j}(s) - a^0_{K_j}(s))_{mn}| \leq \frac{C}{|K_{\delta_j}|}\|\nabla\psi^{n,s}_{K_j} - \nabla\psi^{n,h,s}_{K_j}\|_{L^2(K_{\delta_j})}\|\nabla\overline{\psi}^{m,s}_{K_j} - \nabla\overline{\psi}^{m,h,s}_{K_j}\|_{L^2(K_{\delta_j})}.$$

Using the regularity assumption (**H1**) and standard FE results [19, Sect. 17], we have

$$\|\nabla\psi^{n,s}_{K_j} - \nabla\psi^{n,h,s}_{K_j}\|_{L^2(K_{\delta_j})} \leq Ch^q|\nabla\psi^{n,s}_{K_j}|_{H^{q+1}(K_{\delta_j})} \leq C(h/\varepsilon)^q\sqrt{|K_{\delta_j})|},$$

and similar estimates for $\nabla\overline{\psi}^{m,s}_{K_j}$, which yields $|(\bar{a}^0_{K_j}(s) - a^0_{K_j}(s))_{mn}| \leq C(h/\varepsilon)^{2q}$. □

We can further estimate the modeling error if we make the assumption of locally periodic tensors.

The following estimates on the modeling error r_{MOD} were first presented in [24, 22] (for the estimates (52) and (50)) and in [6] (for the estimates (51)), in the context of linear elliptic homogenization problems. Periodic and Dirichlet micro boundary conditions are discussed.

Lemma 4.7 *Assume* (4),(5) *and (***H2***). Consider the homogenized tensor* $a^0(x,s)$ *and the tensor* $\overline{a}^0_{K_j}(s)$ *defined in* (30) *with parameters* $x = x_{K_j}$ *and* $s = u^H(x_{K_j})$*.*

- *If* $W(K_{\delta_j}) = W^1_{per}(K_{\delta_j})$ *and* $\delta/\varepsilon \in \mathbb{N}$ *then*

$$r_{MOD} \leq C\delta. \tag{50}$$

If in addition, the tensor $a^\varepsilon(x,s)$ *is collocated at* $x = x_{K_j}$ *(i.e. using* (17)*) then*

$$r_{MOD} = 0. \tag{51}$$

- *If* $W(K_{\delta_j}) = H^1_0(K_{\delta_j})$ $(\delta > \varepsilon)$*, then*

$$r_{MOD} \leq C(\delta + \frac{\varepsilon}{\delta}). \tag{52}$$

All above constants C *are independent of* H*,* h*,* ε*,* δ*.*

Proof. The estimates (50), (51), (52) are already known in the context of linear problems [24, 6, 22]. Using the characterization (7), they hold mutatis mutandis for our nonlinear tensor. □

4.3 Uniqueness of the FE-HMM solution

The proof of the uniqueness of the FE-HMM solution of problem (16) relies on the convergence of the Newton method used for the computation of a numerical solution. In fact, our results not only show the uniqueness of a solution of (16) (under appropriate assumptions), but also that the iterative method used in practice to compute an actual solution converges.

For given $z^H, v^H, w^H \in S^\ell_0(\Omega, \mathcal{T}_H)$ we consider the Fréchet derivative ∂B_H obtained by differentiating the nonlinear quantity $B_H(z^H, z^H, w^H)$ with respect to z^H

$$\partial B_H(z^H; v^H, w^H) := B_H(z^H; v^H, w^H) + B'_H(z^H, v^H, w^H), \tag{53}$$

where using Lemma 2.4,

$$B'_H(z^H, v^H, w^H) = \sum_{K\in\mathcal{T}_H}\sum_{j=1}^{J} \omega_{K_j}\frac{d}{ds}a^0_{K_j}(s)|_{s=z^H(x_{K_j})}v^H(x_{K_j})\nabla z^H(x_{K_j}) \cdot \nabla w^H(x_{K_j}). \tag{54}$$

The Newton method for approximating a solution u^H of the nonlinear FE-HMM (16) by a sequence $\{u^H_k\}$ reads in weak form

$$\partial B_H(u^H_k; u^H_{k+1} - u^H_k, w^H) = F_H(w^H) - B_H(u^H_k; u^H_k, w^H), \quad \forall w^H \in S^\ell_0(\Omega, \mathcal{T}_H). \tag{55}$$

In order for B'_H to be well defined, we need, in addition to (4),(5) the assumption (**H3**). We also consider

$$A'_H(z^H, v^H, w^H) = \sum_{K\in\mathcal{T}_H} \sum_{j=1}^{J} \omega_{K_j} \frac{d}{ds}\, a^0(x_{K_j}, s)\big|_{s=z^H(x_{K_j})}\, v^H(x_{K_j})\nabla z^H(x_{K_j}) \cdot \nabla w^H(x_{K_j}). \tag{56}$$

and A_H as defined in (39). Then, by replacing in (55) B_H by A_H and ∂B_H by ∂A_H we obtain the Newton method for solving (40) (standard FEM with numerical integration)

$$\partial A_H(z_k^H; z_{k+1}^H - z_k^H, w^H) = F_H(w^H) - A_H(z_k^H; z_k^H, w^H), \quad \forall w^H \in S_0^\ell(\Omega, \mathcal{T}_H), \tag{57}$$

where $\partial A_H(z^H; v^H, w^H) := A_H(z^H; v^H, w^H) + A'_H(z^H, v^H, w^H)$. The convergence of a Newton method of the type of (57) (single scale nonlinear nonmonotone problem) has been studied in [7]. These results have to be adapted for the Newton method (55) applied to the problem (16). We prove in Lemma 4.11 below that the iteration (55) is well defined for all k and that the sequence of solutions of (55) converges to u^H, the solution of (16), provided that the initial guess $u_0^H \in S_0^\ell(\Omega, \mathcal{T}_H)$ is close enough from u^H. This allows to prove Theorem 3.3, i.e., the uniqueness of a solution u^H of (16). The following quantity will be useful

$$\sigma_H := \sup_{v^H \in S_0^\ell(\Omega,\mathcal{T}_H)} \frac{\|v^H\|_{L^\infty(\Omega)}}{\|v^H\|_{H^1(\Omega)}}.$$

Using (9), one can show the standard estimates[6] $\sigma_H \le C(1 + |\log H|)^{1/2}$ for $d = 2$ and $\sigma_H \le CH^{-1/2}$ for $d = 3$, where C is independent of H. We shall also need the following result.

Lemma 4.8 *Assume that the tensors* a^0, a^ε *satisfy* (5),(36). *Then*

$$\sup_{z^H,v^H,w^H\in S_0^\ell(\Omega,\mathcal{T}_H)} \frac{\left|A_H(z^H, v^H, w^H) - B_H(z^H, v^H, w^H)\right|}{\|v^H\|_{H^1(\Omega)}\|w^H\|_{H^1(\Omega)}} \le Cr_{HMM}, \tag{58}$$

$$\sup_{z^H,v^H,w^H\in S_0^\ell(\Omega,\mathcal{T}_H)} \frac{\left|A'_H(z^H, v^H, w^H) - B'_H(z^H, v^H, w^H)\right|}{\|z^H\|_{W^{1,6}(\Omega)}\|v^H\|_{H^1(\Omega)}\|w^H\|_{H^1(\Omega)}} \le Cr'_{HMM}, \tag{59}$$

where r_{HMM} *and* r'_{HMM} *are defined in* (31),(35), *respectively and where the constant* C *is independent of* H, h, δ.

Proof. The proof of (58) follows the lines of Lemma 4.4. The proof of (59) is nearly identical. Indeed, using Lemma 2.4, the quantity $A'_H(z^H, v^H, w^H) - B'_H(z^H, v^H, w^H)$ is equal to

$$\sum_{K\in\mathcal{T}_H} \sum_{j=1}^{J} \omega_{K_j} \left(\frac{d}{ds}\Big|_{s=z^H(x_{K_j})} \left(a^0(x_{K_j}, s) - a^0_{K_j}(s)\right)\right) v^H(x_{K_j})\nabla z^H(x_{K_j}) \cdot \nabla w^H(x_{K_j}).$$

We deduce the result using the Cauchy-Schwarz inequality (similarly to the proof of Lemma 4.4) and the estimate

$$\|v^H \nabla z^H\|_{L^2(\Omega)} \le \|v^H\|_{L^3(\Omega)}\|\nabla z^H\|_{L^6(\Omega)} \le C\|v^H\|_{H^1(\Omega)}\|z^H\|_{W^{1,6}(\Omega)}$$

which is a consequence of the Hölder inequality. □

[6] These two estimates follow from the inverse inequality $\|v^H\|_{L^\infty(\Omega)} \le CH^{-d/p}\|v^H\|_{L^p(\Omega)}$ and $\|v^H\|_{L^p(\Omega)} \le Cp^{1/2}\|v^H\|_{H^1(\Omega)}$ with $p = |\log H|$ for $d = 2$, and $\|v^H\|_{L^6(\Omega)} \le C\|v^H\|_{H^1(\Omega)}$ for $d = 3$.

