MATHEMATICS OF COMPUTATION
Volume 00, Number 0, Pages 000–000
S 0025-5718(XX)0000-0

ANALYSIS OF THE FINITE ELEMENT HETEROGENEOUS MULTISCALE METHOD FOR QUASILINEAR ELLIPTIC HOMOGENIZATION PROBLEMS

ASSYR ABDULLE AND GILLES VILMART

Abstract. An analysis of the finite element heterogeneous multiscale method for a class of quasilinear elliptic homogenization problems of nonmonotone type is proposed. We obtain optimal convergence results for dimension $d \leq 3$. Our results, which also take into account the microscale discretization, are valid for both simplicial and quadrilateral finite elements. Optimal a-priori error estimates are obtained for the H^1 and L^2 norms, error bounds similar as for linear elliptic problems are derived for the resonance error. Uniqueness of a numerical solution is proved. Moreover, 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.

1. Introduction

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

$$-\nabla \cdot (a^\varepsilon(x, u_\varepsilon(x))\nabla u_\varepsilon(x)) = f(x) \text{ in } \Omega, \tag{1.1}$$

in a domain $\Omega \subset \mathbb{R}^d$, $d \leq 3$, where $a^\varepsilon(x, u)$ is a $d \times d$ tensor. We consider for simplicity the homogeneous Dirichlet boundary conditions $u_\varepsilon = 0$ on $\partial\Omega$, but our analysis could apply to other types of boundary conditions. Such type of problems arise in many applications (e.g., the stationary form of the Richards problem [11], the modeling of the thermal conductivity of the Earth's crust [35], or the heat conduction in composite materials [32]).

Yet, often the multiscale nature of the medium, described in (1.1) through a nonlinear multiscale conductivity tensor $a^\varepsilon(x, u_\varepsilon(x))$, is not taken into account in the modeling due to the high computational cost in solving numerically (1.1) via standard methods resolving the medium's finest scale. Upscaling of equation (1.1) is thus needed for an efficient numerical treatment. Rigorously described by the mathematical homogenization theory [12],[31], 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 (1.1) has been developed in [14, 10, 29] 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 $a^0(x, u_0(x))$

2010 *Mathematics Subject Classification.* 65N30,65M60,74D10,74Q05.

The work of A.A. was supported in part by the Swiss National Science Foundation under Grant 200021 134716/1.

depending nonlinearly on a homogenized solution u_0 (the limit in a certain sense of u_ε as $\varepsilon \to 0$).

While numerical methods for linear elliptic homogenization problems have been studied in many papers - see [3, 24, 26, 8], 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) [26] for nonlinear elliptic problems of the form $-\nabla \cdot (a^\varepsilon(x, u_\varepsilon(x), \nabla u_\varepsilon(x)) = f(x)$ (a tensor nonlinear also with respect to ∇u_ε) have been studied in [27],[26], where a monotonicity assumption has been used to derive *convergence rates*. This assumption leads essentially to problems of the type $-\nabla \cdot (a^\varepsilon(x, \nabla u_\varepsilon(x)) = f(x)$. Following the two-grid discretization framework of [36], an analysis of the finite element heterogeneous multiscale method (FE-HMM) for the problem (1.1) has been proposed in [25] and in [18] for the multiscale finite element method (MsFEM). Simplicial finite elements were considered in both aforementioned work.

Unfortunately, there are several critical issues with the analysis of the FE-HMM for the problem (1.1) in [25] that are addressed in the present paper. In addition, several new results are derived (analysis for quadrilateral FEs, L^2 error estimates, improved resonance error analysis, convergence of the Newton method used to compute the solution). Finally we note that all our results are valid for a fully discrete formulation, taking into account also the micro scale discretization error. For the convenience of the reader we briefly discuss the main issues of the analysis in [25] and briefly discuss our contributions.

The first major issue in [25] resides in the treatment of the variational crime that arises when using the FE-HMM. [1] As an intermediate step, one needs to estimate $|A(u^H; u^H, w^H) - A_H(u^H; u^H, w^H)|$, where $A(u^H; u^H, w^H) = \int_\Omega a^0(x, u^H)\nabla u^H \nabla w^H dx$ is the weak form for the exact problem and A_H a corresponding nonlinear form based on numerical quadrature. In [25, equ. 5.21] the estimate $|A(u^H; u^H, w^H) - A_H(u^H; u^H, w^H)| \leq CH^\ell \|w^H\|_{H^1(\Omega)}$ is used. However, C depends (in a nonlinear way) 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 [25]. Using $W^{1,\infty}$ estimates in [25], recent results [7, Prop. 2] on FEM with numerical quadrature for nonlinear nonmonotone problems and an inverse inequality, it is possible to bound u^H for $\mathcal{P}^1$ and $\mathcal{P}^2$ triangular finite elements. However this argument does not apply for $\mathcal{P}^\ell$-simplicial FEs when $\ell > 2$ and we don't know how to derive such bound in general (notice that our new approach does not rely such bounds).

The second major issue in the analysis of [25] resides in the use [25, Lem. 5.3] of a discrete Green function G_H^z for an error estimate in the $W^{1,\infty}$ norm. The logarithmic bound $\sup_{z\in\overline{\Omega}} \|G_H^z\|_{W^{1,1}(\Omega)} \leq C|\log H|$ (see [25, equ.(5.16)]) is used in the main a priori error estimate result [25, Thm. 5.4]. However, such a logarithmic estimate is not available, to the best of our knowledge, in dimension $d = 3$ for arbitrary bounded convex polyhedral domains. Thus, the results in [25] are not valid for the dimension $d = 3$.

Both aforementioned issues are addressed in our analysis that is valid for $\mathcal{P}^\ell$-simplicial FEs in dimension $1 \leq d \leq 3$. In addition, we also derive several new

[1] We recall that this method couples a macroscopic solver based on FEM with numerical quadrature, with microscopic solvers based on FEM defined on sampling domains that recover locally the missing macroscopic input data.

results: optimal error estimates are derived for $\mathcal{Q}^\ell$-rectangular FEs (for such elements we don't know how to obtain error estimates using the framework in [25] even for the lowest order piecewise bilinear elements), optimal L^2 error estimates are derived for both $\mathcal{P}^\ell$ and $\mathcal{Q}^\ell$ FEs (notice that we cannot simply use the Aubin-Nitche duality argument but need to study linear indefinite elliptic problems with numerical quadrature arising from the linearization of (1.1)), improved convergence rates for the so-called modeling or resonance error r_{MOD} are obtained (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 [25, Thm. 5.5] [2]), the Newton method used in practice to compute a solution of the nonlinear discretized problem is shown to converge providing hence a uniqueness result for our numerical scheme. Finally all our results are derived for a fully discrete FE-HMM scheme, where the errors at both the microscopic and the macroscopic grid are taken into account. The fully discrete error bounds are also optimal in the microscopic convergence rates. Our uniqueness result is also established in this fully discrete setting and it requires a new estimate of the micro error for a modified micro problem (based on the derivative of the effective tensor). Thus, the convergence of the Newton method for sufficiently fine macro and micro meshes is also guaranteed in the fully discrete setting.

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 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.

