A rigorous Multiscale Method for semi-linear elliptic problems

Patrick Henning[1], Axel Målqvist[2 *], Daniel Peterseim[3 †]

November 14, 2012

Abstract

In this paper we propose and analyze a new Multiscale Method for solving semi-linear elliptic problems with heterogeneous and highly variable coefficient functions. For this purpose we construct a generalized finite element basis that spans a low dimensional multiscale space. The basis is assembled by performing localized linear fine-scale computations in small patches that have a diameter of order $H\log(H^{-1})$ where H is the coarse mesh size. Without any assumptions on the type of the oscillations in the coefficients, we give a rigorous proof for a linear convergence of the H^1-error with respect to the coarse mesh size. To solve the arising equations, we propose an algorithm that is based on a damped Newton scheme in the multiscale space.

Keywords finite element method, a priori error estimate, convergence, multiscale method, non-linear, computational homogenization, upscaling

AMS subject classifications 35J15, 65N12, 65N30

[1]Westfälische Wilhelms-Universität Münster, Institut für Numerische und Angewandte Mathematik, Einsteinstr. 62, D-48149 Münster, Germany

[2]Department of Information Technology, Uppsala University, Box 337, SE-751 05 Uppsala, Sweden

[*]A. Målqvist is supported by The Göran Gustafsson Foundation and The Swedish Research Council.

[3]Institut für Mathematik, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin, Germany

[†]D. Peterseim is supported by the DFG Research Center Matheon Berlin through project C33.

1 Introduction

This paper is devoted to the numerical approximation of solutions of semilinear elliptic problems with rapidly oscillating and highly varying coefficient functions. We are concerned with the following type of equations:

$$-\nabla \cdot (A\nabla u) + F(u, \nabla u) = g.$$

Here, we prescribe a (zero-) Dirichlet boundary condition for u, A is a highly variable diffusion matrix and F is a highly variable nonlinear term that typically describes advective and reactive processes. In particular, we have a linear term of second order and nonlinear terms of order 1 and 0. A typical application is the stationary (Kirchhoff transformed) Richards equation that describes the groundwater flow in unsaturated soils (c.f. [1, 2, 3]). The equation reads

$$\nabla \cdot (K\nabla u) - \nabla \cdot (K\, kr(M(u))\vec{g}) = f,$$

where u is the so-called generalized pressure, K is the hydraulic conductivity in the soil, kr the relative permeability depending on the saturation, M is some nonlinearity arising from the Kirchhoff transformation and $\vec{g}$ denotes the gravity vector. If we add an infiltration process, the equation receives an additional nonlinear reaction term.

The numerical treatment of such equations is often complicated and expensive. Due to the high variability of the coefficient functions, one requires extremely fine computational grids that are able to capture all the fine scale oscillations. Using standard methods such as Finite Element or Finite Volume schemes, this results in systems of equations of enormous size and therefore in a tremendous computational demand that can not be handled in a lot of scenarios.

There is a large variety of so-called multiscale methods that are able to overcome this difficulty by decoupling the fine scale computations into local parts. This decreases the computational demand without suffering from a remarkable loss in accuracy. Examples of multiscale methods are the Heterogeneous Multiscale Method (HMM) by E and Engquist [9] and the Multiscale Finite Element Method (MsFEM) proposed by Hou and Wu [15]. Both methods fit into a common framework and are strongly related to numerical homogenization (c.f. [10, 13, 14]). HMM and MsFEM are typically not constructed for a direct approximation of exact solutions but for homogenized solutions and corresponding correctors instead. This implies that they are only able to approximate the exact solution up to a modeling error that

depends on the homogenization setting (c.f. [10]). General proofs of convergence are therefore hard to achieve.

We are concerned with a multiscale method that is based on the concept of the Variational Multiscale Method (VMM) proposed by Hughes et al. [16]. In comparison to HMM and MsFEM, the VMM aims to a direct approximation of the exact solution without suffering from a modeling error remainder arising from homogenization theory. The key idea of the Variational Multiscale Method is to construct a splitting of the original solution space V into the direct sum of a low dimensional space for coarse grid approximations and high dimensional space for fine scale reconstructions. In this work, we consider a modification and extension of this idea that was developed in [18, 21] and that was explicitly proposed in [22]. Here, the splitting is such that we obtain an accurate but low dimensional space V^{ms} (where we are looking for our fine scale approximation instead of an approximation of a coarse part) and a high dimensional residual space V^{f}. The construction of V^{ms} involves the computation of one fine scale problem in a small patch per degree of freedom. Mesh adaptive versions of the VMM with patch size control that can be also applied to this multiscale method were achieved in [18, 19, 20, 23]. The first rigorous proof of convergence was recently obtained in [22] for linear diffusion problems.

In this contribution, we present an efficient way of handling semi-linear elliptic multiscale problems in the modified VMM framework, including a proof of convergence based on the techniques established in [22]. Even though the original problem is nonlinear, the fine scale computations are purely linear problems that can be solved in parallel. The main result of this article is the optimal convergence of the H^1-error between exact solution u and its multiscale approximation u_H^{ms}. We show that, if the patch size is of order $H\log(H^{-1})$, the following error bound holds true:

$$\|u - u_H^{\mathrm{ms}}\|_{H^1(\Omega)} \leq CH.$$

Here, C denotes a generic constant independent of the mesh size of the computational grid or the oscillations of A and F.

The paper is structured as follows. In Section 2 we introduce the setting of this paper, including the assumptions on the considered semi-linear problem. In Section 3 we present and motivate our method and we state the corresponding optimal convergence result. The result itself is proved in Section 4. Finally, to solve the arising nonlinear equations, we propose an algorithm that is based on a damped Newton scheme in the multiscale space. This is done in Section 5.

2 Setting

Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain with polyhedral boundary, let $V := H_0^1(\Omega)$ and let $A \in L^\infty(\Omega, \mathbb{R}^{d\times d}_{\text{sym}})$ denote a matrix valued function with uniformly strictly positive eigenvalues. We assume that the space $H_0^1(\Omega)$ is endowed with the H^1-semi norm given by $|v|_{H^1(\Omega)} := \|\nabla v\|_{L^2(\Omega)}$ (which is equivalent to the common H^1-norm in $H_0^1(\Omega)$). By $\langle\cdot,\cdot\rangle := (\cdot,\cdot)_{L^2(\Omega)}$ we denote the inner product in $L^2(\Omega)$ and $F : \Omega \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ is a nonlinear measurable function.

For a given source term $g \in L^2(\Omega) \subset H^{-1}(\Omega)$ we are concerned with the problem to find $u \in H_0^1(\Omega)$ (i.e. with a homogeneous Dirichlet boundary condition) with

$$\langle A\nabla u, \nabla v\rangle + \langle F(\cdot, u, \nabla u), v\rangle = \langle g, v\rangle \tag{1}$$

for all test functions $v \in H_0^1(\Omega)$. To simplify the notation, we define the operator $B : H_0^1(\Omega) \to H^{-1}(\Omega)$ by

$$\langle B(v), w\rangle_{H^{-1},H_0^1} := \langle A\nabla v, \nabla w\rangle + \langle F(\cdot, v, \nabla v), w\rangle \quad \text{for } v, w \in H_0^1(\Omega),$$

where $\langle\cdot,\cdot\rangle_{H^{-1},H_0^1}$ denote the dual pairing in $H_0^1(\Omega)$. Here, A is a rapidly oscillating, highly heterogeneous diffusion matrix. $F(\cdot, \xi, \zeta)$ is also allowed to be rapidly oscillating, without further assumptions on the type of the oscillations.

We assume that A and F are of the same order of magnitude. This prohibits for instance advection dominated processes, which must be treated with a different approach and requires a different set of multiscale basis functions than the one proposed in this paper. In particular, we will show that if the F-term does not dominate the equation, it is sufficient to construct a multiscale spaced only based on the oscillations of A. This implies that we only need to solve linear problems on the fine scale. For this purpose condition (A4) below is to guarantee that A and F have the same size. If this is not fulfilled, the proposed method needs modifications with respect to the construction of the multiscale basis. Typical examples where condition (A4) is not fulfilled are models for transport of solutes in groundwater where we deal with extremely large Péclet numbers and a corresponding scaling of the advective terms. In this case, the oscillations of F become significant to obtain accurate upscaled and homogenized properties (c.f. [11, 12]).