Lemma 4.9 *Let $\tau > 0$. Under the assumptions of Theorem 3.3, there exist $H_0, \nu, r_0 > 0$ such that if $H \leq H_0$, and $z^H \in S_0^\ell(\Omega, \mathcal{T}_H)$ with*

$$\|z^H\|_{W^{1,6}(\Omega)} \leq \tau, \qquad \sigma_H \|z^H - u_0\|_{H^1(\Omega)} \leq \nu, \quad \textit{and} \quad r_{HMM} + r'_{HMM} \leq r_0$$

where r_{HMM}, r'_{HMM} are defined in (31) *and* (35)*, respectively, then for all linear form G on $S_0^\ell(\Omega, \mathcal{T}_H)$, there exists one and only one solution $v^H \in S_0^\ell(\Omega, \mathcal{T}_H)$ of*

$$\partial B_H(z^H; v^H, w^H) = G(w^H), \quad \forall w^H \in S_0^\ell(\Omega, \mathcal{T}_H).$$

Moreover, v^H satisfies

$$\|v^H\|_{H^1(\Omega)} \leq C \|G\|_{H^{-1}(\Omega)}$$

where we use the notation $\|G\|_{H^{-1}(\Omega)} = \sup_{w^H \in S_0^\ell(\Omega, \mathcal{T}_H)} G(w^H)/\|w^H\|_{H^1(\Omega)}$, and C is a constant independent of H, h and z^H.

Proof. Lemma 4.9 has been proved in [7, Lemma 4.4] for ∂A_H instead of ∂B_H and can be reformulated in terms of the following $\inf - \sup$ inequality: there exist $H_0, \nu > 0$ such that if $H \leq H_0$, $\|z^H\|_{W^{1,6}(\Omega)} \leq \tau$ and $\sigma_H \|z^H - u\|_{H^1(\Omega)} \leq \nu$, then

$$\inf_{v^H \in S_0^\ell(\Omega, \mathcal{T}_H)} \sup_{w^H \in S_0^\ell(\Omega, \mathcal{T}_H)} \frac{\partial A_H(z^H; v^H, w^H)}{\|v^H\|_{H^1(\Omega)} \|w^H\|_{H^1(\Omega)}} \geq K > 0, \tag{60}$$

where K is a constant independent of H and z^H. Using Lemma 4.8 and the inequality $\|z^H\|_{W^{1,6}(\Omega)} \leq \tau$, it follows from (53) that for all $z^H, v^H, w^H \in S_0^\ell(\Omega, \mathcal{T}_H)$,

$$\begin{aligned} \partial B_H(z^H; v^H, w^H) &\geq \partial A_H(z^H; v^H, w^H) - (q_{HMM} + \tau q'_{HMM}) \|v^H\|_{H^1(\Omega)} \|w^H\|_{H^1(\Omega)} \\ &\geq (K - C(r_{HMM} + r'_{HMM}) \|v^H\|_{H^1(\Omega)} \|w^H\|_{H^1(\Omega)}, \end{aligned}$$

where q_{HMM}, q'_{HMM} are the left-hand sides of (58),(59), respectively. We deduce the inf-sup inequality (60) for ∂B_H with $r_{HMM} + r'_{HMM} \leq r_0$ where r_0 is chosen small enough so that $K - Cr_0 > 0$. This concludes the proof. □

In the next lemma we show that $\{u^H\}$ can be bounded in $W^{1,6}(\Omega)$.

Lemma 4.10 *Under the assumptions of Theorem 3.1 and if $r_{HMM} \leq CH$, there exists $\tau > 0$ such that $\|u^H\|_{W^{1,6}(\Omega)} \leq \tau$, where τ is independent of H, h.*

Proof. Using (9) we have the inverse estimate $\|v_H\|_{W^{1,6}(\Omega)} \leq H^{-1} \|v_H\|_{H^1(\Omega)}$ for all $v_H \in S_0^\ell(\Omega, \mathcal{T}_H)$ (see [19, Thm. 17.2]) which yields

$$\begin{aligned} \|u^H\|_{W^{1,6}(\Omega)} &\leq \|u^H - \mathcal{I}_H u_0\|_{W^{1,6}(\Omega)} + \|\mathcal{I}_H u_0\|_{W^{1,6}(\Omega)} \\ &\leq C(H^{-1}(H^\ell + r_{HMM}) + \|u_0\|_{H^2(\Omega)}) \leq \tau, \end{aligned}$$

where $\mathcal{I}_H : C^0(\overline{\Omega}) \to S_0^\ell(\Omega, \mathcal{T}_H)$ denotes the usual nodal interpolant [19, Sect. 12]. □

We can now prove that the Newton method (55) converges at the usual quadratic rate.

Lemma 4.11 *Assume that the hypothesis of Theorem 3.3 hold. Let u^H be a solution of* (16). *Then, there exists $H_0, R_0, \nu > 0$, such that for*

$$r_{HMM} \leq H \leq H_0, \quad r'_{HMM} \leq R_0, \tag{61}$$

and for all $u_0^H \in S_0^\ell(\Omega, \mathcal{T}_H)$ satisfying

$$\sigma_H \|u_0^H - u^H\|_{H^1(\Omega)} \leq \nu \tag{62}$$

the sequence $\{u_k^H\}$ of the Newton method (55) *with initial value u_0^H is well defined, and $e_k = \|u_k^H - u^H\|_{H^1(\Omega)}$ is a decreasing sequence that converges quadratically to* 0 *for $k \to \infty$, i.e.,*

$$e_{k+1} \leq C\sigma_H e_k^2, \tag{63}$$

where C is a constant independent of H, h, k.

The proof is very similar to the one of [21, Theorem 2] (see [7, Theorem 4.5] for details in the context of FEM with numerical quadrature). For completeness we sketch it in the appendix.

We can now prove the claimed uniqueness result.

Proof of Theorem 3.3. Let $u^H, \tilde{u}^H$ be two solutions of (16). We consider the Newton method $\{u_k^H\}$ defined by (55) with the initial guess $u_0^H = \tilde{u}^H$. Using Theorem 3.1, we have $\sigma_H \|\widetilde{u}^H - u^H\|_{H^1(\Omega)} \leq C(\sigma_H H^\ell + \sigma_H r_{HMM})$ and thus $\sigma_H \|\widetilde{u}^H - u^H\|_{H^1(\Omega)}$ satisfies (62) for $r_{HMM} \leq H$ with H small enough. Provided H, r_{HMM}, r'_{HMM} are such that (61) is satisfied, $e_k = \|u_k^H - u^H\|_{H^1(\Omega)}$ converges to 0 for $k \to \infty$ by Lemma 4.11. Since $u_k^H = u_0^H = \tilde{u}^H$, we obtain $u^H = \tilde{u}^H$. □

If we want further to characterize uniqueness in terms of the macro and micro meshes, we need to estimate r_{HMM}, r'_{HMM} in terms of these quantities. This can be done for locally periodic tensors. The quantity r_{HMM} has been estimated in terms of h, ε, δ in Section 4.2. Using similar techniques, the quantity r'_{HMM} defined in (31) can be estimated as described in the following lemma whose proof is postponed to the Appendix.

Lemma 4.12 *Assume that the hypothesis of Corollary 3.4 hold. Then*

$$r'_{HMM} \leq C\Big(\frac{h}{\varepsilon}\Big)^2. \tag{64}$$

If we use the form (17) *instead of* (20) *for the solution u^H of* (16) *then*

$$r'_{HMM} \leq C(\Big(\frac{h}{\varepsilon}\Big)^2 + \delta). \tag{65}$$

Proof the Corollary 3.4. Follows from Theorem 3.3, Lemmas 4.12, 4.6 and 4.7. □

5 Corrector and finescale approximation

We explain in this section how to obtain numerically an approximation of the oscillating solution u^ε of the non-linear problem (1) by using a reconstruction procedure identical to the one presented in [23] for HMM.