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 (1.1), 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 continuous, bounded on $\overline{\Omega} \times \mathbb{R}$, uniformly elliptic, and uniformly Lipschitz continuous with respect to s, with constants independent of the parameter ε. We further assume that $f \in H^{-1}(\Omega)$. Under the above assumptions, for all fixed $\varepsilon > 0$, the weak form of (1.1) has a unique solution $u_\varepsilon \in H^1_0(\Omega)$ (see for example [19, Theorem 11.6]),which satisfies the bound $\|u_\varepsilon\|_{H^1(\Omega)} \leq C\|f\|_{H^{-1}(\Omega)}$. 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 [29]): 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 the solution of the so-called homogenized problem

$$-\nabla \cdot \big(a^0(x, u_0(x))\nabla u_0(x)\big) = f(x) \text{ in } \Omega, \qquad u_0(x) = 0 \text{ on } \partial\Omega, \tag{2.1}$$

[2] Here ε is the size of the period and δ the length, in each spatial direction, of the sampling domains.

and where the tensor $a^0(x,s)$ is called the homogenized tensor. It can be shown [14, Prop. 3.5] that the homogenized tensor is Lipschitz continuous with respect to s, uniformly elliptic, and bounded. Precisely, there exists $\Lambda_1 > 0$ such that

$$|a^0_{mn}(x,s_1) - a^0_{mn}(x,s_2)| \leq \Lambda_1 |s_1 - s_2|, \quad \forall x, \forall s_1, s_2 \in \mathbb{R}, \forall m,n = 1,\ldots,d, \tag{2.2}$$

and there exist $\lambda, \Lambda_0 > 0$ such that

$$\lambda \|\xi\|^2 \leq a^0(x,s)\xi \cdot \xi, \quad \|a^0(x,s)\xi\| \leq \Lambda_0 \|\xi\|, \quad \forall \xi \in \mathbb{R}^d, \forall s \in \mathbb{R}, \forall x \in \Omega. \tag{2.3}$$

Under these assumptions, the homogenized problem (2.1) has also a unique solution $u_0 \in H^1_0(\Omega)$.

We further assume for the analysis that the coefficients of the homogenized tensor are continuous,

$$a^0_{mn} \in C^0(\overline{\Omega} \times \mathbb{R}), \qquad \forall m,n = 1,\ldots,d. \tag{2.4}$$

Let us further mention the following characterization of the homogenized tensor, instrumental to derive the homogenization result. Let $\{a^\varepsilon(\cdot,s)\}$ be the subsequence of tensors considered above, then for all fixed real parameters s, the tensor $a^0(\cdot,s)$ is the homogenized tensor for the linear problem

$$-\nabla \cdot (a^\varepsilon(x,s)\nabla v_\varepsilon(x)) = f(x) \text{ in } \Omega, \quad v_\varepsilon(x) = 0 \text{ on } \partial\Omega. \tag{2.5}$$

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 the weak convergence of u^ε to the solution of (2.1) holds for the whole sequence. The homogenized tensor can be characterized in the following way [10]:

$$a^0(x,s) = \int_Y a(x,y,s)(I + J^T_{\chi(x,y,s)})dy, \qquad \text{for } x \in \Omega, s \in \mathbb{R}, \tag{2.6}$$

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

$$\int_Y a(x,y,s)\nabla_y \chi^i(x,y,s) \cdot \nabla w(y)dy = -\int_Y a(x,y,s)\mathbf{e}_i \cdot \nabla w(y)dy, \quad \forall w \in W^1_{per}(Y), \tag{2.7}$$

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 [24]. The numerical method is based on a macroscopic FEM defined on QF and linear microscopic FEMs recovering the missing macroscopic tensor at the macroscopic quadrature points.

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. For each partition $\mathcal{T}_H$, we define a FE space

$$S^\ell_0(\Omega, \mathcal{T}_H) = \{v^H \in H^1_0(\Omega);\ v^H|_K \in \mathcal{R}^\ell(K),\ \forall K \in \mathcal{T}_H\}, \tag{2.8}$$

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

$$S^q(K_{\delta_j}, \mathcal{T}_h) = \{z^h \in W(K_{\delta_j});\ z^h|_Q, \in \mathcal{R}^q(Q),\ Q \in \mathcal{T}_h\}, \tag{2.9}$$

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

$$W(K_{\delta_j}) = W^1_{per}(K_{\delta_j}) = \{z \in H^1_{per}(K_{\delta_j});\ \int_{K_{\delta_j}} z dx = 0\} \tag{2.10}$$

for a periodic coupling or

$$W(K_{\delta_j}) = H^1_0(K_{\delta_j}) \tag{2.11}$$

for a coupling through Dirichlet boundary conditions. Here $H^1_{per}(K_{\delta_j})$ is defined as the closure in H^1 of $\mathcal{C}^\infty_{per}(K_{\delta_j})$ (the subset of smooth periodic function on K_{δ_j}). The choice of the Sobolev space $W(K_{\delta_j})$ sets the coupling condition between macro and micro solvers. 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

$$w^H_{lin,j}(x) = w^H(x_{K_j}) + (x - x_{K_j}) \cdot \nabla w^H(x_{K_j}) \tag{2.12}$$

at the integration point x_{K_j}. For all real parameters 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

$$\int_{K_{\delta_j}} a^\varepsilon(x, s) \nabla w^{h,s}_{K_j}(x) \cdot \nabla z^h(x) dx = 0 \quad \forall z^h \in S^q(K_{\delta_j}, \mathcal{T}_h). \tag{2.13}$$

2.1.4. *Finite element heterogeneous multiscale method (FE-HMM).*. We have now all the ingredients to define our multiscale method. Find $u^H \in S^\ell_0(\Omega, \mathcal{T}_H)$ such that

$$B_H(u^H; u^H, w^H) = F_H(w^H), \qquad \forall w^H \in S^\ell_0(\Omega, \mathcal{T}_H), \tag{2.14}$$

where
(2.15)
$$B_H(u^H;v^H,w^H) := \sum_{K\in\mathcal{T}_H}\sum_{j=1}^{J}\frac{\omega_{K_j}}{|K_{\delta_j}|}\int_{K_{\delta_j}} a^\varepsilon(x,u^H(x_{K_j}))\nabla v_{K_j}^{h,u^H(x_{K_j})}(x)\cdot\nabla w_{K_j}^{h,u^H(x_{K_j})}(x)dx,$$

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 (2.13) with parameter $s=u^H(x_{K_j})$ (and similarly for $v_{K_j}^{h,u^H(x_{K_j})}$). 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[3] [21, Thm. 4]

$$|F_H(w^H)-F(w^H)| \le CH^\ell\|w^H\|_{H^1(\Omega)},\ \forall w^H\in S_0^\ell(\Omega,\mathcal{T}_H). \tag{2.16}$$

If in addition $f\in W^{\ell+1,p}(\Omega)$, then [21, Thm. 5]

$$|F_H(w^H)-F(w^H)| \le CH^{\ell+1}\Big(\sum_{K\in\mathcal{T}_H}\|w^H\|^2_{H^2(K)}\Big)^{1/2},\ \forall w^H\in S_0^\ell(\Omega,\mathcal{T}_H). \tag{2.17}$$

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
(2.18)
$$\widetilde{B}_H(u^H;v^H,w^H) := \sum_{K\in\mathcal{T}_H}\sum_{j=1}^{J}\frac{\omega_{K_j}}{|K_{\delta_j}|}\int_{K_{\delta_j}} a(x_{K_j},\frac{x}{\varepsilon},u^H(x_{K_j}))\nabla v_{K_j}^{h,u^H(x_{K_j})}(x)\cdot\nabla w_{K_j}^{h,u^H(x_{K_j})}(x)dx,$$

where $w_{K_j}^{h,u^H(x_{K_j})}$ is the solution of the micro problem (2.13) with the 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 (2.15), 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 (2.14). We first recall here a result for the analysis of the FE-HMM, shown in [1], [25] 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.

Lemma 2.1. *Assume that* (**Q1**) *holds and that the tensor* a^ε *satisfies* (2.2),(2.3),(2.4). *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*
(2.19)
$$\lambda\|v^H\|^2_{H^1(\Omega)} \le B_H(z^H;v^H,v^H),\quad |B_H(z^H;v^H,w^H)| \le \Lambda_0\|v^H\|_{H^1(\Omega)}\|w^H\|_{H^1(\Omega)},$$
for all $z^H,v^H,w^H\in S_0^\ell(\Omega,\mathcal{T}_H)$. *Analogous formulas also hold for the modified bilinear form* $\widetilde{B}_H(z^H;\cdot,\cdot)$ *defined in* (2.18).