For the subsequent analytical considerations and in order to guarantee a unique solution of (1), we make the following assumptions:

Assumption 1. *We assume:*

(A1) $A \in L^\infty\left(\Omega, \mathbb{R}^{d\times d}_{\mathrm{sym}}\right)$ *with*

$$\infty > \beta := \|A\|_{L^\infty(\Omega)} = \operatorname*{ess\,sup}_{x\in\Omega} \sup_{\zeta\in\mathbb{R}^d\setminus\{0\}} \frac{A(x)\zeta\cdot\zeta}{|\zeta|^2}.$$

and there exists α *such that*

$$0 < \alpha := \operatorname*{ess\,inf}_{x\in\Omega} \inf_{\zeta\in\mathbb{R}^d\setminus\{0\}} \frac{A(x)\zeta\cdot\zeta}{|\zeta|^2},$$

(A2) There exist $L_1, L_2 \in \mathbb{R}_{>0}$ *such that uniformly for almost every* x *in* Ω*:*

$$\begin{aligned} |F(x,\xi_1,\zeta) - F(x,\xi_2,\zeta)| &\le L_1|\xi_1-\xi_2|, \quad \text{for all } \zeta\in\mathbb{R}^d,\ \xi_1,\xi_2\in\mathbb{R},\\ |F(x,\xi,\zeta_1) - F(x,\xi,\zeta_2)| &\le L_2|\zeta_1-\zeta_2|, \quad \text{for all } \zeta_1,\zeta_2\in\mathbb{R}^d,\ \xi\in\mathbb{R},\\ F(x,0,\zeta) = F(x,\xi,0) &= 0, \quad \text{for all } \zeta\in\mathbb{R}^d,\ \xi\in\mathbb{R}. \end{aligned}$$

(A3) $B : V \to V'$ *is hemicontinuous, i.e. for all* $v_1, v_2, w \in H^1_0(\Omega)$

$$s \mapsto \langle B(v_1+sv_2), w\rangle_{H^{-1},H^1_0} \text{ is a continuous function on } [0,1]$$

and B *is strongly monotone, i.e. there exist* $c_0 > 0$ *so that for all* $u, v \in H^1_0(\Omega)$*:*

$$\langle B(u) - B(v), u - v\rangle_{H^{-1},H^1_0} \ge c_0|u-v|^2_{H^1(\Omega)}. \tag{2}$$

(A4) A and F are of similar size, i.e. O(β)=O(L₁)=O(L₂).

Under assumptions (A1)-(A3), the Browder-Minty theorem yields a unique solution of problem (1).

Remark 1. *Let the* C_p *denote the constant appearing the Poincaré-Friedrichs inequality for* $H^1_0(\Omega)$ *functions. Observe that (A1)-(A3) imply that the solution* $u \in H^1_0(\Omega)$ *of (1) fulfills*

$$\begin{aligned} \|F(u,\nabla u)\|_{L^2(\Omega)} &\le \|F(u,\nabla u) - F(u,0)\|_{L^2(\Omega)} + \|F(u,\nabla u) - F(0,\nabla u)\|_{L^2(\Omega)}\\ &\le (L_1C_p + L_2)|u|_{H^1(\Omega)} \le C_p\frac{L_1C_p+L_2}{c_0}\|g\|_{L^2(\Omega)}. \end{aligned} \tag{3}$$

Note that problem (1) also covers equations such as

$$-\nabla \cdot (\kappa(u) A \nabla u) = F(u, \nabla u),$$

for a strictly positive and sufficiently regular function κ. In this case, the equation can be rewritten as

$$-\nabla \cdot A \nabla u = \tilde{F}(u, \nabla u).$$

In the remainder of this paper, we use the notation $q_1 \lesssim q_2$ if $q_1 \leq C q_2$ where $C > 0$ is a constant that only depends on the shape regularity of the mesh, but not on the mesh size or the validity of assumption (A4). This means that dependencies such as $(L_1 + L_2)\alpha^{-1}$ are always explicitly stated. Dependencies on the contrast $\frac{\beta}{\alpha}$ are allowed to be contained in $\lesssim$.

3 A Rigorous Multiscale Method

In this section we suggest a Rigorous Multiscale Method (RMM) that is based on the concept introduced by Hughes et al. [16] and the specific constructions proposed in [18, 21] for linear problems. The required multiscale (MS) basis functions are obtained with the strategy established in [22].

The main idea of the Variational Multiscale Method is to start from a finite element space $\mathcal{V}_h$ with a highly resolved computational grid and to construct a splitting of this space into the direct sum $\mathcal{V}_h = \mathcal{V}^l \oplus \mathcal{V}^{\mathrm{f}}$ of a low dimensional space $\mathcal{V}^{\mathrm{l}}$ and a 'detail space' $\mathcal{V}^{\mathrm{f}}$ containing all the missing oscillations. Then, a basis of $\mathcal{V}^l$ is assembled and we can compute a Galerkin approximation u_l of u in $\mathcal{V}^{\mathrm{l}}$. However, the success of this approach strongly depends on the choice of $\mathcal{V}^l$. On the one hand, the costs for assembling a basis of $\mathcal{V}^{\mathrm{l}}$ must be kept low. On the other hand, the basis functions somehow need to contain information about fine scale features. For instance, a standard coarse finite element space is cheap to assemble but will fail to yield reliable approximations. On the contrary, the space spanned by high resolution finite element approximations yields perfect approximations, but is as costly as the original problem that we tried to avoid. Therefore, the key is to find an optimal balance between costs and accuracy. In previous works (c.f. [16, 18, 19]) the multiscale basis (MS-basis) of $\mathcal{V}^l$ was constructed involving the full multiscale operator B that corresponds with the left hand side of the original problem. In a fully linear setting, this can be a reasonable choice. However, it gets extremely expensive if B is a nonlinear operator, since it leads to numerous nonlinear equations to solve. Furthermore it is not clear if the constructed set of basis functions leads to good approximations.

One novelty of this work is that we do not involve the full operator B in the construction of the MS-basis, but only the linear diffusive part $\langle A\nabla\cdot, \nabla\cdot\rangle$. Even though the oscillations of F are not captured by the MS-basis, we can show that we are still able to obtain accurate approximations and to preserve the optimal convergence rates.

3.1 Notation and discretization

Let $\mathcal{T}_H$ denote a regular triangulation of Ω and let $H : \overline{\Omega} \to \mathbb{R}_{>0}$ denote the $\mathcal{T}_H$-piecewise constant mesh size function with $H|_T = H_T := \operatorname{diam}(T)$ for all $T \in \mathcal{T}_H$. Additionally, let $\mathcal{T}_h$ be a regular triangulation of Ω that is supposed to be a refinement of $\mathcal{T}_H$. We assume that $\mathcal{T}_h$ is sufficiently small so that all fine scale features of B are captured by the mesh. The mesh size h denotes the maximum diameter of an element of $\mathcal{T}_h$. The corresponding classical (conforming) finite element spaces of piecewise polynomials of degree 1 are given by

$$\begin{aligned} V_H &:= \{v_H \in C^0(\bar{\Omega}) \cap H_0^1(\Omega) \mid \forall T \in \mathcal{T}_H : \ \ (v_H)|_T \text{ is affine}\}, \\ V_h &:= \{v_h \in C^0(\bar{\Omega}) \cap H_0^1(\Omega) \mid \forall K \in \mathcal{T}_h : \ \ (v_h)|_K \text{ is affine}\}. \end{aligned}$$

By J, we denote the dimension of V_H and by $\mathcal{N}_H = \{z_j | \, 1 \leq j \leq J\}$ the set of interior vertices of $\mathcal{T}_H$. For every vertex $z_j \in \mathcal{N}_H$, let $\lambda_j \in V_H$ denote the associated nodal basis function (tent function), i.e. $\lambda_j \in V_H$ with the property $\lambda_j(z_i) = \delta_{ij}$ for all $1 \leq i, j \leq J$.

From now on, we denote by $u_h \in V_h$ the classical finite element approximation of u in the discrete (highly resolved) space V_h, i.e. $u_h \in V_h$ solves

$$\int_\Omega A\nabla u_h \cdot \nabla v_h + F(\cdot, u_h, \nabla u_h)v_h = \int_\Omega g v_h \tag{4}$$

for all $v_h \in V_h$. We assume that V_h resolves the micro structure, i.e. that the error $\|u - u_h\|_{H^1(\Omega)}$ becomes sufficiently small by falling below a given tolerance. For standard finite elements methods the error can be bounded by $C \cdot h^s$ with $s \geq \frac{1}{2}$. However, for regular coefficients, C depends on the derivative of A and F with respect to the spatial variable. If A and F are rapidly oscillating, the derivative becomes very large and h must be very small to equalize the dominance of C. This is only fulfilled, when h resolves the micro structure (we refer to [24] and [25] for some qualitative characterization of this so-called resolution condition). We are therefore dealing with pre-asymptotic effects for the standard methods. The multiscale method that we propose in the subsequent sections will be constructed to converge

to u_h, with a linear speed in H and with a generic constant that is not affected by the fine scale oscillations, i.e. in particular we do not have such pre-asymptotic effects.

3.2 Quasi Interpolation

The key tool in our construction is a linear (quasi-)interpolation operator $\mathfrak{I}_H : V_h \to V_H$ that is continous and surjective. The kernel of this operator is going to be our fine space (or remainder space) V_h^{f}. In [22] a weighted Clément interpolation operator was used. In this work, we do not specify the choice. Instead, we state a set of assumptions that must be fulfilled in order to derive an optimal convergence result for the constructed multiscale method.