We recall that, while the the convergence, up to a subsequence, $u_\varepsilon \to u_0$ is strong in $L^2(\Omega)$, it is only weak in $H^1(\Omega)$. The fine scale oscillation of u^ε introduce usually a $\mathcal{O}(1)$ discrepancy between ∇u^ε and ∇u^0 and the quantity $\|u_\varepsilon - u_0\|_{H^1(\Omega)}$ does not converge to zero in general as $\varepsilon \to 0$. One needs therefore to introduce a corrector $u_{1,\varepsilon}$ to bound the error $\|u^\varepsilon - u_0 - u_{1,\varepsilon}\|_{H^1(\Omega)}$ in terms of ε [12],[29].

We first review the classical construction of such corrector and explain then their numerical approximation using the FE-HMM.

Corrector. Following [14], we consider the following linear homogenization problem: find $\bar{u}^\varepsilon$ such that

$$-\nabla \cdot (a^\varepsilon(x, u_0(x))\nabla \overline{u}_\varepsilon(x)) = f(x) \text{ in } \Omega, \quad \overline{u}_\varepsilon(x) = 0 \text{ on } \partial\Omega. \tag{66}$$

where compared to the non-linear problem (1) the tensor is evaluated at u_0 instead of u^ε, with u_0 the unique solution of (3). Then [14, Sec. 3.2] up to a subsequence, we have for $\varepsilon \to 0$ the weak convergence

$$\overline{u}_\varepsilon \rightharpoonup u_0 \text{ in } H^1(\Omega).$$

We next assume that (**H2**) holds (we have thus the characterization (7) for the homogenized tensor), that $u_0(x)$ is smooth (e.g., $u_0(x) \in C^2(\bar{\Omega})$) and that $a(x, y, s)$ is smooth enough to ensure that the solution χ^i of (8) satisfies $\chi^i \in W^{1,\infty}(\Omega \times Y \times \mathbb{R})$. We then have the following estimate from linear homogenization [29, Sect. 1.4]

$$\|\overline{u}_\varepsilon - u_0 - u_{1,\varepsilon}\|_{H^1(\Omega)} \quad \leq \quad C\sqrt{\varepsilon},$$

where C is independent of ε, and $u_{1,\varepsilon}$ is called a corrector and is defined by

$$u_{1,\varepsilon}(x) = \varepsilon \sum_{j=1}^{d} \chi^j(x, x/\varepsilon, u_0(x)) \frac{\partial u_0(x)}{\partial x_j}. \tag{67}$$

Remark 5.1 *In [14, Sect. 3.4.2], it is shown that any corrector for $\overline{u}_\varepsilon$ is also a corrector for the solution u_ε of the nonlinear problem* (1)*. In our situation, we have*

$$\nabla r_\varepsilon \to 0 \textit{ strongly in } (L^1_{loc}(\Omega))^d \textit{ where } r_\varepsilon(x) := u_\varepsilon(x) - u_0(x) - u_{1,\varepsilon}(x). \tag{68}$$

It would be interesting to derive a rate for the convergence (68)*. For instance in the linear case (i.e. for $a^\varepsilon(x, s)$ independent of s), one has the classical estimate $\|r_\varepsilon\|_{H^1(\Omega)} \leq C\sqrt{\varepsilon}$ (see [29, Sect. 1.4]).*

FE-HMM reconstruction. Consider u^H the solution of (16) in $S^1_0(\Omega, \mathcal{T}_H)$ using the form (20). We also assume that simplicial elements are used. In this case, we have only one quadrature point x_K and one sampling domain K_δ centered at the barycenter of each macro element K. The idea of the numerical reconstruction procedure is to consider the micro function $u^h - u^H$ available on $K_\delta \subset K$ centered around the quadrature point x_K, and to extend it on the whole element K

$$u_{p,\varepsilon}(x)|_K := u^H(x) + (u^h_K - u^H)(x - [x]_{K_\delta}) \text{ for all } x \in K \in \mathcal{T}_H, \tag{69}$$

where for $x \in \mathbb{R}^d$, $[x]_{K_\delta} \in \delta\mathbb{Z}^d$ is such that $x - [x]_{K_\delta} \in K_\delta$.

Remark 5.2 *Notice that the above formulation is equivalent to apply the (standard) FE-HMM post-processing [1],[3] procedure to the linear problem*

$$-\nabla \cdot \left(a^\varepsilon(x, u^H(x))\nabla\widetilde{u}_\varepsilon(x)\right) = f(x) \quad \text{in } \Omega, \quad \widetilde{u}_\varepsilon(x) = 0 \quad \text{on } \partial\Omega, \tag{70}$$

provided one uses a bilinear form collocated at the integration point x_K for the FE-HMM.

Since $u_{p,\varepsilon}(x)$ may be discontinuous on the boundaries of the elements $K \in \mathcal{T}_H$, we define a broken H^1 norm by

$$\|u\|_{\bar{H}^1(\Omega)} = (\sum_{K\in\mathcal{T}_H} \|\nabla u\|^2_{L^2(K)})^{1/2}.$$

Motivated by the convergence (68) shown in [14], we obtain the following theorem.

Theorem 5.3 *Let $\ell = q = 1$. Consider a macro triangulation $\mathcal{T}_H$ with $\mathcal{P}^1$-simplicial elements. Assume that the assumptions of Theorem 3.1 with $\ell = 1$ are satisfied. Assume **(H1)**, **(H2)** with $q = 1$, $\delta/\varepsilon \in \mathbb{N}$, and that a periodic coupling is used in the micro problems of the FE-HMM, i.e. $W(K_\delta) = W^1_{per}(K_\delta)$. We also assume that* (20) *is used. Assume further that $u_0 \in C^2(\overline{\Omega})$. Consider the solution u_ε of* (1) *and the corrector $u_{p,\varepsilon}$ defined in* (69)*. Then*

$$\|u_\varepsilon - u_{p,\varepsilon} - r_\varepsilon\|_{\bar{H}^1(\Omega)} \le C(H + h/\varepsilon + \varepsilon).$$

where C is independent of H, h, ε and r_ε is defined in (68)*.*

Proof. We consider the decomposition

$$u_\varepsilon - u_{p,\varepsilon}(x) - r_\varepsilon(x) = (u_{1,\varepsilon} - \widetilde{u}_{1,\varepsilon}) + (u_0 + \widetilde{u}_{1,\varepsilon} - u_{p,\varepsilon}(x))$$

where $\widetilde{u}_{1,\varepsilon}$ is the corrector associated to the linear problem (70), defined by

$$\widetilde{u}_{1,\varepsilon}(x) = \varepsilon \sum_{j=1}^d \chi^j(x, x/\varepsilon, u^H(x)) \frac{\partial u_0(x)}{\partial x_j}.$$

Using the Lipschitzness of $\partial_{x_i}\chi^j(x, y, \cdot)$ (a consequence of (**H2**)) and $\nabla u_0 \in L^\infty(\Omega)$, we obtain

$$\|u_{1,\varepsilon} - \widetilde{u}_{1,\varepsilon}\|_{H^1(\Omega)} \le C\|u^H - u_0\|_{H^1(\Omega)} \le C(H + (h/\varepsilon)^2).$$

Using Remark 5.2, we deduce using the corrector argument from the linear case theory [1]

$$\|u_0 + \widetilde{u}_{1,\varepsilon} - u_{p,\varepsilon}(x)\|_{\bar{H}^1(\Omega)} \le C(H + h/\varepsilon + \varepsilon).$$

This concludes the proof. □

We deduce from Remark 5.1 and Theorem 5.3 that if H, h/ε, and ε tend simultaneously to zero, then

$$\nabla u_\varepsilon - \bar{\nabla} u_{p,\varepsilon} \to 0 \text{ strongly in } (L^1_{loc}(\Omega))^d$$

where $\bar{\nabla} u_{p,\varepsilon}$ is defined piecewisely as $\bar{\nabla} u_{p,\varepsilon}\big|_K = \nabla u_{p,\varepsilon}\big|_K$ for all $K \in \mathcal{T}_H$.

6 Numerical experiments

In this section, we first present an efficient numerical implementation of the Newton method (55), whose theoretical convergence is shown in Lemma 4.11. We then illustrate numerically that the theoretical a priori convergence rates derived in this paper are optimal for $\mathcal{P}^1$-triangular FEs or $\mathcal{Q}^1$-rectangular FEs.