Notice at this stage that in Lemma 2.1 no structure assumption (as for example local periodicity) is required for the tensor a^ε.

Since the micro problems (2.13) are linear with a uniformly bounded and coercive tensor (2.3), their solutions $w_{K_j}^{h,s}\in S^q(K_{\delta_j},\mathcal{T}_h)$ are always uniquely defined, in

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

particular there is no restriction on the mesh size h. The macro solution u^H of the FE-HMM is a solution of the nonlinear system (2.14) and the existence of a solution u^H of (2.14) follows from a classical fixed point argument.

Theorem 2.2. *Assume that the bilinear form $B_H(z^H;\cdot,\cdot)$, $z^H \in S_0^\ell(\Omega,\mathcal{T}_H)$, defined in* (2.15) *is uniformly elliptic and bounded* (2.19)*, 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* (2.14) *possesses at least one solution $u^H \in S_0^\ell(\Omega,\mathcal{T}_H)$. In addition, $\|u^H\|_{H^1(\Omega)} \leq C\|f\|_{W^{\ell,p}(\Omega)}$ where C is independent of H.*

The proof of Theorem 2.2 follows standard arguments ([22], see also [16]). 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

$$B_H(z^H; S_H z^H, w^H) = F_H(w^H),\ \forall w^H \in S_0^\ell(\Omega,\mathcal{T}_H). \tag{2.20}$$

In contrast, the proof of the uniqueness of a solution u^H is non trivial (see Theorem 3.3).

2.2. **Reformulation of the FE-HMM.** A straightforward computation shows that for all scalars s, the solution $w_{K_j}^{h,s}$ of the linear cell problem (2.13) is given by

$$w_{K_j}^{h,s}(x) = w_{lin,j}^H(x) + \sum_{i=1}^d \psi_{K_j}^{i,h,s}(x)\frac{\partial w_{lin,j}^H}{\partial x_i}, \qquad \text{for } x \in K_{\delta_j}, \tag{2.21}$$

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

$$\int_{K_{\delta_j}} a^\varepsilon(x,s)\nabla\psi_{K_j}^{i,h,s}(x)\cdot\nabla z^h(x)dx = -\int_{K_{\delta_j}} a^\varepsilon(x,s)\mathbf{e}_i\cdot\nabla z^h(x)dx,\ \forall z^h \in S^q(K_{\delta_j},\mathcal{T}_h), \tag{2.22}$$

where $S^q(K_{\delta_j},\mathcal{T}_h)$ is defined in (2.9) with either periodic or Dirichlet boundary conditions.

We also consider for the analysis the following problems (2.23), (2.24), which are analogue to (2.13),(2.22), but without FEM discretization (i.e. with test functions in the space $W(K_{\delta_j})$ defined in (2.10) or (2.11)): find $w_{K_j}^s$ such that $(w_{K_j}^s - w_{lin,j}^H) \in W(K_{\delta_j})$ and

$$\int_{K_{\delta_j}} a^\varepsilon(x,s)\nabla w_{K_j}^s(x)\cdot\nabla z(x)dx = 0 \quad \forall z \in W(K_{\delta_j}). \tag{2.23}$$

Similarly to (2.21), it can be checked that the unique solution of problem (2.23) is given by (2.21) replacing $\psi_{K_j}^{i,h,s}(x)$ with $\psi_{K_j}^{i,s}(x)$, where for each scalar s and for each $\mathbf{e}_i$, $\psi_{K_j}^{i,s}$ are the solutions of the following problem: find $\psi_{K_j}^{i,s} \in W(K_{\delta_j})$ such that

$$\int_{K_{\delta_j}} a^\varepsilon(x,s)\nabla\psi_{K_j}^{i,s}(x)\cdot\nabla z(x)dx = -\int_{K_{\delta_j}} a^\varepsilon(x,s)\mathbf{e}_i\cdot\nabla z(x)dx,\ \forall z \in W(K_{\delta_j}). \tag{2.24}$$

Consider for all scalars s the two tensors

$$a^0_{K_j}(s) := \frac{1}{|K_{\delta_j}|}\int_{K_{\delta_j}} a^\varepsilon(x,s)\left(I + J^T_{\psi^{h,s}_{K_j}(x)}\right)dx, \tag{2.25}$$

$$\bar{a}^0_{K_j}(s) := \frac{1}{|K_{\delta_j}|}\int_{K_{\delta_j}} a^\varepsilon(x,s)\left(I + J^T_{\psi^{s}_{K_j}(x)}\right)dx, \tag{2.26}$$

where $J_{\psi^s_{K_j}(x)}$ and $J_{\psi^{h,s}_{K_j}(x)}$ are $d\times d$ matrices with entries $(J_{\psi^s_{K_j}(x)})_{i\ell} = (\partial\psi^{i,s}_{K_j})/(\partial x_\ell)$ and $(J_{\psi^{h,s}_{K_j}(x)})_{i\ell} = (\partial\psi^{i,h,s}_{K_j})/(\partial x_\ell)$, respectively. The Lemma 2.3 below has been proved in [6],[2] in the context of linear elliptic problems and is a consequence of (2.21) (for $w^{h,s}_{K_j}(x)$ and $w^s_{K_j}(x)$). It permits to interpret (2.14)-(2.15) as a standard FEM applied with a modified tensor.

Lemma 2.3. *Assume that the tensors a^0, a^ε satisfy* (2.3),(2.4). *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:*

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

where $v^{h,s}_{K_j}$, $v^s_{K_j}$ are the solutions of (2.13), (2.23), *respectively, and the tensors $a^0_{K_j}(s)$, $\bar{a}^0_{K_j}(s)$ are defined in* (2.25), (2.26). *Analogous formulas also hold for the terms in the right-hand side of* (2.18), *with $a^\varepsilon(x,s)$ replaced by $a(x_{K_j}, x/\varepsilon, s)$ in the above two identities and in* (2.13),(2.22),(2.23),(2.24),(2.25).

3. Main results

We summarize here the main results of this paper. Given a solution u^H of (2.14) 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 (2.1) and to prove the uniqueness of a numerical solution u^H. We shall consider, for a^0 the homogenized tensor in (2.1) and $a^0_{K_j}$ defined in (2.25), the quantity

$$r_{HMM} := \sup_{K\in\mathcal{T}_H, x_{K_j}\in K, s\in\mathbb{R}} \|a^0(x_{K_j},s) - a^0_{K_j}(s)\|_F, \tag{3.1}$$

where $\|a\|_F = (\sum_{m,n}|a_{mn}|^2)^{1/2}$ denotes the Frobenius norm of a $d\times d$ matrix a.

In what follows we shall assume that the family of triangulations $\{\mathcal{T}_H\}$ satisfies the inverse assumption $H/H_K \le C$ for all $K\in\mathcal{T}_H$ and all $\mathcal{T}_H$. Notice that such an inverse assumption is often assumed for the analysis of FEM for non-linear problems [34, 28, 36, 25, 17]. In our analysis it is used only in the proof of an L^2 estimate (see Lemma 4.2) and for the uniqueness of the numerical solution (Sect. 4.3).

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* (2.1). *Let $\ell \ge 1$. Let $\mu = 0$ or* 1. *Assume (***Q1***), (***Q2***),* (2.16), *and*

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

In addition to (2.2),(2.3),(2.4)*, 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* $m, n = 1 \dots d$*.*

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

$$H \le H_0 \quad \text{and} \quad r_{HMM} \le r_0, \tag{3.2}$$

any solution u^H *of* (2.14) *satisfies*

$$\|u_0 - u^H\|_{H^1(\Omega)} \le C(H^\ell + r_{HMM}) \quad \text{if } \mu = 0, 1, \tag{3.3}$$

$$\|u_0 - u^H\|_{L^2(\Omega)} \le C(H^{\ell+1} + r_{HMM}) \text{ if } \mu = 1 \text{ and (2.17) holds.} \tag{3.4}$$

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 [25, Thm 5.4], Theorem 3.1 is also valid for $d = 3$ and arbitrary high order simplicial and quadrilateral macro FEs. We emphasize that the constants H_0 and r_0 in Theorem 3.1 are independent of H, h, ε, δ. Thus, the framework used to derive Theorem 3.1 allows to re-use results obtained for linear problems to derive a fully discrete error analysis, where the micro FE discretization errors are also taken into account.