Assumption 2 (Assumptions on the interpolation)**.** *We make the following assumptions on the interpolation operator $\mathfrak{I}_H : V_h \to V_H$:*

(A5) $\mathfrak{I}_H \in L(V_h, V_H)$*, i.e.* $\mathfrak{I}_H$ *is linear and continuous,*

(A6) the restriction of $\mathfrak{I}_H$ *to* V_H *is an isomorphism, in particular there holds* $(\mathfrak{I}_H \circ \mathfrak{I}_H^{-1})(v_H) = v_H$ *for all* $v_H \in V_H$*,*

(A7) there exists a generic constant $C_{\mathfrak{I}_H}$*, only depending on the shape regularity of* $\mathcal{T}_H$ *and* $\mathcal{T}_h$*, such that for all* $v_h \in V_h$ *and for all* $T \in \mathcal{T}_H$ *there holds*

$$H_T^{-1}\|v_h - \mathfrak{I}_H(v_h)\|_{L^2(T)} + \|\nabla(v_h - \mathfrak{I}_H(v_h))\|_{L^2(T)} \leq C_{\mathfrak{I}_H}\|\nabla v_h\|_{L^2(\omega_T)}$$

with

$$\omega_T := \bigcup\{K \in \mathcal{T}_H | \, \overline{K} \cap \overline{T} \neq \emptyset\}.$$

(A8) there exists a generic constant $C'_{\mathfrak{I}_H}$*, only depending on the shape regularity of* $\mathcal{T}_H$ *and* $\mathcal{T}_h$*, such that for all* $v_H \in V_H$ *there exists* $v_h \in V_h$ *with*

$$\mathfrak{I}_H(v_h) = v_H, \quad |v_h|_{H^1(\Omega)} \leq C'_{\mathfrak{I}_H}|v_H|_{H^1(\Omega)} \quad \textit{and} \quad \operatorname{supp} v_h \subset \operatorname{supp} v_H.$$

A natural choice for $\mathfrak{I}_H$ might be the Lagrange interpolation operator that fulfills the interpolation estimate in assumption (A7). Furthermore, since the Lagrange operator is a projection, the assumptions (A5), (A6) and (A8) are automatically fulfilled.

Another possibility is to choose $\mathfrak{I}_H$ as a weighted Clément interpolation operator. This construction was proposed in [22]. Given $v \in H_0^1(\Omega)$, $\mathfrak{I}_H v := \sum_{j=1}^{J} v_j \lambda_j$ defines a (weighted) Clément interpolant with nodal values

$$v_j := \left(\int_\Omega v\lambda_j \,\mathrm{d}x\right) / \left(\int_\Omega \lambda_j \,\mathrm{d}x\right) \tag{5}$$

for $1 \leq j \leq J$ (c.f. [5, 6, 7]) and zero in the boundary nodes. Furthermore, there exists the desired generic constant $C_{\mathfrak{I}_H}$ (only depending on the mesh regularity parameter and in particular independent of H_T) such that for all $v \in H_0^1(\Omega)$ and for all $T \in \mathcal{T}_H$ there holds

$$H_T^{-1}\|v - \mathfrak{I}_H v\|_{L^2(T)} + \|\nabla(v - \mathfrak{I}_H v)\|_{L^2(T)} \leq C_{\mathfrak{I}_H}\|\nabla v\|_{L^2(\omega_T)}.$$

We refer to [5, 6] for a proof of this estimate. This gives us (A7). Assumptions (A5) and (A6) are obvious. The validity of (A8) was proved in [22].

Note that in certain applications a particular interpolation operator may be superior over other choices (cf. Remark 4 in [22]).

3.3 Multiscale Splitting and Modified Nodal Basis

In this section, we construct a splitting of the high resultion finite element space V_h into a low dimension multiscale space V^{ms} and some high dimensional remainder space V_h^{f}. From now on, we let $\mathfrak{I}_H : V_h \to V_H$ denote an interpolation operator fulfilling the properties (A5)-(A8). We start with defining V_h^{f} as the kernel of $\mathfrak{I}_H$ in V_h:

$$V_h^{\mathrm{f}} := \{v_h \in V_h \mid \mathfrak{I}_H v_h = 0\}.$$

V_h^{f} represents the features in V_h not captured by V_H. As already mentioned, it can be shown that $(\mathfrak{I}_H)_{|V_H} : V_H \to V_H$ is an isomorphism (see [22]). We therefore get

$$V_h = V_H \oplus V_h^{\mathrm{f}}, \quad \text{where } \underbrace{v_h}_{\in V_h} = \underbrace{\mathfrak{I}_H^{-1}(\mathfrak{I}_H(v_h))}_{\in V_H} + \underbrace{v_h - \mathfrak{I}_H^{-1}(\mathfrak{I}_H(v_h))}_{\in V_h^{\mathrm{f}}}. \tag{6}$$

Here, the property $(\mathfrak{I}_H \circ \mathfrak{I}_H^{-1})(v_H) = v_H$ for all $v_H \in V_H$ implies the equation $\mathfrak{I}_H(v_h - \mathfrak{I}_H^{-1}(\mathfrak{I}_H(v_h))) = \mathfrak{I}_H(v_h) - (\mathfrak{I}_H \circ \mathfrak{I}_H^{-1})(\mathfrak{I}_H(v_h)) = 0$. We still need to modify the splitting of V_h, because V_H is an inappropriate space for a multiscale approximation. We therefore look for the orthogonal complement of V_h^{f} in V_h with respect to the inner product $\langle A\nabla\cdot, \nabla\cdot\rangle_{L^2(\Omega)}$. For this purpose, we define the orthogonal projection $P : V_h \to V_h^{\mathrm{f}}$ as follows. For a given $v_h \in V_h$, $P(v_h) \in V_h^{\mathrm{f}}$ solves

$$\langle A\nabla P(v_h), \nabla w^{\mathrm{f}}\rangle = \langle A\nabla v_h, \nabla w^{\mathrm{f}}\rangle \quad \text{for all } w^{\mathrm{f}} \in V_h^{\mathrm{f}}.$$

Defining the multiscale space $V_{H,h}^{\mathrm{ms}}$ by $V_{H,h}^{\mathrm{ms}} := (V_H - P(V_H))$, this directly leads to an orthogonal splitting:

$$V_h = V_{H,h}^{\mathrm{ms}} \oplus V_h^{\mathrm{f}}, \tag{7}$$

because of

$$V_h = \mathrm{kern}(P) \oplus V_h^{\mathrm{f}} = (V_h - P(V_h)) \oplus V_h^{\mathrm{f}} \overset{(6)}{=} (V_H - P(V_H)) \oplus V_h^{\mathrm{f}} = V_{H,h}^{\mathrm{ms}} \oplus V_h^{\mathrm{f}}.$$

Hence, any function $v_h \in V_h$ can be decomposed into $v_h = v_H^{\mathrm{ms}} + v^{\mathrm{f}}$ with $v_H^{\mathrm{ms}} = \mathfrak{I}_H^{-1}(\mathfrak{I}_H(v_h)) - P(\mathfrak{I}_H^{-1}(\mathfrak{I}_H(v_h)))$ and $v^{\mathrm{f}} = v_h - \mathfrak{I}_H^{-1}(\mathfrak{I}_H(v_h)) + P(\mathfrak{I}_H^{-1}(\mathfrak{I}_H(v_h)))$. Furthermore it holds $\langle A\nabla v_H^{\mathrm{ms}}, \nabla w^{\mathrm{f}} \rangle = 0$ for all $w^{\mathrm{f}} \in V_h^{\mathrm{f}}$. The space $V_{H,h}^{\mathrm{ms}}$ is a multiscale space of the same dimension as the coarse space V_H. However, note that it is only constructed on the basis of the oscillations of A. The oscillations of F are not taken into account. We will show that $V_{H,h}^{\mathrm{ms}}$ still yields the desired approximation properties.

We now introduce a basis of $V_{H,h}^{\mathrm{ms}}$. The image of the nodal basis function $\lambda_j \in V_H$ under the fine scale projection P is denoted by $\phi_j^h = P(\lambda_j) \in V_h^{\mathrm{f}}$, i.e., ϕ_j^h satisfies the corrector problem

$$\langle A\nabla\phi_j^h, \nabla w \rangle = \langle A\nabla\lambda_j, \nabla w \rangle \quad \text{for all } w \in V_h^{\mathrm{f}}. \tag{8}$$

A basis of $V_{H,h}^{\mathrm{ms}}$ is then given by the modified nodal basis

$$\{\lambda_j^{\mathrm{ms}} := \lambda_j - \phi_j^h \mid 1 \leq j \leq J\}. \tag{9}$$

As we can see, solving (8) involves a fine scale computation on the whole domain Ω. However, since the right hand side has small support, we are able to truncate the computations. As we will see in the next section, the correctors show an exponential decay outside of the support if the coarse shape function λ_j.

We define a (preliminary) RMM approximation without truncation.