6.1 Newton method

To solve the non-linear problem (16) with the newton method, we consider a sequence of $\{z_k^H\}$ in $S_0^\ell(\Omega, \mathcal{T}_H)$ and express each function in the FE basis of $S_0^\ell(\Omega, \mathcal{T}_H)$ as $z_k^H = \sum_{i=1}^{M_{macro}} U_k^i \phi_i^H$. We further denote $U_k = (U_k^1, \ldots, U_k^{M_{macro}})^T$. The Newton method (55) translate in terms of matrices as

$$\big(B(z_k^H) + B'(z_k^H)\big)(U_{k+1} - U_k) = -B(z_k^H)U_k + F, \tag{71}$$

where $B(z_k^H), B'(z_k^H)$ are the stiffness matrices associated to the bilinear forms $B_H(z^H;\cdot,\cdot)$, $B'_H(z^H;\cdot,\cdot)$ defined in (17) and (54), respectively. Here, F a vector associated the source term (16), which also contains the boundary data.

Stiffness matrix $B(z_k^H)$. Following the implementation in [5] we consider for each element $K \in \mathcal{T}_H$ the FE basis functions $\{\phi_{K,i}^H\}_{i=1}^{n_K}$ associated with this element and the local contribution $B_K(z_k^H)$ to the stiffness

$$\begin{aligned} (B_K(z_k^H))_{p,q=1}^{n_K} &= \sum_{j=1}^{J} (B_{K,j}(z_k^H))_{p,q=1}^{n_K} \\ &= \sum_{j=1}^{J} \frac{\omega_{K_j}}{|K_{\delta_j}|} \int_{K_{\delta_j}} a^\varepsilon(x, z_k^H(x_{K_j})) \nabla \varphi_{K_j,p}^{h,z^H(x_{K_j})}(x) \cdot \nabla \varphi_{K_j,q}^{h,z^H(x_{K_j})}(x) dx, \end{aligned} \tag{72}$$

where $\varphi_{K_j,p}^{h,z^H(x_{K_j})}, \varphi_{K_j,q}^{h,z^H(x_{K_j})}$ are the solutions of (15) constrained by $\phi_{K,p}^H, \phi_{K,q}^H$, linearized at x_{K_j}, respectively.

Stiffness matrix $B'(z_k^H)$. Differentiating (72), we see that the stiffness matrix $B'(U)$ in (71) associated to the non-symmetric form $B'_H(z^H;\cdot,\cdot)$ defined in (54) is given by the sum of J products of $n_K \times n_K$ matrices

$$B'_K(z_k^H) = \sum_{j=1}^{J} \left(\frac{\partial}{\partial s}(B_{K,j}(s))\bigg|_{s=z^H(x_{K_j})} \right) \Big(U_k(\phi_{K_1}^H(x_{K_j}), \ldots, \phi_{K_{n_K}}^H(x_{K_j})) \Big)$$

for the macro element $K \in \mathcal{T}_H$. Here, the derivative with respect to s of the $n_K \times n_K$ matrix $B_{K,j}(s)$ can be simply approximated by the finite difference

$$\frac{\partial}{\partial s}(B_{K,j}(s)) \approx \frac{B_{K,j}(s+\sqrt{eps}) - B_{K,j}(s)}{\sqrt{eps}},$$

where *eps* is the machine precision. Therefore, the cost of computing together the stiffness matrix $B(z_k^H)$ and $B'(z_k^H)$ is about twice the cost of computing the stiffness matrix $B(z_k^H)$ alone.

Remark 6.1 *The computational cost in the Newton method* (55) *can be further reduced by taking a coarser mesh in the micro problems for the first few Newton iterations for both matrices $B(z_k^H)$ and $B'(z_k^H)$.*

We emphasize that the FE-HMM is embarrassingly parallel as the micro problems are independent one from another. For the numerical tests in Section 6.2, we consider a Matlab implementation of the nonlinear FE-HMM, where the stiffness matrices $B_K(z_k^H)$, $B'_K(z_k^H)$ associated to each element $K \in \mathcal{T}_H$ of the triangulation are computed in parallel (here on 8 processors).

6.2 Numerical examples

In this section, we shall illustrate the sharpness of the H^1 and L^2 a priori error estimates of Sections 3 and 4. First, we consider a simple test problem where the exact homogenized tensor and the exact solution are known analytically. Second, we apply our multiscale method to a steady state model of of Richards equation for porous media flows.

6.2.1 Convergence rates: test problem

We recall that for a tensor of the form $a^\varepsilon(x,s) = a(x, x/\varepsilon, s)$ where $a(x,y,s)$ is periodic with respect to the fast variable y and collocated in the slow variable x (i.e. (17) is used), the H^1 and L^2 errors satisfy (see the second case in (38) with $\ell = q = 1$)

$$\|u^H - u_0\|_{H^1(\Omega)} \le C(H + \hat{h}^2), \qquad \|u^H - u_0\|_{L^2(\Omega)} \le C(H^2 + \hat{h}^2), \tag{73}$$

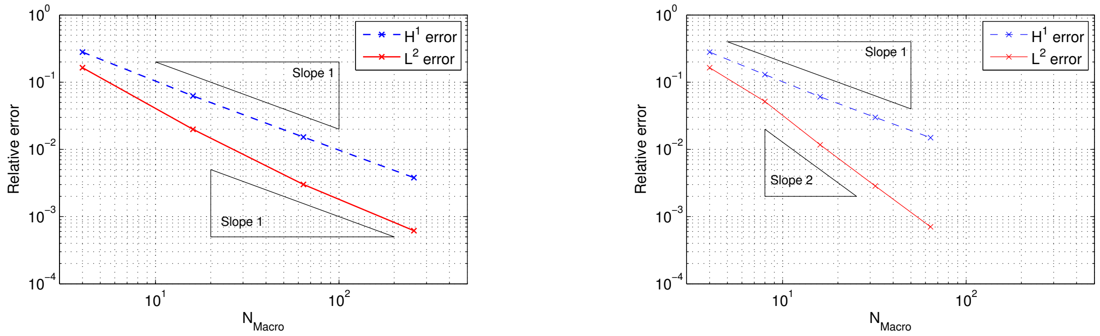
where $\hat{h} := h/\varepsilon$ is the scaled micro mesh size. In the above estimates, periodic boundary conditions are used for (15) and we assume that the micro sampling domains cover one period of the oscillating tensor in each spatial dimension. For rectangular elements, we consider the Gauss quadrature with $J = 4$ nodes $(1/2 \pm \sqrt{3}/6, 1/2 \pm \sqrt{3}/6)$, while for triangular elements, we consider the quadrature formula with $J = 1$ node located at the barycenter. Notice that we obtain similar results when considering either rectangular or triangular elements.

We consider the non-linear problem (1) on the domain $\Omega = (0,1)^2$ with homogeneous Dirichlet boundary conditions and the following anisotropic oscillatory tensor

$$a^\varepsilon(x,s) = \frac{1}{\sqrt{3}} \begin{pmatrix} (2+\sin(2\pi x_1/\varepsilon))(1 + x_1 \sin(\pi s)) & 0 \\ 0 & (2+\sin(2\pi x_2/\varepsilon))(2+\arctan(s)) \end{pmatrix}.$$

The homogenized tensor can be computed analytically and is given by

$$a^0(x,s) = \begin{pmatrix} 1 + x_1 \sin(\pi s) & 0 \\ 0 & 2 + \arctan(s) \end{pmatrix}.$$



(a) Optimal H^1 refinement strategy with $N_{Micro} \sim \sqrt{N_{Macro}}$ where $N_{Micro} = 4, 8, 16, 32$, $N_{Macro} = 4, 16, 64, 256$ respectively.

(b) Optimal L^2 refinement strategy with $N_{Micro} = N_{Macro} = 4, 8, 16, 32, 64$.

Figure 1: Nonlinear homogenization test problem of Sect. 6.2.1. e_{L^2} error (dashed lines) and e_{H^1} error (solid lines) as a function of the size N_{Macro} of the uniform mesh with $M_{Macro} = N_{Macro} \times N_{Macro}$ $\mathcal{Q}^1$-quadrilateral elements.