Remark 3.2. Except for the $W^{1,\infty}$ assumption on u^0 and the smoothness[4] of $s \mapsto a^0(x,s)$ assumed to treat the non-linearity (as in [22] for one-scale nonmonotone standard FEM), the smoothness assumptions of Theorem 3.1 on the homogenized problem are identical to those classically assumed for one-scale linear FEM [21],[20, Sect. 29]. Notice that the H^1 estimate (3.3) and the uniqueness of u^H can be proved straightforwardly provided (3.2) (without assuming $W^{1,\infty}$ regularity of u^0 and quasi-uniform meshes) if $C\lambda^{-1}\Lambda_1\|u^0\|_{H^2(\Omega)} < 1$ (see [7, Theorem 4]).

For the uniqueness result, we shall consider the quantity

$$r'_{HMM} := \sup_{K \in \mathcal{T}_H, x_{K_j} \in K, s \in \mathbb{R}} \left\| \frac{d}{ds} \left(a^0(x_{K_j}, s) - a^0_{K_j}(s) \right) \right\|_F . \tag{3.5}$$

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

(3.6)
$s \in \mathbb{R} \mapsto a^\varepsilon(\cdot, s) \in (W^{1,\infty}(\Omega))^{d \times d}$ is of class C^2 and $|\partial_u^k a^\varepsilon(\cdot, s)|_{W^{1,\infty}(\Omega)} \le C\varepsilon^{-1}$, $k \le 2$,

where C is independent of s and ε.

Theorem 3.3. *Assume that the hypotheses of Theorem 3.1 and* (3.6) *hold. Then, there exist positive constants* H_0, r_0 *such that if*

$$H \le H_0 \qquad \text{and} \qquad H^{-1/2} r_{HMM} + r'_{HMM} \le r_0 \tag{3.7}$$

the solution u^H *of* (2.14) *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.

---

[4]Notice that in the locally periodic case (see assumption (**H2**) below), the C^2 regularity of $s \mapsto a^0(x,s)$ can be shown using assumption (3.6) with the ideas of Lemma 6.1.

Corollary 3.4. In addition to the hypotheses 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 (2.18) is used for the solution u^H of (2.14). Then, there exists positive constants H_0, r_0 such that if

$$H \leq H_0 \qquad \text{and} \qquad H^{-1/2}(h/\varepsilon)^{2q} \leq r_0 \tag{3.8}$$

the solution u^H of (2.14) is unique.

Remark 3.5. Inspecting the proofs of Theorem 3.3 and Corollary 3.4 (postponed to Sect. 4.3) shows that in dimension $d = 2$, the quantity $H^{-1/2}$ in (3.7),(3.8) can be replaced by $(1 + |\log H|)^{1/2}$. In Corollary 3.4, if we use the form (2.15) for the solution u^H of (2.14), to obtain the uniqueness of u^H, we need to assume in addition that δ is small enough ($\varepsilon \leq \delta \leq CH^{1/2}$). Notice also that (3.8) is automatically satisfied if $(h/\varepsilon)^{2q} \leq CH \leq H_0$ with H_0 small enough.

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

$$r_{HMM} \leq \underbrace{\sup_{K\in\mathcal{T}_H, x_{K_j}\in K, s\in\mathbb{R}} \|a^0(x_{K_j}, s) - \bar{a}^0_{K_j}(s))\|_F}_{r_{MOD}} + \underbrace{\sup_{K\in\mathcal{T}_H, x_{K_j}\in K, s\in\mathbb{R}} \|\bar{a}^0_{K_j}(s) - a^0_{K_j}(s))\|_F}_{r_{MIC}}, \tag{3.9}$$

where $\bar{a}^0_{K_j}$ is the tensor defined in (2.26). 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 (2.24) 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 $|a^\varepsilon_{ij}|_{W^{1,\infty}(K)} \leq C\varepsilon^{-1}$, for all parameters s with C independent of ε, s, then classical H^2 regularity results ([33, 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 (2.11) are used. If a^ε_{ij} is locally periodic, we can also use periodic boundary conditions (2.10) and analogous 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 (2.24) 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 $i = 1 \ldots d$. We make the same assumption with the tensor a^ε replaced by $(a^\varepsilon)^T$ in (2.24).

Theorem 3.6. *Under the assumptions of Theorem 3.1 and (***H1***), it holds (for $\mu = 0$ or 1)*

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

where for $\mu = 1$ we also assume (2.17) and we use the notation $H^0(\Omega) = L^2(\Omega)$. Here, C is independent of $H, h, \varepsilon, \delta$.

To estimate further the modeling error r_{MOD} defined in (3.9), we need more structure assumptions on a^ε. Here we assume local periodicity as encoded in the following assumption.

(**H2**) for all $m,n=1,\ldots,d$, 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. *Under the assumptions of Theorem 3.1, (***H1***) and (***H2***), it holds (for $\mu=0$ or 1)*

(3.10)
$$\|u_0-u^H\|_{H^{1-\mu}(\Omega)}\le\begin{cases} C(H^{\ell+\mu}+(\frac{h}{\varepsilon})^{2q}+\delta), & \text{if } W(K_{\delta_j})=W^1_{per}(K_{\delta_j}) \text{ and } \frac{\delta}{\varepsilon}\in\mathbb{N}^*,\\ C(H^{\ell+\mu}+(\frac{h}{\varepsilon})^{2q}), & \text{if } W(K_{\delta_j})=W^1_{per}(K_{\delta_j}) \text{ and } \frac{\delta}{\varepsilon}\in\mathbb{N}^*, \\ & \text{with collocated tensor (see (2.18))},\\ C(H^{\ell+\mu}+(\frac{h}{\varepsilon})^{2q}+\delta+\frac{\varepsilon}{\delta}), & \text{if } W(K_{\delta_j})=H^1_0(K_{\delta_j})\ (\delta>\varepsilon),\end{cases}$$

where for $\mu=1$ we also assume (2.17). The constants C are independent of H,h,ε,δ.

For non periodic problems, we note that the modeling error for linear elliptic problems with random coefficients has been studied in [25, Appendix A]. Related work can be found in [37, 15, 30].

Remark 3.8. While the convergence $u_\varepsilon\to u_0$ (up to a subsequence) is strong in $L^2(\Omega)$, it is only weak in $H^1(\Omega)$ 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 [12],[31] to recover the oscillating solution u_ε. In [14, Sect. 3.4.2], it is shown that any corrector for the linear problem (1.1) where the tensor is evaluated at u_0 instead of u_ε, is also a corrector for the solution u_ε of the nonlinear problem (1.1). In our situation, we have $\nabla r_\varepsilon\to 0$ strongly in $(L^1_{loc}(\Omega))^d$ where $r_\varepsilon(x):=u_\varepsilon(x)-u_0(x)-u_{1,\varepsilon}(x)$.