Definition 2 (RMM approximation without truncation)**.** *The Galerkin approximation $u_{H,h}^{\mathrm{ms}} \in V_{H,h}^{\mathrm{ms}}$ of the exact solution u of problem* (1) *is defined as the solution of*

$$\langle A\nabla u_{H,h}^{\mathrm{ms}}, \nabla v \rangle + \langle F(u_{H,h}^{\mathrm{ms}}, \nabla u_{H,h}^{\mathrm{ms}}), v \rangle = \langle g, v \rangle \quad \textit{for all } v \in V_{H,h}^{\mathrm{ms}}. \tag{10}$$

3.4 Localization

So far, in order to construct a suitable multiscale space, we derived a set of linear fine scale problems (8) that can be solved in parallel. Still, as already mentioned in the previous section, these corrector problems are fine scale equations formulated on the whole domain Ω which makes them almost as expensive as the original problem. However, in [22] it was shown that the correction ϕ_j^h decays with exponential speed outside of the support of the coarse basis function λ_j. We specify this feature as follows. Let $k \in \mathbb{N}_{>0}$. We define nodal patches $\omega_{j,k}$ of k coarse grid layers centered around the node $z_j \in \mathcal{N}_H$ by

$$\begin{aligned} \omega_{j,1} &:= \operatorname{supp} \lambda_j = \cup \left\{ T \in \mathcal{T}_H \mid z_j \in \overline{T} \right\}, \\ \omega_{j,k} &:= \cup \left\{ T \in \mathcal{T}_H \mid \overline{T} \cap \overline{\omega}_{j,k-1} \neq \emptyset \right\} \quad \text{for } k \geq 2. \end{aligned} \tag{11}$$

These are the truncated computational domains for the corrector problems (8). The fast decay is summarized by the following lemma:

Lemma 3 (Decay of the local correctors [22])**.** *Let assumptions (A1) and (A5)-(A8) be fulfilled. Then, for all nodes $z_j \in \mathcal{N}_H$ and for all $k \in \mathbb{N}_{>0}$ there holds the following estimate for the correctors ϕ_j^h:*

$$\|A^{1/2} \nabla \phi_j^h\|_{L^2(\Omega \setminus \omega_{j,k})} \lesssim e^{-(\alpha/\beta)^{1/2} k} \|A^{1/2} \nabla \phi_j^h\|_{L^2(\Omega)}.$$

Remind the definition of '$\lesssim$' at the end of Section 2.

This fast decay motivates an approximation of ϕ_j^h on the truncated nodal patches $\omega_{j,k}$. We therefore define localized fine scale spaces by intersecting V_h^{f} with those functions that vanish outside the patch $\omega_{j,k}$, i.e.

$$V_h^{\mathrm{f}}(\omega_{j,k}) := \{ v \in V_h^{\mathrm{f}} \mid v_{|\Omega \setminus \omega_{j,k}} = 0 \}$$

for a given node $z_j \in \mathcal{N}_H$. The solutions $\phi_{j,k}^h \in V_h^{\mathrm{f}}(\omega_{j,k})$ of

$$\langle A \nabla \phi_{j,k}^h, \nabla w \rangle = \langle A \nabla \lambda_j, \nabla w \rangle \quad \text{for all } w \in V_h^{\mathrm{f}}(\omega_{j,k}), \tag{12}$$

are approximations of ϕ_j^h from (8) with local support and therefore cheap to solve. We define localized multiscale finite element spaces by

$$V_{H,h}^{\mathrm{ms},k} = \operatorname{span}\{ \lambda_{j,k}^{\mathrm{ms}} := \lambda_j - \phi_{j,k}^h \mid 1 \leq j \leq J \} \subset V_h. \tag{14}$$

We can now define a RMM approximation including truncation:

Definition 4 (RMM approximation with truncation)**.** *The Galerkin approximation* $u_{H,h}^{\mathrm{ms},k} \in V_{H,h}^{\mathrm{ms},k}$ *of the exact solution* u *of problem* (1) *is defined as the solution of*

$$\langle A\nabla u_{H,h}^{\mathrm{ms},k}, \nabla v\rangle + \langle F(u_{H,h}^{\mathrm{ms},k}, \nabla u_{H,h}^{\mathrm{ms},k}), v\rangle = \langle g, v\rangle \quad \text{for all } v \in V_{H,h}^{\mathrm{ms},k}. \tag{15}$$

Note, that changing F and g does not change the multiscale basis $\{\lambda_{j,k}^{\mathrm{ms}} \mid 1 \leq j \leq J\}$. Once $V_{H,h}^{\mathrm{ms},k}$ is computed, it can be reused for various data functions F and g. This makes the new problems cheap to solve.

3.5 A-priori error estimate

We are now prepared to state the main result of this contribution, namely the optimal convergence of the method for the case that the local patches $\omega_{j,k}$ have a diameter of order $H\log(H^{-1})$:

Theorem 5. *Let* $u \in H_0^1(\Omega)$ *denote the exact solution given by problem (1), let* $u_h \in V_h$ *denote the corresponding finite element approximation in the Lagrange space with a highly resolved computational grid (i.e. the solution of (4)) and let* $u_{H,h}^{\mathrm{ms},k} \in V_{H,h}^{\mathrm{ms},k}$ *be the solution of our proposed multiscale method with truncation (i.e. the solution of (15)). If assumptions (A1)-(A8) are fulfilled and if the diameter of the local patches* $\omega_{j,k}$ *fulfills* $O(\mathrm{diam}(\omega_{j,k})) \gtrsim H\log(\|H^{-1}\|_{L^\infty(\Omega)})$ *(*$1 \leq j \leq J$*), then the following a-priori error estimate holds true:*

$$\|u - u_{H,h}^{\mathrm{ms},k}\|_{H^1(\Omega)} \leq C\left(\|H\|_{L^\infty(\Omega)} + \|u - u_h\|_{H^1(\Omega)}\right).$$

Here, C *denotes a generic constant with the property* $C = O(1)$*. In particular,* C *is independent of the oscillations of* A *and* F *and only depends on the shape regularity of* $\mathcal{T}_H$ *and* $\mathcal{T}_h$*. A suitable choice of* $m \in \mathbb{N}$ *with* $k = m \cdot \log(\|H^{-1}\|_{L^\infty(\Omega)})$ *depends on the contrast* $\frac{\beta}{\alpha}$*. The larger the contrast, the bigger should be* m*.*

A proof of Theorem 5 is presented in the subsequent section. In particular, the result is a conclusion from Theorem 8 which is stated in Section 4 below. In Theorem 8 we also give details on the generic constant C. We will see that it essentially depends on $\frac{(L_1+L_2)}{\alpha}$. Recall that L_1 and L_2 denote the Lipschitz constants of F (c.f. (A2)) and that α is the smallest eigenvalue of A. This shows the significance of assumption (A4). For instance, consider the scenario of a pollutant being transported by groundwater flow. In this case, A describes the hydraulic conductivity which changes its properties on a scale of size ϵ. On the other hand, F describes the gravity driven flow

that is scaled with the so called Péclet number. However, in the described scenario the Péclet number is of order ϵ^{-1} (c.f. Bourlioux and Majda [4]) implying that $\mathrm{O}(L_1) = \epsilon^{-1}$. So the generic constant C is of order ϵ^{-1}. This means that we need $H < \epsilon$, i.e. we still need to resolve the micro structure with the coarse grid $\mathcal{T}_H$ producing the same costs as the original problem. On the other hand, if $H \gg \epsilon$ the estimate stated in Theorem 5 is of no value, because the right hand side remains large.

4 Error Analysis

This section is devoted to the proof of Theorem 5. In particular, we state a detailed version of the result (see Theorem 8 below), where we specify the occurring constants. The proof is splitted into several lemmata. We start with an a-priori error estimate for a RMM approximation without truncation:

Lemma 6. *Let $u_h \in V_h$ denote the highly resolved finite element approximation defined via equation (4) and let $u_{H,h}^{\mathrm{ms}} \in V_{H,h}^{\mathrm{ms}}$ denote the RMM approximation given by equation (10). Under assumptions (A1)-(A3) and (A5)-(A8), the following a-priori error estimate holds true:*

$$\begin{aligned} &|u_h - u_H^{\mathrm{ms}}|_{H^1(\Omega)} \\ &\quad\lesssim\ \tilde{C}_0\big(\|Hg\|_{L^2(\Omega)} + \|H\|_{L^\infty(\Omega)} C_p \frac{L_1 C_p + L_2}{c_0} \|g\|_{L^2(\Omega)}\big), \end{aligned}$$

where

$$\tilde{C}_0 := \left(\frac{\beta + \|H\|_{L^\infty(\Omega)}(L_1 C_p + L_2)}{c_0 \cdot \alpha}\right).$$

Proof. Due to (7), we know that there exist $\tilde{u}_{H,h}^{\mathrm{ms}} \in V_{H,h}^{\mathrm{ms}}$ and $\tilde{u}_h^{\mathrm{f}} \in V_h^{\mathrm{f}}$, such that

$$u_h = \tilde{u}_{H,h}^{\mathrm{ms}} + \tilde{u}_h^{\mathrm{f}}.$$

We use the Galerkin orthogonality obtained from the equations (4) and (10) to conclude for all $v \in V_{H,h}^{\mathrm{ms}}$:

$$\langle A\nabla(u_h - u_{H,h}^{\mathrm{ms}}), \nabla v\rangle + \langle F(u_h, \nabla u_h), v\rangle - \langle F(u_{H,h}^{\mathrm{ms}}, \nabla u_{H,h}^{\mathrm{ms}}), v\rangle = 0. \quad (16)$$

In particular $v = u_{H,h}^{\mathrm{ms}} - \tilde{u}_{H,h}^{\mathrm{ms}} \in V_{H,h}^{\mathrm{ms}}$ is an admissible test function in (16).