The source $f(x)$ in (1) is adjusted analytically so that the homogenized solution u_0 is

$$u_0(x) = 8\sin(\pi x_1)x_2(1 - x_2), \tag{74}$$

The H^1 and L^2 relative errors between the exact homogenized solution u_0 and the FE-HMM solution u^H can be estimated by quadrature with

$$\begin{aligned} e_{L^2}^2 &:= \|u_0\|_{L^2(\Omega)}^{-2} \sum_{K\in\mathcal{T}_H} \sum_{j=1}^{J} \omega_{K_j} |u^H(x_{K_j}) - u_0(x_{K_j})|^2, \\ e_{H^1}^2 &:= \|\nabla u_0\|_{L^2(\Omega)}^{-2} \sum_{K\in\mathcal{T}_H} \sum_{j=1}^{J} \omega_{K_j} \|\nabla u^H(x_{K_j}) - \nabla u_0(x_{K_j})\|^2, \end{aligned}$$

so that
$$e_{L^2} \approx \frac{\|u_0 - u^H\|_{L^2(\Omega)}}{\|u_0\|_{L^2(\Omega)}}, \qquad e_{H^1} \approx \frac{\|\nabla(u_0 - u^H)\|_{L^2(\Omega)}}{\|\nabla u_0\|_{L^2(\Omega)}}.$$

We now let $\varepsilon = 10^{-2}$. We emphasize that ε is needed for the algorithm but its precise value is not important, as for locally periodic problem solved with periodic micro boundary conditions, the convergence rate and the computational cost are *independent* of ε (see (73)). We consider a sequence of uniform macro partitions $\mathcal{T}_H$ with meshsize $H = 1/N_{Macro}$ and $N_{Macro} = 4, 6, 8, \ldots, 256$. We choose $\mathcal{Q}^1$-rectangular elements with size $H = 1/N_{Macro}$ in the experiments below; the results are silimar for $\mathcal{P}^1$-triangular elements.

In Figure 1 the H^1 an L^2 relative errors between the exact homogenized solution and the FE-HMM solutions are shown for the above sequence of partitions using a simultaneous refinement of H and $\hat{h}$ according to $\hat{h} \sim H$ (L^2 norm) and $\hat{h} \sim \sqrt{H}$ (H^1 norm). We observe the expected (optimal) convergence rates (73) in agreement with Theorem 3.1.

We next show that the ratio between the macro and micro meshes is sharp. For that, we refine the macromesh H while keeping fixed the micro mesh size. This is illustrated in Figure 2, where we plot the H^1 an L^2 relative errors as a function of $H = 1/N_{Macro}$. Five

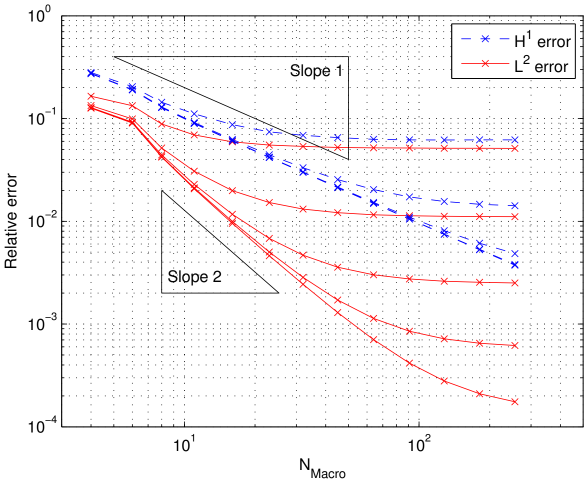


Figure 2: Nonlinear homogenization test problem of Sect. 6.2.1. e_{L^2} error (dashed lines) and e_{H^1} error (solid lines) as a function of the size N_{Macro} of the uniform mesh with $M_{Macro} = N_{Macro} \times N_{Macro}$ $\mathcal{Q}^1$-quadrilateral elements. The lines correspond respectively to $N_{Micro} = 4, 8, 16, 32, 64$.

sizes of micro meshes are chosen with size $\hat{h}, = 1/N_{Micro}$ and $N = Micro = 4, 8, 16, 32$. We observe that for small values of $H = 1/N_{Macro}$, the error due to the macro domain discretization is dominant. [7] For large values of $N_{Macro} = 1/H$, the error due to the micro domains discretization becomes dominant and the H^1 and L^2 errors becomes independent of N_{Macro} (horizontal lines). We observe that when $N_{Micro} = 1/\hat{h}$ is multiplied by 2, both the H^1 and L^2 errors are divided by 4, which corroborates Theorem 3.6: the micro error has size $\mathcal{O}(\hat{h}^2)$. This experiments illustrate that *simultaneous* refinement of macro and micro meshes (at the right ratio) is needed for optimal convergence rates with minimal computational cost.

6.2.2 Richards equation for multiscale porous media

We consider the Richards equation for describing the fluid pressure $u(x, t)$ in an unsaturated porous medium, with multiscale permeability tensor K^ε and volumetric water content Θ^ε,

$$\frac{\partial \Theta^\varepsilon(u_\varepsilon(x))}{\partial t} - \nabla \cdot (K^\varepsilon(u_\varepsilon(x))\nabla u_\varepsilon(x))) + \frac{\partial K^\varepsilon(u_\varepsilon(x))}{\partial x_2} = f(x) \quad \text{in } \Omega,$$

where x_2 is the vertical coordinate, and f corresponds to possible sources or sinks. We choose an exponential model for the permeability tensor K^ε similar to the one in [16, Sect. 5.1],

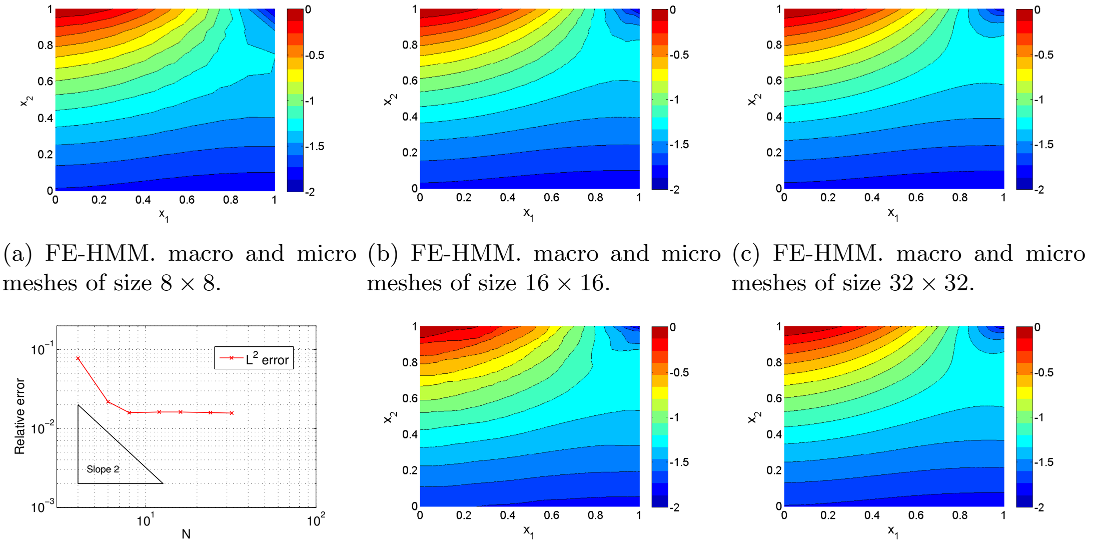
$$K^\varepsilon(x, s) = \alpha^\varepsilon(x) e^{\alpha^\varepsilon(x)s} \qquad \text{where } \alpha^\varepsilon(x) = \frac{1/117.4}{(2 + 1.8\sin(2\pi(2x_2/\varepsilon - x_1/\varepsilon)))}. \tag{75}$$

For our numerical simulation, we consider the steady state $\partial\Theta^\varepsilon(u_\varepsilon)/\partial t = 0$.

$$-\nabla \cdot (K^\varepsilon(u_\varepsilon(x))\nabla(u_\varepsilon(x) - x_2)) = 0 \quad \text{in } \Omega = (0, 1)^2, \tag{76}$$

where for simplicity we set $f(x) \equiv 0$. Notice that (76) can be cast in the form (1) by considering the change of variable $v_\varepsilon(x) = u_\varepsilon(x) - x_2$. We add mixed boundary conditions

[7] Notice that the curves for the H^1 error are nearly identical for $N_{Micro} = 32, 64$.



(a) FE-HMM. macro and micro meshes of size 8 × 8.

(b) FE-HMM. macro and micro meshes of size 16 × 16.

(c) FE-HMM. macro and micro meshes of size 32 × 32.

(d) FE-HMM. L^2 relative error.

(e) FEM. mesh size: 32 × 32 (unresolved).

(f) FEM. mesh size: 1024 × 1024 (finescale).