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 proof of Theorem 3.1. Consider for $z^H,v^H,w^H\in S^\ell_0(\Omega,\mathcal{T}_H)$,

$$A_H(z^H;v^H,w^H):=\sum_{K\in\mathcal{T}_H}\sum_{j=1}^{J}\omega_{K,j}a^0(x_{K_j},z^H(x_{K_j}))\nabla v^H(x_{K_j})\cdot\nabla w^H(x_{K_j}), \tag{4.1}$$

where $a^0(x,s)$ is the tensor defined in (2.1) and let $\widetilde{u}^H_0$ be a solution of

$$A_H(\widetilde{u}^H_0;\widetilde{u}^H_0,w^H)=F_H(w^H),\qquad \forall w^H\in S^\ell_0(\Omega,\mathcal{T}_H). \tag{4.2}$$

The problem (4.2) is a standard FEM with numerical quadrature for the problem (2.1). 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

$$Q_H(z^H):=\sup_{w^H\in S^\ell_0(\Omega,\mathcal{T}_H)}\frac{|A_H(z^H,z^H,w^H)-F_H(w^H)|}{\|w^H\|_{H^1(\Omega)}},\quad \forall z^H\in S^\ell_0(\Omega,\mathcal{T}_H). \tag{4.3}$$

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] (following the lines of Lemma 4, Lemma 6 and Theorem 3 in [7], respectively) and are therefore omitted.

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

$$\|u_0 - z^H\|_{H^1(\Omega)} \leq C(H^\ell + Q_H(z^H) + \|u_0 - z^H\|_{L^2(\Omega)}), \tag{4.4}$$

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

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

$$\|u_0 - z^H\|_{L^2(\Omega)} \leq C(H^{\ell+\mu} + Q_H(z^H) + \|u_0 - z^H\|^2_{H^1(\Omega)}), \tag{4.5}$$

where C is independent of H and z^H.

Lemma 4.3. *Under the assumptions of Theorem 3.1, 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, $\|u_0 - z^H\|_{L^2(\Omega)} \to 0$ for $H \to 0$.*

We next notice that $Q_H(z^H)$ can be bounded in terms of r_{HMM} defined in (3.1).

Lemma 4.4. *Assume that the tensors a^0, a^ε satisfy (2.3),(2.4). Then*

$$Q_H(z^H) \leq C r_{HMM} \|z^H\|_{H^1(\Omega)} + \sup_{w^H \in S_0^\ell(\Omega, \mathcal{T}_H)} \frac{|B_H(z^H, z^H, w^H) - F_H(w^H)|}{\|w^H\|_{H^1(\Omega)}}, \tag{4.6}$$

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

Proof. The proof is a consequence of the inequality

$$|A_H(z^H, z^H, w^H) - F_H(w^H)| \leq |(A_H - B_H)(z^H; z^H, w^H)| + |B_H(z^H, z^H, w^H) - F_H(w^H)|.$$

Using Lemma 2.3 and the Cauchy-Schwarz inequality, the first term is the above right-hand side can be bounded by

$$|(A_H - B_H)(z^H; z^H, w^H)| \leq C \sup_{K \in \mathcal{T}_H, x_{K_j} \in K s \in \mathbb{R}} \|a^0(x_{K_j}, s)) - a^0_{K_j}(s))\|_F \|z^H\|_{H^1(\Omega)} \|w^H\|_{H^1(\Omega)} \tag{4.7}$$

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 (2.14). Then $Q_H(u^H) \leq C r_{HMM}$, where $Q_H(u^H)$ is defined in (4.3) and the constant C is independent of H, h, δ.

Proof. Follows from Lemma 4.4 and Theorem 2.2. □

Proof of Theorem 3.1. We apply Lemmas 4.1, 4.2, 4.3 with $z^H = u^H$, the solution of (2.14). Let $\mu = 0$. This yields, for all H small enough

$$\|u^H - u_0\|_{H^1(\Omega)} \leq C(H^\ell + r_{HMM} + \|u^H - u_0\|_{L^2(\Omega)}), \tag{4.8}$$

$$\|u^H - u_0\|_{L^2(\Omega)} \leq C(H^\ell + r_{HMM} + \|u^H - u_0\|^2_{H^1(\Omega)}), \tag{4.9}$$

$$\|u^H - u_0\|_{L^2(\Omega)} \to 0 \text{ for } (H, r_{HMM}) \to 0, \tag{4.10}$$

where we recall that $Q_H(u^H) \leq C r_{HMM}$. Substituting (4.9) into (4.8), 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

$$(1 - C\|u^H - u_0\|_{H^1(\Omega)})\|u^H - u_0\|_{H^1(\Omega)} \leq C(H^\ell + r_{HMM}). \tag{4.11}$$

Using (4.8) and (4.10), 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 (4.11), independently of the choice of the particular solution u^H. This concludes the proof of (3.3) for H and r_{HMM} small enough. For $\mu = 1$, the inequality (4.9) 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 (3.3) yields (3.4). □

4.2. **Explicit convergence rates for the micro and modeling error.** In this section we give the proofs of Theorems 3.6 and 3.7. For that, we need to quantify r_{HMM} defined in (3.1) and involved in Theorem 3.1. In view of the decomposition (3.9) 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 [23]. 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, a^ε satisfy* (2.3),(2.4) *and (***H1***). Then $r_{MIC} \leq C(h/\varepsilon)^{2q}$, where C is independent of H, h, δ, ε.*

Proof. From Lemma 2.3 and (2.22),(2.24), we deduce

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

where $\overline{\psi}^{n,i}_{K_j}$, $i = 1, \ldots, d$ denote the solutions of (2.24) with $a^\varepsilon(x, s)$ replaced by $a^\varepsilon(x, s)^T$. Using (2.22), (2.24), 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 (2.22) with $a^\varepsilon(x, s)$ replaced by $a^\varepsilon(x, s)^T$), and we obtain

$$(\bar{a}^0_{K_j}(s) - a^0_{K_j}(s))_{mn} = \frac{-1}{|K_{\delta_j}|}\int_{K_{\delta_j}} a^\varepsilon(x, s)\left(\nabla\psi^{n,s}_{K_j} - \nabla\psi^{n,h,s}_{K_j}\right)\cdot(\nabla\overline{\psi}^{m,s}_{K_j} - \nabla\overline{\psi}^{m,h,s}_{K_j})dx \tag{4.12}$$

Using the regularity assumption (**H1**) and standard FE results [20, 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 analogous estimates for $\nabla\overline{\psi}^{m,s}_{K_j}$. This combined with (4.12) and the Cauchy-Schwarz inequality concludes the proof. □

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 [25, 23] (for the estimates $r_{MOD} \leq C(\delta + \varepsilon/\delta)$ and $r_{MOD} \leq C\delta$) and in [6] (for the estimates $r_{MOD} = 0$), in the context of linear elliptic homogenization problems. Periodic and Dirichlet micro boundary conditions are discussed.

Lemma 4.7. *Assume* (2.2),(2.3),(2.4), *and* (**H2**). *Consider the homogenized tensor* $a^0(x,s)$ *and the tensor* $\overline{a}^0_{K_j}(s)$ *defined in* (2.26) *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* $r_{MOD} \leq C\delta$.
- *If in addition, the tensor* $a^\varepsilon(x,s)$ *is collocated at* $x = x_{K_j}$ *(i.e. using* (2.15)*) then* $r_{MOD} = 0$.
- *If* $W(K_{\delta_j}) = H^1_0(K_{\delta_j})$ $(\delta > \varepsilon)$, *then* $r_{MOD} \leq C(\delta + \varepsilon/\delta)$.