Together with $\mathfrak{I}_H(\tilde{u}_h^{\mathrm{f}}) = 0$, this yields:

$$
\begin{aligned}
& c_0|u_h - u_{H,h}^{\mathrm{ms}}|^2_{H^1(\Omega)} \\
& \overset{(2)}{\leq} \quad \langle A\nabla(u_h - u_{H,h}^{\mathrm{ms}}), \nabla(u_h - u_{H,h}^{\mathrm{ms}})\rangle \\
& \qquad + \langle F(u_h, \nabla u_h) - F(u_{H,h}^{\mathrm{ms}}, \nabla u_{H,h}^{\mathrm{ms}}), u_h - u_{H,h}^{\mathrm{ms}}\rangle \\
& \overset{(16)}{=} \quad \langle A\nabla(u_h - u_{H,h}^{\mathrm{ms}}), \nabla(u_h - \tilde{u}_{H,h}^{\mathrm{ms}})\rangle \\
& \qquad + \langle F(u_h, \nabla u_h) - F(u_{H,h}^{\mathrm{ms}}, \nabla u_{H,h}^{\mathrm{ms}}), u_h - \tilde{u}_{H,h}^{\mathrm{ms}}\rangle \\
& = \quad \langle A\nabla(u_h - u_{H,h}^{\mathrm{ms}}), \nabla\tilde{u}_h^{\mathrm{f}}\rangle + \langle F(u_h, \nabla u_h) - F(u_{H,h}^{\mathrm{ms}}, \nabla u_h), \tilde{u}_h^{\mathrm{f}} - \mathfrak{I}_H(\tilde{u}_h^{\mathrm{f}})\rangle \\
& \qquad + \langle F(u_{H,h}^{\mathrm{ms}}, \nabla u_h) - F(u_{H,h}^{\mathrm{ms}}, \nabla u_{H,h}^{\mathrm{ms}}), \tilde{u}_h^{\mathrm{f}} - \mathfrak{I}_H(\tilde{u}_h^{\mathrm{f}})\rangle \\
& \lesssim \quad \beta|u_h - u_{H,h}^{\mathrm{ms}}|_{H^1(\Omega)}|\tilde{u}_h^{\mathrm{f}}|_{H^1(\Omega)} \\
& \qquad + \|H\|_{L^\infty(\Omega)}(L_1\|u_h - u_{H,h}^{\mathrm{ms}}\|_{L^2(\Omega)} + L_2|u_h - u_{H,h}^{\mathrm{ms}}|_{H^1(\Omega)})|\tilde{u}_h^{\mathrm{f}}|_{H^1(\Omega)} \\
& \lesssim \quad (\beta + \|H\|_{L^\infty(\Omega)}(L_1 C_p + L_2)) \cdot |u_h - u_{H,h}^{\mathrm{ms}}|_{H^1(\Omega)} \cdot |\tilde{u}_h^{\mathrm{f}}|_{H^1(\Omega)}.
\end{aligned}
$$

With $\langle A\nabla\tilde{u}_{H,h}^{\mathrm{ms}}, \nabla\tilde{u}_h^{\mathrm{f}}\rangle = 0$ and with $\mathfrak{I}_H(v_{\mathrm{f}}) = 0$ for all $v_{\mathrm{f}} \in V^{\mathrm{f}}$ we get

$$
\begin{aligned}
& \alpha|\tilde{u}_h^{\mathrm{f}}|^2_{H^1(\Omega)} \leq \langle A\nabla\tilde{u}_h^{\mathrm{f}}, \nabla\tilde{u}_h^{\mathrm{f}}\rangle \\
& \quad = \quad \langle A\nabla u_h, \nabla\tilde{u}_h^{\mathrm{f}}\rangle = \langle g, \tilde{u}_h^{\mathrm{f}}\rangle - \langle F(u_h, \nabla u_h), \tilde{u}_h^{\mathrm{f}}\rangle \\
& \quad = \quad \langle g, \tilde{u}_h^{\mathrm{f}} - \mathfrak{I}_H(\tilde{u}_h^{\mathrm{f}})\rangle - \langle F(u_h, \nabla u_h), \tilde{u}_h^{\mathrm{f}} - \mathfrak{I}_H(\tilde{u}_h^{\mathrm{f}})\rangle \\
& \quad \overset{(3)}{\lesssim} \quad \big(\|Hg\|_{L^2(\Omega)} + \|H\|_{L^\infty(\Omega)} C_p \frac{L_1 C_p + L_2}{c_0}\|g\|_{L^2(\Omega)}\big) \cdot |\tilde{u}_h^{\mathrm{f}}|_{H^1(\Omega)}.
\end{aligned}
$$

The theorem follows by combing the results. □

The next lemma is a conclusion from the previous one:

Lemma 7. *Let $u_h \in V_h$ denote the fine scale approximation obtained from equation (4) and let $u_{H,h}^{\mathrm{ms},k} \in V_{H,h}^{\mathrm{ms},k}$ denote the solution of problem (15) (fully discrete RMM with truncation). If the assumptions (A1)-(A3) and (A5)-(A8) hold true we obtain the following estimate:*

$$
\begin{aligned}
& |u_h - u_{H,h}^{\mathrm{ms},k}|_{H^1(\Omega)} \\
& \quad \lesssim \quad \tilde{C}_2\|g\|_{L^2(\Omega)}\|H\|_{L^\infty(\Omega)} + \tilde{C}_3 \min_{v_{H,h}^{\mathrm{ms},k} \in V_{H,h}^{\mathrm{ms},k}} \|A^{\frac{1}{2}}\nabla(u_{H,h}^{\mathrm{ms}} - v_{H,h}^{\mathrm{ms},k})\|_{L^2(\Omega)},
\end{aligned}
$$

where

$$\tilde{C}_1 := (\beta + (L_1 C_p + L_2) C_p) \cdot \left(\frac{\beta + \|H\|_{L^\infty(\Omega)} (L_1 C_p + L_2)}{c_0^2 \cdot \alpha} \right),$$
$$\tilde{C}_2 := \tilde{C}_1 + \tilde{C}_1 \cdot C_p \frac{L_1 C_p + L_2}{c_0},$$
$$\tilde{C}_3 := \frac{1 + \alpha^{-\frac{1}{2}} (L_1 C_p + L_2) C_p}{c_0}.$$

Proof. Let $v_{H,h}^{\mathrm{ms},k} \in V_{H,h}^{\mathrm{ms},k}$ denote an arbitrary element. Using the Galerkin orthogonality obtained from (4) and (15), we start in the same way as in the proof of Lemma 6 to get:

$$\begin{aligned}
& c_0 |u_h - u_{H,h}^{\mathrm{ms},k}|_{H^1(\Omega)}^2 \\
& \overset{(2)}{\leq} \langle A \nabla (u_h - u_{H,h}^{\mathrm{ms},k}), \nabla (u_h - u_{H,h}^{\mathrm{ms},k}) \rangle \\
& \qquad + \langle F(u_h, \nabla u_h) - F(u_{H,h}^{\mathrm{ms},k}, \nabla u_{H,h}^{\mathrm{ms},k}), u_h - u_{H,h}^{\mathrm{ms},k} \rangle \\
& \overset{(16)}{=} \langle A \nabla (u_h - u_{H,h}^{\mathrm{ms},k}), \nabla (u_h - u_{H,h}^{\mathrm{ms}}) + \nabla (u_{H,h}^{\mathrm{ms}} - v_{H,h}^{\mathrm{ms},k}) \rangle \\
& \qquad + \langle F(u_h, \nabla u_h) - F(u_{H,h}^{\mathrm{ms},k}, \nabla u_{H,h}^{\mathrm{ms},k}), (u_h - u_{H,h}^{\mathrm{ms}}) + (u_{H,h}^{\mathrm{ms}} - v_{H,h}^{\mathrm{ms},k}) \rangle \\
& \leq (\beta + (L_1 C_p + L_2) C_p) |u_h - u_{H,h}^{\mathrm{ms},k}|_{H^1(\Omega)} |u_h - u_{H,h}^{\mathrm{ms}}|_{H^1(\Omega)} \\
& \qquad + (1 + \alpha^{-\frac{1}{2}} (L_1 C_p + L_2) C_p) |u_h - u_{H,h}^{\mathrm{ms},k}|_{H^1(\Omega)} \|A^{\frac{1}{2}} \nabla (u_{H,h}^{\mathrm{ms}} - v_{H,h}^{\mathrm{ms},k})\|_{L^2(\Omega)}.
\end{aligned}$$

Dividing by $|u_h - u_{H,h}^{\mathrm{ms},k}|_{H^1(\Omega)}$ and estimating $|u_h - u_{H,h}^{\mathrm{ms}}|_{H^1(\Omega)}$ with Lemma 6 yields the result. □

Combining the results of Lemma 3 and Lemma 7 yields the main result of this contribution:

Theorem 8. *Let $u_h \in V_h$ be solution of (4) and let $u_{H,h}^{\mathrm{ms},k} \in V_{H,h}^{\mathrm{ms},k}$ be the solution of (15). If the assumptions (A1)-(A3) and (A5)-(A8) hold true and if the number of layers k fulfills $k \gtrsim \log(\|H^{-1}\|_{L^\infty(\Omega)})$, then it holds*

$$|u_h - u_{H,h}^{\mathrm{ms},k}|_{H^1(\Omega)} \lesssim \tilde{C} \|H\|_{L^\infty(\Omega)} \|g\|_{L^2(\Omega)},$$

where

$$\tilde{C} := \tilde{C}_2 + C_p \frac{\beta}{c_0} \tilde{C}_3$$

and with $\tilde{C}_2$ and $\tilde{C}_3$ as in Lemma 7.