Figure 3: Richards problem (75)-(76). Top pictures: level curves of the FE-HMM solutions with $N_{Macro} = N_{Micro}$. Fig. (d): L^2 relative error for $u^H - u_\varepsilon$ versus $N = N_{Macro} = N_{Micro}$ (optimal L^2 refinement strategy). Figs. (e)-(f): level curves of the standard FEM solutions.

of Dirichlet and Neumann types. We put Neumann conditions on the left, right and bottom boundaries of the domain (n denotes the vector normal to the boundary) and a Dirichlet condition on the top boundary. Precisely, we take

$$\begin{aligned} &u_\varepsilon(x) = -1.9x_1^2 \quad \text{on } \partial\Omega_D = [0,1] \times \{1\}, \\ &n \cdot (K^\varepsilon(u_\varepsilon(x))\nabla(u_\varepsilon(x) - x_2)) = 0 \quad \text{on } \partial\Omega_N = \{0,1\} \times [0,1] \cup [0,1] \times \{0\}. \end{aligned}$$

We refine the macro and micro meshes according to the optimal strategy as seen in the above test problem. The numerical results are compared to a resolved standard FE solution for the fine scale problem where $\varepsilon = 10^{-2}$ using $\sim 10^6$ degrees of freedom, and plotted in Fig. 3(f). As we compare the fine scale solution with the FE-HMM solution (without reconstruction) we restrict ourselves to comparison in the L^2 norm. From the results in Sections 3 and 4 we know that

$$\|u^H - u_\varepsilon\|_{L^2(\Omega)} \leq C(H^2 + \hat{h}^2) + \eta_\varepsilon$$

where $\eta_\varepsilon := \|u_0 - u_\varepsilon\|_{L^2(\Omega)} \to 0$ for $\varepsilon \to 0$. We first see in Figure 3(d) the expected convergence rate for the L^2 error when macro and micro meshes are refined at the same speed $N_{Macro} = N_{Micro} = N$, and the horizontal line corresponds to the term η_ε, which numerically appears to be of the size[8] of ε. In Figures 3(a)-(c), we plot the level curves of the FE-HMM solution for problem (3), where we consider uniform $N \times N$ macro meshes with couples of $\mathcal{P}^1$-triangular FEs, and uniform $N \times N$ micro meshes with $\mathcal{Q}^1$-rectangular FEs. For comparison, we also plot the standard FEM solution of (1) with a coarse 32 × 32 mesh (unresolved) and a finescale solution on a fine 1024 × 1024 mesh. We observe that the

[8] Recall that for linear homogenization problems, one has $\|u_0 - u_\varepsilon\|_{L^2(\Omega)} \leq C\varepsilon$ [29, Sect. 1.4].

unresolved FEM does not yield a qualitative correct result. In contrast, the FE-HMM permits to capture the correct behavior of the resolved solution at a much lower computational cost.

7 Appendix

We provide in this appendix a proof of Lemmas 4.11 and 4.12. We start with Lemma 4.11. As mentioned earlier we only sketch its proof.

Proof of Lemma 4.11. The proof is a consequence of Lemma 4.9 which can be applied in the special case $z^H = u^H$ thanks to Lemma 4.10 and the estimate $\sigma_H \|u^H - u^0\|_{H^1(\Omega)} \leq C\sigma_H H^\ell \leq \nu$ for all H small enough (using (33)).

Given the Newton interation u_k^H, we show by induction on k the following two statements:

(i) the next Newton iteration u_{k+1}^H exists and is uniquely defined by (55),

(ii) $\sigma_H e_{k+1} \leq \nu$ and the estimates (63) holds.

Since the linear system (55) is of finite dimension, to prove the point (i) it is sufficient to show that he homogeneous problem $\partial B(u_k^H; v^H, w^H) = 0$ has only the trivial solution $v^H = 0$ for all $w^H \in S_0^\ell(\Omega, \mathcal{T}_H)$. Using the Hölder inequality and the Lipschitzness of $a_K^0(s)$ and $\partial a_K^0(s)/\partial s$ with respect to s, one can show for all $w^H \in S_0^\ell(\Omega, \mathcal{T}_H)$,

$$\begin{aligned}
\partial B(u^H; v^H, w^H) &= \partial B(u^H; v^H, w^H) - \partial B(u_k^H; v^H, w^H) \\
&\leq C\|u^H - u_k^H\|_{L^\infty(\Omega)} \|w^H\|_{H^1(\Omega)} (\|v^H\|_{H^1(\Omega)} + \|v^H\|_{L^3(\Omega)} \|u^H\|_{W^{1,6}(\Omega)}) \\
&+ C\|v^H\|_{L^\infty(\Omega)} \|u^H - u_k^H\|_{H^1(\Omega)} \|w^H\|_{H^1(\Omega)} \\
&\leq C\sigma_H \|u^H - u_k^H\|_{H^1(\Omega)} \|v^H\|_{H^1(\Omega)} \|w^H\|_{H^1(\Omega)}.
\end{aligned}$$

Using $\partial B(u_k^H; v^H, w^H) = 0$, we deduce from Lemma 4.9 and the induction hypothesis $\sigma_H e_k \leq \nu$ that $\|v^H\|_{H^1(\Omega)} \leq C\nu \|v^H\|_{H^1(\Omega)}$, which implies $v^H = 0$ if ν is chosen small enough so that $C\nu < 1$. The proof of (i) is achieved. For the proof of (ii), using (55), one can show the following estimates

$$\begin{aligned}
\partial B_H(u^H; u_{k+1}^H - u^H, w^H) &= \{\partial B_H(u^H; u_{k+1}^H - u_k^H, w^H) - \partial B_H(u_k^H; u_{k+1}^H - u_k^H, w^H)\} \\
&+ \{\partial B_H(u^H; u_k^H - u^H, w^H) + B_H(u^H; u^H, w^H) \\
&\quad - B_H(u_k^H; u_k^H, w^H)\} \\
&\leq C\sigma_H (e_k e_{k+1} + e_k^2) \|w^H\|_{H^1(\Omega)}
\end{aligned}$$

for all $w^H \in S_0^\ell(\Omega, \mathcal{T}_H)$, where we used the C^2 regularity with respect to s of the tensor $a_K^0(s)$, two integrations by parts (see [21, Theorem 2]) and the Hölder inequality. We notice that the C^2 regularity of $a_K^0(s)$ and the boundedness of $\partial a_K^0(s)/\partial s$, $k \leq 2$ can be shown from (36) using the idea of the proof of Lemma 7.1 (see below). Using again Lemma 4.9 with $z^H = u^H$, we deduce $e_{k+1} \leq C\sigma_H(e_k e_{k+1} + e_k^2)$, which yields $(1 - C\nu)e_{k+1} \leq C\sigma_H e_k^2$. This gives (63) for ν small enough and the estimates $\sigma_H e_{k+1} \leq \nu$ follows immediately from (63) and $C\sigma_H e_k \leq C\nu \leq 1$. We thus obtain (ii). □

We shall now prove Lemma 4.12. For that, we will often use the following inequality (77). Given a closed subspace H of $W(K_{\delta_j})$, let ψ_i, $i=1,2$ be the solutions of

$$\int_{K_{\delta_j}} a_i(x)\nabla\psi_i(x)\cdot\nabla z(x)dx = -\int_{K_{\delta_j}} f_i(x)\cdot\nabla z(x)dx,\ \forall z\in H,$$

where $a_1,a_2\in L^\infty(K_{\delta_j})^{d\times d}$ are elliptic and bounded tensors and $f_1,f_2\in L^2(K_{\delta_j})^d$. A short computation shows

$$\|\nabla\psi_1-\nabla\psi_2\|_{L^2(K_{\delta_j})}\le\lambda^{-1}\sup_{x\in K_{\delta_j}}\|a_1(x)-a_2(x)\|_F\|f_2\|_{L^2(K_{\delta_j})}+\|f_1-f_2\|_{L^2(K_{\delta_j})},\tag{77}$$

where λ is the minimum of the ellipticity constants of a_1,a_2. We also need a regularity result for the solutions of (25).