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

Proof. All the estimates of Lemma 4.7 are already known in the context of linear problems [25, 6, 23]. Using the characterization (2.6), 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 (2.14) 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 (2.14) (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

$$\partial B_H(z^H; v^H, w^H) := B_H(z^H; v^H, w^H) + B'_H(z^H; v^H, w^H), \tag{4.13}$$

where using Lemma 2.3,

$$B'_H(z^H; v^H, w^H) = \sum_{K\in\mathcal{T}_H}\sum_{j=1}^J \omega_{K_j}\frac{d}{ds}a^0_{K_j}(s)|_{s=z^H(x_{K_j})}v^H(x_{K_j})\nabla z^H(x_{K_j})\cdot\nabla w^H(x_{K_j}). \tag{4.14}$$

The Newton method for approximating a solution u^H of the nonlinear FE-HMM (2.14) by a sequence $\{u^H_k\}$ reads in weak form

$$\partial B_H(u^H_k; u^H_{k+1} - u^H_k, w^H) = F_H(w^H) - B_H(u^H_k; u^H_k, w^H), \quad \forall w^H \in S^\ell_0(\Omega, \mathcal{T}_H). \tag{4.15}$$

In order for B'_H to be well defined, we need, in addition to (2.2),(2.3),(2.4), the assumption (3.6). We also consider

$$A'_H(z^H; v^H, w^H) = \sum_{K\in\mathcal{T}_H}\sum_{j=1}^J \omega_{K_j}\frac{d}{ds}\, a^0(x_{K_j}, s)\big|_{s=z^H(x_{K_j})}\, v^H(x_{K_j})\nabla z^H(x_{K_j})\cdot\nabla w^H(x_{K_j}). \tag{4.16}$$

and A_H as defined in (4.1). Then, by replacing in (4.15) B_H by A_H and ∂B_H by ∂A_H we obtain the Newton method for solving (4.2) (standard FEM with numerical integration)

$$\partial A_H(z^H_k; z^H_{k+1} - z^H_k, w^H) = F_H(w^H) - A_H(z^H_k; z^H_k, w^H), \quad \forall w^H \in S^\ell_0(\Omega, \mathcal{T}_H), \tag{4.17}$$

where $\partial A_H(z^H; v^H, w^H) := A_H(z^H; v^H, w^H) + A'_H(z^H; v^H, w^H)$. We prove in Lemma 4.11 below that the iteration (4.15) is well defined for all k and that the sequence of solutions of (4.15) converges to u^H, the solution of (2.14), provided that the initial guess $u_0^H \in S_0^\ell(\Omega, \mathcal{T}_H)$ is close enough to u^H. This allows to prove Theorem 3.3, i.e., the uniqueness of a solution u^H of (2.14). The following quantity will be useful

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

One can show (provided quasi-uniform meshes) the standard estimates[5] $\sigma_H \leq C(1 + |\log H|)^{1/2}$ for $d = 2$ and $\sigma_H \leq CH^{-1/2}$ 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* (2.3),(3.6)*. Then*

$$\sup_{z^H, v^H, w^H \in S_0^\ell(\Omega,\mathcal{T}_H)} \frac{\left|A_H(z^H; v^H, w^H) - B_H(z^H; v^H, w^H)\right|}{\|v^H\|_{H^1(\Omega)} \|w^H\|_{H^1(\Omega)}} \leq C r_{HMM}, \tag{4.18}$$

$$\sup_{z^H, v^H, w^H \in S_0^\ell(\Omega,\mathcal{T}_H)} \frac{\left|A'_H(z^H; v^H, w^H) - B'_H(z^H; v^H, w^H)\right|}{\|z^H\|_{W^{1,6}(\Omega)} \|v^H\|_{H^1(\Omega)} \|w^H\|_{H^1(\Omega)}} \leq C r'_{HMM}, \tag{4.19}$$

where r_{HMM} and r'_{HMM} are defined in (3.1),(3.5)*, respectively and where the constant C is independent of H, h, δ.*

Proof. The proof of (4.18) was given in (4.7). The proof of (4.19) is nearly identical. Indeed, using Lemma 2.3, 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 $\|v^H \nabla z^H\|_{L^2(\Omega)} \leq \|v^H\|_{L^3(\Omega)} \|\nabla z^H\|_{L^6(\Omega)} \leq C\|v^H\|_{H^1(\Omega)} \|z^H\|_{W^{1,6}(\Omega)}$. □

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*

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

where r_{HMM}, r'_{HMM} are defined in (3.1) *and* (3.5)*, respectively, then for all linear forms 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

$$\|v^H\|_{H^1(\Omega)} \leq C \sup_{w^H \in S_0^\ell(\Omega,\mathcal{T}_H)} \frac{G(w^H)}{\|w^H\|_{H^1(\Omega)}}$$

where C is a constant independent of H, h and z^H.

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

Proof. Lemma 4.9 has been proved in [7, Lemma 7] 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

$$\inf_{v^H \in S_0^\ell(\Omega,\mathcal{T}_H)} \sup_{w^H \in S_0^\ell(\Omega,\mathcal{T}_H)} \frac{\partial A_H(z^H; v^H, w^H)}{\|v^H\|_{H^1(\Omega)} \|w^H\|_{H^1(\Omega)}} \geq K > 0, \tag{4.20}$$

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 (4.13) 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 (4.18),(4.19), respectively. We deduce the inf-sup inequality (4.20) 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 $\{u^H\}$ is 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 the quasi-uniform mesh assumption, 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 [20, Thm. 17.2]) which yields

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

where $\mathcal{I}_H : C^0(\overline{\Omega}) \to S_0^\ell(\Omega, \mathcal{T}_H)$ denotes the usual nodal interpolant [20, Sect. 12]. □

We can now prove that the Newton method (4.15) converges at the usual quadratic rate.

Lemma 4.11. *Assume that the hypotheses of Theorem 3.3 hold. Let u^H be a solution of (2.14). There exists $H_0, r_0, \nu > 0$, such that if (3.7) holds and $u_0^H \in S_0^\ell(\Omega, \mathcal{T}_H)$ satisfies*

$$\sigma_H \|u_0^H - u^H\|_{H^1(\Omega)} \leq \nu, \tag{4.21}$$

then the sequence $\{u_k^H\}$ of the Newton method (4.15) with initial value u_0^H is well defined and $\|u_{k+1}^H - u^H\|_{H^1(\Omega)} \leq C\sigma_H \|u_k^H - u^H\|^2_{H^1(\Omega)}$, where C is a constant independent of H, h, k.

Proof. The proof of Lemma 4.11 follows closely the lines of the proof of [22, Theorem 2] (see [7, Theorem 6] in the context of FEM with numerical quadrature), where we use the C^2 regularity of the tensor $a_K^0(s)$ with respect to s, and the boundedness of $\partial^k a_K^0(s)/\partial s^k$, $k \leq 2$ which can be shown from (3.6), using the idea of the proof of Lemma 6.1 (see Appendix). The main ingredient of the proof is Lemma 4.9 which can be applied in the special case $z^H = u^H$ thanks to Lemma 4.10 and the estimate (using (3.3),(3.7) and $\sigma_H \leq CH^{1/2}$)

$$\sigma_H \|u^H - u^0\|_{H^1(\Omega)} \leq C\sigma_H (H^\ell + r_{HMM}) \leq C(\sigma_H H^\ell + \sigma_H r_0) \leq \nu$$

for all H small enough and r_0 chosen small enough in (3.7). We omit the details. □

We can now prove the claimed uniqueness result.

Proof of Theorem 3.3. Let $u^H, \tilde{u}^H$ be two solutions of (2.14). We consider the Newton method $\{u_k^H\}$ defined by (4.15) 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 (4.21) for H_0, r_0 small enough using (3.7). By Lemma 4.11, for ν small enough, $e_k = \|u_k^H - u^H\|_{H^1(\Omega)}$ converges to 0 for $k \to \infty$. Using (2.14), we have $u_k^H = u_0^H = \tilde{u}^H$ for all k, which yields $u^H = \tilde{u}^H$. □

If we want further to characterize the 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 (3.1) can be estimated as described in the following lemma whose proof is postponed to the Appendix.

Lemma 4.12. *Assume that the hypotheses of Corollary 3.4 hold, then $r'_{HMM} \leq C(h/\varepsilon)^2$. If we use the form* (2.15) *instead of* (2.18) *for the solution u^H of* (2.14)*, then $r'_{HMM} \leq C((h/\varepsilon)^2 + \delta)$.*

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

5. Numerical experiments

In this section, we first present an efficient numerical implementation of the Newton method (4.15), 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.