Proof. We define $w_{H,h}^{\mathrm{ms},k} \in V_{H,k}^{\mathrm{ms}}$ by

$$w_{H,h}^{\mathrm{ms},k} := \sum_{j=1}^{J} u_{H,h}^{\mathrm{ms}}(z_j)\lambda_{j,k}^{\mathrm{ms}} = \sum_{j=1}^{J} u_{H,h}^{\mathrm{ms}}(z_j)(\lambda_j - \phi_{j,k}^h)$$

where $u_{H,h}^{\mathrm{ms}}(z_j)$, $j = 1, 2, \ldots, J$, are the coefficients in the basis representation of $u_{H,h}^{\mathrm{ms}}$ from Definition 2. Hence,

$$\begin{aligned} &\min_{v_{H,h}^{\mathrm{ms},k} \in V_{H,h}^{\mathrm{ms},k}} \|A^{\frac{1}{2}}\nabla(u_{H,h}^{\mathrm{ms}} - v_{H,h}^{\mathrm{ms},k})\|_{L^2(\Omega)}^2 \\ &\leq \|A^{\frac{1}{2}}\nabla(u_{H,h}^{\mathrm{ms}} - w_{H,h}^{\mathrm{ms},k})\|_{L^2(\Omega)}^2 \\ &\lesssim \sum_{j=1}^{J} k^d u_{H,h}^{\mathrm{ms}}(z_j)^2 \|A^{1/2}\nabla(\phi_j^h - \phi_{j,k}^h)\|_{L^2(\Omega)}^2. \end{aligned} \tag{17}$$

For details on the last step, we refer to Lemma 18 in [22]. Due to the Galerkin orthogonality for the corrector problems we get

$$\|A^{1/2}\nabla(\phi_j^h - \phi_{j,k}^h)\|_{L^2(\Omega)}^2 \lesssim \|A^{1/2}\nabla\phi_j^h\|_{L^2(\Omega\setminus\omega_{j,k-1})}^2.$$

We refer to the first part of the proof of Lemma 8 in [22] for details. The application of Lemma 3, (8) and some inverse inequality yield

$$\begin{aligned} \|A^{1/2}\nabla(\phi_j^h - \phi_{j,k}^h)\|_{L^2(\Omega)}^2 &\lesssim e^{-2(\alpha/\beta)^{1/2}k}\|A^{1/2}\nabla\phi_j^h\|_{L^2(\Omega)}^2 \\ &\leq e^{-2(\alpha/\beta)^{1/2}k}\|A^{1/2}\nabla\lambda_j^h\|_{L^2(\Omega)}^2 \\ &\leq \beta e^{-2(\alpha/\beta)^{1/2}k} H^{-2}\|\lambda_j^h\|_{L^2(\Omega)}^2 \end{aligned}$$

By choosing $k = m \cdot \log(\|H^{-1}\|_{L^\infty(\Omega)})$ with $m \in \mathbb{N}$, we can achieve an arbitrary fast polynomial convergence of this term in H (this will also cancel the k^d term). However, we bound this by a linear convergence since this is fastest rate that we can obtain for the whole error. Finally, the combination of this estimate and (17) plus $\sum_{j=1}^{J} u_{H,h}^{\mathrm{ms}}(z_j)^2\|\lambda_j^h\|_{L^2(\Omega)}^2 \lesssim \|\mathfrak{I}_H u_{H,h}^{\mathrm{ms}}\|_{L^2(\Omega)}^2 \lesssim \|\nabla u_{H,h}^{\mathrm{ms}}\|_{L^2(\Omega)}^2 \leq C_p^2 c_0^{-2}\|g\|_{L^2(\Omega)}^2$ yields the assertion. □

5 The Multiscale Newton scheme

In this section we discuss a solution algorithm for handling the nonlinear multiscale problem (15). For this purpose, we consider a damped Newton's

method in the multiscale space $V_{H,h}^{\mathrm{ms},k}$. Remind the considered problem: we are looking for $u \in H_0^1(\Omega)$ with

$$\langle B(u), v\rangle_{H^{-1},H_0^1} = \langle g, v\rangle \quad \text{for all } v \in H_0^1(\Omega),$$

where we introduced the notation

$$\langle B(v), w\rangle_{H^{-1},H_0^1} := \langle A\nabla v, \nabla w\rangle + \langle F(\cdot, v, \nabla v), w\rangle.$$

Here, $B : H_0^1(\Omega) \to H^{-1}(\Omega)$ is a hemicontinuous and strongly monotone operator due to assumption (A3). As already mentioned, under these assumptions, the Browder-Minty theorem yields a unique solution of the above problem. However, we will need an additional assumption on F to guarantee that the Newton scheme converges:

Assumption 3. *Let $DF(x,\cdot,\cdot)$ denote the Jacobian matrix of $F(x,\cdot,\cdot)$.*

(A9) We assume that there exists some constant $L_D \geq 0$ so that for almost every x in Ω and for all $(\xi,\zeta) \in \mathbb{R}\times\mathbb{R}^d$

$$|DF(x,\xi_1,\zeta_1) - DF(x,\xi_2,\zeta_2)| \leq L_D|(\xi_1,\zeta_1) - (\xi_2,\zeta_2)|,$$

i.e. $F(x,\cdot,\cdot) \in W^{2,\infty}(\mathbb{R}\times\mathbb{R}^d)$.

For clarity of the presentation we will leave out several indices within this section. In particular, we make use of the following notation:

Definition 9. *For simplicity, we define*

$$V^{\mathrm{ms}} := V_{H,h}^{\mathrm{ms},k} \quad \textit{with basis } \lambda_j^{\mathrm{ms}} := \lambda_{j,k}^{\mathrm{ms}} = \lambda_j - \phi_{j,k}^h \textit{ for } 1 \leq j \leq J.$$

Furthermore, we denote $u^{\mathrm{ms}} := u_{H,h}^{\mathrm{ms},k}$. Additionally, let

$$\partial_1 F(x,\xi,\zeta) := \partial_\xi F(x,\xi,\zeta) \quad \textit{and} \quad \partial_2 F(x,\xi,\zeta) := \partial_\zeta F(x,\xi,\zeta).$$

We now describe the Newton strategy in detail. The fully discrete multiscale problem to solve reads:

$$\text{find } u^{\mathrm{ms}} \in V^{\mathrm{ms}}: \quad \langle A\nabla u^{\mathrm{ms}}, \nabla\lambda_j^{\mathrm{ms}}\rangle + \langle F(\cdot, u^{\mathrm{ms}}, \nabla u^{\mathrm{ms}}), \lambda_j^{\mathrm{ms}}\rangle - \langle g, \lambda_j^{\mathrm{ms}}\rangle = 0$$

for all $1 \leq j \leq J$. Again, using Browder-Minty, u^{ms} exists and is unique. Accordingly, we get the following well posed algebraic version of the problem:

$$\text{find } \bar{\alpha} \in \mathbb{R}^J: \quad G(\bar{\alpha}) = 0$$

and where $G : \mathbb{R}^J \to \mathbb{R}^J$ is given by

$$\begin{aligned}&(G(\alpha))_l \qquad\qquad (18)\\ :=\ &\sum_{j=1}^{J} \alpha_j \langle A\nabla\lambda_j^{\mathrm{ms}}, \nabla\lambda_l^{\mathrm{ms}}\rangle + \langle F(\cdot, \sum_{j=1}^{J}\alpha_j\lambda_j^{\mathrm{ms}}, \sum_{j=1}^{J}\alpha_j\nabla\lambda_j^{\mathrm{ms}}), \lambda_l^{\mathrm{ms}}\rangle - \langle g, \lambda_l^{\mathrm{ms}}\rangle.\end{aligned}$$

We have the relation $u^{\mathrm{ms}} = \sum_{j=1}^{J} \bar{\alpha}_j \lambda_j^{\mathrm{ms}}$. Before we can apply the Newton method to (18), we need to ensure that the iterations of the scheme are well defined. The following lemma ensures this:

Lemma 10. *Let $(X, \|\cdot\|_X)$ denote a Hilbert space with dual space X'. Let furthermore $B : X \to X'$ be a hemicontinuous and strongly monotone operator on X, i.e. there exists $c_0 > 0$ so that*

$$\begin{aligned}&\langle B(v) - B(w), v - w\rangle_X \geq c_0\|v - w\|_X^2 \quad \textit{for all } v, w \in X \textit{ and}\\ &\quad s \mapsto \langle B(u + sv), w\rangle_X\end{aligned}$$

is a continuous function on $[0, 1]$ for all $u, v, w \in X$. Let X_N denote a finite dimensional subspace with basis $\{\psi_1, ..., \psi_N\}$ and let $b : \mathbb{R}^N \to V_N$ define the linear bijection with $b(\alpha) := \sum_{i=1}^{N} \alpha_i\psi_i$. If $G(\alpha) := b^{-1}(A(b(\alpha)))$, then the Jacobi matrix $DG(\alpha) \in \mathbb{R}^{n\times n}$ has only positive eigenvalues.