Lemma 7.1 *Assume that a^ε is uniformly elliptic and satisfies (36) with $k=1$. Consider the solution $\psi^{i,s}_{K_j}$ of (25). Then, the map $s\mapsto\psi^{i,s}_{K_j}\in H^1(K_{\delta_j})$ is of class C^1 and satisfies*

$$\frac{\partial}{\partial s}\psi^{i,s}_{K_j}=\phi^{i,s}_{K_j},\qquad\frac{\partial}{\partial s}\nabla\psi^{i,s}_{K_j}=\nabla\phi^{i,s}_{K_j},\tag{78}$$

where for all $z\in W(K_{\delta_j})$,

$$\int_{K_{\delta_j}} a^\varepsilon(x,s)\nabla\phi^{i,s}_{K_j}(x)\cdot\nabla z(x)dx=-\int_{K_{\delta_j}}\partial_u a^\varepsilon(x,s)(\nabla\psi^{i,s}_{K_j}(x)+\mathbf{e}_i)\cdot\nabla z(x)dx.\tag{79}$$

A similar statement holds also for the FEM discretization $\psi^{i,h,s}_{K_j}$ defined in (28), where $\frac{\partial}{\partial s}\psi^{i,h,s}_{K_j}=\phi^{i,h,s}_{K_j}$ satisfies (78) and (79) with $\psi^{i,s}_{K_j},\phi^{i,s}_{K_j}$ and z replaced by $\psi^{i,h,s}_{K_j},\phi^{i,h,s}_{K_j}$ and $z^h\in S^q(K_{\delta_j},\mathcal{T}_h)$ respectively.

Proof. We consider twice the problem (28) with parameters s and $s+\Delta s$, respectively. We deduce from (77) with $H=W(K_{\delta_j})$, and the smoothness of $s\mapsto a^\varepsilon(x,s)$ that

$$\|\psi^{i,s+\Delta s}_{K_j}(x)-\psi^{i,s}_{K_j}(x)\|_{H^1(K_{\delta_j})}\to 0\quad\text{for}\quad\Delta s\to 0.$$

Consider now the identity

$$\begin{aligned}&\int_{K_{\delta_j}} a^\varepsilon(x,s)\nabla(\psi^{i,s+\Delta s}_{K_j}(x)-\psi^{i,s}_{K_j}(x))\cdot\nabla z(x)dx\\&=-\int_{K_{\delta_j}}(a^\varepsilon(x,s+\Delta s)-a^\varepsilon(x,s))(\nabla\psi^{i,s+\Delta s}_{K_j}(x)+\mathbf{e}_i)\cdot\nabla z(x)dx\end{aligned}\tag{80}$$

Dividing (80) by Δs and subtracting (79), we deduce from (77)

$$\begin{aligned}&\|(\psi^{i,s+\Delta s}_{K_j}(x)-\psi^{i,s}_{K_j}(x))/\Delta s-\phi^{i,s}_{K_j}(x)\|_{H^1(K_{\delta_j})}\\&\le C\Big\|\Big((a^\varepsilon(x,s+\Delta s)-a^\varepsilon(x,s))/\Delta s-\partial_u a^\varepsilon(x,s)\Big)(\nabla\psi^{i,s+\Delta s}_{K_j}(x)+\mathbf{e}_i)\|_{L^2(\Omega)}\\&+C\|\partial_u a^\varepsilon(x,s)\nabla(\psi^{i,s+\Delta s}_{K_j}(x)-\psi^{i,s}_{K_j}(x))\|_{L^2(\Omega)}\to 0\quad\text{for}\quad\Delta s\to 0,\end{aligned}$$

which shows that $\frac{\partial}{\partial s}\psi^{i,s}_{K_j}(x)$ exists and that (78),(79) hold. Using again the property (77), we obtain similarly the continuity of $s\mapsto\phi^{i,s}_{K_j}\in H^1(K_{\delta_j})$. This concludes the proof for $\psi^{i,s}_{K_j}$. The proof for $\psi^{i,h,s}_{K_j}$ is nearly identical, using the property (77) with $H=S^q(K_{\delta_j},\mathcal{T}_h)$ □

Proof of Lemma 4.12. We first prove the estimate (64). We set $x = x_{K_j}$ in (7). A change of variable $y \to x_{K_j} + x/\varepsilon$ shows that

$$(a^0(x_{K_j}, s))_{mn} = \frac{1}{|K_{\delta_j}|} \int_{K_{\delta_j}} a(x_K, x/\varepsilon, s)(\mathbf{e}_n + \nabla\chi^n(x_K, x/\varepsilon, s)) \cdot \mathbf{e}_m \tag{81}$$

where $\chi^n(x_K, x/\varepsilon, s)$ solves for all $z \in W(K_{\delta_j})$,

$$\int_{K_{\delta_j}} a(x_K, x/\varepsilon, s)\nabla\chi^n(x_K, x/\varepsilon, s) \cdot \nabla z(x)dx = -\int_{K_{\delta_j}} a(x_K, x/\varepsilon, s)\mathbf{e}_n \cdot \nabla z(x)dx, \tag{82}$$

As the tensor a^ε is (locally) periodic and $\delta/\varepsilon \in \mathbb{N}$, if we collocate a^ε in (30) and in (7) at $x = x_{K_j}$, we obtain $a^0(x_{K_j}, s) = \overline{a}^0_{K_j}(s)$ and $\psi^{n,s}_{K_j}(x) = \varepsilon\chi^n(x_{K_j}, x/\varepsilon, s)$.

We consider the elliptic system $-\nabla \cdot (A\nabla\Xi) = \nabla \cdot F_i$ formed by problems (28)-(79), where

$$A = \begin{pmatrix} a(x_{K_j}, x/\varepsilon, s) & 0 \\ \partial_u a(x_{K_j}, x/\varepsilon, s) & a(x_{K_j}, x/\varepsilon, s) \end{pmatrix}, \ F = \begin{pmatrix} a(x_{K_j}, x/\varepsilon, s)\mathbf{e}_n & 0 \\ 0 & \partial_u a(x_{K_j}, x/\varepsilon, s)\mathbf{e}_n \end{pmatrix}$$

and $\Xi = (\psi^{n,s}_{K_j}, \phi^{n,s}_{K_j})^T$. It follows form well known H^2 regularity results [13, Sect. 3.4-3.6] that $\phi^{n,s}_{K_j}, \psi^{n,s}_{K_j} \in H^2(K_{\delta_j})$ and $\|\phi^{n,s}_{K_j}\|_{H^2(K_{\delta_j})} + \|\psi^{n,s}_{K_j}\|_{H^2(K_{\delta_j})} \le C\varepsilon^{-1}\sqrt{|K_{\delta_j}|}$. From standard FEM results [19, Sect. 17], we deduce that the corresponding FEM discretization $(\psi^{m,h,s}_{K_j}, \phi^{m,h,s}_{K_j})$ satisfies

$$\begin{aligned} \|\nabla\psi^{n,s}_{K_j} - \nabla\psi^{n,h,s}_{K_j}\|_{L^2(K_{\delta_j})} &\le Ch\|\psi^{n,s}_{K_j}\|_{H^2(K_{\delta_j})} \le C(h/\varepsilon)\sqrt{|K_{\delta_j}|}, \\ \|\nabla\phi^{n,s}_{K_j} - \nabla\phi^{n,h,s}_{K_j}\|_{L^2(K_{\delta_j})} &\le Ch\|\phi^{n,s}_{K_j}\|_{H^2(K_{\delta_j})} \le C(h/\varepsilon)\sqrt{|K_{\delta_j}|}. \end{aligned}$$

Now, using Lemma 7.1 and differentiating the identity (49) with respect to s, we deduce from the Cauchy-Schwarz inequality

$$\begin{aligned} |\frac{d}{ds}(\bar{a}^0_{K_j}(s) - a^0_{K_j}(s))_{mn}| &\le \frac{1}{|K_{\delta_j}|}\Big(\|\nabla\psi^{n,s}_{K_j} - \nabla\psi^{n,h,s}_{K_j}\|_{L^2(K_{\delta_j})}\|\nabla\overline{\psi}^{m,s}_{K_j} - \nabla\overline{\psi}^{m,h,s}_{K_j}\|_{L^2(K_{\delta_j})} \\ &+ \|\nabla\phi^{n,s}_{K_j} - \nabla\phi^{n,h,s}_{K_j}\|_{L^2(K_{\delta_j})}\|\nabla\overline{\psi}^{m,s}_{K_j} - \nabla\overline{\psi}^{m,h,s}_{K_j}\|_{L^2(K_{\delta_j})} \\ &+ \|\nabla\psi^{n,s}_{K_j} - \nabla\psi^{n,h,s}_{K_j}\|_{L^2(K_{\delta_j})}\|\nabla\overline{\phi}^{m,s}_{K_j} - \nabla\overline{\phi}^{m,h,s}_{K_j}\|_{L^2(K_{\delta_j})}\Big) \\ &\le C(h/\varepsilon)^2, \end{aligned}$$

where we used similar FEM estimates (as obtained for $\psi^{n,h,s}_{K_j}, \phi^{n,h,s}_{K_j}$) for $\overline{\psi}^{m,h,s}_{K_j}, \overline{\phi}^{m,h,s}_{K_j}$. This concludes the proof of (64).