5.1. Newton method implementation.

To solve the non-linear problem (2.14) 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 (4.15) translate in terms of matrices as

$$\big(B(z_k^H) + B'(z_k^H)\big)(U_{k+1} - U_k) = -B(z_k^H) U_k + F, \tag{5.1}$$

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 (2.15) and (4.14), respectively. Here, F is a vector associated the source term (2.14), which also contains the boundary data.

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 matrix $(B_K(z_k^H))_{p,q=1}^{n_K} = \sum_{j=1}^J (B_{K,j}(z_k^H))_{p,q=1}^{n_K}$ with

$$(B_{K,j}(z_k^H))_{p,q=1}^{n_K} = \frac{\omega_{K_j}}{|K_{\delta_j}|} \int_{K_{\delta_j}} a^\varepsilon(x, z_k^H(x_{K_j})) \nabla \varphi_{K_j,p}^{h,z^H(x_{K_j})}(x) \cdot \nabla \varphi_{K_j,q}^{h,z^H(x_{K_j})}(x) dx, \tag{5.2}$$

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 (2.13) constrained by $\phi_{K,p}^H, \phi_{K,q}^H$, linearized at x_{K_j}, respectively.

Differentiating (5.2), we see that the stiffness matrix $B'(U)$ in (5.1) associated to the non-symmetric form $B'_H(z^H; \cdot, \cdot)$ defined in (4.14) is given by the sum of J

products of $n_K \times n_K$ matrices

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

where the column vector $U_{K,k}$ of size n_K gives the components of z^H in the basis $\{\phi_{K,i}^H\}_{i=1}^{n_K}$ of 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 the stiffness matrices for both $B(z_k^H)$ and $B'(z_k^H)$ is about twice the cost of computing the stiffness matrix $B(z_k^H)$ alone.

5.2. **Numerical illustration of theoretical convergence rates.** We shall now illustrate the sharpness of the H^1 and L^2 error estimates of Sections 3 and 4. Rectangular macro $\mathcal{Q}^1$ elements (Gauss quadrature with $J = 4$ nodes $(1/2 \pm \sqrt{3}/6, 1/2 \pm \sqrt{3}/6)$) will be used and $\delta/\varepsilon \in \mathbb{N}^*$ (we emphasize that similar result can be obtained with triangular $\mathcal{P}^1$ macro elements).

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 (see (2.18)) we have

$$\|u^H - u_0\|_{H^1(\Omega)} \leq C(H + \hat{h}^2), \qquad \|u^H - u_0\|_{L^2(\Omega)} \leq C(H^2 + \hat{h}^2), \tag{5.3}$$

where $\hat{h} := h/\varepsilon$ is the scaled micro mesh size.

We consider the non-linear problem (1.1) on the domain $\Omega = (0,1)^2$ with homogeneous Dirichlet boundary conditions and the following anisotropic 2×2 diagonal 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 the diagonal matrix

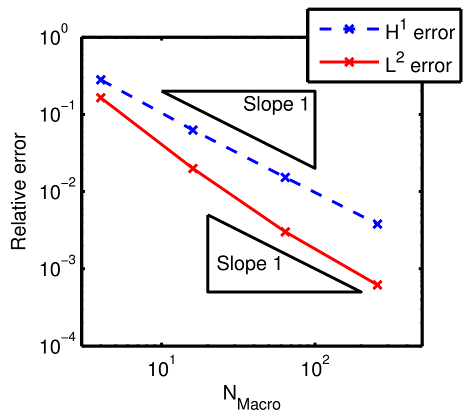
$$a^0(x,s) = \begin{pmatrix} 1 + x_1 \sin(\pi s) & 0 \\ 0 & 2 + \arctan(s) \end{pmatrix}$$

The source $f(x)$ in (1.1) is adjusted analytically so that the homogenized solution is $u_0(x) = 8\sin(\pi x_1)x_2(1-x_2)$, The H^1 and L^2 relative errors between the exact homogenized solution u_0 and the FE-HMM solution u^H,

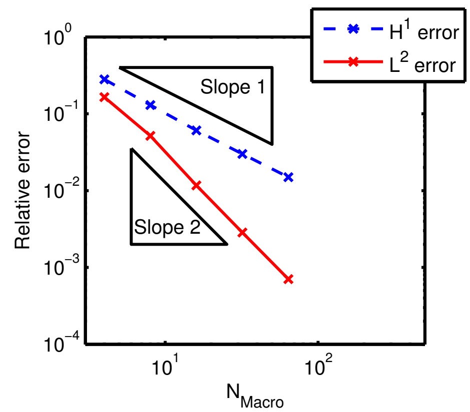
$$e_{L^2} = \|u_0 - u^H\|_{L^2(\Omega)}/\|u_0\|_{L^2(\Omega)}, \qquad e_{H^1} = \|\nabla(u_0 - u^H)\|_{L^2(\Omega)}/\|\nabla u_0\|_{L^2(\Omega)},$$

can be estimated by quadrature with $\|u_0 - u^H\|_{L^2(\Omega)}^2 \approx \sum_{K\in\mathcal{T}_H} \sum_{j=1}^{J} \omega_{K_j} |u^H(x_{K_j}) - u_0(x_{K_j})|^2$, and similarly for $\|\nabla(u_0 - u^H)\|_{L^2(\Omega)}$. We consider a sequence of uniform macro partitions $\mathcal{T}_H$ with meshsize $H = 1/N_{Macro}$ and $N_{Macro} = 4, 6, 8, \dots, 256$.

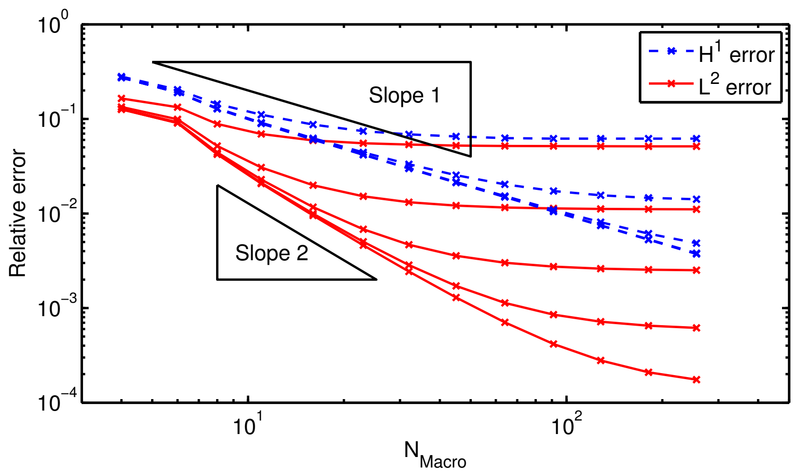
In Figure 1(a),(b) the H^1 an L^2 relative errors between the exact homogenized solution and the FE-HMM solutions are shown for the above sequence of partitions using a simultaneous refinement of H and $\hat{h}$ according to $\hat{h} \sim H$ (L^2 norm) and $\hat{h} \sim \sqrt{H}$ (H^1 norm). We observe the expected (optimal) convergence rates (5.3) in agreement with Theorem 3.1.



(a) Optimal H^1 refinement strategy with $N_{Micro} \sim \sqrt{N_{Macro}}$.



(b) Optimal L^2 refinement strategy with $N_{Micro} = N_{Macro}$.



(c) Computation with fixed $N_{Micro} = 4, 8, 16, 32, 64$.

FIGURE 1. Convergence rates: e_{L^2} error (dashed lines) and e_{H^1} error (solid lines).