Proof. Let B' denote the Fréchet the derivative of B, given by

$$B'(u)(v) = \lim_{s\to 0} \frac{B(u + sv) - B(u)}{s} \quad \text{for } u, v \in X.$$

This and the strong monotonicity yield:

$$\begin{aligned}\langle B'(u)(v), v\rangle_{H^{-1},H_0^1} &= \lim_{s\to 0} \frac{(B(u + sv) - B(u))(v)}{s}\\ &= \lim_{s\to 0} \frac{1}{s^2}(B(u + sv) - B(u))(u + sv - u) \qquad (19)\\ &\geq \lim_{s\to 0} \frac{1}{s^2} c_0\|sv\|^2 = c_0\|v\|^2.\end{aligned}$$

Next, observe that b induces an inner product on $\mathbb{R}^N$ by $(\alpha_1, \alpha_2)_b := \langle b(\alpha_1), b(\alpha_2)\rangle_X$. Let $\alpha := b^{-1}(u)$ then we get

$$\begin{aligned}B'(u)(\psi_i) &= \lim_{s\to 0} \frac{B(u + s\psi_i)}{B(u)} s\\ &= \lim_{s\to 0} \frac{(b \circ b^{-1})(B(\sum_{j=1}^{N}(\alpha_j + \delta_{ij})\psi_j) - (b \circ b^{-1})(B(\sum_{j=1}^{N}\alpha_j\psi_j))}{s}\\ &= b\left(\lim_{s\to 0} \frac{G(\alpha_j + se_j) - G(\alpha_j)}{s}\right)\\ &= b(D_\alpha G(\alpha)e_j).\end{aligned}$$

Using this, we get for arbitrary $\xi \in \mathbb{R}^N$ and $v_\xi := b(\xi)$:

$$\begin{aligned}(D_\alpha G(\alpha)\xi, \xi)_b &= \sum_{i,j}^{N} \xi_i \xi_j (D_\alpha G(\alpha) e_i, e_j)_b \\ &= \sum_{i,j}^{N} \xi_i \xi_j (b(D_\alpha G(\alpha) e_i), b(e_j))_X \\ &= \sum_{i,j}^{N} \xi_i \xi_j (B'(u)(\psi_i), \psi_j)_X \\ &= (B'(u)(v_\xi), v_\xi)_X \overset{(19)}{\geq} c_0 \|v_\xi\|_X^2 = c_0 \|\xi\|_b^2.\end{aligned}$$

Since all norms in $\mathbb{R}^N$ are equivalent we have the desired result. □

Now, we can apply the damped Newton method for solving the nonlinear algebraic equation $G(\bar{\alpha}) = 0$. If $D_\alpha G$ denotes the Jacobian matrix of G, we get the following iteration scheme:

$$\alpha^{(n+1)} := \alpha^{(n)} + \triangle\alpha^{(n)},$$

where $\triangle\alpha^{(n)}$ solves

$$D_\alpha G(\alpha^{(n)}) \triangle\alpha^{(n)} = -G(\alpha^{(n)}). \tag{20}$$

Here, $D_\alpha G$ is given by:

$$\begin{aligned}D_{\alpha_i} \left(G(\alpha)\right)_l := \langle A \nabla \lambda_i^{\mathrm{ms}}, \nabla \lambda_l^{\mathrm{ms}} \rangle + \langle \partial_1 F(\cdot, \sum_{j=1}^{J} \alpha_j \lambda_j^{\mathrm{ms}}, \sum_{j=1}^{J} \alpha_j \nabla \lambda_j^{\mathrm{ms}}) \lambda_i^{\mathrm{ms}}, \lambda_l^{\mathrm{ms}} \rangle \\ + \langle \partial_2 F(\cdot, \sum_{j=1}^{J} \alpha_j \lambda_j^{\mathrm{ms}}, \sum_{j=1}^{J} \alpha_j \nabla \lambda_j^{\mathrm{ms}}) \cdot \nabla \lambda_i^{\mathrm{ms}}, \lambda_l^{\mathrm{ms}} \rangle.\end{aligned}$$

Lemma 10 ensures that equation (20) has a unique solution $\triangle\alpha^{(n)}$, i.e. that the Newton iteration is well posed. Since $G \in C^1(\mathbb{R}^N)$ has a nonsingular Jacobian matrix $D_\alpha G$ (due to Lemma 10) and since we have Lipschitz-continuity of $D_\alpha G$ (due to Assumption 3), we have the that Newton scheme converges quadratically as long as the starting value is close enough to the exact solution (c.f. [8]). However, this means that we can only guarantee local convergence of the method. In order to ensure global convergence, we can use a simple damping strategy, i.e. we are looking for a damping parameter $\zeta \in (0, 1]$ so that $\alpha^{(n+1)} := \alpha^{(n)} + \zeta \triangle\alpha^{(n)}$ with the property $G(\alpha^{(n+1)}) <$

$G(\alpha^{(n)})$. Under the same assumptions (i.e. (A1)-(A3) and (A9)), we get that the damped Newton scheme converges, i.e. there exists a (damping) interval $(0, \zeta_0]$ and for any $\zeta \in (0, \zeta_0]$ we get the above 'damping property'. This is an easy observation if we look at the function $h(\zeta) := |G(\alpha^{(n)} + \zeta \triangle \alpha^{(n)})|^2$ which fulfills $h(0) > 0$ and $h'(0) = -2G(\alpha^{(n)}) \cdot G(\alpha^{(n)}) < 0$. In summary we have globally linear convergence of the method (using damping) and locally (i.e. in an environment of the solution) even quadratic convergence using the classical Newton scheme without damping.

With these considerations, we can state the full algorithm below. Recall that $\mathcal{N}_H$ denotes the set of interior vertices of $\mathcal{T}_H$ and for $z_j \in \mathcal{N}_H$, $\lambda_j \in V_H$ denotes the corresponding nodal basis function.

Note that in the presented algorithm, each iteration starts with the damping parameter $\lambda_n = 1$ and we do not use damping parameters from previous iterations. The advantage is that we automatically get quadratic convergence of the Newton scheme as soon as we leave the region where damping is required. Therefore, damping is only used when really necessary.

Proposition 11. *We use the notation stated in Definition 9. Let $u \in H_0^1(\Omega)$ denote the solution of (1), let $u_h \in V_h$ denote the solution of (4) and let $u^{\mathrm{ms}} \in V^{\mathrm{ms}}$ denote the solution of (15)). Furthermore, we let $u^{\mathrm{ms},(n)} := u_{H,h}^{\mathrm{ms},k,(n)}$ define the n'th iterate from the damped Newton Variational Multiscale Method stated in the algorithm. Under assumptions (A1)-(A9), the Newton step (20) is well posed, yields an unique solution and $u^{\mathrm{ms},(n)}$ converges at least linearly to u^{ms}. If furthermore $k \gtrsim \log(\|H^{-1}\|_{L^\infty(\Omega)})$, the following a-priori error estimates hold true:*

$$\|u - u^{\mathrm{ms}}\|_{H^1(\Omega)} \leq C \left(\|H\|_{L^\infty(\Omega)} + \|u - u_h\|_{H^1(\Omega)} \right)$$

where C is a generic constant with the property $C = O(1)$ (see Theorem 5 and 8 for details) and

$$\|u^{\mathrm{ms}} - u^{\mathrm{ms},(n)}\|_{H^1(\Omega)} \leq L_n(H) \|u^{\mathrm{ms}} - u^{\mathrm{ms},(n-1)}\|_{H^1(\Omega)}.$$

Here, we have $L_n(H) < 1$.

If $u^{\mathrm{ms},(n-1)}$ is sufficiently close to u^{ms}, we even get quadratic convergence of the Newton scheme, i.e.:

$$\|u^{\mathrm{ms}} - u^{\mathrm{ms},(n)}\|_{H^1(\Omega)} \leq L_n(H) \|u^{\mathrm{ms}} - u^{\mathrm{ms},(n-1)}\|_{H^1(\Omega)}^2.$$

with

$$L_n(H) \leq \frac{\|(D_\alpha G)^{-1}\|_{L^\infty(\mathbb{R}^N)}}{L},$$

Algorithm: dampedNewtonRMM($abstol$, $reltol$, $\alpha^{(0)}$, k)

In parallel **foreach** $z_j \in \mathcal{N}_H$ **do**
  compute $\phi^h_{j,k} \in V^{\mathrm{f}}_h(\omega_{j,k})$ with

$$\langle A\nabla\phi^h_{j,k}, \nabla w\rangle = \langle A\nabla\lambda_j, \nabla w\rangle \quad \text{for all } w \in V^{\mathrm{f}}_h(\omega_{j,k}).$$

end
Set $V^{\mathrm{ms},k}_{H,h} := \{\lambda_j - \phi^h_{j,k} \mid 1 \le j \le J\}$. Set $\lambda^{\mathrm{ms}}_{j,k} = \lambda_j - \phi^h_{j,k}$.