We now focus on the proof of the estimate (65) where the formulation (17) is used. We notice that the Lipchitzness of the tensors $a(x, y, s)$, $\partial_u a(x, y, s)$ with respect to $x \in K_{\delta_j}$ yields for $k = 0, 1$,

$$\sup_{x \in K_{\delta_j}, s \in \mathbb{R}} \|\partial^k_u a(x, x/\varepsilon, s) - \partial^k_u a(x_{K_j}, x/\varepsilon, s)\|_F \le C\delta$$

Using the inequality (77) with $H = S^q(K_{\delta_j}, \mathcal{T}_h)$, this perturbation of the tensors $a, \partial_u a$ induces a perturbation of $\psi^{n,h,s}_{K_j}$ and $\phi^{n,h,s}_{K_j}$ of size $\le C\delta\sqrt{|K_{\delta_j}|}$, which yields

$$r'_{MOD} \le C\left(\left(\frac{h}{\varepsilon}\right)^2 + \delta\right)$$

and this concludes the proof of (65). □

References

[1] A. Abdulle, *On a-priori error analysis of Fully Discrete Heterogeneous Multiscale FEM*, SIAM Multiscale Model. Simul., 4, no. 2, (2005), 447–459.

[2] A. Abdulle, *Discontinuous Galerkin finite element heterogeneous multiscale method for elliptic problems with multiple scales*, to appear in Math. Comp.

[3] A. Abdulle, *The finite element heterogeneous multiscale method: a computational strategy for multiscale PDEs,* GAKUTO Int. Ser. Math. Sci. Appl., 31 (2009), 135–184.

[4] A. Abdulle, *A priori and a posteriori analysis for numerical homogenization: a unified framework*, to appear Ser. Contemporary Applied Mathematics, CAM, World Sci. Publishing, Singapore.

[5] A. Abdulle and A. Nonnenmacher, *A short and versatile finite element multiscale code for homogenization problems*, Comput. Methods Appl. Mech. Engrg. 198 (2009), 2839–2859.

[6] A. Abdulle and C. Schwab, *Heterogeneous multiscale FEM for diffusion problem on rough surfaces*, SIAM Multiscale Model. Simul., 3, no. 1 (2005), 195–220.

[7] A. Abdulle and G. Vilmart, *A priori error estimates for finite element methods with numerical quadrature for nonmonotone nonlinear elliptic problems,* submitted for publication. `http://infoscience.epfl.ch/record/152102`

[8] N. André and M. Chipot, *Uniqueness and nonuniqueness for the approximation of quasilinear elliptic equations*, SIAM J. Numer. Anal. 33 (5) (1996), 1981–1994.

[9] M. Artola and G. Duvaut, *Homogénéisation d'une classe de problèmes non linéaires,* C. R. Acad. Sci. Paris Sér. A-B 288 (1979), no. 16, 775–778.

[10] M. Artola and G. Duvaut, *Un résultat d'homogénéisation pour une classe de problèmes de diffusion non linéaires stationnaires*, Ann. Fac. Sci. Toulouse Math. (5) 4 (1982), no. 1, 1–28.

[11] J. Bear and Y. Bachmat, *Introduction to modelling of transport phenomena in porous media*, Kluwer Academic, Dordrecht, The Netherlands, 1991.

[12] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structure*, North Holland, Amsterdam, 1978.

[13] L. Bers, F. John, and M. Schechter, *Partial differential equations*, Lectures in Applied Mathematics, Proceedings of the Summer Seminar, Boulder, CO, 1957.

[14] L. Boccardo and F. Murat, Homogénéisation de problèmes quasi-linéaires, Publ. IRMA, Lille., 3 (1981), no. 7, 1351.

[15] S. Brenner and R. Scott, *The mathematical theory of finite element methods.* Third edition. Texts in Applied Mathematics, 15. Springer, New York, 2008.

[16] Z. Chen, W. Deng, and H. Ye, *Upscaling of a class of nonlinear parabolic equations for the flow transport in heterogeneous porous media*, Commun. Math. Sci. 3 (2005), no. 4, 493515.

[17] Z. Chen and T. Y. Savchuk, *Analysis of the multiscale finite element method for nonlinear and random homogenization problems,* SIAM J. Numer. Anal. 46 (2008), 260–279.

[18] M. Chipot, *Elliptic equations: an introductory course*, Birkhäuser Advanced Texts: Basler Lehrbücher. Birkhäuser Verlag, Basel, 2009.

[19] P.G. Ciarlet, *Basic error estimates for elliptic problems*, Handb. Numer. Anal., Vol. 2, North-Holland, Amsterdam (1991), 17–351.

[20] P.G. Ciarlet and P.A. Raviart, *The combined effect of curved boundaries and numerical integration in isoparametric finite element method*, in A. K Aziz (Ed), Math. Foundation of the FEM with Applications to PDE, Academic Press, New York, NY, (1972), 409–474.

[21] J. Douglas, Jr. and T. Dupont, *A Galerkin method for a nonlinear Dirichlet problem*, Math. Comp., 29 (131) (1975), 689–696.

[22] R. Du and P. Ming, *Heterogeneous multiscale finite element method with novel numerical integration schemes* Commun. Math. Sci., 8(4) (2010), 863–885.

[23] W. E and B. Engquist, *The Heterogeneous multi-scale methods*, Commun. Math. Sci., 1 (2003), 87–132.

[24] W. E, P. Ming and P. Zhang, *Analysis of the heterogeneous multiscale method for elliptic homogenization problems*, J. Amer. Math. Soc. 18 (2005), no. 1, 121–156.

[25] Y. Efendiev and T. Y. Hou, *Multiscale finite element methods. Theory and applications,* Surveys and Tutorials in the Applied Mathematical Sciences, 4, Springer, New York, 2009.

[26] Y. R. Efendiev, T. Hou, and V. Ginting, Multiscale nite element methods for nonlinear problems and their applications, Commun. Math. Sci., 2 (2004), 553–589.

[27] M. Feistauer, M. Křížek, and V. Sobotíková, *An analysis of finite element variational crimes for a nonlinear elliptic problem of a nonmonotone type*, East-West J. Numer. Math. 1 (4) (1993), 267–285.

[28] N. Fusco, and G. Moscariello, *On the homogenization of quasilinear divergence structure operators*, Ann. Mat. Pura Appl. (4) 146 (1987), 1–13.

[29] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals,* Springer-Verlag, 1994.

[30] A. Karageorghisa and D. Lesnicb, *Steady-state nonlinear heat conduction in composite materials using the method of fundamental solutions,* Comput. Methods Appl. Mech. Engrg. 197 (2008), no. 33-40, 3122–3137.

[31] O.A. Ladyzhenskaya, *The boundary value problems of mathematical physics*, Applied Mathematical Sciences, 49, Springer-Verlag New York Inc., 1985.

[32] A. M. Meirmanov, *The Stefan problem*, De Gruyter expositions in Mathematics 3, Berlin, 1992.

[33] J. A. Nitsche, *On L_∞-convergence of finite element approximations to the solution of a nonlinear boundary value problem*, Topics in numerical analysis, III (Proc. Roy. Irish Acad. Conf., Trinity Coll., Dublin, 1976), Academic Press, London-New York (1977) 317–325.

[34] J. Poussin and J. Rappaz, *Consistency, stability, a priori and a posteriori errors for Petrov-Galerkin methods applied to nonlinear problems*, Numer. Math., 69 (2) (1994), 213–231.

[35] A.G. Whittington, A.M. Hofmeister, and P.I. Nabelek, *Temperature-dependent thermal diffusivity of the Earths crust and implications for magmatism,* Nature 458 (2009), 319-321.

[36] J. Xu, *Two-grid discretization techniques for linear and nonlinear PDE*, SIAM J. Numer. Anal., 33, 5 (1996), 1759–1777.

Section de Mathématiques, École Polytechnique Fédérale de Lausanne,
Station 8, 1015 Lausanne, Switzerland

E-mail addresses: `Assyr.Abdulle@epfl.ch, Gilles.Vilmart@epfl.ch`