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 ($N = Micro = 4, 8, 16, 32, 64$). This is illustrated in Figure 1(c), where we plot the H^1 an L^2 relative errors as a function of $H = 1/N_{Macro}$. It is observed that optimal macro convergence rates are obtained only if macro and micro meshes are refined *simultaneously.*

6. Appendix

We provide in this appendix a proof of Lemma 4.12 which is a crucial ingredient for the proof of Corollary 3.4 on the uniqueness of the numerical solution u^H for small enough macro and micro mesh sizes H, h. For that, we will often use the following inequality (6.1). Given a closed subspace H of $W(K_{\delta_j})$, let $\psi_i \in H$, $i = 1, 2$ 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

$$\|\nabla\psi_1 - \nabla\psi_2\|_{L^2(K_{\delta_j})} \leq \lambda^{-1} \sup_{x\in K_{\delta_j}} \|a_1(x) - a_2(x)\|_F \|f_2\|_{L^2(K_{\delta_j})} + \|f_1 - f_2\|_{L^2(K_{\delta_j})}, \tag{6.1}$$

where λ is the minimum of the ellipticity constants of a_1, a_2. We also need a regularity result for the solutions of (2.22).

Lemma 6.1. *Assume that a^ε is uniformly elliptic and satisfies* (3.6) *with $k = 1$. Consider the solution $\psi^{i,s}_{K_j}$ of* (2.22)*. Then, the map $s \mapsto \psi^{i,s}_{K_j} \in H^1(K_{\delta_j})$ is of class C^1 and satisfies*

$$\frac{\partial}{\partial s}\psi^{i,s}_{K_j} = \phi^{i,s}_{K_j}, \qquad \frac{\partial}{\partial s}\nabla\psi^{i,s}_{K_j} = \nabla\phi^{i,s}_{K_j}, \tag{6.2}$$

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

$$\int_{K_{\delta_j}} a^\varepsilon(x,s)\nabla\phi^{i,s}_{K_j}(x)\cdot\nabla z(x)dx = -\int_{K_{\delta_j}} \partial_u a^\varepsilon(x,s)(\nabla\psi^{i,s}_{K_j}(x) + \mathbf{e}_i)\cdot\nabla z(x)dx. \tag{6.3}$$

An analogous statement holds also for the FEM discretization $\psi^{i,h,s}_{K_j}$ defined in (2.24)*, where $\frac{\partial}{\partial s}\psi^{i,h,s}_{K_j} = \phi^{i,h,s}_{K_j}$ satisfies* (6.2) *and* (6.3) *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 (2.24) with parameters s and $s+\Delta s$, respectively. We deduce from (6.1) with $H = W(K_{\delta_j})$, and the smoothness of $s \mapsto a^\varepsilon(x,s)$ that $\|\psi^{i,s+\Delta s}_{K_j}(x) - \psi^{i,s}_{K_j}(x)\|_{H^1(K_{\delta_j})} \to 0$ for $\Delta s \to 0$. Consider now the identity

$$\int_{K_{\delta_j}} a^\varepsilon(x,s)\nabla(\psi^{i,s+\Delta s}_{K_j} - \psi^{i,s}_{K_j})\cdot\nabla z dx = -\int_{K_{\delta_j}} (a^\varepsilon(x,s+\Delta s) - a^\varepsilon(x,s))(\nabla\psi^{i,s+\Delta s}_{K_j} + \mathbf{e}_i)\cdot\nabla z dx \tag{6.4}$$

Dividing (6.4) by Δs, subtracting (6.3) and taking $\Delta s \to 0$, we deduce from (6.1) that $\frac{\partial}{\partial s}\psi^{i,s}_{K_j}(x)$ exists and that (6.2),(6.3) hold. Using again the property (6.1), 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 (6.1) with $H = S^q(K_{\delta_j}, \mathcal{T}_h)$ □

Proof of Lemma 4.12. We start with the first estimate. We set $x = x_{K_j}$ in (2.6). A change of variable $y \to x_{K_j} + x/\varepsilon$ shows that

$$(a^0(x_{K_j}, s))_{mn} = \frac{1}{|K_{\delta_j}|}\int_{K_{\delta_j}} a(x_K, x/\varepsilon, s)(\mathbf{e}_n + \nabla\chi^n(x_K, x/\varepsilon, s))\cdot\mathbf{e}_m \tag{6.5}$$

where $\chi^n(x_K, x/\varepsilon, s)$ solves for all $z \in W(K_{\delta_j})$,
(6.6)
$$\int_{K_{\delta_j}} a(x_K, x/\varepsilon, s)\nabla\chi^n(x_K, x/\varepsilon, s)\cdot\nabla z(x)dx = -\int_{K_{\delta_j}} a(x_K, x/\varepsilon, s)\mathbf{e}_n\cdot\nabla z(x)dx,$$

As the tensor a^ε is (locally) periodic and $\delta/\varepsilon \in \mathbb{N}^*$, if we collocate a^ε in (2.26) and in (2.6) 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 $-\nabla\cdot(A\nabla\Xi) = \nabla\cdot F_n$ formed by the augmented problem (2.24)-(6.3), where

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

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 [20, Sect. 17], we deduce that the corresponding FEM discretization $(\psi^{m,h,s}_{K_j}, \phi^{m,h,s}_{K_j})$ satisfies

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

and similarly for $\psi^{n,h,s}_{K_j}$ in place of $\phi^{n,h,s}_{K_j}$. Now, using Lemma 6.1 and differentiating the identity (4.12) with respect to s, we deduce from the Cauchy-Schwarz inequality $|\frac{d}{ds}(\overline{a}^0_{K_j}(s) - a^0_{K_j}(s))_{mn}| \le C(h/\varepsilon)^2$, 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 $r'_{HMM} \le C(h/\varepsilon)^2$. Consider now the case where the formulation (2.15) 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 (6.1) 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 concludes the proof. □

References

[1] A. Abdulle, *On a-priori error analysis of Fully Discrete Heterogeneous Multiscale FEM*, SIAM Multiscale Model. Simul., 4(2) (2005), 447–459.

[2] A. Abdulle, *Discontinuous Galerkin finite element heterogeneous multiscale method for elliptic problems with multiple scales*, Math. Comp., 81 (2012), 687–713.

[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*, Ser. Contemp. Appl. Math. CAM, 16 (2011), 280–305.

[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(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,* Numer. Math., 121 (2012) 397-431.

[8] A. Abdulle, W. E, B. Engquist, and E. Vanden-Eijnden, *The heterogeneous multiscale method*, Acta Numer., 21 (2012), 1–87.

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

[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., 4(5) (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(7) (1981), 13–51.

[15] A. Bourgeat and A. Piatnitski, *Approximations of effective coefficients in stochastic homogenization*, Ann. I. H. Poincaré, 40 (2004), 153–165.

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

[17] 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(4) (2005), 493–515.

[18] 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.

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

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

[21] 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.

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

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

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

[25] W. E, P. Ming and P. Zhang, *Analysis of the heterogeneous multiscale method for elliptic homogenization problems*, J. Amer. Math. Soc., 18(1) (2005), 121–156.

[26] 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.

[27] 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.

[28] 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.

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

[30] A. Gloria. *Numerical approximation of effective coefficients in stochastic homogenization of discrete elliptic equations*, M2AN Math. Model. Numer. Anal., 46 (2012), 1–38.

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

[32] A. Karageorghis and D. Lesnic, *Steady-state nonlinear heat conduction in composite materials using the method of fundamental solutions,* Comput. Methods Appl. Mech. Engrg., 197(33-40) (2008), 3122–3137.

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

[34] 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.

[35] A.G. Whittington, A.M. Hofmeister, and P.I. Nabelek, *Temperature-dependent thermal diffusivity of the Earth's 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.

[37] V.V. Yurinskiĭ. *Averaging of symmetric diffusion in random medium*, Sibirskii Matematicheskii Zhurnal, 27 (1986), 167–180.

ANMC, Section de Mathématiques, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

E-mail address: Assyr.Abdulle@epfl.ch

ANMC, Section de Mathématiques, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

Current address: École Normale Supérieure de Cachan, Antenne de Bretagne, INRIA Rennes, IRMAR, CNRS, UEB, av. Robert Schuman, F-35170 Bruz, France

E-mail address: Gilles.Vilmart@bretagne.ens-cachan.fr