Set $\alpha^{(n)} := \alpha^{(0)}$. Set $u^{\mathrm{ms},k,(n)}_{H,h} := \sum_{j=1}^J \alpha^{(n)}_j \lambda^{\mathrm{ms}}_{j,k}$. Set

$$(G(\alpha))_i := \sum_{j=1}^J \alpha_j \langle A\nabla\lambda^{\mathrm{ms}}_{j,k}, \nabla\lambda^{\mathrm{ms}}_{i,k}\rangle + \langle F(\cdot, \sum_{j=1}^J \alpha_j\lambda^{\mathrm{ms}}_{j,k}, \sum_{j=1}^J \alpha_j \nabla\lambda^{\mathrm{ms}}_{j,k}) - g, \lambda^{\mathrm{ms}}_{i,k}\rangle.$$

Set $tol := |G(\alpha^{(0)})|_2 \cdot reltol + abstol$.

while $|G(\alpha^{(n)})|_2 > tol$ **do**
  Set $u^{\mathrm{ms},k,(n)}_{H,h} := \sum_{j=1}^J \alpha^{(n)}_j \lambda^{\mathrm{ms}}_{j,k}$.
  Define the entries of the stiffness matrix $M^{(n)}$ by:

$$\begin{aligned} M^{(n)}_{il} \;:=\; & \langle A\nabla\lambda^{\mathrm{ms}}_{l,k}, \nabla\lambda^{\mathrm{ms}}_{i,k}\rangle + \langle \partial_1 F(\cdot, u^{\mathrm{ms},k,(n)}_{H,h}, \nabla u^{\mathrm{ms},k,(n)}_{H,h})\lambda^{\mathrm{ms}}_{l,k}, \lambda^{\mathrm{ms}}_{i,k}\rangle \\ & + \langle \partial_2 F(\cdot, u^{\mathrm{ms},k,(n)}_{H,h}, \nabla u^{\mathrm{ms},k,(n)}_{H,h}) \cdot \nabla\lambda^{\mathrm{ms}}_{l,k}, \lambda^{\mathrm{ms}}_{i,k}\rangle. \end{aligned}$$

  Define the entries of the right hand side by:

$$F^{(n)}_i := \langle g, \lambda^{\mathrm{ms}}_{i,k}\rangle - \langle A\nabla u^{\mathrm{ms},k,(n)}_{H,h}, \nabla\lambda^{\mathrm{ms}}_{i,k}\rangle - \langle F(\cdot, u^{\mathrm{ms},k,(n)}_{H,h}, \nabla u^{\mathrm{ms},k,(n)}_{H,h}), \lambda^{\mathrm{ms}}_{i,k}\rangle.$$

  Find $(\triangle\alpha)^{(n+1)} \in \mathbb{R}^J$, with

$$M^{(n)}(\triangle\alpha)^{(n+1)} = F^{(n)}.$$

  Set $\lambda_n := 1$. Set $\alpha^{(n+1)} := \alpha^{(n)} + \lambda_n \triangle\alpha^{(n)}$.
  while $G(\alpha^{(n+1)}) \ge G(\alpha^{(n)})$ **do**
    Set $\lambda_n := \frac{1}{2}\lambda_n$. Set $\alpha^{(n+1)} := \alpha^{(n)} + \lambda_n \triangle\alpha^{(n)}$.
  end
  Set $\alpha^{(n)} := \alpha^{(n+1)}$. Set $tol := |G(\alpha^{(n)})|_2 \cdot reltol + abstol$.
end

Set $u^{\mathrm{ms},k,(n)}_{H,h} := \sum_{j=1}^J \alpha^{(n)}_j \lambda^{\mathrm{ms}}_{j,k}$.

where L denotes the Lipschitz-constant of $D_\alpha G$. As indicated, $L_n(H)$ typically depends on the mesh size. However, in some cases of semi-linear problems, it is possible to bound $L_n(H)$ independent of the triangulation (c.f. [17]). In particular, if $F(x, u, \nabla u) = F(x, u)$ (i.e. no dependency on ∇u) we get that $L_n(H) = L_n$ independent of the underlying mesh. The proof can be obtained analogously to the proof of Proposition 4.1 in [17]. The proof fails for general $F(x, u, \nabla u)$.

Remark 12. *Note that the proposed method only requires the computation of the multiscale basis $\{\lambda_j^{\mathrm{ms}} | \, 1 \leq j \leq J\}$ once at the beginning. For each iteration step of the damped Newton scheme, (20) is a cheap (low dimensional) linear problem that can reuse the initially computed multiscale basis. If the multiscale basis was computed using the nonlinear term F, local corrector problems would have to be solved for each Newton step newly, making the whole procedure expensive.*

References

[1] H. W. Alt and S. Luckhaus. Quasilinear elliptic-parabolic differential equations. *Math. Z.*, 183(3):311–341, 1983.

[2] H. Berninger. Non-overlapping domain decomposition for the Richards equation via superposition operators. *Lect. Notes Comput. Sci. Eng.*, 70: Springer, Berlin, 169–176, 2009.

[3] H. Berninger, R. Kornhuber and O. Sander. On nonlinear Dirichlet-Neumann algorithms for jumping nonlinearities. *Domain decomposition methods in science and engineering XVI*, Lect. Notes Comput. Sci. Eng. (55), Springer, Berlin: 489–496 (2007).

[4] A. Bourlioux and A. J. Majda, An elementary model for the validation of flamelet approximations in non-premixed turbulent combustion. *Combust. Theory Model.* 4(2):189–210, 2000.

[5] C. Carstensen, Quasi-interpolation and a posteriori error analysis in finite element methods. *M2AN Math. Model. Numer. Anal.*, 33(6):1187–1202, 1999.

[6] C. Carstensen and R. Verfürth. Edge residuals dominate a posteriori error estimates for low order finite element methods. *SIAM J. Numer. Anal.*, 36(5):1571–1587 (electronic), 1999.

[7] Ph. Clément. Approximation by finite element functions using local regularization. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér., RAIRO Analyse Numérique*, 9(R-2):77–84, 1975.

[8] J. E. Dennis Jr. and R. B. Schnabel Numerical Methods for Unconstrained Optimization and Nonlinear Equations. SIAM Classics in Applied Mathematics (1996).

[9] W. E and B. Engquist. The heterogeneous multiscale methods. *Commun. Math. Sci.*, 1(1):87–132, 2003.

[10] A. Gloria. An analytical framework for the numerical homogenization of monotone elliptic operators and quasiconvex energies. *SIAM Multiscale Model. Simul.*, 5(3):996–1043 (electronic), 2006.

[11] P. Henning and M. Ohlberger. The heterogeneous multiscale finite element method for advection-diffusion problems with rapidly oscillating coefficients and large expected drift. *Netw. Heterog. Media*, 5(4):711–744, 2010.

[12] P. Henning and M. Ohlberger. A Note on Homogenization of Advection-Diffusion Problems with Large Expected Drift. *Z. Anal. Anwend.*, 30(3):319–339, 2011.

[13] P. Henning and M. Ohlberger. A-posteriori error estimation for a heterogeneous multiscale method for monotone operators and beyond a periodic setting. *FB 10 , Universität Münster*, Preprint 04/11 - N, 2011.

[14] P. Henning. Convergence of MsFEM approximations for elliptic, nonperiodic homogenization problems. *Netw. Heterog. Media*, 7(3):503–524, 2012.

[15] T. Hou and X.-H. Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.*, 134(1):169–189, 1997.

[16] T.J.R. Hughes, G.R. Feijóo, L. Mazzei and J.-B. Quincy. The variational multiscale method - a paradigm for computational mechanics. *Comput. Methods Appl. Mech. Engrg.*, 166: 3–24, 1998.

[17] J. Karátson. Characterizing Mesh Independent Quadratic Convergence of Newton's Method for a Class of Elliptic Problems. *J. Math. Anal.*, 44(3):12791303, 2012.

[18] M. G. Larson and A. Målqvist. Adaptive variational multiscale methods based on a posteriori error estimation: energy norm estimates for elliptic problems. *Comput. Methods Appl. Mech. Engrg.*, 196(21-24):2313–2324, 2007.

[19] M. G. Larson and A. Målqvist. An adaptive variational multiscale method for convection-diffusion problems. *Comm. Numer. Methods Engrg.*, 25(1):65–79, 2009.

[20] M. G. Larson and A. Målqvist. A mixed adaptive variational multiscale method with applications in oil reservoir simulation. *Math. Models Methods Appl. Sci.*, 19(7):1017–1042, 2009.

[21] A. Målqvist. Multiscale methods for elliptic problems. *Multiscale Model. Simul.*, 9(3):1064–1086, 2011.

[22] A. Målqvist and D. Peterseim. Localization of Elliptic Multiscale Problems. arXiv:1110.0692v3

[23] J. M. Nordbotten. Adaptive variational multiscale methods for multiphase flow in porous media. *SIAM Multiscale Model. Simul.*, 7(3):1455–1473, 2008.

[24] D. Peterseim and S.A. Sauter. Finite Elements for Elliptic Problems with Highly Varying, Non-Periodic Diffusion Matrix. *SIAM Multiscale Model. Simul.*, 10(3):665–695, 2012.

[25] D. Peterseim. Robustness of Finite Element Simulations in Densely Packed Random Particle Composites. *Netw. Heterog. Media*, 7(1):113-126, 2012.